# Linear Algebra And Analysis 

Kuttler klkuttler@gmail.com

February 7, 2024

## CONTENTS

1 Some Prerequisite Topics ..... 1
1.1 Sets and Set Notation ..... 1
1.2 The Schroder Bernstein Theorem ..... 2
1.3 Equivalence Relations ..... 6
1.4 Well Ordering and Induction ..... 6
1.5 The Complex Numbers and Fields ..... 8
1.6 Polar Form of Complex Numbers ..... 11
1.7 Roots of Complex Numbers ..... 12
1.8 The Quadratic Formula ..... 13
1.9 The Complex Exponential ..... 14
1.10 The Fundamental Theorem of Algebra ..... 15
1.11 Ordered Fields ..... 16
1.12 Division of Numbers ..... 18
1.13 Polynomials ..... 20
1.14 The Method of Partial Fractions ..... 25
1.15 Finite Fields ..... 27
1.16 Some Topics From Analysis ..... 29
1.17 lim sup and lim inf ..... 29
1.18 Exercises ..... 32
I Linear Algebra For Its Own Sake ..... 37
2 Systems of Linear Equations ..... 39
2.1 Elementary Operations ..... 39
2.2 Gauss Elimination ..... 40
2.3 When are Two Polynomials Relatively Prime? ..... 44
2.4 Exercises ..... 45
3 Vector Spaces ..... 51
3.1 Linear Combinations of Vectors, Independence ..... 53
3.2 Subspaces ..... 56
3.3 Exercises ..... 60
3.4 Polynomials and Fields ..... 63
3.4.1 The Algebraic Numbers and Minimum Polynomial ..... 69
3.4.2 Lindermannn Weierstrass Theorem ..... 73
3.5 Exercises ..... 74

## CONTENTS

4 Matrices ..... 77
4.1 Properties of Matrix Multiplication ..... 79
4.2 Finding the Inverse of a Matrix ..... 82
4.3 Linear Relations and Row Operations ..... 85
4.4 Block Multiplication of Matrices ..... 88
4.5 Elementary Matrices ..... 91
4.6 Exercises ..... 95
5 Linear Transformations ..... 101
$5.1 \mathscr{L}(V, W)$ as a Vector Space ..... 101
5.2 The Matrix of a Linear Transformation ..... 103
5.3 Rotations About a Given Vector* ..... 111
5.4 Exercises ..... 113
6 Direct Sums and Block Diagonal Matrices ..... 119
6.1 A Theorem of Sylvester, Direct Sums ..... 124
6.2 Finding the Minimum Polynomial ..... 129
6.3 Eigenvalues, Eigenvectors ..... 132
6.4 Diagonalizability ..... 135
6.5 A Formal Derivative and Diagonalizability ..... 140
6.6 Exercises ..... 142
7 Canonical Forms ..... 151
7.1 Reduction to Diagonal Matrix ..... 151
7.2 Quotients ..... 153
7.3 Cyclic Decomposition ..... 155
7.4 A Direct Sum Decomposition ..... 156
7.5 Uniqueness ..... 158
7.6 Canonical Forms ..... 160
7.7 Exercises ..... 164
8 Determinants ..... 171
8.1 The Function sgn ..... 171
8.2 The Definition of the Determinant ..... 173
8.3 A Symmetric Definition ..... 175
8.4 Basic Properties of the Determinant ..... 176
8.4.1 Binet Cauchy Formula ..... 177
8.5 Expansion Using Cofactors ..... 178
8.6 A Formula for the Inverse ..... 179
8.6.1 Cramer's Rule ..... 180
8.6.2 An Identity of Cauchy ..... 181
8.7 Rank of a Matrix ..... 183
8.8 Summary of Determinants ..... 185
8.9 The Cayley Hamilton Theorem ..... 185
8.10 Exercises ..... 188
9 Some Items Which Resemble Linear Algebra ..... 195
9.1 The Symmetric Polynomial Theorem ..... 195
9.2 Transcendental Numbers ..... 199
9.3 The Fundamental Theorem of Algebra ..... 208
9.4 More on Algebraic Field Extensions ..... 211
9.4.1 The Galois Group ..... 217
9.4.2 Normal Field Extensions ..... 221
9.4.3 Normal Subgroups and Quotient Groups ..... 221
9.4.4 Separable Polynomials ..... 223
9.4.5 Intermediate Fields and Normal Subgroups ..... 225
9.4.6 Permutations ..... 227
9.4.7 Solvable Groups ..... 231
9.4.8 Solvability by Radicals ..... 234
9.5 A Few Generalizations ..... 237
9.5.1 The Normal Closure of a Field Extension ..... 237
9.5.2 Conditions for Separability ..... 238
II Linear Algebra as Baby Functional Analysis ..... 243
10 Normed Linear Spaces ..... 245
10.1 Metric Spaces ..... 245
10.1.1 Limits ..... 245
10.1.2 Cauchy Sequences, Completeness ..... 248
10.1.3 Closure of a Set ..... 250
10.1.4 Continuous Functions ..... 251
10.1.5 Separable Metric Spaces ..... 252
10.1.6 Compact Sets in Metric Space ..... 253
10.1.7 Lipschitz Continuity and Contraction Maps ..... 255
10.1.8 Convergence Of Functions ..... 257
10.2 Connected Sets ..... 259
10.3 Subspaces Spans And Bases ..... 261
10.4 Inner Product and Normed Linear Spaces ..... 263
10.4.1 The Inner Product in $\mathbb{F}^{n}$ ..... 263
10.4.2 General Inner Product Spaces ..... 264
10.4.3 Normed Vector Spaces ..... 265
10.5 Tietze Extension Theorem ..... 266
10.5.1 The $p$ Norms ..... 268
10.5.2 Orthonormal Bases ..... 270
10.6 Equivalence Of Norms ..... 272
10.7 Norms On $\mathscr{L}(X, Y)$ ..... 274
10.8 Limits Of A Function ..... 276
10.9 Exercises ..... 279
11 Limits of Vectors and Matrices ..... 285
11.1 Regular Markov Matrices ..... 285
11.2 Migration Matrices ..... 289
11.3 Absorbing States ..... 289
11.4 Positive Matrices ..... 293
11.5 Functions Of Matrices ..... 300
11.6 Exercises ..... 304

## CONTENTS

12 Inner Product Spaces, Least Squares ..... 307
12.1 Orthogonal Projections ..... 307
12.2 Formula for Distance to a Subspace ..... 312
12.3 Riesz Representation Theorem, Adjoint Map ..... 313
12.4 Least Squares ..... 317
12.5 Fredholm Alternative ..... 318
12.6 The Determinant and Volume ..... 319
12.7 Finding an Orthogonal Basis ..... 320
12.8 Exercises ..... 321
13 Matrices and the Inner Product ..... 329
13.1 Schur's Theorem, Hermitian Matrices ..... 329
13.2 Quadratic Forms ..... 333
13.3 The Estimation Of Eigenvalues ..... 334
13.4 Advanced Theorems ..... 336
13.5 Exercises ..... 339
13.6 Cauchy's Interlacing Theorem, Eigenvalues ..... 348
13.7 The Right Polar Factorization ..... 351
13.8 The Square Root ..... 354
13.9 An Application To Statistics ..... 355
13.10 Simultaneous Diagonalization ..... 357
13.11 Fractional Powers ..... 362
13.12 Roots of Positive Linear Maps ..... 363
13.12.1 The Product of Positive Self Adjoint Operators ..... 363
13.12.2 Roots of Positive Self Adjoint Operators ..... 364
13.13 Spectral Theory of Self Adjoint Operators ..... 366
13.14 Positive and Negative Linear Transformations ..... 370
13.15 The Singular Value Decomposition ..... 372
13.16 Approximation In The Frobenius Norm ..... 374
13.17 Least Squares And Singular Value Decomposition ..... 377
13.18 The Moore Penrose Inverse ..... 377
13.19 The Spectral Norm And The Operator Norm ..... 381
13.20 The Positive Part Of A Hermitian Matrix ..... 382
13.21 Exercises ..... 383
14 Analysis Of Linear Transformations ..... 389
14.1 The Condition Number ..... 389
14.2 The Spectral Radius ..... 390
14.3 Series And Sequences Of Linear Operators ..... 393
14.4 Iterative Methods For Linear Systems ..... 398
14.5 Exercises ..... 403
15 Numerical Methods, Eigenvalues ..... 411
15.1 The Power Method For Eigenvalues ..... 411
15.1.1 The Shifted Inverse Power Method ..... 414
15.1.2 The Explicit Description Of The Method ..... 415
15.2 Automation With Matlab ..... 415
15.2.1 Complex Eigenvalues ..... 418
15.2.2 Rayleigh Quotients and Estimates for Eigenvalues ..... 419
15.3 The $Q R$ Algorithm ..... 423
15.3.1 Basic Properties And Definition ..... 423
15.3.2 The Case Of Real Eigenvalues ..... 426
15.3.3 The $Q R$ Algorithm In The General Case ..... 430
15.3.4 Upper Hessenberg Matrices ..... 435
15.4 Exercises ..... 438
16 Approximation of Functions and the Integral ..... 441
16.1 Weierstrass Approximation Theorem ..... 441
16.2 Functions of Many Variables ..... 443
16.3 A Generalization with Tietze Extension Theorem ..... 446
16.4 An Approach to the Integral ..... 446
16.5 The Müntz Theorems ..... 452
16.6 Exercises ..... 454
A Homological Methods* ..... 455
A. 1 Singular Simplices and Boundaries ..... 456
A. 2 The Homology Groups ..... 458
A. 3 Homotopy ..... 462
A. 4 The Boundary Map on Geometric Simplices ..... 466
A. 5 The Subdivision Operation ..... 467
A. 6 Exact Sequences ..... 474
A. 7 Computing Homology Groups ..... 479
A. 8 The Homology Groups of Spheres ..... 480
A. 9 Brouwer Fixed Points ..... 482
A. 10 Topological Degree on Spheres ..... 484
A. 11 Functions Defined on a Subset of $S^{n}$ ..... 486
A. 12 The Degree on Open Subsets of $\mathbb{R}^{n}$ ..... 491
A. 13 Jordan Separation Theorem ..... 494
A. 14 Analysis and the Degree ..... 498
B The Hausdorff Maximal Theorem ..... 503
B. 1 The Hamel Basis ..... 505
B. 2 Exercises ..... 506

## CONTENTS

Copyright © 2018, You are welcome to use this, including copying it for use in classes or referring to it on line but not to publish it for money.

## CONTENTS

## Preface

Is Linear Algebra part of Modern Algebra or is it baby Functional Analysis? It depends somewhat on your interests. I tend to lean toward the baby functional analysis, but algebraic ideas are certainly important and some would give strong arguments that these ideas are of most significance. Certainly it is all about linear transformations however you look at it, and the canonical forms are completely algebraic in nature. I have therefore, chosen to present the subject in two parts, the first being Linear Algebra as a part of Algebra with very little if any reference to Analysis. It involves general fields of scalars and makes no reference or minimal reference to completeness. This is all about polynomials as formal objects and the division algorithm. After this, the more analytical aspects of this subject are considered, inner products, numerical methods, applications to differential equations and so forth. The field of scalars will be the real or complex numbers. Some analysis ideas do in fact creep in to the first part, but they are generally fairly rudimentary, occur as examples, and will have been seen in calculus. This book is not meant to be read before a calculus course.

It may be that increased understanding is obtained by this kind of presentation in which that which is purely algebraic is presented first. This also involves emphasizing the minimum polynomial more than the characteristic polynomial and postponing the determinant. In each part, I have included a few related topics which are similar to ideas found in linear algebra or which have linear algebra as a fundamental part.

The book is a re written version of an earlier book. It also includes several topics not in this other book. I have tried to introduce rings and modules in the context of a presentation of the canonical forms.

## CONTENTS

## Chapter 1

## Some Prerequisite Topics

The reader should be familiar with most of the topics in this chapter. However, it is often the case that set notation is not familiar and so a short discussion of this is included first. Complex numbers are then considered in somewhat more detail. Many of the applications of linear algebra require the use of complex numbers, so this is the reason for this introduction. Then polynomials and finite fields are discussed briefly to emphasize that linear algebra works for any field of scalars, not just the field of real and complex numbers.

### 1.1 Sets and Set Notation

A set is just a collection of things called elements. Often these are also referred to as points in calculus. For example $\{1,2,3,8\}$ would be a set consisting of the elements $1,2,3$, and 8. To indicate that 3 is an element of $\{1,2,3,8\}$, it is customary to write $3 \in\{1,2,3,8\}$. $9 \notin\{1,2,3,8\}$ means 9 is not an element of $\{1,2,3,8\}$. Sometimes a rule specifies a set. For example you could specify a set as all integers larger than 2 . This would be written as $S=\{x \in \mathbb{Z}: x>2\}$. This notation says: the set of all integers, $x$, such that $x>2$.

If $A$ and $B$ are sets with the property that every element of $A$ is an element of $B$, then $A$ is a subset of $B$. For example, $\{1,2,3,8\}$ is a subset of $\{1,2,3,4,5,8\}$, in symbols, $\{1,2,3,8\} \subseteq\{1,2,3,4,5,8\}$. It is sometimes said that " $A$ is contained in $B$ " or even " $B$ contains $A$ ". The same statement about the two sets may also be written as $\{1,2,3,4,5,8\} \supseteq$ $\{1,2,3,8\}$.

The union of two sets is the set consisting of everything which is an element of at least one of the sets, $A$ or $B$. As an example of the union of two sets $\{1,2,3,8\} \cup\{3,4,7,8\}=$ $\{1,2,3,4,7,8\}$ because these numbers are those which are in at least one of the two sets. In general

$$
A \cup B \equiv\{x: x \in A \text { or } x \in B\}
$$

Be sure you understand that something which is in both $A$ and $B$ is in the union. It is not an exclusive or.

The intersection of two sets, $A$ and $B$ consists of everything which is in both of the sets. Thus $\{1,2,3,8\} \cap\{3,4,7,8\}=\{3,8\}$ because 3 and 8 are those elements the two sets have in common. In general,

$$
A \cap B \equiv\{x: x \in A \text { and } x \in B\} .
$$

The symbol $[a, b]$ where $a$ and $b$ are real numbers, denotes the set of real numbers $x$, such that $a \leq x \leq b$ and $[a, b)$ denotes the set of real numbers such that $a \leq x<b$. $(a, b)$ consists of the set of real numbers $x$ such that $a<x<b$ and ( $a, b]$ indicates the set of numbers $x$ such that $a<x \leq b .[a, \infty)$ means the set of all numbers $x$ such that $x \geq a$ and $(-\infty, a]$ means the set of all real numbers which are less than or equal to $a$. These sorts of sets of real numbers are called intervals. The two points $a$ and $b$ are called endpoints of the interval. Other intervals such as $(-\infty, b)$ are defined by analogy to what was just explained. In general, the curved parenthesis indicates the end point it sits next to is not included while the square parenthesis indicates this end point is included. The reason that there will always be a curved parenthesis next to $\infty$ or $-\infty$ is that these are not real numbers. Therefore, they cannot be included in any set of real numbers.

A special set which needs to be given a name is the empty set also called the null set, denoted by $\emptyset$. Thus $\emptyset$ is defined as the set which has no elements in it. Mathematicians like to say the empty set is a subset of every set. The reason they say this is that if it were
not so, there would have to exist a set $A$, such that $\emptyset$ has something in it which is not in $A$. However, $\emptyset$ has nothing in it and so the least intellectual discomfort is achieved by saying $\emptyset \subseteq A$.

If $A$ and $B$ are two sets, $A \backslash B$ denotes the set of things which are in $A$ but not in $B$. Thus

$$
A \backslash B \equiv\{x \in A: x \notin B\}
$$

Set notation is used whenever convenient.
To illustrate the use of this notation relative to intervals consider three examples of inequalities. Their solutions will be written in the notation just described.

Example 1.1.1 Solve the inequality $2 x+4 \leq x-8$
$x \leq-12$ is the answer. This is written in terms of an interval as $(-\infty,-12]$.
Example 1.1.2 Solve the inequality $(x+1)(2 x-3) \geq 0$.
The solution is $x \leq-1$ or $x \geq \frac{3}{2}$. In terms of set notation this is denoted by $(-\infty,-1] \cup$ $\left[\frac{3}{2}, \infty\right)$.

Example 1.1.3 Solve the inequality $x(x+2) \geq-4$.
This is true for any value of $x$. It is written as $\mathbb{R}$ or $(-\infty, \infty)$.
Something is in the Cartesian product of a set whose elements are sets if it consists of a single thing taken from each set in the family. Thus $(1,2,3) \in\{1,4, .2\} \times\{1,2,7\} \times$ $\{4,3,7,9\}$ because it consists of exactly one element from each of the sets which are separated by $\times$. Also, this is the notation for the Cartesian product of finitely many sets. If $\mathscr{S}$ is a set whose elements are sets, $\prod_{A \in \mathscr{S}} A$ signifies the Cartesian product.

The Cartesian product is the set of choice functions, a choice function being a function which selects exactly one element of each set of $\mathscr{S}$. You may think the axiom of choice, stating that the Cartesian product of a nonempty family of nonempty sets is nonempty, is innocuous but there was a time when many mathematicians were ready to throw it out because it implies things which are very hard to believe, things which never happen without the axiom of choice.

### 1.2 The Schroder Bernstein Theorem

It is very important to be able to compare the size of sets in a rational way. The most useful theorem in this context is the Schroder Bernstein theorem which is the main result to be presented in this section. The Cartesian product is discussed above. The next definition reviews this and defines the concept of a function.

Definition 1.2.1 Let $X$ and $Y$ be sets.

$$
X \times Y \equiv\{(x, y): x \in X \text { and } y \in Y\}
$$

A relation is defined to be a subset of $X \times Y$. A function $f$, also called a mapping, is a relation which has the property that if $(x, y)$ and $\left(x, y_{1}\right)$ are both elements of the $f$, then $y=y_{1}$. The domain of $f$ is defined as

$$
D(f) \equiv\{x:(x, y) \in f\}
$$

written as $f: D(f) \rightarrow Y$. Another notation which is used is the following

$$
f^{-1}(y) \equiv\{x \in D(f): f(x)=y\}
$$

This is called the inverse image.
It is probably safe to say that most people do not think of functions as a type of relation which is a subset of the Cartesian product of two sets. A function is like a machine which takes inputs, $x$ and makes them into a unique output, $f(x)$. Of course, that is what the above definition says with more precision. An ordered pair, $(x, y)$ which is an element of the function or mapping has an input, $x$ and a unique output $y$,denoted as $f(x)$ while the name of the function is $f$. "mapping" is often a noun meaning function. However, it also is a verb as in " $f$ is mapping $A$ to $B$ ". That which a function is thought of as doing is also referred to using the word "maps" as in: $f$ maps $X$ to $Y$. However, a set of functions may be called a set of maps so this word might also be used as the plural of a noun. There is no help for it. You just have to suffer with this nonsense.

The following theorem which is interesting for its own sake will be used to prove the Schroder Bernstein theorem, proved by Dedekind in 1887. The proof given here is like the version in Hewitt and Stromberg [21].

Theorem 1.2.2 Let $f: X \rightarrow Y$ and $g: Y \rightarrow X$ be two functions. Then there exist sets $A, B, C, D$, such that

$$
\begin{gathered}
A \cup B=X, C \cup D=Y, A \cap B=\emptyset, C \cap D=\emptyset, \\
f(A)=C, g(D)=B .
\end{gathered}
$$

The following picture illustrates the conclusion of this theorem.


Proof: Consider the empty set, $\emptyset \subseteq X$. If $y \in Y \backslash f(\emptyset)$, then $g(y) \notin \emptyset$ because $\emptyset$ has no elements. Also, if $A, B, C$, and $D$ are as described above, $A$ also would have this same property that the empty set has. However, $A$ is probably larger. Therefore, say $A_{0} \subseteq X$ satisfies $\mathscr{P}$ if whenever $y \in Y \backslash f\left(A_{0}\right), g(y) \notin A_{0}$.

$$
\mathscr{A} \equiv\left\{A_{0} \subseteq X: A_{0} \text { satisfies } \mathscr{P}\right\}
$$

Let $A=\cup \mathscr{A}$. If $y \in Y \backslash f(A)$, then for each $A_{0} \in \mathscr{A}, y \in Y \backslash f\left(A_{0}\right)$ and so $g(y) \notin A_{0}$. Since $g(y) \notin A_{0}$ for all $A_{0} \in \mathscr{A}$, it follows $g(y) \notin A$. Hence $A$ satisfies $\mathscr{P}$ and is the largest subset of $X$ which does so. Now define

$$
C \equiv f(A), D \equiv Y \backslash C, B \equiv X \backslash A
$$

It only remains to verify that $g(D)=B$. It was just shown that $g(D) \subseteq B$.
Suppose $x \in B=X \backslash A$. Then $A \cup\{x\}$ does not satisfy $\mathscr{P}$ and so there exists $y \in$ $Y \backslash f(A \cup\{x\}) \subseteq D$ such that $g(y) \in A \cup\{x\}$. But $y \notin f(A)$ and so since $A$ satisfies $\mathscr{P}$, it follows $g(y) \notin A$. Hence $g(y)=x$ and so $x \in g(D)$. Hence $g(D)=B$.

Theorem 1.2.3 (Schroder Bernstein) If $f: X \rightarrow Y$ and $g: Y \rightarrow X$ are one to one, then there exists $h: X \rightarrow Y$ which is one to one and onto.

Proof:Let $A, B, C, D$ be the sets of Theorem 1.2.2 and define

$$
h(x) \equiv\left\{\begin{array}{c}
f(x) \text { if } x \in A \\
g^{-1}(x) \text { if } x \in B
\end{array}\right.
$$

Then $h$ is the desired one to one and onto mapping.
Recall that the Cartesian product may be considered as the collection of choice functions.

Definition 1.2.4 Let $I$ be a set and let $X_{i}$ be a set for each $i \in I . f$ is a choice function written as $f \in \prod_{i \in I} X_{i}$ if $f(i) \in X_{i}$ for each $i \in I$.

The axiom of choice says that if $X_{i} \neq \emptyset$ for each $i \in I$, for $I$ a set, then $\prod_{i \in I} X_{i} \neq \emptyset$.
Sometimes the two functions, $f$ and $g$ are onto but not one to one. It turns out that with the axiom of choice, a similar conclusion to the above may be obtained.

Corollary 1.2.5 If $f: X \rightarrow Y$ is onto and $g: Y \rightarrow X$ is onto, then there exists $h: X \rightarrow Y$ which is one to one and onto.

Proof: For each $y \in Y, f^{-1}(y) \equiv\{x \in X: f(x)=y\} \neq \emptyset$. Therefore, by the axiom of choice, there exists $f_{0}^{-1} \in \prod_{y \in Y} f^{-1}(y)$ which is the same as saying that for each $y \in Y$, $f_{0}^{-1}(y) \in f^{-1}(y)$. Similarly, there exists $g_{0}^{-1}(x) \in g^{-1}(x)$ for all $x \in X$. Then $f_{0}^{-1}$ is one to one because if $f_{0}^{-1}\left(y_{1}\right)=f_{0}^{-1}\left(y_{2}\right)$, then $y_{1}=f\left(f_{0}^{-1}\left(y_{1}\right)\right)=f\left(f_{0}^{-1}\left(y_{2}\right)\right)=y_{2}$. Similarly $g_{0}^{-1}$ is one to one. Therefore, by the Schroder Bernstein theorem, there exists $h: X \rightarrow Y$ which is one to one and onto.

Definition 1.2.6 $A$ set $S$, is finite if there exists a natural number $n$ and a map $\theta$ which maps $\{1, \cdots, n\}$ one to one and onto $S$. $S$ is infinite if it is not finite. A set $S$, is called countable if there exists a map $\theta$ mapping $\mathbb{N}$ one to one and onto $S$.(When $\theta$ maps a set $A$ to a set $B$, this will be written as $\theta: A \rightarrow B$ in the future.) Here $\mathbb{N} \equiv\{1,2, \cdots\}$, the natural numbers. $S$ is at most countable if there exists a map $\theta: \mathbb{N} \rightarrow S$ which is onto.

The property of being at most countable is often referred to as being countable because the question of interest is normally whether one can list all elements of the set, designating a first, second, third etc. in such a way as to give each element of the set a natural number. The possibility that a single element of the set may be counted more than once is often not important.

Theorem 1.2.7 If $X$ and $Y$ are both at most countable, then $X \times Y$ is also at most countable. If either $X$ or $Y$ is countable, then $X \times Y$ is also countable.

Proof:It is given that there exists a mapping $\eta: \mathbb{N} \rightarrow X$ which is onto. Define $\eta(i) \equiv x_{i}$ and consider $X$ as the set $\left\{x_{1}, x_{2}, x_{3}, \cdots\right\}$. Similarly, consider $Y$ as the set $\left\{y_{1}, y_{2}, y_{3}, \cdots\right\}$. It follows the elements of $X \times Y$ are included in the following rectangular array.

$$
\begin{array}{ccccc}
\left(x_{1}, y_{1}\right) & \left(x_{1}, y_{2}\right) & \left(x_{1}, y_{3}\right) & \cdots & \leftarrow \text { Those which have } x_{1} \text { in first slot. } \\
\left(x_{2}, y_{1}\right) & \left(x_{2}, y_{2}\right) & \left(x_{2}, y_{3}\right) & \cdots & \leftarrow \text { Those which have } x_{2} \text { in first slot. } \\
\left(x_{3}, y_{1}\right) & \left(x_{3}, y_{2}\right) & \left(x_{3}, y_{3}\right) & \cdots & \leftarrow \text { Those which have } x_{3} \text { in first slot. }
\end{array}
$$

Follow a path through this array as follows.


Thus the first element of $X \times Y$ is $\left(x_{1}, y_{1}\right)$, the second element of $X \times Y$ is $\left(x_{1}, y_{2}\right)$, the third element of $X \times Y$ is $\left(x_{2}, y_{1}\right)$ etc. This assigns a number from $\mathbb{N}$ to each element of $X \times Y$. Thus $X \times Y$ is at most countable.

It remains to show the last claim. Suppose without loss of generality that $X$ is countable. Then there exists $\alpha: \mathbb{N} \rightarrow X$ which is one to one and onto. Let $\beta: X \times Y \rightarrow \mathbb{N}$ be defined by $\beta((x, y)) \equiv \alpha^{-1}(x)$. Thus $\beta$ is onto $\mathbb{N}$. By the first part there exists a function from $\mathbb{N}$ onto $X \times Y$. Therefore, by Corollary 1.2.5, there exists a one to one and onto mapping from $X \times Y$ to $\mathbb{N}$.

Theorem 1.2.8 If $X$ and $Y$ are at most countable, then $X \cup Y$ is at most countable. If either $X$ or $Y$ are countable, then $X \cup Y$ is countable.

Proof:As in the preceding theorem,

$$
X=\left\{x_{1}, x_{2}, x_{3}, \cdots\right\}, Y=\left\{y_{1}, y_{2}, y_{3}, \cdots\right\}
$$

Consider the following array consisting of $X \cup Y$ and path through it.


Thus the first element of $X \cup Y$ is $x_{1}$, the second is $x_{2}$ the third is $y_{1}$ the fourth is $y_{2}$ etc.
Consider the second claim. By the first part, there is a map from $\mathbb{N}$ onto $X \times Y$. Suppose without loss of generality that $X$ is countable and $\alpha: \mathbb{N} \rightarrow X$ is one to one and onto. Then define $\beta(y) \equiv 1$, for all $y \in Y$, and $\beta(x) \equiv \alpha^{-1}(x)$. Thus, $\beta$ maps $X \times Y$ onto $\mathbb{N}$ and this shows there exist two onto maps, one mapping $X \cup Y$ onto $\mathbb{N}$ and the other mapping $\mathbb{N}$ onto $X \cup Y$. Then Corollary 1.2.5 yields the conclusion.

Note that by induction this shows that if you have any finite set whose elements are countable sets, then the union of these is countable.

### 1.3 Equivalence Relations

There are many ways to compare elements of a set other than to say two elements are equal or the same. For example, in the set of people let two people be equivalent if they have the same weight. This would not be saying they were the same person, just that they weigh the same. Often such relations involve considering one characteristic of the elements of a set and then saying the two elements are equivalent if they are the same as far as the given characteristic is concerned.

Definition 1.3.1 Let $S$ be a set. $\sim$ is an equivalence relation on $S$ if it satisfies the following axioms.

1. $x \sim x$ for all $x \in S$. (Reflexive)
2. If $x \sim y$ then $y \sim x$. (Symmetric)
3. If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)

Definition 1.3.2 $[x]$ denotes the set of all elements of $S$ which are equivalent to $x$ and $[x]$ is called the equivalence class determined by $x$ or just the equivalence class of $x$.

With the above definition one can prove the following simple theorem.
Theorem 1.3.3 Let $\sim$ be an equivalence relation defined on a set, $S$ and let $\mathscr{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x]=[y]$ or it is not true that $x \sim y$ and $[x] \cap[y]=\emptyset$.

Proof: If $x \sim y$, then if $z \in[y]$, you have $x \sim y$ and $y \sim z$ so $x \sim z$ which shows that $[y] \subseteq[x]$. Similarly, $[x] \subseteq[y]$. If it is not the case that $x \sim y$, then there can be no intersection of $[x]$ and $[y]$ because if $z$ were in this intersection, then $x \sim z, z \sim y$ so $x \sim y$.

### 1.4 Well Ordering and Induction

Mathematical induction and well ordering are two extremely important principles in math. They are often used to prove significant things which would be hard to prove otherwise.

Definition 1.4.1 A set is well ordered if every nonempty subset $S$, contains a smallest element $z$ having the property that $z \leq x$ for all $x \in S$.

Axiom 1.4.2 Any set of integers larger than a given number is well ordered.
In particular, the natural numbers defined as $\mathbb{N} \equiv\{1,2, \cdots\}$ is well ordered.
The above axiom implies the principle of mathematical induction. The symbol $\mathbb{Z}$ denotes the set of all integers. Note that if $a$ is an integer, then there are no integers between $a$ and $a+1$.

Theorem 1.4.3 (Mathematical induction) A set $S \subseteq \mathbb{Z}$, having the property that $a \in S$ and $n+1 \in S$ whenever $n \in S$ contains all integers $x \in \mathbb{Z}$ such that $x \geq a$.

Proof: Let $T$ consist of all integers larger than or equal to $a$ which are not in $S$. The theorem will be proved if $T=\emptyset$. If $T \neq \emptyset$ then by the well ordering principle, there would have to exist a smallest element of $T$, denoted as $b$. It must be the case that $b>a$ since by definition, $a \notin T$. Thus $b \geq a+1$, and so $b-1 \geq a$ and $b-1 \notin S$ because if $b-1 \in S$, then $b-1+1=b \in S$ by the assumed property of $S$. Therefore, $b-1 \in T$ which contradicts the choice of $b$ as the smallest element of $T$. ( $b-1$ is smaller.) Since a contradiction is obtained by assuming $T \neq \emptyset$, it must be the case that $T=\emptyset$ and this says that every integer at least as large as $a$ is also in $S$.

Mathematical induction is a very useful device for proving theorems about the integers.
Example 1.4.4 Prove by induction that $\sum_{k=1}^{n} k^{2}=\frac{n(n+1)(2 n+1)}{6}$.
-
By inspection, if $n=1$ then the formula is true. The sum yields 1 and so does the formula on the right. Suppose this formula is valid for some $n \geq 1$ where $n$ is an integer. Then

$$
\sum_{k=1}^{n+1} k^{2}=\sum_{k=1}^{n} k^{2}+(n+1)^{2}=\frac{n(n+1)(2 n+1)}{6}+(n+1)^{2}
$$

The step going from the first to the second line is based on the assumption that the formula is true for $n$. This is called the induction hypothesis. Now simplify the expression in the second line,

$$
\frac{n(n+1)(2 n+1)}{6}+(n+1)^{2}
$$

This equals

$$
(n+1)\left(\frac{n(2 n+1)}{6}+(n+1)\right)
$$

and

$$
\frac{n(2 n+1)}{6}+(n+1)=\frac{6(n+1)+2 n^{2}+n}{6}=\frac{(n+2)(2 n+3)}{6}
$$

Therefore,

$$
\sum_{k=1}^{n+1} k^{2}=\frac{(n+1)(n+2)(2 n+3)}{6}=\frac{(n+1)((n+1)+1)(2(n+1)+1)}{6}
$$

showing the formula holds for $n+1$ whenever it holds for $n$. This proves the formula by mathematical induction.

Example 1.4.5 Show that for all $n \in \mathbb{N}, \frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2 n-1}{2 n}<\frac{1}{\sqrt{2 n+1}}$.
If $n=1$ this reduces to the statement that $\frac{1}{2}<\frac{1}{\sqrt{3}}$ which is obviously true. Suppose then that the inequality holds for $n$. Then

$$
\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2 n-1}{2 n} \cdot \frac{2 n+1}{2 n+2}<\frac{1}{\sqrt{2 n+1}} \frac{2 n+1}{2 n+2}=\frac{\sqrt{2 n+1}}{2 n+2}
$$

The theorem will be proved if this last expression is less than $\frac{1}{\sqrt{2 n+3}}$. This happens if and only if

$$
\left(\frac{1}{\sqrt{2 n+3}}\right)^{2}=\frac{1}{2 n+3}>\frac{2 n+1}{(2 n+2)^{2}}
$$

which occurs if and only if $(2 n+2)^{2}>(2 n+3)(2 n+1)$ and this is clearly true which may be seen from expanding both sides. This proves the inequality.

Lets review the process just used. If $S$ is the set of integers at least as large as 1 for which the formula holds, the first step was to show $1 \in S$ and then that whenever $n \in S$, it follows $n+1 \in S$. Therefore, by the principle of mathematical induction, $S$ contains $[1, \infty) \cap \mathbb{Z}$, all positive integers. In doing an inductive proof of this sort, the set $S$ is normally not mentioned. One just verifies the steps above. First show the thing is true for some $a \in \mathbb{Z}$ and then verify that whenever it is true for $m$ it follows it is also true for $m+1$. When this has been done, the theorem has been proved for all $m \geq a$.

### 1.5 The Complex Numbers and Fields

Recall that a real number is a point on the real number line. Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus $(a, b)$ identifies a point whose $x$ coordinate is $a$ and whose $y$ coordinate is $b$. In dealing with complex numbers, such a point is written as $a+i b$. For example, in the following picture, I have graphed the point $3+2 i$. You see it corresponds to the point in the plane whose coordinates are $(3,2)$.

$$
f_{, \quad,} \cdot 3+2 i
$$

and addition are defined in the most obvious way subject to the convention that $i^{2}=$ -1 . Thus,

$$
(a+i b)+(c+i d)=(a+c)+i(b+d)
$$

and

$$
\begin{aligned}
(a+i b)(c+i d) & =a c+i a d+i b c+i^{2} b d \\
& =(a c-b d)+i(b c+a d)
\end{aligned}
$$

Every non zero complex number $a+i b$, with $a^{2}+b^{2} \neq 0$, has a unique multiplicative inverse.

$$
\frac{1}{a+i b}=\frac{a-i b}{a^{2}+b^{2}}=\frac{a}{a^{2}+b^{2}}-i \frac{b}{a^{2}+b^{2}}
$$

You should prove the following theorem, assuming $\mathbb{R}$ is a field.
Theorem 1.5.1 The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms. These are the following list of properties. In this list, $\mathbb{F}$ is the symbol for a field.

1. $x+y=y+x$, (commutative law for addition)
2. There exists 0 such that $x+0=x$ for all $x$, (additive identity).
3. For each $x \in \mathbb{F}$, there exists $-x \in \mathbb{F}$ such that $x+(-x)=0$, (existence of additive inverse).
4. $(x+y)+z=x+(y+z),($ associative law for addition $)$.
5. $x y=y x$, (commutative law for multiplication). You could write this as $x \times y=y \times x$.
6. $(x y) z=x(y z),($ associative law for multiplication $)$.
7. There exists 1 such that $1 x=x$ for all $x$,(multiplicative identity).
8. For each $x \neq 0$, there exists $x^{-1}$ such that $x x^{-1}=1$.(existence of multiplicative inverse).
9. $x(y+z)=x y+x z$. (distributive law $)$.

The symbol $x-y$ means $x+(-y)$. We call this subtraction of $y$ from $x$. The symbol $x / y$ for $y \neq 0$ means $x\left(y^{-1}\right)$. This is called division. When you have a field $\mathbb{F}$ some things follow right away from the above axioms.

Theorem 1.5.2 Let $\mathbb{F}$ be a field. This means it satisfies the axioms of the above theorem. Then the following hold.

$$
\begin{array}{lll}
\text { 1. } 0 \text { is unique } & 2 .-x \text { is unique } & 3.0 x=0 \\
\text { 4. }(-1) x=-x & \text { 3. } x^{-1} \text { is unique } &
\end{array}
$$

Proof: Consider the first claim. Suppose $\hat{0}$ is another additive identity. Then

$$
\hat{0}=\hat{0}+0=0
$$

and so sure enough, there is only one such additive identity. Consider uniqueness of $-x$ next. Suppose $y$ is also an additive inverse. Then

$$
-x=-x+0=-x+(x+y)=(-x+x)+y=0+y=y
$$

so the additive inverse is unique also.

$$
0 x=(0+0) x=0 x+0 x
$$

Now add $-0 x$ to both sides to conclude that $0=0 x$. Next

$$
0=(1+-1) x=x+(-1) x
$$

and by uniqueness of $-x$, this implies $(-1) x=-x$ as claimed. Finally, if $x \neq 0$ and $y$ is a multiplicative inverse,

$$
x^{-1}=1 x^{-1}=(y x) x^{-1}=y\left(x x^{-1}\right)=y 1=y
$$

so $y=x^{-1}$.
Note that if $0=1$, you would have from the above that all the elements of the field are 0 . This is of no interest at all so we typically assume $0 \neq 1$. Something which satisfies
these axioms is called a field. Linear algebra is all about fields, although in this book, the field of most interest will be the field of complex numbers or the field of real numbers. You have seen in earlier courses that the real numbers also satisfy the above axioms. For a proof of this well accepted fact and construction of the real numbers, see Hobson [22] or my single variable advanced calculus book. Other books which do this are Hewitt and Stromberg [21] or Rudin [36]. There are two ways to show this, one due to Cantor and the other by Dedikind. Both are in Hobson, my book follows Cantor and so does the one by Hewitt and Stromberg. Rudin presents the other method.

An important construction regarding complex numbers is the complex conjugate denoted by a horizontal line above the number. It is defined as follows.

$$
\overline{a+i b} \equiv a-i b
$$

What it does is reflect a given complex number across the $x$ axis. Algebraically, the following formula is easy to obtain.

$$
\begin{aligned}
(\overline{a+i b})(a+i b) & =(a-i b)(a+i b) \\
& =a^{2}+b^{2}-i(a b-a b)=a^{2}+b^{2}
\end{aligned}
$$

Definition 1.5.3 Define the absolute value of a complex number as follows.

$$
|a+i b| \equiv \sqrt{a^{2}+b^{2}}
$$

Thus, denoting by $z$ the complex number $z=a+i b$,

$$
|z|=(z \bar{z})^{1 / 2}
$$

Also from the definition, if $z=x+i y$ and $w=u+i v$ are two complex numbers, then $|z w|=|z||w|$. You should verify this.

Notation 1.5.4 Recall the following notation. $\sum_{j=1}^{n} a_{j} \equiv a_{1}+\cdots+a_{n}$. There is also $a$ notation which is used to denote a product. $\prod_{j=1}^{n} a_{j} \equiv a_{1} a_{2} \cdots a_{n}$.

The triangle inequality holds for the absolute value for complex numbers just as it does for the ordinary absolute value.

Proposition 1.5.5 Let $z, w$ be complex numbers. Then the triangle inequality holds.

$$
|z+w| \leq|z|+|w|, \quad \| z|-|w|| \leq|z-w| .
$$

Proof: Let $z=x+i y$ and $w=u+i v$. First note that

$$
z \bar{w}=(x+i y)(u-i v)=x u+y v+i(y u-x v)
$$

and so $|x u+y v| \leq|z \bar{w}|=|z||w|$.

$$
\begin{gathered}
|z+w|^{2}=(x+u+i(y+v))(x+u-i(y+v)) \\
=(x+u)^{2}+(y+v)^{2}=x^{2}+u^{2}+2 x u+2 y v+y^{2}+v^{2} \\
\leq|z|^{2}+|w|^{2}+2|z||w|=(|z|+|w|)^{2},
\end{gathered}
$$

so this shows the first version of the triangle inequality. To get the second,

$$
z=z-w+w, w=w-z+z
$$

and so by the first form of the inequality

$$
|z| \leq|z-w|+|w|,|w| \leq|z-w|+|z|
$$

and so both $|z|-|w|$ and $|w|-|z|$ are no larger than $|z-w|$ and this proves the second version because $||z|-|w||$ is one of $|z|-|w|$ or $|w|-|z|$.

With this definition, it is important to note the following. Be sure to verify this. It is not too hard but you need to do it.

Remark 1.5.6 $:$ Let $z=a+i b$ and $w=c+i d$. Then $|z-w|=\sqrt{(a-c)^{2}+(b-d)^{2}}$. Thus the distance between the point in the plane determined by the ordered pair $(a, b)$ and the ordered pair $(c, d)$ equals $|z-w|$ where $z$ and $w$ are as just described.

For example, consider the distance between $(2,5)$ and $(1,8)$. From the distance formula this distance equals $\sqrt{(2-1)^{2}+(5-8)^{2}}=\sqrt{10}$. On the other hand, letting $z=2+i 5$ and $w=1+i 8, z-w=1-i 3$ and so $(z-w)(\overline{z-w})=(1-i 3)(1+i 3)=10$ so $|z-w|=\sqrt{10}$, the same thing obtained with the distance formula.

### 1.6 Polar Form of Complex Numbers

Complex numbers, are often written in the so called polar form which is described next. Suppose $z=x+i y$ is a complex number. Then

$$
x+i y=\sqrt{x^{2}+y^{2}}\left(\frac{x}{\sqrt{x^{2}+y^{2}}}+i \frac{y}{\sqrt{x^{2}+y^{2}}}\right) .
$$

Now note that

$$
\left(\frac{x}{\sqrt{x^{2}+y^{2}}}\right)^{2}+\left(\frac{y}{\sqrt{x^{2}+y^{2}}}\right)^{2}=1
$$

and so $\left(\frac{x}{\sqrt{x^{2}+y^{2}}}, \frac{y}{\sqrt{x^{2}+y^{2}}}\right)$ is a point on the unit circle. Therefore, there exists a unique angle $\theta \in[0,2 \pi)$ such that

$$
\cos \theta=\frac{x}{\sqrt{x^{2}+y^{2}}}, \sin \theta=\frac{y}{\sqrt{x^{2}+y^{2}}}
$$

The polar form of the complex number is then $r(\cos \theta+i \sin \theta)$ where $\theta$ is this angle just described and $r=\sqrt{x^{2}+y^{2}} \equiv|z|$.


### 1.7 Roots of Complex Numbers

A fundamental identity is the formula of De Moivre which follows.
Theorem 1.7.1 Let $r>0$ be given. Then if $n$ is a positive integer,

$$
[r(\cos t+i \sin t)]^{n}=r^{n}(\cos n t+i \sin n t)
$$

Proof: It is clear the formula holds if $n=1$. Suppose it is true for $n$.

$$
[r(\cos t+i \sin t)]^{n+1}=[r(\cos t+i \sin t)]^{n}[r(\cos t+i \sin t)]
$$

which by induction equals

$$
\begin{gathered}
=r^{n+1}(\cos n t+i \sin n t)(\cos t+i \sin t) \\
=r^{n+1}((\cos n t \cos t-\sin n t \sin t)+i(\sin n t \cos t+\cos n t \sin t)) \\
=r^{n+1}(\cos (n+1) t+i \sin (n+1) t)
\end{gathered}
$$

by the formulas for the cosine and sine of the sum of two angles.
Corollary 1.7.2 Let $z$ be a non zero complex number. Then there are always exactly $k k^{\text {th }}$ roots of $z$ in $\mathbb{C}$.

Proof: Let $z=x+i y$ and let $z=|z|(\cos t+i \sin t)$ be the polar form of the complex number. By De Moivre's theorem, a complex number $r(\cos \alpha+i \sin \alpha)$, is a $k^{\text {th }}$ root of $z$ if and only if $r^{k}(\cos k \alpha+i \sin k \alpha)=|z|(\cos t+i \sin t)$. This requires $r^{k}=|z|$ and so $r=$ $|z|^{1 / k}$ and also both $\cos (k \alpha)=\cos t$ and $\sin (k \alpha)=\sin t$. This can only happen if $k \alpha=$ $t+2 l \pi$ for $l$ an integer. Thus $\alpha=\frac{t+2 l \pi}{k}, l \in \mathbb{Z}$ and so the $k^{t h}$ roots of $z$ are of the form $|z|^{1 / k}\left(\cos \left(\frac{t+2 l \pi}{k}\right)+i \sin \left(\frac{t+2 l \pi}{k}\right)\right), l \in \mathbb{Z}$. Since the cosine and sine are periodic of period $2 \pi$, there are exactly $k$ distinct numbers which result from this formula.

Example 1.7.3 Find the three cube roots of $i$.
First note that $i=1\left(\cos \left(\frac{\pi}{2}\right)+i \sin \left(\frac{\pi}{2}\right)\right)$. Using the formula in the proof of the above corollary, the cube roots of $i$ are

$$
1\left(\cos \left(\frac{(\pi / 2)+2 l \pi}{3}\right)+i \sin \left(\frac{(\pi / 2)+2 l \pi}{3}\right)\right)
$$

where $l=0,1,2$. Therefore, the roots are

$$
\cos \left(\frac{\pi}{6}\right)+i \sin \left(\frac{\pi}{6}\right), \cos \left(\frac{5}{6} \pi\right)+i \sin \left(\frac{5}{6} \pi\right), \cos \left(\frac{3}{2} \pi\right)+i \sin \left(\frac{3}{2} \pi\right)
$$

Thus the cube roots of $i$ are $\frac{\sqrt{3}}{2}+i\left(\frac{1}{2}\right), \frac{-\sqrt{3}}{2}+i\left(\frac{1}{2}\right)$, and $-i$.
The ability to find $k^{t h}$ roots can also be used to factor some polynomials.
Example 1.7.4 Factor the polynomial $x^{3}-27$.

First find the cube roots of 27. By the above procedure using De Moivre's theorem, these cube roots are $3,3\left(\frac{-1}{2}+i \frac{\sqrt{3}}{2}\right)$, and $3\left(\frac{-1}{2}-i \frac{\sqrt{3}}{2}\right)$. Therefore, $x^{3}-27=$

$$
(x-3)\left(x-3\left(\frac{-1}{2}+i \frac{\sqrt{3}}{2}\right)\right)\left(x-3\left(\frac{-1}{2}-i \frac{\sqrt{3}}{2}\right)\right)
$$

Note also $\left(x-3\left(\frac{-1}{2}+i \frac{\sqrt{3}}{2}\right)\right)\left(x-3\left(\frac{-1}{2}-i \frac{\sqrt{3}}{2}\right)\right)=x^{2}+3 x+9$ and so

$$
x^{3}-27=(x-3)\left(x^{2}+3 x+9\right)
$$

where the quadratic polynomial $x^{2}+3 x+9$ cannot be factored without using complex numbers.

Note that even though the polynomial $x^{3}-27$ has all real coefficients, it has some complex zeros, $\frac{-1}{2}+i \frac{\sqrt{3}}{2}$ and $\frac{-1}{2}-i \frac{\sqrt{3}}{2}$. These zeros are complex conjugates of each other. It is always this way. You should show this is the case. To see how to do this, see Problems 17 and 18 below.

Another fact for your information is the fundamental theorem of algebra. This theorem says that any polynomial of degree at least 1 having any complex coefficients always has a root in $\mathbb{C}$. This is sometimes referred to by saying $\mathbb{C}$ is algebraically complete. Gauss is usually credited with giving a proof of this theorem in 1797 but many others worked on it and the first completely correct proof was due to Argand in 1806. For more on this theorem, you can google fundamental theorem of algebra and look at the interesting Wikipedia article on it. Proofs of this theorem usually involve the use of techniques from calculus even though it is really a result in algebra. A proof and plausibility explanation is given later.

### 1.8 The Quadratic Formula

The quadratic formula $x=\frac{-b \pm \sqrt{b^{2}-4 a c}}{2 a}$ gives the solutions $x$ to $a x^{2}+b x+c=0$ where $a, b, c$ are real numbers. It holds even if $b^{2}-4 a c<0$. This is easy to show from the above. There are exactly two square roots to this number $b^{2}-4 a c$ from the above methods using De Moivre's theorem. These roots are of the form

$$
\sqrt{4 a c-b^{2}}\left(\cos \left(\frac{\pi}{2}\right)+i \sin \left(\frac{\pi}{2}\right)\right)=i \sqrt{4 a c-b^{2}}
$$

and

$$
\sqrt{4 a c-b^{2}}\left(\cos \left(\frac{3 \pi}{2}\right)+i \sin \left(\frac{3 \pi}{2}\right)\right)=-i \sqrt{4 a c-b^{2}}
$$

Thus the solutions, according to the quadratic formula are still given correctly by the above formula.

Do these solutions predicted by the quadratic formula continue to solve the quadratic equation? Yes, they do. You only need to observe that when you square a square root of a complex number $z$, you recover $z$. Thus

$$
a\left(\frac{-b+\sqrt{b^{2}-4 a c}}{2 a}\right)^{2}+b\left(\frac{-b+\sqrt{b^{2}-4 a c}}{2 a}\right)+c
$$

$$
\begin{aligned}
& =a\left(\frac{1}{2 a^{2}} b^{2}-\frac{1}{a} c-\frac{1}{2 a^{2}} b \sqrt{b^{2}-4 a c}\right)+b\left(\frac{-b+\sqrt{b^{2}-4 a c}}{2 a}\right)+c \\
= & \left(-\frac{1}{2 a}\left(b \sqrt{b^{2}-4 a c}+2 a c-b^{2}\right)\right)+\frac{1}{2 a}\left(b \sqrt{b^{2}-4 a c}-b^{2}\right)+c=0
\end{aligned}
$$

Similar reasoning shows directly that $\frac{-b-\sqrt{b^{2}-4 a c}}{2 a}$ also solves the quadratic equation.
What if the coefficients of the quadratic equation are actually complex numbers? Does the formula hold even in this case? The answer is yes. This is a hint on how to do Problem 28 below, a special case of the fundamental theorem of algebra, and an ingredient in the proof of some versions of this theorem.

Example 1.8.1 Find the solutions to $x^{2}-2 i x-5=0$.
Formally, from the quadratic formula, these solutions are

$$
x=\frac{2 i \pm \sqrt{-4+20}}{2}=\frac{2 i \pm 4}{2}=i \pm 2 .
$$

Now you can check that these really do solve the equation. In general, this will be the case. See Problem 28 below.

### 1.9 The Complex Exponential

It was shown above that every complex number is of the form $r(\cos \theta+i \sin \theta)$ where $r \geq 0$. Laying aside the zero complex number, this shows that every non zero complex number is of the form $e^{\alpha}(\cos \beta+i \sin \beta)$. We write this in the form $e^{\alpha+i \beta}$. Having done so, does it follow that the expression preserves the most important property of the function $t \rightarrow e^{(\alpha+i \beta) t}$ for $t$ real, that

$$
\left(e^{(\alpha+i \beta) t}\right)^{\prime}=(\alpha+i \beta) e^{(\alpha+i \beta) t} ?
$$

By the definition just given which does not contradict the usual definition in case $\beta=0$ and the usual rules of differentiation in calculus,

$$
\begin{aligned}
\left(e^{(\alpha+i \beta) t}\right)^{\prime} & =\left(e^{\alpha t}(\cos (\beta t)+i \sin (\beta t))\right)^{\prime} \\
& =e^{\alpha t}[\alpha(\cos (\beta t)+i \sin (\beta t))+(-\beta \sin (\beta t)+i \beta \cos (\beta t))]
\end{aligned}
$$

Now consider the other side. From the definition it equals

$$
\begin{gathered}
(\alpha+i \beta)\left(e^{\alpha t}(\cos (\beta t)+i \sin (\beta t))\right)=e^{\alpha t}[(\alpha+i \beta)(\cos (\beta t)+i \sin (\beta t))] \\
=e^{\alpha t}[\alpha(\cos (\beta t)+i \sin (\beta t))+(-\beta \sin (\beta t)+i \beta \cos (\beta t))]
\end{gathered}
$$

which is the same thing. This is of fundamental importance in differential equations. It shows that there is no change in going from real to complex numbers for $\omega$ in the consideration of the problem $y^{\prime}=\omega y, y(0)=1$. The solution is always $e^{\omega t}$. The formula just discussed, that

$$
e^{\alpha}(\cos \beta+i \sin \beta)=e^{\alpha+i \beta}
$$

is Euler's formula.

### 1.10 The Fundamental Theorem of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in $\mathbb{C}$ has a zero in $\mathbb{C}$. If $\mathbb{C}$ is replaced by $\mathbb{R}$, this is not true because of the example, $x^{2}+1=0$. This theorem is a very remarkable result and notwithstanding its title, all the most straightforward proofs depend on either analysis or topology. It was first mostly proved by Gauss in 1797. The first complete proof was given by Argand in 1806. The proof given here follows Rudin [36]. See also Hardy [19] for a similar proof, more discussion and references. The shortest proof is found in the theory of complex analysis. First I will give an informal explanation of this theorem which shows why it is reasonable to believe in the fundamental theorem of algebra.

Theorem 1.10.1 Let $p(z)=a_{n} z^{n}+a_{n-1} z^{n-1}+\cdots+a_{1} z+a_{0}$ where each $a_{k}$ is a complex number and $a_{n} \neq 0, n \geq 1$. Then there exists $w \in \mathbb{C}$ such that $p(w)=0$.

To begin with, here is the informal explanation. Dividing by the leading coefficient $a_{n}$, there is no loss of generality in assuming that the polynomial is of the form

$$
p(z)=z^{n}+a_{n-1} z^{n-1}+\cdots+a_{1} z+a_{0}
$$

If $a_{0}=0$, there is nothing to prove because $p(0)=0$. Therefore, assume $a_{0} \neq 0$. From the polar form of a complex number $z$, it can be written as $|z|(\cos \theta+i \sin \theta)$. Thus, by DeMoivre's theorem,

$$
z^{n}=|z|^{n}(\cos (n \theta)+i \sin (n \theta))
$$

It follows that $z^{n}$ is some point on the circle of radius $|z|^{n}$
Denote by $C_{r}$ the circle of radius $r$ in the complex plane which is centered at 0 . Then if $r$ is sufficiently large and $|z|=r$, the term $z^{n}$ is far larger than the rest of the polynomial. It is on the circle of radius $|z|^{n}$ while the other terms are on circles of fixed multiples of $|z|^{k}$ for $k \leq n-1$. Thus, for $r$ large enough, $A_{r}=\left\{p(z): z \in C_{r}\right\}$ describes a closed curve which misses the inside of some circle having 0 as its center. It won't be as simple as suggested in the following picture, but it will be a closed curve thanks to De Moivre's theorem and the observation that the cosine and sine are periodic. Now shrink $r$. Eventually, for $r$ small enough, the non constant terms are negligible and so $A_{r}$ is a curve which is contained in some circle centered at $a_{0}$ which has 0 on the outside.

$r$ small


Thus it is reasonable to believe that for some $r$ during this shrinking process, the set $A_{r}$ must hit 0 . It follows that $p(z)=0$ for some $z$.

For example, consider the polynomial $x^{3}+x+$ $1+i$. It has no real zeros. However, you could let $z=r(\cos t+i \sin t)$ and insert this into the polynomial. Thus you would want to find a point where

$$
(r(\cos t+i \sin t))^{3}+r(\cos t+i \sin t)+1+i=0+0 i
$$

Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

$$
r^{3} \cos ^{3} t-3 r^{3} \cos t \sin ^{2} t+r \cos t+1+i\left(3 r^{3} \cos ^{2} t \sin t-r^{3} \sin ^{3} t+r \sin t+1\right)
$$

Thus you need to have both the real and imaginary parts equal to 0 . In other words, you need to have $(0,0)=$

$$
\left(r^{3} \cos ^{3} t-3 r^{3} \cos t \sin ^{2} t+r \cos t+1,3 r^{3} \cos ^{2} t \sin t-r^{3} \sin ^{3} t+r \sin t+1\right)
$$

for some value of $r$ and $t$. First here is a graph of this parametric function of $t$ for $t \in[0,2 \pi]$ on the left, when $r=4$. Note how the graph misses the origin $0+i 0$. In fact, the closed curve is in the exterior of a circle which has the point $0+i 0$ on its inside.




Next is the graph when $r=.5$. Note how the closed curve is included in a circle which has $0+i 0$ on its outside. As you shrink $r$ you get closed curves. At first, these closed curves enclose $0+i 0$ and later, they exclude $0+i 0$. Thus one of them should pass through this point. In fact, consider the curve which results when $r=1.386$ which is the graph on the right. Note how for this value of $r$ the curve passes through the point $0+i 0$. Thus for some $t, 1.386(\cos t+i \sin t)$ is a solution of the equation $p(z)=0$ or very close to one.

Now here is a short rigorous proof for those who have studied analysis. The needed analysis will be presented later in the book. You need the extreme value theorem for example.

Proof: Suppose the nonconstant polynomial $p(z)=a_{0}+a_{1} z+\cdots+a_{n} z^{n}, a_{n} \neq 0$, has no zero in $\mathbb{C}$. Since $\lim _{|z| \rightarrow \infty}|p(z)|=\infty$, there is a $z_{0}$ with

$$
\left|p\left(z_{0}\right)\right|=\min _{z \in \mathbb{C}}|p(z)|>0
$$

Then let $q(z)=\frac{p\left(z+z_{0}\right)}{p\left(z_{0}\right)}$. This is also a polynomial which has no zeros and the minimum of $|q(z)|$ is 1 and occurs at $z=0$. Since $q(0)=1$, it follows $q(z)=1+a_{k} z^{k}+r(z)$ where $r(z)$ is of the form

$$
r(z)=a_{m} z^{m}+a_{m+1} z^{m+1}+\ldots+a_{n} z^{n} \text { for } m>k
$$

Choose a sequence, $z_{n} \rightarrow 0$, such that $a_{k} z_{n}^{k}<0$. For example, let $-a_{k} z_{n}^{k}=(1 / n)$ so $z_{n}=$ $\left(-a_{k}\right)^{1 / k}\left(\frac{1}{n}\right)^{1 / k}$ and Then

$$
\begin{aligned}
\left|q\left(z_{n}\right)\right| & =\left|1+a_{k} z^{k}+r(z)\right| \leq 1-1 / n+\left|r\left(z_{n}\right)\right| \\
& \leq 1-\frac{1}{n}+\frac{1}{n} \sum_{j=m}^{n}\left|a_{j}\right|\left|a_{k}\right|^{1 / k}\left(\frac{1}{n}\right)^{(j-k) / k}<1
\end{aligned}
$$

for all $n$ large enough because the sum is smaller than 1 whenever $n$ is large enough, showing $\left|q\left(z_{n}\right)\right|<1$ whenever $n$ is large enough. This is a contradiction to $|q(z)| \geq 1$.

### 1.11 Ordered Fields

To do linear algebra, you need a field which is something satisfying the axioms listed in Theorem 1.5.1. This is generally all that is needed to do linear algebra but for the sake of completeness, the concept of an ordered field is considered here. The real numbers also have an order defined on them. This order may be defined by reference to the positive real
numbers, those to the right of 0 on the number line, denoted by $\mathbb{R}^{+}$. More generally, for a field, one could consider an order if there is such a "positive cone" called the positive numbers such that the following axioms hold.

Axiom 1.11.1 The sum of two positive real numbers is positive.
Axiom 1.11.2 The product of two positive real numbers is positive.
Axiom 1.11.3 For a given real number $x$ one and only one of the following alternatives holds. Either $x$ is positive, $x=0$, or $-x$ is positive.

An example of this is the field of rational numbers.
Definition 1.11.4 $x<y$ exactly when $y+(-x) \equiv y-x \in \mathbb{R}^{+}$. In the usual way, $x<y$ is the same as $y>x$ and $x \leq y$ means either $x<y$ or $x=y$. The symbol $\geq$ is defined similarly.

Theorem 1.11.5 The following hold for the order defined as above.

1. If $x<y$ and $y<z$ then $x<z$ (Transitive law).
2. If $x<y$ then $x+z<y+z$ (addition to an inequality).
3. If $x \leq 0$ and $y \leq 0$, then $x y \geq 0$.
4. If $x>0$ then $x^{-1}>0$.
5. If $x<0$ then $x^{-1}<0$.
6. If $x<y$ then $x z<y z$ if $z>0$, (multiplication of an inequality).
7. If $x<y$ and $z<0$, then $x z>z y$ (multiplication of an inequality).
8. Each of the above holds with $>$ and $<$ replaced by $\geq$ and $\leq$ respectively except for 4 and 5 in which we must also stipulate that $x \neq 0$.
9. For any $x$ and $y$, exactly one of the following must hold. Either $x=y, x<y$, or $x>y$ (trichotomy).

Proof: First consider 1, the transitive law. Suppose $x<y$ and $y<z$. Why is $x<z$ ? In other words, why is $z-x \in \mathbb{R}^{+}$? It is because $z-x=(z-y)+(y-x)$ and both $z-y, y-x \in$ $\mathbb{R}^{+}$. Thus by 1.11.1 above, $z-x \in \mathbb{R}^{+}$and so $z>x$.

Next consider 2, addition to an inequality. If $x<y$ why is $x+z<y+z ?$ it is because

$$
\begin{aligned}
(y+z)+-(x+z) & =(y+z)+(-1)(x+z) \\
& =y+(-1) x+z+(-1) z \\
& =y-x \in \mathbb{R}^{+}
\end{aligned}
$$

Next consider 3. If $x \leq 0$ and $y \leq 0$, why is $x y \geq 0$ ? First note there is nothing to show if either $x$ or $y$ equal 0 so assume this is not the case. By 1.11.3-x>0 and $-y>0$. Therefore, by 1.11.2 and what was proved about $-x=(-1) x,(-x)(-y)=(-1)^{2} x y \in \mathbb{R}^{+}$. Is $(-1)^{2}=1$ ? If so the claim is proved. But $-(-1)=(-1)^{2}$ and $-(-1)=1$ because $-1+1=0$.

Next consider 4. If $x>0$ why is $x^{-1}>0$ ? By 1.11.3 either $x^{-1}=0$ or $-x^{-1} \in \mathbb{R}^{+}$. It can't happen that $x^{-1}=0$ because then you would have to have $1=0 x$ and as was shown earlier, $0 x=0$. Therefore, consider the possibility that $-x^{-1} \in \mathbb{R}^{+}$. This can't work either because then you would have

$$
(-1) x^{-1} x=(-1)(1)=-1
$$

and it would follow from 1.11 .2 that $-1 \in \mathbb{R}^{+}$. But this is impossible because if $x \in \mathbb{R}^{+}$, then $(-1) x=-x \in \mathbb{R}^{+}$and contradicts 1.11 .3 which states that either $-x$ or $x$ is in $\mathbb{R}^{+}$but not both.

Next consider 5. If $x<0$, why is $x^{-1}<0$ ? As before, $x^{-1} \neq 0$. If $x^{-1}>0$, then as before, $-x\left(x^{-1}\right)=-1 \in \mathbb{R}^{+}$which was just shown not to occur.

Next consider 6. If $x<y$ why is $x z<y z$ if $z>0$ ? This follows because $y z-x z=$ $z(y-x) \in \mathbb{R}^{+}$since both $z$ and $y-x \in \mathbb{R}^{+}$.

Next consider 7. If $x<y$ and $z<0$, why is $x z>z y$ ? This follows because $z x-z y=$ $z(x-y) \in \mathbb{R}^{+}$by what was proved in 3 .

The last two claims are obvious and left for you.

### 1.12 Division of Numbers

First of all, recall the Archimedean property of the real numbers which says that if $x$ is any real number, and if $a>0$ then there exists a positive integer $n$ such that $n a>x$. Geometrically, it is essentially the following: For any $a>0$, the succession of disjoint intervals $[0, a),[a, 2 a),[2 a, 3 a), \cdots$ includes all nonnegative real numbers. Here is a picture of some of these intervals.


Then the version of the Euclidean algorithm presented here says that, for an arbitrary nonnegative real number $b$, it is in exactly one interval $[p a,(p+1) a)$ where $p$ is some nonnegative integer. This seems obvious from the picture, but here is a proof.

Theorem 1.12.1 Suppose $0<a$ and let $b \geq 0$. Then there exists a unique integer $p$ and real number $r$ such that $0 \leq r<a$ and $b=p a+r$.

Proof: Let $S \equiv\{n \in \mathbb{N}: a n>b\}$. By the Archimedean property this set is nonempty. Let $p+1$ be the smallest element of $S$. Then $p a \leq b$ because $p+1$ is the smallest in $S$. Therefore, $r \equiv b-p a \geq 0$. If $r \geq a$ then $b-p a \geq a$ and so $b \geq(p+1) a$ contradicting $p+1 \in S$. Therefore, $r<a$ as desired.

To verify uniqueness of $p$ and $r$, suppose $p_{i}$ and $r_{i}, i=1,2$, both work and $r_{2}>r_{1}$. Then a little algebra shows $p_{1}-p_{2}=\frac{r_{2}-r_{1}}{a} \in(0,1)$.Thus $p_{1}-p_{2}$ is an integer between 0 and 1 , and there are no such integers. The case that $r_{1}>r_{2}$ cannot occur either by similar reasoning. Thus $r_{1}=r_{2}$ and it follows that $p_{1}=p_{2}$.

Note that if $a, b$ are integers, then so is $r$.
Corollary 1.12.2 The same conclusion is reached if $b<0$.
Proof: In this case, $S \equiv\{n \in \mathbb{N}: a n>-b\}$. Let $p+1$ be the smallest element of $S$. Then $p a \leq-b<(p+1) a$ and so $(-p) a \geq b>-(p+1) a$. Let $r \equiv b+(p+1)$. Then
$b=-(p+1) a+r$ and $a>r \geq 0$. As to uniqueness, say $r_{i}$ works and $r_{1}>r_{2}$. Then you would have

$$
b=p_{1} a+r_{1}, b=p_{2} a+r_{2}
$$

and $p_{2}-p_{1}=\frac{r_{1}-r_{2}}{a} \in(0,1)$ which is impossible because $p_{2}-p_{1}$ is an integer. Hence $r_{1}=r_{2}$ and so also $p_{1}=p_{2}$.

Corollary 1.12.3 Suppose $a, b \neq 0$, then there exists $r$ such that $|r|<|a|$ and for some $p$ an integer, $b=a p+r$.

Proof: This is done in the above except for the case where $a<0$. So suppose this is the case. Then $b=p(-a)+r$ where $r$ is positive and $0 \leq r<-a=|a|$. Thus $b=(-p) a+r$ such that $0 \leq|r|<|a|$.

This theorem is called the Euclidean algorithm when $a$ and $b$ are integers and this is the case of most interest here. Note that if $a, b$ are integers, then so is $r$. Note that

$$
7=2 \times 3+1, \quad 7=3 \times 3-2,|1|<3,|-2|<3
$$

so in this last corollary, the $p$ and $r$ are not unique.
The following definition describes what is meant by a prime number and also what is meant by the word "divides".

Definition 1.12.4 The number, a divides the number, $b$ if in Theorem 1.12.1, $r=0$. That is there is zero remainder. The notation for this is $a \mid b$, read a divides $b$ and $a$ is called a factor of $b$. A prime number is a number at least 2 which has the property that the only numbers which divide it are itself and 1. The greatest common divisor of two positive integers, $m, n$ is that number, $p$ which has the property that $p$ divides both $m$ and $n$ and also if $q$ divides both $m$ and $n$, then $q$ divides $p$. Two integers are relatively prime if their greatest common divisor is one. The greatest common divisor of $m$ and $n$ is denoted as $(m, n)$.

There is a phenomenal and amazing theorem which relates the greatest common divisor to the smallest number in a certain set. Suppose $m, n$ are two positive integers. Then if $x, y$ are integers, so is $x m+y n$. Consider all integers which are of this form. Some are positive such as $1 m+1 n$ and some are not. The set $S$ in the following theorem consists of exactly those integers of this form which are positive. Then the greatest common divisor of $m$ and $n$ will be the smallest number in $S$. This is what the following theorem says.

Theorem 1.12.5 Let m,n be two positive integers and define

$$
S \equiv\{x m+y n \in \mathbb{N}: x, y \in \mathbb{Z}\}
$$

Then the smallest number in $S$ is the greatest common divisor, denoted by $(m, n)$.
Proof: First note that both $m$ and $n$ are in $S$ so it is a nonempty set of positive integers. By well ordering, there is a smallest element of $S$, called $p=x_{0} m+y_{0} n$. Either $p$ divides $m$ or it does not. If $p$ does not divide $m$, then by Theorem 1.12.1, $m=p q+r$ where $0<r<p$. Thus $m=\left(x_{0} m+y_{0} n\right) q+r$ and so, solving for $r$,

$$
r=m\left(1-x_{0}\right)+\left(-y_{0} q\right) n \in S
$$

However, this is a contradiction because $p$ was the smallest element of $S$. Thus $p \mid m$. Similarly $p \mid n$.

Now suppose $q$ divides both $m$ and $n$. Then $m=q x$ and $n=q y$ for integers, $x$ and $y$. Therefore,

$$
p=m x_{0}+n y_{0}=x_{0} q x+y_{0} q y=q\left(x_{0} x+y_{0} y\right)
$$

showing $q \mid p$. Therefore, $p=(m, n)$.
This amazing theorem will now be used to prove a fundamental property of prime numbers which leads to the fundamental theorem of arithmetic, the major theorem which says every integer can be factored as a product of primes.

Theorem 1.12.6 If $p$ is a prime and $p \mid a b$ then either $p \mid$ a or $p \mid b$.
Proof: Suppose $p$ does not divide $a$. Then since $p$ is prime, the only factors of $p$ are 1 and $p$ so follows $(p, a)=1$ and therefore, there exists integers, $x$ and $y$ such that $1=a x+y p$. Multiplying this equation by $b$ yields $b=a b x+y b p$.Since $p \mid a b, a b=p z$ for some integer z. Therefore,

$$
b=a b x+y b p=p z x+y b p=p(x z+y b)
$$

and this shows $p$ divides $b$.
The following is the fundamental theorem of arithmetic.
Theorem 1.12.7 (Fundamental theorem of arithmetic) Let $a \in \mathbb{N} \backslash\{1\}$. Then $a=\prod_{i=1}^{n} p_{i}$ where $p_{i}$ are all prime numbers. Furthermore, this prime factorization is unique except for the order of the factors.

Proof: If $a$ equals a prime number, the prime factorization clearly exists. In particular the prime factorization exists for the prime number 2. Assume this theorem is true for all $a \leq n-1$. If $n$ is a prime, then it has a prime factorization. On the other hand, if $n$ is not a prime, then there exist two integers $k$ and $m$ such that $n=k m$ where each of $k$ and $m$ are less than $n$. Therefore, each of these is no larger than $n-1$ and consequently, each has a prime factorization. Thus so does $n$. It remains to argue the prime factorization is unique except for order of the factors.

Suppose $\prod_{i=1}^{n} p_{i}=\prod_{j=1}^{m} q_{j}$ where the $p_{i}$ and $q_{j}$ are all prime, there is no way to reorder the $q_{k}$ such that $m=n$ and $p_{i}=q_{i}$ for all $i$, and $n+m$ is the smallest positive integer such that this happens. Then by Theorem 1.12.6, $p_{1} \mid q_{j}$ for some $j$. Since these are prime numbers this requires $p_{1}=q_{j}$. Reordering if necessary it can be assumed that $q_{j}=q_{1}$. Then dividing both sides by $p_{1}=q_{1}, \prod_{i=1}^{n-1} p_{i+1}=\prod_{j=1}^{m-1} q_{j+1}$. Since $n+m$ was as small as possible for the theorem to fail, it follows that $n-1=m-1$ and the prime numbers, $q_{2}, \cdots, q_{m}$ can be reordered such that $p_{k}=q_{k}$ for all $k=2, \cdots, n$. Hence $p_{i}=q_{i}$ for all $i$ because it was already argued that $p_{1}=q_{1}$, and this results in a contradiction.

### 1.13 Polynomials

It will be very important to be able to work with polynomials in certain parts of linear algebra to be presented later. Polynomials are a lot like integers. The notion of division is important for polynomials in the same way that it is for integers.

Definition 1.13.1 A polynomial is an expression of the form $a_{n} \lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+$ $a_{0}, a_{n} \neq 0$ where the $a_{i}$ come from a field of scalars. Two polynomials are equal means that the coefficients match for each power of $\lambda$. The degree of a polynomial is the largest power of $\lambda$. Thus the degree of the above polynomial is $n$. Addition of polynomials is defined in
the usual way as is multiplication of two polynomials. The leading term in the above polynomial is $a_{n} \lambda^{n}$. The coefficient of the leading term is called the leading coefficient. It is called a monic polynomial if $a_{n}=1$.

Note that the degree of the zero polynomial is not defined in the above. Multiplication of polynomials has an important property.

Lemma 1.13.2 If $f(\boldsymbol{\lambda}) g(\boldsymbol{\lambda})=0$, then either $f(\boldsymbol{\lambda})=0$ or $g(\boldsymbol{\lambda})=0$. That is, there are no nonzero divisors of 0 . If $f(\lambda)$ is a monic polynomial and $f(\lambda) g(\lambda)=0$, then $g(\lambda)=0$ and if $f(\boldsymbol{\lambda}) g_{1}(\boldsymbol{\lambda})=f(\boldsymbol{\lambda}) g_{2}(\boldsymbol{\lambda})$, then $g_{1}(\boldsymbol{\lambda})=g_{2}(\boldsymbol{\lambda})$.

Proof: Let $f(\boldsymbol{\lambda})$ have degree $n$ and $g(\boldsymbol{\lambda})$ degree $m$. If $m+n=0$, it is easy to see that the conclusion holds. Suppose the conclusion holds for $m+n \leq M$ and suppose $m+n=M+1$. Then $f(\lambda) g(\lambda)=$

$$
\begin{align*}
& \left(a_{0}+a_{1} \lambda+\cdots+a_{n-1} \lambda^{n-1}+a_{n} \lambda^{n}\right)\left(b_{0}+b_{1} \lambda+\cdots+b_{m-1} \lambda^{m-1}+b_{m} \lambda^{m}\right) \\
= & (\overbrace{a(\lambda)+a_{n} \lambda^{n}}^{f(\lambda)})(\overbrace{b(\lambda)+b_{m} \lambda^{m}}^{g(\lambda)})  \tag{1.1}\\
= & a(\lambda) b(\lambda)+b_{m} \lambda^{m} a(\lambda)+a_{n} \lambda^{n} b(\lambda)+a_{n} b_{m} \lambda^{n+m}
\end{align*}
$$

Either $a_{n}=0$ or $b_{m}=0$. Suppose $b_{m}=0$. It is similar if $a_{n}=0$. Then from 1.1,

$$
\left(a(\lambda)+a_{n} \lambda^{n}\right) b(\lambda)=0 .
$$

By induction, one of these polynomials in the product is 0 . If $b(\lambda) \neq 0$, then $a(\lambda)+a_{n} \lambda^{n}=$ 0 so $a_{n}=0$ and $a(\lambda)=0$ so $f(\lambda) \equiv a(\lambda)+a_{n} \lambda^{n}=0$. If $b(\lambda)=0$, then, since $b_{m}=0$, $g(\lambda) \equiv b(\lambda)+b_{m} \lambda^{m}=0$. The remaining claims are clear from this because $f(\lambda) \neq 0$ if $f(\lambda)$ is monic.

Lemma 1.13.3 Let $f(\lambda)$ and $g(\lambda) \neq 0$ be polynomials. Then there exist polynomials, $q(\lambda)$ and $r(\lambda)$ such that

$$
f(\lambda)=q(\lambda) g(\lambda)+r(\lambda)
$$

where the degree of $r(\lambda)$ is less than the degree of $g(\lambda)$ or $r(\lambda)=0$. These polynomials $q(\lambda)$ and $r(\lambda)$ are unique.

Proof: Suppose that $f(\boldsymbol{\lambda})-q(\boldsymbol{\lambda}) g(\boldsymbol{\lambda})$ is never equal to 0 for any $q(\boldsymbol{\lambda})$. If it is, then the conclusion follows. Now suppose

$$
\begin{equation*}
r(\boldsymbol{\lambda})=f(\boldsymbol{\lambda})-q(\boldsymbol{\lambda}) g(\boldsymbol{\lambda}) \tag{*}
\end{equation*}
$$

where the degree of $r(\lambda)$ is as small as possible. Let it be $m$. Suppose $m \geq n$ where $n$ is the degree of $g(\lambda)$. Say $r(\lambda)=b \lambda^{m}+a(\lambda)$ where $a(\lambda)$ is 0 or has degree less than $m$ while $g(\lambda)=\hat{b} \lambda^{n}+\hat{a}(\lambda)$ where $\hat{a}(\lambda)$ is 0 or has degree less than $n$. Then

$$
r(\lambda)-\frac{b}{\hat{b}} \lambda^{m-n} g(\lambda)=b \lambda^{m} \stackrel{r(\lambda)}{+a}(\lambda)-\left(b \lambda^{m}+\frac{b}{\hat{b}} \lambda^{m-n} \hat{a}(\lambda)\right)=a(\lambda)-\tilde{a}(\lambda),
$$

a polynomial having degree less than $m$. Therefore from the above,

$$
a(\lambda)-\tilde{a}(\lambda)=\overbrace{(f(\lambda)-q(\lambda) g(\lambda))}^{=r(\lambda)}-\frac{b}{\hat{b}} \lambda^{m-n} g(\lambda)=f(\lambda)-\hat{q}(\lambda) g(\lambda)
$$

which is of the same form as $*$ having smaller degree. However, $m$ was as small as possible. Hence $m<n$ after all.

As to uniqueness, if you have $r(\lambda), \hat{r}(\boldsymbol{\lambda}), q(\boldsymbol{\lambda}), \hat{q}(\boldsymbol{\lambda})$ which work, then you would have

$$
(\hat{q}(\lambda)-q(\lambda)) g(\lambda)=r(\lambda)-\hat{r}(\lambda)
$$

Now if the polynomial on the right is not zero, then neither is the one on the left. Hence this would involve two polynomials which are equal although their degrees are different. This is impossible. Hence $r(\lambda)=\hat{r}(\boldsymbol{\lambda})$ and so, the above lemma shows $\hat{q}(\boldsymbol{\lambda})=q(\boldsymbol{\lambda})$.

Definition 1.13.4 A polynomial $f$ is said to divide a polynomial $g$ if $g(\boldsymbol{\lambda})=f(\boldsymbol{\lambda}) r(\boldsymbol{\lambda})$ for some polynomial $r(\lambda)$. Let $\left\{\phi_{i}(\lambda)\right\}$ be a finite set of polynomials. The greatest common divisor will be the monic polynomial $q(\lambda)$ such that $q(\lambda)$ divides each $\phi_{i}(\lambda)$ and if $p(\lambda)$ divides each $\phi_{i}(\lambda)$, then $p(\lambda)$ divides $q(\lambda)$. The finite set of polynomials $\left\{\phi_{i}\right\}$ is said to be relatively prime if their greatest common divisor is 1. A polynomial $f(\lambda)$ is irreducible if there is no polynomial with coefficients in $\mathbb{F}$ which divides it except nonzero scalar multiples of $f(\boldsymbol{\lambda})$ and constants. In other words, it is not possible to write $f(\lambda)=a(\lambda) b(\lambda)$ where each of $a(\lambda), b(\lambda)$ have degree less than the degree of $f(\lambda)$ unless one of $a(\lambda), b(\lambda)$ is $a$ constant.

Proposition 1.13.5 The greatest common divisor is unique.
Proof: Suppose both $q(\boldsymbol{\lambda})$ and $q^{\prime}(\boldsymbol{\lambda})$ work. Then $q(\boldsymbol{\lambda})$ divides $q^{\prime}(\boldsymbol{\lambda})$ and the other way around and so

$$
q^{\prime}(\lambda)=q(\lambda) l(\lambda), q(\lambda)=l^{\prime}(\lambda) q^{\prime}(\lambda)
$$

Therefore, the two must have the same degree. Hence $l^{\prime}(\lambda), l(\lambda)$ are both constants. However, this constant must be 1 because both $q(\lambda)$ and $q^{\prime}(\lambda)$ are monic.

Theorem 1.13.6 Let $\left\{\phi_{i}(\lambda)\right\}$ be polynomials, not all of which are zero polynomials. Then there exists a greatest common divisor and it equals the monic polynomial $\psi(\lambda)$ of smallest degree such that there exist polynomials $r_{i}(\lambda)$ satisfying

$$
\psi(\lambda)=\sum_{i=1}^{p} r_{i}(\lambda) \phi_{i}(\lambda) .
$$

Proof: Let $S$ denote the set of monic polynomials of the form $\sum_{i=1}^{p} r_{i}(\lambda) \phi_{i}(\lambda)$. where $r_{i}(\lambda)$ is a polynomial. Then $S \neq \emptyset$ because some $\phi_{i}(\lambda) \neq 0$. Then let the $r_{i}$ be chosen such that the degree of the expression $\sum_{i=1}^{p} r_{i}(\lambda) \phi_{i}(\lambda)$ is as small as possible. Letting $\psi(\lambda)$ equal this sum, it remains to verify it is the greatest common divisor. First, does it divide each $\phi_{i}(\lambda)$ ? Suppose it fails to divide $\phi_{1}(\lambda)$. Then by Lemma 1.13.3

$$
\phi_{1}(\lambda)=\psi(\lambda) l(\lambda)+r(\lambda)
$$

where degree of $r(\lambda)$ is less than that of $\psi(\lambda)$. Then dividing $r(\lambda)$ by the leading coefficient if necessary and denoting the result by $\psi_{1}(\lambda)$, it follows the degree of $\psi_{1}(\lambda)$ is less than the degree of $\psi(\lambda)$ and $\psi_{1}(\lambda)$ equals for some $a \in \mathbb{F}$

$$
\begin{gathered}
\psi_{1}(\lambda)=\left(\phi_{1}(\lambda)-\psi(\lambda) l(\lambda)\right) a \\
=\left(\phi_{1}(\lambda)-\sum_{i=1}^{p} r_{i}(\lambda) \phi_{i}(\lambda) l(\lambda)\right) a \\
=\left(\left(1-r_{1}(\lambda)\right) \phi_{1}(\lambda)+\sum_{i=2}^{p}\left(-r_{i}(\lambda) l(\lambda)\right) \phi_{i}(\lambda)\right) a
\end{gathered}
$$

This is one of the polynomials in $S$. Therefore, $\psi(\lambda)$ does not have the smallest degree after all because the degree of $\psi_{1}(\lambda)$ is smaller. This is a contradiction. Therefore, $\psi(\lambda)$ divides $\phi_{1}(\lambda)$. Similarly it divides all the other $\phi_{i}(\lambda)$.

If $p(\lambda)$ divides all the $\phi_{i}(\lambda)$, then it divides $\psi(\lambda)$ because of the formula for $\psi(\lambda)$ which equals $\sum_{i=1}^{p} r_{i}(\lambda) \phi_{i}(\lambda)$. Thus $\psi(\lambda)$ satisfies the condition to be the greatest common divisor. This shows the greatest common divisor exists and equals the above description of it.

Lemma 1.13.7 Suppose $\phi(\lambda)$ and $\psi(\lambda)$ are monic polynomials which are irreducible and not equal. Then they are relatively prime.

Proof: Suppose $\eta(\lambda)$ is a nonconstant polynomial. If $\eta(\lambda)$ divides $\phi(\lambda)$, then since $\phi(\lambda)$ is irreducible, $\phi(\lambda)=\eta(\lambda) \tilde{a}$ for some constant $\tilde{a}$. Thus $\eta(\lambda)$ equals $a \phi(\lambda)$ for some $a \in \mathbb{F}$. If $\eta(\lambda)$ divides $\psi(\lambda)$ then it must be of the form $b \psi(\lambda)$ for some $b \in \mathbb{F}$ and so it follows

$$
\begin{gathered}
\eta(\lambda)=a \phi(\lambda)=b \psi(\lambda), \\
\psi(\lambda)=\frac{a}{b} \phi(\lambda)
\end{gathered}
$$

but both $\psi(\lambda)$ and $\phi(\lambda)$ are monic polynomials which implies $a=b$ and so $\psi(\lambda)=\phi(\lambda)$. This is assumed not to happen. It follows the only polynomials which divide both $\psi(\lambda)$ and $\phi(\lambda)$ are constants and so the two polynomials are relatively prime. Thus a polynomial which divides them both must be a constant, and if it is monic, then it must be 1 . Thus 1 is the greatest common divisor.

Lemma 1.13.8 Let $\psi(\lambda)$ be an irreducible monic polynomial not equal to 1 which divides

$$
\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}, k_{i} \text { a positive integer }
$$

where each $\phi_{i}(\lambda)$ is an irreducible monic polynomial not equal to 1 . Then $\psi(\lambda)$ equals some $\phi_{i}(\lambda)$.

Proof : Say $\psi(\lambda) l(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}$. Suppose $\psi(\lambda) \neq \phi_{i}(\lambda)$ for all $i$. Then by Lemma 1.13.7, there exist polynomials $m_{i}(\lambda), n_{i}(\lambda)$ such that

$$
\begin{aligned}
1 & =\psi(\lambda) m_{i}(\lambda)+\phi_{i}(\lambda) n_{i}(\lambda) \\
\phi_{i}(\lambda) n_{i}(\lambda) & =1-\psi(\lambda) m_{i}(\lambda)
\end{aligned}
$$

Hence, $\left(\phi_{i}(\lambda) n_{i}(\lambda)\right)^{k_{i}}=\left(1-\psi(\lambda) m_{i}(\lambda)\right)^{k_{i}}$ and so letting $n(\lambda)=\prod_{i} n_{i}(\lambda)^{k_{i}}$,

$$
\begin{gathered}
n(\lambda) l(\lambda) \psi(\lambda)=\prod_{i=1}^{p}\left(n_{i}(\lambda) \phi_{i}(\lambda)\right)^{k_{i}}=\prod_{i=1}^{p}\left(1-\psi(\lambda) m_{i}(\lambda)\right)^{k_{i}} \\
=1+g(\lambda) \psi(\lambda)
\end{gathered}
$$

for a suitable polynomial $g(\lambda)$. You just separate out the term $1^{k_{i}}=1$ in that product and then all terms that are left have a $\psi(\lambda)$ as a factor. Hence

$$
(n(\lambda) l(\lambda)-g(\lambda)) \psi(\lambda)=1
$$

which is impossible because $\psi(\lambda)$ is not equal to 1 .
Of course, since coefficients are in a field, you can drop the stipulation that the polynomials are monic and replace the conclusion with: $\psi(\lambda)$ is a multiple of some $\phi_{i}(\lambda)$.

The following is the analog of the fundamental theorem of arithmetic for polynomials.
Theorem 1.13.9 Let $f(\lambda)$ be a nonconstant polynomial with coefficients in $\mathbb{F}$. Then there is some $a \in \mathbb{F}$ such that $f(\lambda)=a \prod_{i=1}^{n} \phi_{i}(\lambda)$ where $\phi_{i}(\lambda)$ is an irreducible nonconstant monic polynomial and repeats are allowed. Furthermore, this factorization is unique in the sense that any two of these factorizations have the same nonconstant factors in the product, possibly in different order and the same constant a. Every subset of $\left\{\phi_{i}(\lambda), i=1, \ldots, n\right\}$ having at least two elements is relatively prime.

Proof: That such a factorization exists is obvious. If $f(\lambda)$ is irreducible, you are done. Factor out the leading coefficient. If not, then $f(\lambda)=a \phi_{1}(\lambda) \phi_{2}(\lambda)$ where these are monic polynomials. Continue doing this with the $\phi_{i}$ and eventually arrive at a factorization of the desired form.

It remains to argue the factorization is unique except for order of the factors. Suppose

$$
a \prod_{i=1}^{n} \phi_{i}(\lambda)=b \prod_{i=1}^{m} \psi_{i}(\lambda)
$$

where the $\phi_{i}(\lambda)$ and the $\psi_{i}(\lambda)$ are all irreducible monic nonconstant polynomials and $a, b \in \mathbb{F}$. If $n>m$, then by Lemma 1.13.8, each $\psi_{i}(\lambda)$ equals one of the $\phi_{j}(\lambda)$. By the above cancellation lemma, Lemma 1.13.2, you can cancel all these $\psi_{i}(\lambda)$ with appropriate $\phi_{j}(\lambda)$ and obtain a contradiction because the resulting polynomials on either side would have different degrees. Similarly, it cannot happen that $n<m$. It follows $n=m$ and the two products consist of the same polynomials. Then it follows $a=b$. If you have such a subset of the $\phi_{i}(\lambda)$, the monic polynomial of smallest degree which divides them all must be 1 because none of the $\phi_{i}(\lambda)$ divide any other since they are all irreducible.

The following corollary will be well used. This corollary seems rather believable but does require a proof.

Corollary 1.13.10 Let $q(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}$ where the $k_{i}$ are positive integers and the $\phi_{i}(\lambda)$ are irreducible distinct monic polynomials. Suppose also that $p(\lambda)$ is a monic polynomial which divides $q(\lambda)$. Then

$$
p(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{r_{i}}
$$

where $r_{i}$ is a nonnegative integer no larger than $k_{i}$.

Proof: Using Theorem 1.13.9, let $p(\lambda)=b \prod_{i=1}^{s} \psi_{i}(\lambda)^{r_{i}}$ where the $\psi_{i}(\lambda)$ are each irreducible and monic and $b \in \mathbb{F}$. Since $p(\lambda)$ is monic, $b=1$. Then there exists a polynomial $g(\lambda)$ such that

$$
p(\lambda) g(\lambda)=g(\lambda) \prod_{i=1}^{s} \psi_{i}(\lambda)^{r_{i}}=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}
$$

Hence $g(\lambda)$ must be monic. Therefore,

$$
p(\boldsymbol{\lambda}) g(\boldsymbol{\lambda})=\overbrace{\prod_{i=1}^{s} \psi_{i}(\boldsymbol{\lambda})^{r_{i}}}^{p(\lambda)} \prod_{j=1}^{l} \eta_{j}(\boldsymbol{\lambda})=\prod_{i=1}^{p} \phi_{i}(\boldsymbol{\lambda})^{k_{i}}
$$

for $\eta_{j}$ monic and irreducible. By uniqueness, each $\psi_{i}$ equals one of the $\phi_{j}(\lambda)$ and the same holding true of the $\eta_{i}(\lambda)$. Therefore, $p(\lambda)$ is of the desired form because you can cancel the $\eta_{j}(\lambda)$ from both sides.

### 1.14 The Method of Partial Fractions

A very useful method is the method of partial fractions having to do with rational functions, quotients of polynomials. In applications known to me, these are usually thought of as functions of $\lambda$ and this is what we like to call such quotients, but everything is based only on the usual algebraic manipulations for polynomials.

Proposition 1.14.1 Suppose $r(\lambda)=\frac{a(\lambda)}{p(\lambda)^{m}}$ where $a(\lambda)$ is a polynomial and $p(\lambda)$ is a polynomial of degree at least 1. Then

$$
r(\lambda)=q(\lambda)+\sum_{k=1}^{m} \frac{b_{k}(\lambda)}{p(\lambda)^{k}}, \text { where degree of } b_{k}(\lambda)<\text { degree of } p(\lambda) \text { or } b_{k}(\lambda)=0
$$

Proof: Suppose first that $m=1$. If the degree of $a(\lambda)$ is larger than the degree of $p(\lambda)$, then do the division algorithm to write $a(\lambda)=p(\lambda) q(\lambda)+\hat{a}(\lambda)$ where the degree of $\hat{a}(\lambda)$ is less than the degree of $p(\lambda)$ or else $\hat{a}(\lambda)=0$. Thus the expression reduces to

$$
\frac{p(\lambda) q(\lambda)+m(\lambda)}{p(\lambda)}=q(\lambda)+\frac{\hat{a}(\lambda)}{p(\lambda)}
$$

and now it is in the desired form. Thus the Proposition is true if $m=1$. Suppose it is true for $m-1 \geq 1$. Then there is nothing to show if the degree of $a(\lambda)$ is less than the degree of $p(\lambda)$, so assume the degree of $a(\lambda)$ is larger than the degree of $p(\lambda)$. Then use the division algorithm as above and write

$$
\frac{a(\lambda)}{p(\lambda)^{m}}=\frac{p(\lambda) q(\lambda)+\hat{a}(\lambda)}{p(\lambda)^{m}}
$$

where the degree of $\hat{a}(\lambda)$ is less than the degree of $p(\lambda)$ or else is 0 . Then the above equals

$$
\frac{a(\lambda)}{p(\lambda)^{m}}=\frac{q(\lambda)}{p(\lambda)^{m-1}}+\frac{\hat{a}(\lambda)}{p(\lambda)^{m}}
$$

and by induction on the first term on the right, this proves the proposition.

With this, the general partial fractions theorem is next. From Theorem 1.13.9, every polynomial $q(\lambda)$ has a factorization of the form $\prod_{i=1}^{M} p_{i}(\lambda)^{m_{i}}$ where the $p_{i}(\lambda)$ are irreducible, meaning they cannot be factored further. Thus the polynomials $p_{i}(\lambda)$ are distinct and relatively prime as is every subset having at least two of these $p_{i}(\boldsymbol{\lambda})$.

Proposition 1.14.2 Let $\frac{a(\lambda)}{b(\lambda)}$ be any rational function. Then it is of the form

$$
\frac{a(\lambda)}{\prod_{i=1}^{M} p_{i}(\lambda)^{m_{i}}}
$$

where the $p_{i}(\lambda)$ are distinct irreducible polynomials, meaning they can't be factored any further as described in the chapter and each $m_{i}$ is a nonnegative integer.

Then there are polynomials $q(\lambda)$ and $n_{k i}(\lambda)$ with the degree of $n_{k i}(\lambda)$ less than the degree of $p_{i}(\lambda)$ or $n_{k i}(\lambda)=0$, such that

$$
\begin{equation*}
\frac{a(\lambda)}{b(\lambda)}=q(\lambda)+\sum_{i=1}^{M} \sum_{k=1}^{m_{i}} \frac{n_{k i}(\lambda)}{p_{i}(\lambda)^{k}} \tag{1.2}
\end{equation*}
$$

Proof: Suppose first that $\sum_{i=1}^{M} m_{i}=1$. Then the rational function is of the form $\frac{a(\lambda)}{p(\lambda)}$ and this can be placed in the desired form by an application of the division algorithm as above. Suppose now that this proposition is true if $\sum_{i=1}^{M} m_{i} \leq n$ for some $n \geq 1$ and suppose you have

$$
\frac{a(\lambda)}{b(\lambda)}=\frac{a(\lambda)}{\prod_{j=1}^{M} p_{j}(\lambda)^{m_{j}}}, \quad \sum_{j=1}^{M} m_{j}=n+1, \text { each } m_{j} \geq 0
$$

If some $m_{j}=n+1$, then one obtains the situation of Proposition 1.14.1. Therefore, it suffices to assume that no $m_{j}=n+1$ so there are at least two $m_{j}$ which are nonzero.

Every subset of the $\left\{p_{1}(\boldsymbol{\lambda}), p_{2}(\boldsymbol{\lambda}), \ldots, p_{M}(\boldsymbol{\lambda})\right\}$ having at least two $p_{i}(\boldsymbol{\lambda})$ is relatively prime because these polynomials are all irreducible. Therefore, there are polynomials $b_{i}(\lambda)$ such that $b_{i}(\lambda)=0$ if $m_{i}=0$ and $\sum_{i=1}^{M} b_{i}(\lambda) p_{i}(\lambda)=1$. Then multiply by this to obtain

$$
\frac{a(\lambda)}{b(\lambda)}=\frac{a(\lambda)}{\prod_{j=1}^{M} p_{j}(\lambda)^{m_{j}}}=\frac{a(\lambda) \sum_{i=1}^{M} b_{i}(\lambda) p_{i}(\lambda)}{\prod_{j=1}^{M} p_{j}(\lambda)^{m_{j}}}=\sum_{i=1}^{M} \frac{a(\lambda) b_{i}(\lambda) p_{i}(\lambda)}{\prod_{j=1}^{M} p_{j}(\lambda)^{m_{j}}}
$$

Now in the $i^{t h}$ term of the sum, the $p_{i}(\lambda)$ in the top cancels with exactly one of the factors in the bottom or else the term is 0 . It follows that the original $\frac{a(\lambda)}{b(\lambda)}$ is of the form $\sum_{i=1}^{N} \frac{\hat{a}_{i}(\lambda)}{\prod_{j=1}^{M} p_{j}(\lambda)^{m_{i j}}}$ where $\sum_{j=1}^{M} m_{i j} \leq n$. By induction applied to each of the terms in this sum, one obtains $\frac{a(\lambda)}{b(\lambda)}$ equal to an expression of the form in 1.2.

Proposition 1.14.3 The partial fractions expansion is unique.
Proof: Subtracting, you get

$$
q(\lambda)-\hat{q}(\lambda)=\sum_{i=1}^{M} \sum_{k=1}^{m_{i}} \frac{\hat{n}_{k i}(\lambda)-n_{k i}(\lambda)}{p_{i}(\lambda)^{k}}
$$

and so, the left side is 0 since otherwise, you could multiply by the product of the $p_{i}(\lambda)^{k}$ and get equality of two polynomials of different degree. Hence you have

$$
\begin{equation*}
\sum_{i=1}^{M} \sum_{k=1}^{m_{i}} \frac{n_{k i}(\lambda)}{p_{i}(\lambda)^{k}}=\sum_{i=1}^{M} \sum_{k=1}^{m_{i}} \frac{\hat{n}_{k i}(\lambda)}{p_{i}(\lambda)^{k}} \tag{*}
\end{equation*}
$$

Now multiply both sides by $\prod_{i \neq j} p_{i}(\lambda)^{m_{i}} p_{j}(\lambda)^{m_{j}-1}$. Then

$$
P(\lambda)+\frac{n_{m_{j} j}(\lambda)}{p_{j}(\lambda)}=\hat{P}(\lambda)+\frac{\hat{n}_{m_{j} j}(\lambda)}{p_{j}(\lambda)}
$$

then by the same argument, $P(\lambda)=\hat{P}(\lambda)$ and now $\frac{n_{m_{j}}(\lambda)}{p_{j}(\lambda)}=\frac{\hat{n}_{m_{j} j}(\lambda)}{p_{j}(\lambda)}$ and so

$$
p_{j}(\lambda)\left(n_{m_{j} j}(\lambda)-\hat{n}_{m_{j} j}(\lambda)\right)=0
$$

and from Lemma 1.13.2, since $p_{j}(\lambda) \neq 0$, it follows that $n_{m_{j} j}(\lambda)-\hat{n}_{m_{j} j}(\lambda)=0$. Thus in * the $m_{j}$ can be replaced with $m_{j}-1$.

Continue this way, to show that $n_{k j}(\lambda)-\hat{n}_{k j}(\lambda)=0$ for each $k$. Since $j$ is arbitrary, this shows uniqueness.

### 1.15 Finite Fields

The emphasis of the first part of this book will be on what can be done on the basis of algebra alone. Linear algebra only needs a field of scalars along with some axioms involving an Abelian group of vectors and there are infinitely many examples of fields, including some which are finite. Since it is good to have examples in mind, I will present the finite fields of residue classes modulo a prime number in this little section. Then, when linear algebra is developed in the first part of the book and reference is made to a field of scalars, you should think that it is possible that the field might be this field of residue classes.

Here is the construction of the finite fields $\mathbb{Z}_{p}$ for $p$ a prime.
Definition 1.15.1 Let $\mathbb{Z}^{+}$denote the set of nonnegative integers, $\mathbb{Z}^{+}=\{0,1,2,3, \cdots\}$. Also let $p$ be a prime number. We will say that two integers, $a, b$ are equivalent and write $a \sim b$ if $a-b$ is divisible by $p$. Thus they are equivalent if $a-b=p x$ for some integer $x$.

Proposition 1.15.2 The relation $\sim$ is an equivalence relation. Denoting by $\bar{n}$ the equivalence class determined by $n \in \mathbb{N}$, the following are well defined operations.

$$
\begin{aligned}
\bar{n}+\bar{m} & \equiv \overline{n+m} \\
\bar{n} \bar{m} & \equiv \overline{n m}
\end{aligned}
$$

which makes the set $\mathbb{Z}_{p}$ consisting of $\{\overline{0}, \overline{1}, \cdots, \overline{p-1}\}$ into a field.
Proof: First note that for $n \in \mathbb{Z}^{+}$there always exists $r \in\{0,1, \cdots, p-1\}$ such that $\bar{n}=\bar{r}$. This is clearly true because if $n \in \mathbb{Z}^{+}$, then $n=m p+r$ for $r<p$, this by the Euclidean algorithm. Thus $\bar{r}=\bar{n}$. Now suppose that $\bar{n}_{1}=\bar{n}$ and $\bar{m}_{1}=\bar{m}$. Is it true that $\overline{n_{1}+m_{1}}=\overline{n+m}$ ? Is it true that $(n+m)-\left(n_{1}+m_{1}\right)$ is a multiple of $p$ ? Of course since
$n_{1}-n$ and $m_{1}-m$ are both multiples of $p$. Similarly, is $\overline{n_{1} m_{1}}=\overline{n m}$ ? Is $n m-n_{1} m_{1}$ a multiple of $p$ ? Of course this is so because

$$
\begin{aligned}
n m-n_{1} m_{1} & =n m-n_{1} m+n_{1} m-n_{1} m_{1} \\
& =m\left(n-n_{1}\right)+n_{1}\left(m-m_{1}\right)
\end{aligned}
$$

which is a multiple of $p$. Thus the operations are well defined. It follows that all of the field axioms hold except possibly the existence of a multiplicative inverse and an additive inverse. First consider the question of an additive inverse. A typical thing in $\mathbb{Z}_{p}$ is of the form $\bar{r}$ where $0 \leq r \leq p-1$. Then consider $(\overline{p-r})$. By definition, $\bar{r}+\overline{p-r}=\bar{p}=\overline{0}$ and so the additive inverse exists.

Now consider the existence of a multiplicative inverse. This is where $p$ is prime is used. Say $\bar{n} \neq \overline{0}$. That is, $n$ is not a multiple of $p, 0 \leq n<p$. Then since $p$ is prime, $n, p$ are relatively prime and so there are integers $x, y$ such that $1=x n+y p$. Choose $m \geq 0$ such that $p m+x>0, p m+y>0$. Then

$$
1+p m n+p m p=(p m+x) n+(p m+y) p
$$

It follows that $\overline{1+p m n+p^{2} m}=\overline{1}, \overline{1}=\overline{(p m+x)} \bar{n}$ and so $\overline{(p m+x)}$ is the multiplicative inverse of $\bar{n}$.

Thus $\mathbb{Z}_{p}$ is a finite field, known as the field of residue classes modulo $p$.
Something else which is often considered is a commutative ring with unity.
Definition 1.15.3 A commutative ring with unity is just a field except it lacks the property that nonzero elements have a multiplicative inverse. It has all other properties. In this book, this will be referred to simply as a commutative ring. I will assume that commutative rings always have 1. Thus the axioms of a commutative ring with unity are as follows:

Axiom 1.15.4 Here are the axioms for a commutative ring with unity.

1. $x+y=y+x$ (commutative law for addition)
2. There exists 0 such that $x+0=x$ for all $x$, (additive identity).
3. For each $x \in \mathbb{F}$, there exists $-x \in \mathbb{F}$ such that $x+(-x)=0$, (existence of additive inverse).
4. $(x+y)+z=x+(y+z),($ associative law for addition $)$.
5. $x y=y x$, (commutative law for multiplication). You could write this as $x \times y=y \times x$.
6. $(x y) z=x(y z),($ associative law for multiplication).
7. There exists 1 such that $1 x=x$ for all $x$,(multiplicative identity).
8. $x(y+z)=x y+x z .($ distributive law $)$.

An example of such a thing is $\mathbb{Z}_{m}$ where $m$ is not prime, also the ordinary integers. However, the integers are also an integral domain.

Definition 1.15.5 A commutative ring with unity is called an integral domain if, in addition to the above, whenever $a b=0$, it follows that either $a=0$ or $b=0$.

### 1.16 Some Topics From Analysis

Recall from calculus that if $A$ is a nonempty set, $\sup _{a \in A} f(a)$ denotes the least upper bound of $f(A)$ or if this set is not bounded above, it equals $\infty$. Also $\inf _{a \in A} f(a)$ denotes the greatest lower bound of $f(A)$ if this set is bounded below and it equals $-\infty$ if $f(A)$ is not bounded below. Thus to say $\sup _{a \in A} f(a)=\infty$ is just a way to say that $A$ is not bounded above.

Definition 1.16.1 Let $f(a, b) \in[-\infty, \infty]$ for $a \in A$ and $b \in B$ where $A, B$ are sets which means that $f(a, b)$ is either a number, $\infty$, or $-\infty$. The symbol, $+\infty$ is interpreted as a point out at the end of the number line which is larger than every real number. Of course there is no such number. That is why it is called $\infty$. The symbol, $-\infty$ is interpreted similarly. Then $\sup _{a \in A} f(a, b)$ means $\sup \left(S_{b}\right)$ where $S_{b} \equiv\{f(a, b): a \in A\}$.

Note that if $\left\{a_{n}\right\}$ is an increasing sequence of real numbers,

$$
\sup _{n}\left\{a_{n}\right\}=\lim _{n \rightarrow \infty} a_{n}
$$

if $\sup _{n}\left\{a_{n}\right\}<\infty$ and also if we define $\lim _{n \rightarrow \infty} a_{n} \equiv \infty$ if $\sup _{n}\left\{a_{n}\right\}=\infty$.
Unlike limits, you can take the sup in different orders.
Lemma 1.16.2 Let $f(a, b) \in[-\infty, \infty]$ for $a \in A$ and $b \in B$ where $A, B$ are sets. Then

$$
\sup _{a \in A} \sup _{b \in B} f(a, b)=\sup _{b \in B} \sup _{a \in A} f(a, b) .
$$

Proof: Note that for all $a, b$,

$$
f(a, b) \leq \sup _{b \in B} \sup _{a \in A} f(a, b)
$$

and therefore, for all $a, \sup _{b \in B} f(a, b) \leq \sup _{b \in B} \sup _{a \in A} f(a, b)$. Therefore,

$$
\sup _{a \in A} \sup _{b \in B} f(a, b) \leq \sup _{b \in B} \sup _{a \in A} f(a, b) .
$$

Repeat the same argument interchanging $a$ and $b$, to get the conclusion of the lemma.

### 1.17 lim sup and lim inf

The nice thing about limsup and liminf is that they always exist, unlike the limit of a sequence. Recall how in calculus, there is no limit of $(-1)^{n}$. First here is a simple lemma and definition.

Definition 1.17.1 Denote by $[-\infty, \infty]$ the real line along with symbols $\infty$ and $-\infty$. It is understood that $\infty$ is larger than every real number and $-\infty$ is smaller than every real number. Then if $\left\{A_{n}\right\}$ is an increasing sequence of points of $[-\infty, \infty], \lim _{n \rightarrow \infty} A_{n}$ is defined to equal $\infty$ if the only upper bound of the set $\left\{A_{n}\right\}$ is $\infty$. If $\left\{A_{n}\right\}$ is bounded above by a real number, then $\lim _{n \rightarrow \infty} A_{n}$ is defined in the usual way and equals the least upper bound of $\left\{A_{n}\right\}$. If $\left\{A_{n}\right\}$ is a decreasing sequence of points of $[-\infty, \infty], \lim _{n \rightarrow \infty} A_{n}$ equals $-\infty$ if the only lower bound of the sequence $\left\{A_{n}\right\}$ is $-\infty$. If $\left\{A_{n}\right\}$ is bounded below by a real number, then $\lim _{n \rightarrow \infty} A_{n}$ is defined in the usual way and equals the greatest lower bound of $\left\{A_{n}\right\}$. More simply, if $\left\{A_{n}\right\}$ is increasing, $\lim _{n \rightarrow \infty} A_{n} \equiv \sup \left\{A_{n}\right\}$ and if $\left\{A_{n}\right\}$ is decreasing then $\lim _{n \rightarrow \infty} A_{n} \equiv \inf \left\{A_{n}\right\}$.

Before discussing limsup and liminf, here is a very useful observation about double sums.

Theorem 1.17.2 Let $a_{i j} \geq 0$. Then

$$
\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i j}=\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{i j}
$$

Proof: First note there is no trouble in defining these sums because the $a_{i j}$ are all nonnegative. If a sum diverges, it only diverges to $\infty$ and so $\infty$ is the value of the sum. Next note that

$$
\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{i j} \geq \sup _{n} \sum_{j=r i=r}^{\infty} \sum_{i=}^{n} a_{i j}
$$

because $\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{i j} \geq \sum_{j=r}^{\infty} \sum_{i=r}^{n} a_{i j}$ for each $n$. Therefore,

$$
\begin{aligned}
& \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{i j} \geq \sup _{n} \sum_{j=r}^{\infty} \sum_{i=r}^{n} a_{i j}=\sup _{n} \lim _{m \rightarrow \infty} \sum_{j=r}^{m} \sum_{i=r}^{n} a_{i j} \\
& =\sup _{n} \lim _{m \rightarrow \infty} \sum_{i=r}^{n} \sum_{j=r}^{m} a_{i j}=\sup _{n} \sum_{i=r}^{n} \lim _{m \rightarrow \infty} \sum_{j=r}^{m} a_{i j} \\
& =\sup _{n} \sum_{i=r}^{n} \sum_{j=r}^{\infty} a_{i j}=\lim _{n \rightarrow \infty} \sum_{i=r}^{n} \sum_{j=r}^{\infty} a_{i j}=\sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{i j}
\end{aligned}
$$

Interchanging the $i$ and $j$ in the above argument proves the theorem.
Lemma 1.17.3 Let $\left\{a_{n}\right\}$ be a sequence of real numbers and $U_{n} \equiv \sup \left\{a_{k}: k \geq n\right\}$. Then $\left\{U_{n}\right\}$ is a decreasing sequence. Also if $L_{n} \equiv \inf \left\{a_{k}: k \geq n\right\}$, then $\left\{L_{n}\right\}$ is an increasing sequence. Therefore, $\lim _{n \rightarrow \infty} L_{n}$ and $\lim _{n \rightarrow \infty} U_{n}$ both exist.

Proof: Let $W_{n}$ be an upper bound for $\left\{a_{k}: k \geq n\right\}$. Then since these sets are getting smaller, it follows that for $m<n, W_{m}$ is an upper bound for $\left\{a_{k}: k \geq n\right\}$. In particular if $W_{m}=U_{m}$, then $U_{m}$ is an upper bound for $\left\{a_{k}: k \geq n\right\}$ and so $U_{m}$ is at least as large as $U_{n}$, the least upper bound for $\left\{a_{k}: k \geq n\right\}$. The claim that $\left\{L_{n}\right\}$ is decreasing is similar.

From the lemma, the following definition makes sense.
Definition 1.17.4 Let $\left\{a_{n}\right\}$ be any sequence of points of $[-\infty, \infty]$

$$
\begin{aligned}
{\lim \sup _{n \rightarrow \infty}}^{a_{n}} & \equiv \lim _{n \rightarrow \infty} \sup \left\{a_{k}: k \geq n\right\} \\
\lim _{n \rightarrow \infty} \inf _{n \rightarrow \infty} & \equiv \lim _{n \rightarrow \infty} \inf \left\{a_{k}: k \geq n\right\}
\end{aligned}
$$

Now the following shows the relation of liminf and limsup to the limit.
Theorem 1.17.5 Suppose $\left\{a_{n}\right\}$ is a sequence of real numbers and that $\limsup _{n \rightarrow \infty} a_{n}$ and $\liminf _{n \rightarrow \infty} a_{n}$ are both real numbers. Then $\lim _{n \rightarrow \infty} a_{n}$ exists if and only if $\liminf _{n \rightarrow \infty} a_{n}=$ $\limsup _{n \rightarrow \infty} a_{n}$ and in this case,

$$
\lim _{n \rightarrow \infty} a_{n}=\lim \inf _{n \rightarrow \infty} a_{n}=\lim \sup _{n \rightarrow \infty} a_{n}
$$

Proof: First note that

$$
\sup \left\{a_{k}: k \geq n\right\} \geq \inf \left\{a_{k}: k \geq n\right\}
$$

and so,

$$
\limsup _{n \rightarrow \infty} a_{n} \equiv \lim _{n \rightarrow \infty} \sup \left\{a_{k}: k \geq n\right\} \geq \lim _{n \rightarrow \infty} \inf \left\{a_{k}: k \geq n\right\} \equiv \lim _{n \rightarrow \infty} \inf _{n} a_{n}
$$

Suppose first that $\lim _{n \rightarrow \infty} a_{n}$ exists and is a real number $a$. Then from the definition of a limit, there exists $N$ corresponding to $\varepsilon / 6$ in the definition. Hence, if $m, n \geq N$, then

$$
\left|a_{n}-a_{m}\right| \leq\left|a_{n}-a\right|+\left|a-a_{n}\right|<\frac{\varepsilon}{6}+\frac{\varepsilon}{6}=\frac{\varepsilon}{3}
$$

From the definition of $\sup \left\{a_{k}: k \geq N\right\}$, there exists $n_{1} \geq N$ such that

$$
\sup \left\{a_{k}: k \geq N\right\} \leq a_{n_{1}}+\varepsilon / 3
$$

Similarly, there exists $n_{2} \geq N$ such that

$$
\inf \left\{a_{k}: k \geq N\right\} \geq a_{n_{2}}-\varepsilon / 3
$$

It follows that

$$
\sup \left\{a_{k}: k \geq N\right\}-\inf \left\{a_{k}: k \geq N\right\} \leq\left|a_{n_{1}}-a_{n_{2}}\right|+\frac{2 \varepsilon}{3}<\varepsilon
$$

Since the sequence, $\left\{\sup \left\{a_{k}: k \geq N\right\}\right\}_{N=1}^{\infty}$ is decreasing and $\left\{\inf \left\{a_{k}: k \geq N\right\}\right\}_{N=1}^{\infty}$ is increasing, it follows that

$$
0 \leq \lim _{N \rightarrow \infty} \sup \left\{a_{k}: k \geq N\right\}-\lim _{N \rightarrow \infty} \inf \left\{a_{k}: k \geq N\right\} \leq \varepsilon
$$

Since $\varepsilon$ is arbitrary, this shows

$$
\begin{equation*}
\lim _{N \rightarrow \infty} \sup \left\{a_{k}: k \geq N\right\}=\lim _{N \rightarrow \infty} \inf \left\{a_{k}: k \geq N\right\} \tag{1.3}
\end{equation*}
$$

Next suppose 1.3 and both equal $a \in \mathbb{R}$. Then

$$
\lim _{N \rightarrow \infty}\left(\sup \left\{a_{k}: k \geq N\right\}-\inf \left\{a_{k}: k \geq N\right\}\right)=0
$$

Since $\sup \left\{a_{k}: k \geq N\right\} \geq \inf \left\{a_{k}: k \geq N\right\}$ it follows that for every $\varepsilon>0$, there exists $N$ such that

$$
\sup \left\{a_{k}: k \geq N\right\}-\inf \left\{a_{k}: k \geq N\right\}<\varepsilon
$$

and for every $N$,

$$
\inf \left\{a_{k}: k \geq N\right\} \leq a \leq \sup \left\{a_{k}: k \geq N\right\}
$$

Thus if $n \geq N,\left|a-a_{n}\right|<\varepsilon$ which implies that $\lim _{n \rightarrow \infty} a_{n}=a$. In case

$$
a=\infty=\lim _{N \rightarrow \infty} \sup \left\{a_{k}: k \geq N\right\}=\lim _{N \rightarrow \infty} \inf \left\{a_{k}: k \geq N\right\}
$$

then if $r \in \mathbb{R}$ is given, there exists $N$ such that $\inf \left\{a_{k}: k \geq N\right\}>r$ which is to say that $\lim _{n \rightarrow \infty} a_{n}=\infty$. The case where $a=-\infty$ is similar except you use $\sup \left\{a_{k}: k \geq N\right\}$.

The significance of limsup and liminf, in addition to what was just discussed, is contained in the following theorem which follows quickly from the definition.

Theorem 1.17.6 Suppose $\left\{a_{n}\right\}$ is a sequence of points of $[-\infty, \infty]$. Let

$$
\lambda=\lim \sup _{n \rightarrow \infty} a_{n} .
$$

Then if $b>\lambda$, it follows there exists $N$ such that whenever $n \geq N, a_{n} \leq b$. If $c<\lambda$, then $a_{n}>c$ for infinitely many values of $n$. Let $\gamma=\liminf _{n \rightarrow \infty} a_{n}$. Then if $d<\gamma$, it follows there exists $N$ such that whenever $n \geq N, a_{n} \geq d$. If $e>\gamma$, it follows $a_{n}<e$ for infinitely many values of $n$.

The proof of this theorem is left as an exercise for you. It follows directly from the definition and it is the sort of thing you must do yourself. Here is one other simple proposition.

Proposition 1.17.7 Let $\lim _{n \rightarrow \infty} a_{n}=a>0$. Then

$$
\limsup _{n \rightarrow \infty} a_{n} b_{n}=a \lim \sup _{n \rightarrow \infty} b_{n} .
$$

Proof: This follows from the definition. Let $\lambda_{n}=\sup \left\{a_{k} b_{k}: k \geq n\right\}$. For all $n$ large enough, $a_{n}>a-\varepsilon$ where $\varepsilon$ is small enough that $a-\varepsilon>0$. Therefore,

$$
\lambda_{n} \geq \sup \left\{b_{k}: k \geq n\right\}(a-\varepsilon)
$$

for all $n$ large enough. Then

$$
\begin{aligned}
\limsup _{n \rightarrow \infty} a_{n} b_{n} & =\lim _{n \rightarrow \infty} \lambda_{n} \equiv \limsup _{n \rightarrow \infty} a_{n} b_{n} \\
& \geq \lim _{n \rightarrow \infty}\left(\sup \left\{b_{k}: k \geq n\right\}(a-\varepsilon)\right) \\
& =(a-\varepsilon) \lim \sup _{n \rightarrow \infty} b_{n}
\end{aligned}
$$

Similar reasoning shows $\limsup \sin _{n \rightarrow \infty} a_{n} b_{n} \leq(a+\varepsilon) \limsup _{n \rightarrow \infty} b_{n}$. Since $\varepsilon>0$ is arbitrary, the conclusion follows.

### 1.18 Exercises

1. Prove by induction that $\sum_{k=1}^{n} k^{3}=\frac{1}{4} n^{4}+\frac{1}{2} n^{3}+\frac{1}{4} n^{2}$.
2. Prove by induction that whenever $n \geq 2, \sum_{k=1}^{n} \frac{1}{\sqrt{k}}>\sqrt{n}$.
3. Prove by induction that $1+\sum_{i=1}^{n} i(i!)=(n+1)$ !.
4. The binomial theorem states $(x+y)^{n}=\sum_{k=0}^{n}\binom{n}{k} x^{n-k} y^{k}$ where

$$
\binom{n+1}{k}=\binom{n}{k}+\binom{n}{k-1} \text { if } k \in[1, n],\binom{n}{0} \equiv 1 \equiv\binom{n}{n}
$$

Prove the binomial theorem by induction. Next show that

$$
\binom{n}{k}=\frac{n!}{(n-k)!k!}, 0!\equiv 1
$$

5. Let $z=5+i 9$. Find $z^{-1}$.
6. Let $z=2+i 7$ and let $w=3-i 8$. Find $z w, z+w, z^{2}$, and $w / z$.
7. Give the complete solution to $x^{4}+16=0$.
8. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16 .
9. If $z$ is a complex number, show there exists $\omega$ a complex number with $|\omega|=1$ and $\omega z=|z|$.
10. De Moivre's theorem says $[r(\cos t+i \sin t)]^{n}=r^{n}(\cos n t+i \sin n t)$ for $n$ a positive integer. Does this formula continue to hold for all integers $n$, even negative integers? Explain.
11. You already know formulas for $\cos (x+y)$ and $\sin (x+y)$ and these were used to prove De Moivre's theorem. Now using De Moivre's theorem, derive a formula for $\sin (5 x)$ and one for $\cos (5 x)$.
12. If $z$ and $w$ are two complex numbers and the polar form of $z$ involves the angle $\theta$ while the polar form of $w$ involves the angle $\phi$, show that in the polar form for $z w$ the angle involved is $\theta+\phi$. Also, show that in the polar form of a complex number $z, r=|z|$.
13. Factor $x^{3}+8$ as a product of linear factors.
14. Write $x^{3}+27$ in the form $(x+3)\left(x^{2}+a x+b\right)$ where $x^{2}+a x+b$ cannot be factored any more using only real numbers.
15. Completely factor $x^{4}+16$ as a product of linear factors.
16. Factor $x^{4}+16$ as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.
17. If $z, w$ are complex numbers prove $\overline{z w}=\overline{z w}$ and then show by induction that

$$
\overline{\prod_{j=1}^{n} z_{j}}=\prod_{j=1}^{n} \overline{z_{j}}
$$

Also verify that $\overline{\sum_{k=1}^{m} z_{k}}=\sum_{k=1}^{m} \overline{z_{k}}$. In words this says the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.
18. Suppose $p(x)=a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ where all the $a_{k}$ are real numbers. Suppose also that $p(z)=0$ for some $z \in \mathbb{C}$. Show it follows that $p(\bar{z})=0$ also.
19. Show that $1+i, 2+i$ are the only two zeros to

$$
p(x)=x^{2}-(3+2 i) x+(1+3 i)
$$

so the zeros do not necessarily come in conjugate pairs if the coefficients are not real.
20. I claim that $1=-1$. Here is why. $-1=i^{2}=\sqrt{-1} \sqrt{-1}=\sqrt{(-1)^{2}}=\sqrt{1}=1$. This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?
21. De Moivre's theorem is really a grand thing. I plan to use it now for rational exponents, not just integers. $1=1^{(1 / 4)}=(\cos 2 \pi+i \sin 2 \pi)^{1 / 4}=\cos (\pi / 2)+i \sin (\pi / 2)=$ $i$. Therefore, squaring both sides it follows $1=-1$ as in the previous problem. What does this tell you about De Moivre's theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?
22. Review Problem 10 at this point. Now here is another question: If $n$ is an integer, is it always true that $(\cos \theta-i \sin \theta)^{n}=\cos (n \theta)-i \sin (n \theta)$ ? Explain.
23. Suppose you have any polynomial in $\cos \theta$ and $\sin \theta$. By this I mean an expression of the form $\sum_{\alpha=0}^{m} \sum_{\beta=0}^{n} a_{\alpha \beta} \cos ^{\alpha} \theta \sin ^{\beta} \theta$ where $a_{\alpha \beta} \in \mathbb{C}$. Can this always be written in the form $\sum_{\gamma=-(n+m)}^{m+n} b_{\gamma} \cos \gamma \theta+\sum_{\tau=-(n+m)}^{n+m} c_{\tau} \sin \tau \theta$ ? Explain.
24. Show that $\mathbb{C}$ cannot be considered an ordered field. Hint: Consider $i^{2}=-1$.
25. Suppose $p(x)=a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ is a polynomial and it has $n$ zeros,

$$
z_{1}, z_{2}, \cdots, z_{n}
$$

listed according to multiplicity. ( $z$ is a root of multiplicity $m$ if the polynomial $f(x)=$ $(x-z)^{m}$ divides $p(x)$ but $(x-z) f(x)$ does not.) Show that

$$
p(x)=a_{n}\left(x-z_{1}\right)\left(x-z_{2}\right) \cdots\left(x-z_{n}\right) .
$$

26. Give the solutions to the following quadratic equations having real coefficients.
(a) $x^{2}-2 x+2=0$
(d) $x^{2}+4 x+9=0$
(b) $3 x^{2}+x+3=0$
(c) $x^{2}-6 x+13=0$
(e) $4 x^{2}+4 x+5=0$
27. Give the solutions to the following quadratic equations having complex coefficients. Note how the solutions do not come in conjugate pairs as they do when the equation has real coefficients.
(a) $x^{2}+2 x+1+i=0$
(d) $x^{2}-4 i x-5=0$
(b) $4 x^{2}+4 i x-5=0$
(c) $4 x^{2}+(4+4 i) x+1+2 i=0$
(e) $3 x^{2}+(1-i) x+3 i=0$
28. Prove the fundamental theorem of algebra for quadratic polynomials having coefficients in $\mathbb{C}$. That is, show that an equation of the form $a x^{2}+b x+c=0$ where $a, b, c$ are complex numbers, $a \neq 0$ has a complex solution. Hint: Consider the fact, noted earlier that the expressions given from the quadratic formula do in fact serve as solutions.
29. Prove the Euclidean algorithm: If $m, n$ are positive integers, then there exist integers $q, r \geq 0$ such that $r<m$ and $n=q m+r$ Hint: You might try considering

$$
S \equiv\{n-k m: k \in \mathbb{N} \text { and } n-k m<0\}
$$

and picking the smallest integer in $S$ or something like this. It was done in the chapter, but go through it yourself.
30. Recall that two polynomials are equal means that the coefficients of corresponding powers of $\lambda$ are equal. Thus a polynomial equals 0 if and only if all coefficients equal 0 . In calculus we usually think of a polynomial as 0 if it sends every value of $x$ to 0 . Suppose you have the following polynomial $\overline{1} x^{2}+\overline{1} x$ where it is understood to be a polynomial in $\mathbb{Z}_{2}$. Thus it is not the zero polynomial. Show, however, that this equals zero for all $x \in \mathbb{Z}_{2}$ so we would be tempted to say it is zero if we use the conventions of calculus.
31. Prove Wilson's theorem. This theorem states that if $p$ is a prime, then $(p-1)!+1$ is divisible by $p$. Wilson's theorem was first proved by Lagrange in the 1770's. Hint: Check directly for $p=2,3$. Show that $\overline{p-1}=-\overline{1}$ and that if $a \in\{2, \cdots, p-2\}$, then $(\bar{a})^{-1} \neq \bar{a}$. Thus a residue class $\bar{a}$ and its multiplicative inverse for $a \in\{2, \cdots, p-2\}$ occur in pairs. Show that this implies that the residue class of $(p-1)$ ! must be -1 . From this, draw the conclusion.
32. Show that in the arithmetic of $\mathbb{Z}_{p},(\bar{x}+\bar{y})^{p}=(\bar{x})^{p}+(\bar{y})^{p}$, a well known formula among students.
33. Consider $(\bar{a}) \in \mathbb{Z}_{p}$ for $p$ a prime, and suppose $(\bar{a}) \neq \overline{1}, \overline{0}$. Fermat's little theorem says that $(\bar{a})^{p-1}=\overline{1}$. In other words $(a)^{p-1}-1$ is divisible by $p$. Prove this. Hint: Show that there must exist $r \geq 1, r \leq p-1$ such that $(\bar{a})^{r}=\overline{1}$. To do so, consider $\overline{1},(\bar{a}),(\bar{a})^{2}, \cdots$. Then these all have values in $\{\overline{1}, \overline{2}, \cdots, \overline{p-1}\}$, and so there must be a repeat in $\left\{\overline{1},(\bar{a}), \cdots,(\bar{a})^{p-1}\right\}$, say $p-1 \geq l>k$ and $(\bar{a})^{l}=(\bar{a})^{k}$. Then tell why $(\bar{a})^{l-k}-\overline{1}=0$. Let $r$ be the first positive integer such that $(\bar{a})^{r}=\overline{1}$. Let $G=$ $\left\{\overline{1},(\bar{a}), \cdots,(\bar{a})^{r-1}\right\}$. Show that every residue class in $G$ has its multiplicative inverse in $G$. In fact, $(\bar{a})^{k}(\bar{a})^{r-k}=\overline{1}$. Also verify that the entries in $G$ must be distinct. Now consider the sets $\bar{b} G \equiv\left\{\bar{b}(\bar{a})^{k}: k=0, \cdots, r-1\right\}$ where $\bar{b} \in\{\overline{1}, \overline{2}, \cdots, \overline{p-1}\}$. Show that two of these sets are either the same or disjoint and that they all consist of $r$ elements. Explain why it follows that $p-1=l r$ for some positive integer $l$ equal to the number of these distinct sets. Then explain why $(\bar{a})^{p-1}=(\bar{a})^{l r}=\overline{1}$.
34. Let $p(x)$ and $q(x)$ be polynomials. Then by the division algorithm, there exist polynomials $l(x), r(x)$ equal to 0 or having degree smaller than $p(x)$ such that

$$
q(x)=p(x) l(x)+r(x)
$$

If $k(x)$ is the greatest common divisor of $p(x)$ and $q(x)$, explain why $k(x)$ must divide $r(x)$. Then argue that $k(x)$ is also the greatest common divisor of $p(x)$ and $r(x)$. Now repeat the process for the polynomials $p(x)$ and $r(x)$. This time, the remainder term will have degree smaller than $r(x)$. Keep doing this and eventually the remainder must be 0 . Describe an algorithm based on this which will determine the greatest common divisor of two polynomials.
35. Consider $\mathbb{Z}_{m}$ where $m$ is not a prime. Show that although this will not be a field, it is a commutative ring with unity.
36. This and the next few problems are to illustrate the utility of the limsup. A sequence of numbers $\left\{x_{n}\right\}$ in $\mathbb{C}$ is called a Cauchy sequence if for every $\varepsilon>0$ there exists $m$ such that if $k, l \geq m$, then $\left|x_{k}-x_{l}\right|<\varepsilon$. The complex numbers are said to be complete because any Cauchy sequence converges. This is one form of the completeness axiom. Using this axiom, show that $\sum_{k=0}^{\infty} r^{k} \equiv \lim _{n \rightarrow \infty} \sum_{k=0}^{n} r^{k}=\frac{1}{1-r}$ whenever $r \in \mathbb{C}$ and $|r|<1$. Hint: You need to do a computation with the sum and show that the partial sums form a Cauchy sequence.
37. Show that if $\sum_{j=1}^{\infty}\left|c_{j}\right|$ converges, meaning that $\lim _{n \rightarrow \infty} \sum_{j=1}^{n}\left|c_{j}\right|$ exists, then $\sum_{j=1}^{\infty} c_{j}$ also converges, meaning $\lim _{n \rightarrow \infty} \sum_{j=1}^{n} c_{j}$ exists, this for $c_{j} \in \mathbb{C}$. Recall from calculus, this says that absolute convergence implies convergence.
38. Show that if $\sum_{j=1}^{\infty} c_{j}$ converges, meaning $\lim _{n \rightarrow \infty} \sum_{j=1}^{n} c_{j}$ exists, then it must be the case that $\lim _{n \rightarrow \infty} c_{n}=0$.
39. If $\limsup \sin _{k \rightarrow \infty}\left|a_{k}\right|^{1 / k}<1$, then $\sum_{k=1}^{\infty}\left|a_{k}\right|$ converges, while if $\limsup \operatorname{sum}_{n \rightarrow \infty}\left|a_{n}\right|^{1 / n}>$ 1 , then the series diverges spectacularly because $\lim _{n \rightarrow \infty}\left|c_{n}\right|$ fails to equal 0 and in fact has a subsequence which converges to $\infty$. Show this. Also show that if $\limsup _{n \rightarrow \infty}\left|a_{n}\right|^{1 / n}=1$, the test fails because there are examples where the series can converge and examples where the series diverges. This is an improved version of the root test from calculus. It is improved because limsup always exists. Hint: For the last part, consider $\sum_{n} \frac{1}{n}$ and $\sum_{n} \frac{1}{n^{2}}$. Review calculus to see why the first diverges and the second converges.
40. Consider a power series $\sum_{n=0}^{\infty} a_{n} x^{n}$. Derive a condition for the radius of convergence using limsup $\operatorname{sim}_{n \rightarrow \infty}\left|a_{n}\right|^{1 / n}$. Recall that the radius of convergence $R$ is such that if $|x|<$ $R$, then the series converges and if $|x|>R$, the series diverges and if $|x|=R$ is it not known whether the series converges. In this problem, assume only that $x \in \mathbb{C}$.
41. Show that if $a_{n}$ is a sequence of real numbers, then

$$
\lim \inf _{n \rightarrow \infty}\left(-a_{n}\right)=-\lim \sup _{n \rightarrow \infty} a_{n}
$$

## Part I

## Linear Algebra For Its Own Sake

## Chapter 2

## Systems of Linear Equations

This part of the book is about linear algebra itself, as a part of algebra. Some geometric and analytic concepts do creep in, but it is primarily about algebra. It involves general fields and has very little to do with limits and completeness although some geometry is included, but not much. Numbers are elements of a field.

### 2.1 Elementary Operations

In this chapter, the main interest is in fields of scalars consisting of $\mathbb{R}$ or $\mathbb{C}$, but everything is applied to arbitrary fields. Consider the following example.

Example 2.1.1 Find $x$ and $y$ such that

$$
\begin{equation*}
x+y=7 \text { and } 2 x-y=8 . \tag{2.1}
\end{equation*}
$$

The set of ordered pairs, $(x, y)$ which solve both equations is called the solution set.
You can verify that $(x, y)=(5,2)$ is a solution to the above system. The interesting question is this: If you were not given this information to verify, how could you determine the solution? You can do this by using the following basic operations on the equations, none of which change the set of solutions of the system of equations.

Definition 2.1.2 Elementary operations are those operations consisting of the following.

1. Interchange the order in which the equations are listed.
2. Multiply any equation by a nonzero number.
3. Replace any equation with itself added to a multiple of another equation.

Example 2.1.3 To illustrate the third of these operations on this particular system, consider the following.

$$
\begin{gathered}
x+y=7 \\
2 x-y=8
\end{gathered}
$$

The system has the same solution set as the system

$$
\begin{gathered}
x+y=7 \\
-3 y=-6
\end{gathered}
$$

To obtain the second system, take the second equation of the first system and add -2 times the first equation to obtain $-3 y=-6$. Now, this clearly shows that $y=2$ and so it follows from the other equation that $x+2=7$ and so $x=5$.

Of course a linear system may involve many equations and many variables. The solution set is still the collection of solutions to the equations. In every case, the above operations of Definition 2.1.2 do not change the set of solutions to the system of linear equations.

Theorem 2.1.4 Given two equations involving the variables, $\left(x_{1}, \cdots, x_{n}\right)$.

$$
\begin{equation*}
E_{1}=f_{1}, E_{2}=f_{2} \tag{2.2}
\end{equation*}
$$

where $E_{1}$ and $E_{2}$ are expressions

$$
\begin{aligned}
& E_{1}=a_{1} x_{1}+\cdots+a_{n} x_{n} \\
& E_{2}=b_{1} x_{1}+\cdots+b_{n} x_{n}
\end{aligned}
$$

involving the variables and $f_{1}$ and $f_{2}$ are constants where the $a_{i}, b_{i}, f_{1}, f_{2}$ are in a field $\mathbb{F}$. (In the above example there are only two variables, $x$ and $y$ and $E_{1}=x+y$ while $E_{2}=2 x-y$.) Then the system $E_{1}=f_{1}, E_{2}=f_{2}$ has the same solution set as

$$
\begin{equation*}
E_{1}=f_{1}, E_{2}+a E_{1}=f_{2}+a f_{1} \tag{2.3}
\end{equation*}
$$

Also the system $E_{1}=f_{1}, E_{2}=f_{2}$ has the same solutions as the system, $E_{2}=f_{2}, E_{1}=f_{1}$. The system $E_{1}=f_{1}, E_{2}=f_{2}$ has the same solution as the system $E_{1}=f_{1}, a E_{2}=a f_{2}$ provided $a \neq 0$.

Proof: If $\left(x_{1}, \cdots, x_{n}\right)$ solves $E_{1}=f_{1}, E_{2}=f_{2}$ then it solves the first equation in $E_{1}=$ $f_{1}, E_{2}+a E_{1}=f_{2}+a f_{1}$. Also, it satisfies $a E_{1}=a f_{1}$ and so, since it also solves $E_{2}=f_{2}$ it must solve $E_{2}+a E_{1}=f_{2}+a f_{1}$. Therefore, if $\left(x_{1}, \cdots, x_{n}\right)$ solves $E_{1}=f_{1}, E_{2}=f_{2}$ it must also solve $E_{2}+a E_{1}=f_{2}+a f_{1}$. On the other hand, if it solves the system $E_{1}=f_{1}$ and $E_{2}+a E_{1}=f_{2}+a f_{1}$, then $a E_{1}=a f_{1}$ and so you can subtract these equal quantities from both sides of $E_{2}+a E_{1}=f_{2}+a f_{1}$ to obtain $E_{2}=f_{2}$ showing that it satisfies $E_{1}=f_{1}, E_{2}=f_{2}$.

The second assertion of the theorem which says that the system $E_{1}=f_{1}, E_{2}=f_{2}$ has the same solution as the system, $E_{2}=f_{2}, E_{1}=f_{1}$ is seen to be true because it involves nothing more than listing the two equations in a different order. They are the same equations.

The third assertion of the theorem which says $E_{1}=f_{1}, E_{2}=f_{2}$ has the same solution as the system $E_{1}=f_{1}, a E_{2}=a f_{2}$ provided $a \neq 0$ is verified as follows: If $\left(x_{1}, \cdots, x_{n}\right)$ is a solution of $E_{1}=f_{1}, E_{2}=f_{2}$, then it is a solution to $E_{1}=f_{1}, a E_{2}=a f_{2}$ because the second system only involves multiplying the equation, $E_{2}=f_{2}$ by $a$. If $\left(x_{1}, \cdots, x_{n}\right)$ is a solution of $E_{1}=f_{1}, a E_{2}=a f_{2}$, then upon multiplying $a E_{2}=a f_{2}$ by the number $1 / a$, you find that $E_{2}=f_{2}$.

Stated simply, the above theorem shows that the elementary operations do not change the solution set of a system of equations.

### 2.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an augmented matrix. For example the suppose you want to find the solution for $x, y, z$ in $\mathbb{Z}_{5}$ to the system

$$
\begin{aligned}
x+\overline{3} y+z & =\overline{0}, \overline{2} x+y+\overline{3} z=\overline{3} \\
\overline{2} y+z & =\overline{4}
\end{aligned}
$$

To simplify, write the coefficients without the bar but do the arithmetic in $\mathbb{Z}_{5}$.

$$
\left(\begin{array}{llll}
1 & 3 & 1 & 0 \\
2 & 1 & 3 & 3 \\
0 & 2 & 1 & 4
\end{array}\right)
$$

It has exactly the same information as the original system but here the columns correspond to the variables and the rows correspond to the equations in the system.

To solve the system, we can use Gauss elimination in the usual way. The solution set is not changed by using the row operations. Take $3=-2$ times the top equation and add to the second.

$$
\left(\begin{array}{llll}
1 & 3 & 1 & 0 \\
0 & 0 & 1 & 3 \\
0 & 2 & 1 & 4
\end{array}\right)
$$

Now switch the bottom two rows.

$$
\left(\begin{array}{llll}
1 & 3 & 1 & 0 \\
0 & 2 & 1 & 4 \\
0 & 0 & 1 & 3
\end{array}\right)
$$

Then take 4 times the bottom row and add to the top two.

$$
\left(\begin{array}{llll}
1 & 3 & 0 & 2 \\
0 & 2 & 0 & 1 \\
0 & 0 & 1 & 3
\end{array}\right)
$$

Next multiply the second row by 3

$$
\left(\begin{array}{llll}
1 & 3 & 0 & 2 \\
0 & 1 & 0 & 3 \\
0 & 0 & 1 & 3
\end{array}\right)
$$

Now take 2 times the second row and add to the top.

$$
\left(\begin{array}{llll}
1 & 0 & 0 & 3 \\
0 & 1 & 0 & 3 \\
0 & 0 & 1 & 3
\end{array}\right)
$$

Therefore, the solution is $x=y=z=3$. How do you know when to stop? You certainly should stop doing row operations if you have gotten a matrix in row reduced echelon form described next. The leading entry of a row is the first nonzero row encountered when starting at the left entry and moving from left to right along the row.

## Definition 2.2.1 An augmented matrix is in row reduced echelon form if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.
3. All entries in a column above and below a leading entry are zero.
4. Each leading entry is a 1, the only nonzero entry in its column.

Echelon form means that the leading entries of successive rows fall from upper left to lower right.

Example 2.2.2 Here are some matrices which are in row reduced echelon form.

$$
\left(\begin{array}{llllll}
1 & 0 & 0 & 5 & 8 & 0 \\
0 & 0 & 1 & 2 & 7 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array}\right),\left(\begin{array}{llll}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Example 2.2.3 Here are matrices in echelon form which are not in row reduced echelon form but which are in echelon form.

$$
\left(\begin{array}{llllll}
1 & 0 & 6 & 5 & 8 & 2 \\
0 & 0 & 2 & 2 & 7 & 3 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array}\right),\left(\begin{array}{llll}
1 & 3 & 5 & 4 \\
0 & 2 & 0 & 7 \\
0 & 0 & 3 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Example 2.2.4 Here are some matrices which are not in echelon form.

$$
\left(\begin{array}{cccc}
0 & 0 & 0 & 0 \\
1 & 2 & 3 & 3 \\
0 & 1 & 0 & 2 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{array}\right),\left(\begin{array}{ccc}
1 & 2 & 3 \\
2 & 4 & -6 \\
4 & 0 & 7
\end{array}\right),\left(\begin{array}{cccc}
0 & 2 & 3 & 3 \\
1 & 5 & 0 & 2 \\
7 & 5 & 0 & 1 \\
0 & 0 & 1 & 0
\end{array}\right) .
$$

The following is the algorithm for obtaining a matrix which is in row reduced echelon form.

## Algorithm 2.2.5

This algorithm tells how to start with a matrix and do row operations on it in such a way as to end up with a matrix in row reduced echelon form.

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.
2. Use row operations to zero out the entries below the first pivot position.
3. Ignore the row containing the most recent pivot position identified and the rows above it. Repeat steps 1 and 2 to the remaining sub-matrix, the rectangular array of numbers obtained from the original matrix by deleting the rows you just ignored. Repeat the process until there are no more rows to modify. The matrix will then be in echelon form.
4. Moving from right to left, use the nonzero elements in the pivot positions to zero out the elements in the pivot columns which are above the pivots.
5. Divide each nonzero row by the value of the leading entry. The result will be a matrix in row reduced echelon form.

Sometimes there is no solution to a system of equations. When this happens, the system is said to be inconsistent.

Here is another example based on the use of row operations.
Example 2.2.6 Give the complete solution to the system of equations, $3 x-y-5 z=9$, $y-10 z=0$, and $-2 x+y=-6$.

The augmented matrix of this system is

$$
\left(\begin{array}{cccc}
3 & -1 & -5 & 9 \\
0 & 1 & -10 & 0 \\
-2 & 1 & 0 & -6
\end{array}\right)
$$

After doing row operations, to obtain row reduced echelon form,

$$
\left(\begin{array}{cccc}
1 & 0 & -5 & 3 \\
0 & 1 & -10 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

The equations corresponding to this reduced echelon form are $y=10 z$ and $x=3+5 z$. Apparently $z$ can equal any number. Lets call this number $t$. ${ }^{1}$ Therefore, the solution set of this system is $x=3+5 t, y=10 t$, and $z=t$ where $t$ is completely arbitrary. The system has an infinite set of solutions which are given in the above simple way. This is what it is all about, finding the solutions to the system.

In summary,
Definition 2.2.7 A system of linear equations is a list of equations,

$$
\begin{gathered}
a_{11} x_{1}+a_{12} x_{2}+\cdots+a_{1 n} x_{n}=b_{1} \\
a_{21} x_{1}+a_{22} x_{2}+\cdots+a_{2 n} x_{n}=b_{2} \\
\vdots \\
a_{m 1} x_{1}+a_{m 2} x_{2}+\cdots+a_{m n} x_{n}=b_{m}
\end{gathered}
$$

where $a_{i j}$ are numbers, and $b_{j}$ is a number. The above is a system of $m$ equations in the $n$ variables, $x_{1}, x_{2} \cdots, x_{n}$. Nothing is said about the relative size of $m$ and $n$. Written more simply in terms of summation notation, the above can be written in the form

$$
\sum_{j=1}^{n} a_{i j} x_{j}=f_{i}, i=1,2,3, \cdots, m
$$

It is desired to find $\left(x_{1}, \cdots, x_{n}\right)$ solving each of the equations listed.

[^0]As illustrated above, such a system of linear equations may have a unique solution, no solution, or infinitely many solutions and these are the only three cases which can occur for any linear system. Furthermore, you do exactly the same things to solve any linear system. You write the augmented matrix and do row operations until you get a simpler system in which it is possible to see the solution, usually obtaining a matrix in echelon or reduced echelon form. All is based on the observation that the row operations do not change the solution set. You can have more equations than variables, fewer equations than variables, etc. It doesn't matter. You always set up the augmented matrix and go to work on it.

Definition 2.2.8 A system of linear equations is called consistent if there exists a solution. It is called inconsistent if there is no solution.

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent would mean there is no choice of the variables which can satisfy each of the conditions.

### 2.3 When are Two Polynomials Relatively Prime?

Suppose you have two polynomials having coefficients in a field of scalars $\mathbb{F}$. How can you tell if they are relatively prime? One way is outlined in an earlier excercise. Here is another. By the method of partial fractions if $p(x), q(x)$ are relatively prime polynomials of degree at least 1 , then it follows from the partial fractions theorem, Proposition 1.14.2 that there is a partial fractions expansion of the following form.

$$
\frac{1}{p(x) q(x)}=\frac{a(x)}{p(x)}+\frac{b(x)}{q(x)}
$$

where the degree of $a(x)$ is smaller than the degree of $p(x)$ and the degree of $b(x)$ is smaller than the degree of $q(x)$. Conversely, if there is such a partial fractions expansion, then $1=a(x) q(x)+b(x) p(x)$ and these two polynomials are relatively prime. Checking the existence of such a partial fractions expansion is a simple example of finding a solution to a linear system of equations. Thus this is a question which can be resolved without having to factor the polynomials and instead uses the method of row operations to resolve the question.

Example 2.3.1 Consider $x^{2}-3 x+2$ and $x^{4}-4 x^{2}+4$. Then these are relatively prime if and only if there is such a partial fractions expansion just described. We would need

$$
\frac{1}{\left(x^{2}-3 x+2\right)\left(x^{4}-4 x^{2}+4\right)}=\frac{a x+b}{x^{2}-3 x+2}+\frac{c+d x+e x^{2}+f x^{3}}{x^{4}-4 x^{2}+4}
$$

Now multiply and write

$$
\begin{aligned}
1= & (a x+b)\left(x^{4}-4 x^{2}+4\right)+\left(c+d x+e x^{2}+f x^{3}\right)\left(x^{2}-3 x+2\right) \\
= & (a+f) x^{5}+(b-3 f+e) x^{4}+(d-4 a+2 f-3 e) x^{3} \\
& +(c-4 b-3 d+2 e) x^{2}+(4 a-3 c+2 d) x+(4 b+2 c)
\end{aligned}
$$

and so you would need to solve the following system of equations

$$
\begin{gathered}
a+f=0, b-3 f+e=0, d-4 a+2 f-3 e=0 \\
c-4 b-3 d+2 e=0,4 a-3 c+2 d=0,4 b+2 c=1
\end{gathered}
$$

Solving the system of equations, $a=-\frac{3}{4}, b=\frac{7}{4}, c=-3, d=-3, f=\frac{3}{4}, e=\frac{1}{2}$. It follows that these two polynomials are relatively prime. Note how it was not necessary to factor them to find out this information.

Example 2.3.2 Consider $x-1$ and $x^{2}-1$. These are clearly not relatively prime. Consider the above technique.

If they were relatively prime, then there would be a partial fractions expansion of the form $\frac{1}{(x-1)\left(x^{2}-1\right)}=\frac{a}{x-1}+\frac{b x+c}{x^{2}-1}$ and so, multiplying gives

$$
\begin{aligned}
& 1=a\left(x^{2}-1\right)+(b x+c)(x-1) \\
& =(a+b) x^{2}+(c-b) x-(a+c)
\end{aligned}
$$

Thus you would need to solve $a+b=0, c-b=0,-(a+c)=1$. However, there is no solution so we know these two are not relatively prime. Now again, I didn't need to factor these to draw this conclusion. Thus this gives another way to tell whether two polynomials are relatively prime. It turns out that being able to do this is useful, as is shown later.

### 2.4 Exercises

1. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\square$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$
\left(\begin{array}{ccccc|c}
\square & * & * & * & * & * \\
0 & \square & * & * & 0 & * \\
0 & 0 & \square & * & * & * \\
0 & 0 & 0 & 0 & \square & *
\end{array}\right)
$$

2. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\square$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$
\left(\begin{array}{ccccc|c}
\square & * & * & * & * & * \\
0 & \square & 0 & * & 0 & * \\
0 & 0 & 0 & \square & * & * \\
0 & 0 & 0 & 0 & \square & *
\end{array}\right)
$$

3. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\square$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If
consistent, is the solution unique?

$$
\left(\begin{array}{ccccc:c}
\square & * & * & * & * & * \\
0 & \boldsymbol{■} & * & * & 0 & * \\
0 & 0 & 0 & 0 & \boldsymbol{\square} & 0 \\
0 & 0 & 0 & 0 & * & \boldsymbol{\square}
\end{array}\right)
$$

4. Suppose a system of equations has fewer equations than variables. Must such a system be consistent? If so, explain why and if not, give an example which is not consistent.
5. If a system of equations has more equations than variables, can it have a solution? If so, give an example and if not, tell why not.
6. Find $h$ such that $\left(\begin{array}{cc|c}2 & h & 4 \\ 3 & 6 & 7\end{array}\right)$ is the augmented matrix of an inconsistent matrix.
7. Find $h$ such that $\left(\begin{array}{cc|c}1 & h & 3 \\ 2 & 4 & 6\end{array}\right)$ is the augmented matrix of a consistent matrix.
8. Find $h$ such that $\left(\begin{array}{cc|c}1 & 1 & 4 \\ 3 & h & 12\end{array}\right)$ is the augmented matrix of a consistent matrix.
9. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions. $\left(\begin{array}{cc|c}1 & h & 2 \\ 2 & 4 & k\end{array}\right)$.
10. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions. $\left(\begin{array}{ll|l}1 & 2 & 2 \\ 2 & h & k\end{array}\right)$.
11. Find the solution in $\mathbb{Z}_{5}$ to the following system of equations.

$$
\begin{gathered}
x+2 y+z-w=2 \\
x-y+z+w=1 \\
2 x+y-z=1 \\
4 x+2 y+z=0
\end{gathered}
$$

12. Find the solution to the following system in $\mathbb{Z}_{5}$

$$
\begin{gathered}
x+2 y+z-w=2 \\
x-y+z+w=0 \\
2 x+y-z=1 \\
4 x+2 y+z=3
\end{gathered}
$$

13. Find the general solution of the system whose augmented matrix is

$$
\left(\begin{array}{llll}
1 & 2 & 0 & 2 \\
1 & 3 & 4 & 2 \\
1 & 0 & 2 & 1
\end{array}\right)
$$

Find solutions in $\mathbb{Z}_{7}$.
14. Find the general solution of the system whose augmented matrix is

$$
\left(\begin{array}{llll}
1 & 2 & 0 & 2 \\
2 & 0 & 1 & 1 \\
3 & 2 & 1 & 3
\end{array}\right)
$$

in $\mathbb{Z}_{7}$.
15. Find the general solution in $\mathbb{Z}_{3}$ of the system whose augmented matrix is

$$
\left(\begin{array}{llll}
2 & 1 & 0 & 1 \\
1 & 0 & 1 & 2
\end{array}\right)
$$

16. Solve the system whose augmented matrix is

$$
\left(\begin{array}{llllll}
1 & 0 & 2 & 1 & 1 & 2 \\
0 & 1 & 0 & 1 & 2 & 1 \\
1 & 2 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 2 & 2
\end{array}\right)
$$

in $\mathbb{Z}_{3}$
17. Find the general solution of the system whose augmented matrix is

$$
\left(\begin{array}{cccccc}
1 & 0 & 2 & 1 & 1 & 2 \\
0 & 1 & 0 & 1 & 2 & 1 \\
0 & 2 & 0 & 0 & 1 & 3 \\
1 & -1 & 2 & 2 & 2 & 0
\end{array}\right)
$$

Find the solutions to this one in $\mathbb{Z}_{5}$.
18. Give the complete solution to the system of equations, $7 x+14 y+15 z=22,2 x+$ $4 y+3 z=5$, and $3 x+6 y+10 z=13$.
19. Give the complete solution to the system of equations, $3 x-y+4 z=6, y+8 z=0$, and $-2 x+y=-4$.
20. Give the complete solution to the system of equations, $9 x-2 y+4 z=-17,13 x-$ $3 y+6 z=-25$, and $-2 x-z=3$.
21. Give the complete solution to the system of equations, $65 x+84 y+16 z=546,81 x+$ $105 y+20 z=682$, and $84 x+110 y+21 z=713$.
22. Give the complete solution to the system of equations, $8 x+2 y+3 z=-3,8 x+3 y+$ $3 z=-1$, and $4 x+y+3 z=-9$.
23. Give the complete solution to the system of equations, $-8 x+2 y+5 z=18,-8 x+$ $3 y+5 z=13$, and $-4 x+y+5 z=19$.
24. Give the complete solution to the system of equations, $3 x-y-2 z=3, y-4 z=0$, and $-2 x+y=-2$.
25. Give the complete solution to the system of equations, $-9 x+15 y=66,-11 x+18 y=$ $79,-x+y=4$, and $z=3$.
26. Give the complete solution to the system of equations, $-19 x+8 y=-108,-71 x+$ $30 y=-404,-2 x+y=-12,4 x+z=14$.
27. Consider the system $-5 x+2 y-z=0$ and $-5 x-2 y-z=0$. Both equations equal zero and so $-5 x+2 y-z=-5 x-2 y-z$ which is equivalent to $y=0$. Thus $x$ and $z$ can equal anything. But when $x=1, z=-4$, and $y=0$ are plugged in to the equations, it doesn't work. Why?
28. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. Four times the weight of Gaston plus the weight of Siegfried equals 290 pounds. Brunhilde would balance all three of the others. Find the weights of the four sisters.
29. The steady state temperature, $u$ in a plate solves Laplace's equation, $\Delta u=0$. One way to approximate the solution which is often used is to divide the plate into a square mesh and require the temperature at each node to equal the average of the temperature at the four adjacent nodes. This procedure is justified by the mean value property of harmonic functions. In the following picture, the numbers represent the observed temperature at the indicated nodes. Your task is to find the temperature at the interior nodes, indicated by $x, y, z$, and $w$. One of the equations is $z=\frac{1}{4}(10+0+w+x)$.

30. Consider the following diagram of four circuits.


Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as $\Omega$. The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by $I_{1}, I_{2}, I_{3}, I_{4}$ and it is understood that the motion is in the counter clockwise direction. If $I_{k}$ ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop. In the above diagram, the top left circuit should give the equation

$$
2 I_{2}-2 I_{1}+5 I_{2}-5 I_{3}+3 I_{2}=5
$$

For the circuit on the lower left, you should have

$$
4 I_{1}+I_{1}-I_{4}+2 I_{1}-2 I_{2}=-10
$$

Write equations for each of the other two circuits and then give a solution to the resulting system of equations. You might use a computer algebra system to find the solution. It might be more convenient than doing it by hand.
31. Consider the following diagram of three circuits.


Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as $\Omega$. The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by $I_{1}, I_{2}, I_{3}$ and it is understood that the motion is in the counter clockwise direction. If $I_{k}$ ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that
The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop. Find $I_{1}, I_{2}, I_{3}$.
32. Determine whether $x^{3}-x^{2}-x+1$ and $x^{3}-x^{2}+x-1$ are relatively prime. They are obviously not. However, use the technique of partial fractions to verify this or use the earlier method in an earlier problem for finding the greatest common divisor.

## Chapter 3

## Vector Spaces

It is time to consider the idea of an abstract vector space which is something which has two operations satisfying the following vector space axioms.

Definition 3.0.1 A vector space is an Abelian group of "vectors" satisfying the axioms of an Abelian group,

$$
v+w=w+v
$$

the commutative law of addition,

$$
(v+w)+z=v+(w+z),
$$

the associative law for addition,

$$
v+0=v,
$$

the existence of an additive identity,

$$
v+(-v)=0
$$

the existence of an additive inverse, along with a field of "scalars" $\mathbb{F}$ which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)

$$
\begin{align*}
\alpha(v+w) & =\alpha v+\alpha v,  \tag{3.1}\\
(\alpha+\beta) v & =\alpha v+\beta v,  \tag{3.2}\\
\alpha(\beta v) & =\alpha \beta(v),  \tag{3.3}\\
1 v & =v . \tag{3.4}
\end{align*}
$$

For example, any field is a vector space having field of scalars equal to the field itself. The field of scalars is often $\mathbb{R}$ or $\mathbb{C}$ and the vector space will be called real or complex depending on whether the field is $\mathbb{R}$ or $\mathbb{C}$. However, other fields are also possible. For example, one could use the field of rational numbers or even the field of the integers mod $p$ for $p$ a prime. A vector space is also called a linear space. These axioms do not tell us anything about what is being considered. Nevertheless, one can prove some fundamental properties just based on these vector space axioms.

Proposition 3.0.2 In any vector space, 0 is unique, $-x$ is unique, $0 x=0$, and $(-1) x=-x$.
Proof: Suppose $0^{\prime}$ is also an additive identity. Then for 0 the additive identity in the axioms,

$$
0^{\prime}=0^{\prime}+0=0
$$

Next suppose $x+y=0$. Then add $-x$ to both sides.

$$
-x=-x+(x+y)=(-x+x)+y=0+y=y
$$

Thus if $y$ acts like the additive inverse, it is the additive inverse.

$$
0 x=(0+0) x=0 x+0 x
$$

Now add $-0 x$ to both sides. This gives $0=0 x$. Finally,

$$
(-1) x+x=(-1) x+1 x=(-1+1) x=0 x=0
$$

By the uniqueness of the additive inverse shown earlier, $(-1) x=-x$.
If you are interested in considering other fields, you should have some examples other than $\mathbb{C}, \mathbb{R}, \mathbb{Q}$. Some of these are discussed in the following exercises. If you are happy with only considering $\mathbb{R}$ and $\mathbb{C}$, skip these exercises. Here is an important example which gives the typical vector space.

Example 3.0.3 Let $\Omega$ be a nonempty set and define $V$ to be the set of functions defined on $\Omega$. Letting $a, b, c$ be scalars coming from a field $\mathbb{F}$ and $f, g, h$ functions, the vector operations are defined as

$$
\begin{aligned}
(f+g)(x) & \equiv f(x)+g(x) \\
(a f)(x) & \equiv a(f(x))
\end{aligned}
$$

Then this is an example of a vector space. Note that the set where the functions have their values can be any vector space having field of scalars $\mathbb{F}$.

To verify this, check the axioms.

$$
(f+g)(x)=f(x)+g(x)=g(x)+f(x)=(g+f)(x)
$$

Since $x$ is arbitrary, $f+g=g+f$.

$$
\begin{aligned}
& ((f+g)+h)(x) \equiv(f+g)(x)+h(x)=(f(x)+g(x))+h(x) \\
= & f(x)+(g(x)+h(x))=(f(x)+(g+h)(x))=(f+(g+h))(x)
\end{aligned}
$$

and so $(f+g)+h=f+(g+h)$. Let 0 denote the function which is given by $0(x)=0$. Then this is an additive identity because

$$
(f+0)(x)=f(x)+0(x)=f(x)
$$

and so $f+0=f$. Let $-f$ be the function which satisfies $(-f)(x) \equiv-f(x)$. Then

$$
(f+(-f))(x) \equiv f(x)+(-f)(x) \equiv f(x)+-f(x)=0
$$

Hence $f+(-f)=0$.

$$
((a+b) f)(x) \equiv(a+b) f(x)=a f(x)+b f(x) \equiv(a f+b f)(x)
$$

and so $(a+b) f=a f+b f$.

$$
\begin{gathered}
(a(f+g))(x) \equiv a(f+g)(x) \equiv a(f(x)+g(x)) \\
=a f(x)+b g(x) \equiv(a f+b g)(x)
\end{gathered}
$$

and so $a(f+g)=a f+b g$.

$$
((a b) f)(x) \equiv(a b) f(x)=a(b f(x)) \equiv(a(b f))(x)
$$

so $(a b f)=a(b f)$. Finally $(1 f)(x) \equiv 1 f(x)=f(x)$ so $1 f=f$.
As above, $\mathbb{F}$ will be a field. It illustrates the important example of $\mathbb{F}^{n}$, a vector space with field of scalars $\mathbb{F}$. It is a case of the above general consideration involving functions. Indeed, you simply let $\Omega=\{1,2, \cdots, n\}$. We write such a function $f:\{1,2, \cdots, n\} \rightarrow \mathbb{F}$ in as an ordered list of numbers $(f(1), \cdots, f(n))$. The definition, incorporating the usual notation is as follows.

Definition 3.0.4 Define $\mathbb{F}^{n} \equiv\left\{\left(x_{1}, \cdots, x_{n}\right): x_{j} \in \mathbb{F}\right.$ for $\left.j=1, \cdots, n\right\}$.

$$
\left(x_{1}, \cdots, x_{n}\right)=\left(y_{1}, \cdots, y_{n}\right)
$$

if and only iffor all $j=1, \cdots, n, x_{j}=y_{j}$. When $\left(x_{1}, \cdots, x_{n}\right) \in \mathbb{F}^{n}$, it is conventional to denote $\left(x_{1}, \cdots, x_{n}\right)$ by the single bold face letter $\boldsymbol{x}$. The numbers, $x_{j}$ are called the coordinates. Elements in $\mathbb{F}^{n}$ are called vectors. The set

$$
\{(0, \cdots, 0, t, 0, \cdots, 0): t \in \mathbb{R}\}
$$

for $t$ in the $i^{\text {th }}$ slot is called the $i^{\text {th }}$ coordinate axis. The point $\mathbf{0} \equiv(0, \cdots, 0)$ is called the origin. Note that this can be considered as the set of $\mathbb{F}$ valued functions defined on $(1,2, \cdots, n)$. When the ordered list $\left(x_{1}, \cdots, x_{n}\right)$ is considered, it is just a way to say that $f(1)=x_{1}, f(2)=x_{2}$ and so forth. Thus it is a case of the typical example of a vector space mentioned above.

### 3.1 Linear Combinations of Vectors, Independence

The fundamental idea in linear algebra is the following notion of a linear combination.
Definition 3.1.1 Let $x_{1}, \cdots, x_{n}$ be vectors in a vector space. A finite linear combination of these vectors is a vector which is of the form $\sum_{j=1}^{n} a_{j} x_{j}$ where the $a_{j}$ are scalars. In short, it is a sum of scalars times vectors. span $\left(x_{1}, \cdots, x_{n}\right)$ denotes the set of all linear combinations of the vectors $x_{1}, \cdots, x_{n}$. More generally, if $S$ is any set of vectors, $\operatorname{span}(S)$ consists of all finite linear combinations of vectors from $S$.

Definition 3.1.2 Let $(V, \mathbb{F})$ be a vector space and its field of scalars. Then $S \subseteq V$ is said to be linearly independent if whenever $\left\{v_{1}, \cdots, v_{m}\right\} \subseteq V$ with the $v_{i}$ distinct, then there is only one way to have a linear combination $\sum_{i=1}^{n} c_{i} v_{i}=0$ and this is to have each $c_{i}=0$. More succinctly, if $\sum_{i=1}^{n} c_{i} v_{i}=0$ then each $c_{i}=0$. A set $S \subseteq V$ is linearly dependent if it is not linearly independent. That is, there is some subset of $S\left\{v_{1}, \cdots, v_{n}\right\}$ and scalars $c_{i}$ not all zero such that $\sum_{i=1}^{n} c_{i} v_{i}=0$.

The following is a useful equivalent description of what it means to be independent.
Proposition 3.1.3 A set of vectors $S$ is independent if and only if no vector is a linear combination of the others.

Proof: $\Rightarrow$ Suppose $S$ is linearly independent. Could you have for some

$$
\left\{u_{1}, \cdots, u_{r}\right\} \subseteq S
$$

$u_{i}=\sum_{j \neq i} c_{j} u_{j}$ ? No. This is not possible because if the above holds, then you would have $0=(-1) u_{i}+\sum_{j \neq i} c_{j} u_{j}$ in contradiction to the assumption that $\left\{u_{1}, \cdots, u_{r}\right\}$ is linearly independent.
$\Leftarrow$ Suppose now that no vector in $S$ is a linear combination of the others. Suppose $\sum_{i=1}^{n} c_{i} u_{i}=0$ where each $u_{i} \in S$. It is desired to show that whenever this happens, each $c_{i}=0$. Could any of the $c_{i}$ be non zero? No. If $c_{k} \neq 0$, then you would have $\sum_{i=1}^{n} \frac{c_{i}}{c_{k}} u_{i}=0$ and so $u_{k}=\sum_{i \neq k}-\frac{c_{i}}{c_{k}} u_{i}$ showing that one can obtain $u_{k}$ as a linear combination of the other vectors after all. It follows that all $c_{i}=0$ and so $\left\{u_{1}, \cdots, u_{r}\right\}$ is linearly independent.

Example 3.1.4 Determine whether the real valued functions defined on $\mathbb{R}$ given by the polynomials

$$
x^{2}+2 x+1, x^{2}+2 x, x^{2}+x+1
$$

are independent with field of scalars $\mathbb{R}$.
Suppose $a\left(x^{2}+2 x+1\right)+b\left(x^{2}+2 x\right)+c\left(x^{2}+x+1\right)=0$ then differentiate both sides to obtain $a(2 x+2)+b(2 x+2)+c(2 x+1)=0$. Now differentiate again. This yields $2 a+$ $2 b+2 c=0$. In the second equation, let $x=-1$. Then $-c=0$ so $c=0$. Thus

$$
\begin{array}{r}
a\left(x^{2}+2 x+1\right)+b\left(x^{2}+2 x\right)=0 \\
a+b=0
\end{array}
$$

Now let $x=0$ in the top equation to find that $a=0$. Then from the bottom equation, it follows that $b=0$ also. Thus the three functions are linearly independent.

The main theorem is the following, called the replacement or exchange theorem. It uses the argument of the second half of the above proposition repeatedly.

Theorem 3.1.5 Let $\left\{u_{1}, \cdots, u_{r}\right\},\left\{v_{1}, \cdots, v_{s}\right\}$ be subsets of a vector space $V$ with field of scalars $\mathbb{F}$ and suppose $\left\{u_{1}, \cdots, u_{r}\right\}$ is linearly independent and each $u_{i} \in \operatorname{span}\left(v_{1}, \cdots, v_{s}\right)$. Then $r \leq s$. In words, linearly independent sets are no longer than spanning sets.

Proof: Say $r>s$. By assumption, $u_{1}=\sum_{i} b_{i} v_{i}$. Not all of the $b_{i}$ can equal 0 because if this were so, you would have $u_{1}=0$ which would violate the assumption that $\left\{u_{1}, \cdots, u_{r}\right\}$ is linearly independent. You could write

$$
1 u_{1}+0 u_{2}+\cdots+0 u_{r}=0
$$

since $u_{1}=0$. Thus some $v_{i}$ say $v_{i_{1}}$ is a linear combination of the vector $u_{1}$ along with the $v_{j}$ for $j \neq i$. It follows that the span of $\left\{u_{1}, v_{1}, \cdots, \hat{v}_{i_{1}}, \cdots, v_{n}\right\}$ includes each of the $u_{i}$ where the hat indicates that $v_{i_{1}}$ has been omitted from the list of vectors. Now suppose each $u_{i}$ is in

$$
\operatorname{span}\left(u_{1} \cdots, u_{k}, v_{1}, \cdots, \hat{v}_{i_{1}}, \cdots, \hat{v}_{i_{k}} \cdots, v_{s}\right)
$$

where the vectors $\hat{v}_{i_{1}}, \cdots, \hat{v}_{i_{k}}$ have been omitted for $k \leq s$. Then there are scalars $c_{i}$ and $d_{i}$ such that

$$
u_{k+1}=\sum_{i=1}^{k} c_{i} u_{i}+\sum_{j \notin\left\{i_{1}, \cdots, i_{k}\right\}} d_{j} v_{j}
$$

By the assumption that $\left\{u_{1}, \cdots, u_{r}\right\}$ is linearly independent, not all of the $d_{j}$ can equal 0 . Why? Therefore, there exists $i_{k+1} \notin\left\{i_{1}, \cdots, i_{k}\right\}$ such that $d_{i_{k}} \neq 0$. Hence one can solve for $v_{i_{k+1}}$ as a linear combination of $\left\{u_{1}, \cdots, u_{r}\right\}$ and the $v_{j}$ for $j \notin\left\{i_{1}, \cdots, i_{k}, i_{k+1}\right\}$. Thus we can replace this $v_{i_{k+1}}$ by a linear combination of these vectors, and so the $u_{j}$ are in

$$
\operatorname{span}\left(u_{1}, \cdots, u_{k}, u_{k+1}, v_{1}, \cdots, \hat{v}_{i_{1}}, \cdots, \hat{v}_{i_{k}}, \hat{v}_{i_{k+1}}, \cdots, v_{s}\right)
$$

Continuing this replacement process, it follows that since $r>s$, one can eliminate all of the vectors $\left\{v_{1}, \cdots, v_{s}\right\}$ and obtain that the $u_{i}$ are contained in span $\left(u_{1}, \cdots, u_{s}\right)$. But then you would have $u_{s+1} \in \operatorname{span}\left(u_{1}, \cdots, u_{s}\right)$ which is impossible since these vectors $\left\{u_{1}, \cdots, u_{r}\right\}$ are linearly independent. It follows that $r \leq s$.

Next is the definition of dimension and basis of a vector space.

Definition 3.1.6 Let $V$ be a vector space with field of scalars $\mathbb{F}$. A subset $S$ of $V$ is a basis for $V$ means that

1. $\operatorname{span}(S)=V$
2. $S$ is linearly independent.

The plural of basis is bases. It is this way to avoid hissing when referring to it.
The dimension of a vector space is the number of vectors in a basis. A vector space is finite dimensional if it equals the span of some finite set of vectors.

Lemma 3.1.7 Let $S$ be a linearly independent set of vectors in a vector space $V$. Suppose $v \notin \operatorname{span}(S)$. Then $\{S, v\}$ is also a linearly independent set of vectors.

Proof: Suppose $\left\{u_{1}, \cdots, u_{n}, v\right\}$ is a finite subset of $S$ and $a v+\sum_{i=1}^{n} b_{i} u_{i}=0$ where $a, b_{1}, \cdots, b_{n}$ are scalars. Does it follow that each of the $b_{i}$ equals zero and that $a=0$ ? If so, then this shows that $\{S, v\}$ is indeed linearly independent. First note that $a=0$ since if not, you could write $v=\sum_{i=1}^{n}-\frac{b_{i}}{a} u_{i}$ contrary to the assumption that $v \notin \operatorname{span}(S)$. Hence you have $a=0$ and also $\sum_{i} b_{i} u_{i}=0$. But $S$ is linearly independent and so by assumption each $b_{i}=0$.

Proposition 3.1.8 Let $V$ be a finite dimensional nonzero vector space with field of scalars $\mathbb{F}$. Then it has a basis and also any two bases have the same number of vectors so the above definition of a basis is well defined.

Proof: Pick $u_{1} \neq 0$. If $\operatorname{span}\left(u_{1}\right)=V$, then this is a basis. If not, there exists $u_{2} \notin$ $\operatorname{span}\left(u_{1}\right)$. Then by Lemma 3.1.7, $\left\{u_{1}, u_{2}\right\}$ is linearly independent. If $\operatorname{span}\left(u_{1}, u_{2}\right)=V$, stop. You have a basis. Otherwise, there exists $u_{3} \notin \operatorname{span}\left(u_{1}, u_{2}\right)$. Then by Lemma 3.1.7, $\left\{u_{1}, u_{2}, u_{3}\right\}$ is linearly independent. Continue this way. Eventually the process yields $\left\{u_{1}, \cdots, u_{n}\right\}$ which is linearly independent and $\operatorname{span}\left(u_{1}, \cdots, u_{n}\right)=V$. Otherwise there would exist a linearly independent set of $k$ vectors for all $k$. However, by assumption, there is a finite set of vectors $\left\{v_{1}, \cdots, v_{s}\right\}$ such that $\operatorname{span}\left(v_{1}, \cdots, v_{s}\right)=V$. Therefore, $k \leq s$. Thus there is a basis for $V$.

If $\left\{v_{1}, \cdots, v_{s}\right\},\left\{u_{1}, \cdots, u_{r}\right\}$ are two bases, then since they both span $V$ and are both linearly independent, it follows from Theorem 3.1.5 that $r \leq s$ and $s \leq r$.

As a specific example, consider $\mathbb{F}^{n}$ as the vector space. As mentioned above, these are the mappings from $(1, \cdots, n)$ to the field $\mathbb{F}$. It was shown in Example 3.0.3 that this is indeed a vector space with field of scalars $\mathbb{F}$. We usually think of this $\mathbb{F}^{n}$ as the set of ordered $n$ tuples

$$
\left\{\left(x_{1}, \cdots, x_{n}\right): x_{i} \in \mathbb{F}\right\}
$$

with addition and scalar mutiplication defined as

$$
\begin{gathered}
\left(x_{1}, \cdots, x_{n}\right)+\left(\hat{x}_{1}, \cdots, \hat{x}_{n}\right)=\left(x_{1}+\hat{x}_{1}, \cdots, x_{n}+\hat{x}_{n}\right) \\
\alpha\left(x_{1}, \cdots, x_{n}\right)=\left(\alpha x_{1}, \cdots, \alpha x_{n}\right)
\end{gathered}
$$

Also, when referring to vectors in $\mathbb{F}^{n}$, it is customary to denote them as bold faced letters. It is more convenient to write these vectors in $\mathbb{F}^{n}$ as columns of numbers rather than as rows as done earlier. Thus

$$
\boldsymbol{x}=\left(\begin{array}{c}
x_{1} \\
\vdots \\
x_{n}
\end{array}\right) \equiv\left(\begin{array}{lll}
x_{1} & \cdots & x_{n}
\end{array}\right)^{T}
$$

Observation 3.1.9 $\mathbb{F}^{n}$ has dimension $n$. To see this, note that a basis is $\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{n}$ where

$$
e_{i} \equiv\left(\begin{array}{lllll}
0 & \cdots & 1 & \cdots & 0
\end{array}\right)^{T}
$$

the vector in $\mathbb{F}^{n}$ which has a 1 in the $i^{\text {th }}$ position and a zero everywhere else.
To see this, note that

$$
\left(\begin{array}{c}
x_{1} \\
x_{2} \\
\vdots \\
x_{n}
\end{array}\right)=\sum_{i=1}^{n} x_{i} \boldsymbol{e}_{i}
$$

and that if $\mathbf{0}=\sum_{i=1}^{n} x_{i} \boldsymbol{e}_{i}$ then

$$
\left(\begin{array}{c}
x_{1} \\
x_{2} \\
\vdots \\
x_{n}
\end{array}\right)=\left(\begin{array}{c}
0 \\
0 \\
\vdots \\
0
\end{array}\right)
$$

so each $x_{i}$ is zero. Thus this set of vectors is a spanning set and is linearly independent so it is a basis. There are $n$ of these vectors and so the dimension of $\mathbb{F}^{n}$ is indeed $n$.

There is a fundamental observation about linear combinations of vectors in $\mathbb{F}^{n}$ which is stated next.

Theorem 3.1.10 Let $a_{1}, \cdots, a_{n}$ be vectors in $\mathbb{F}^{m}$ where $m<n$. Then there exist scalars $x_{1}, \cdots, x_{n}$ not all equal to zero such that $x_{1} \boldsymbol{a}_{1}+\cdots+x_{n} \boldsymbol{a}_{n}=\mathbf{0}$.

Proof: If the conclusion were not so, then by definition, $\left\{a_{1}, \cdots, a_{n}\right\}$ would be independent. However, there is a spanning set with only $m$ vectors, namely $\left\{\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{m}\right\}$ contrary to Theorem 3.1.5. Since these vectors cannot be independent, they must be dependent which is the conclusion of the theorem.

### 3.2 Subspaces

The notion of a subspace is of great importance in applications. Here is what is meant by a subspace.

Definition 3.2.1 Let $V$ be a vector space with field of scalars $\mathbb{F}$. Then let $W \subseteq V, W \neq \emptyset$. That is, $W$ is a non-empty subset of $V$. Then $W$ is a subspace of $V$ if whenever $\alpha, \beta$ are scalars and $u, v$ are vectors in $W$, it follows that $\alpha u+\beta v \in W$. In words, $W$ is closed with respect to linear combinations.

The fundamental result about subspaces is that they are themselves vector spaces.
Theorem 3.2.2 Let $W$ be a non-zero subset of $V$ a vector space with field of scalars $\mathbb{F}$. Then it is a subspace if and only if it is itself a vector space with field of scalars $\mathbb{F}$.

Proof: Suppose $W$ is a subspace. Why is it a vector space? To be a vector space, the operations of addition and scalar multiplication must satisfy the axioms for a vector space. However, all of these are obvious because it is a subset of $V$. The only thing which is not obvious is whether 0 is in $W$ and whether $-u \in W$ whenever $u$ is. But these follow right away from Proposition 3.0.2 because if $u \in W,(-1) u=-u \in W$ by the fact that $W$ is closed with respect to linear combinations, in particular multiplication by the scalar -1 . Similarly, take $u \in W$. Then $0=0 u \in W$. As to + being an operation on $W$, this also follows because for $u, v \in W, u+v \in W$. Thus if it is a subspace, it is indeed a vector space.

Conversely, suppose it is a vector space. Then by definition, it is closed with respect to linear combinations and so it is a subspace.

This leads to the following simple result.
Proposition 3.2.3 Let $W$ be a nonzero subspace of a finite dimensional vector space $V$ with field of scalars $\mathbb{F}$. Then $W$ is also a finite dimensional vector space.

Proof: Suppose $\operatorname{span}\left(v_{1}, \cdots, v_{n}\right)=V$. Using the same construction of Proposition 3.1.8, the same process must stop after $k \leq n$ steps since otherwise one could obtain a linearly independent set of vectors with more vectors in it than a spanning set. Thus it has a basis with no more than $n$ vectors.

Example 3.2.4 Show that $W=\left\{(x, y, z) \in \mathbb{R}^{3}: x-2 y-z=0\right\}$ is a subspace of $\mathbb{R}^{3}$. Find a basis for it.

You have from the equation that $x=2 y+z$ and so any vector in this set is of the form

$$
\left(\begin{array}{c}
2 y+z \\
y \\
z
\end{array}\right): y, z \in \mathbb{R}
$$

Conversely, any vector which is of the above form satisfies the condition to be in $W$. Therefore, $W$ is of the form

$$
y\left(\begin{array}{l}
2 \\
1 \\
0
\end{array}\right)+z\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)
$$

where $y, z$ are scalars. Hence it equals the span of the two vectors in $\mathbb{R}^{3}$ in the above. Are the two vectors linearly independent? If so, they will be a basis. Suppose then that

$$
y\left(\begin{array}{l}
2 \\
1 \\
0
\end{array}\right)+z\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

Then from the second position, $y=0$. It follows then that $z=0$ also and so the two vectors form a linearly independent set. Hence a basis for $W$ is

$$
\left\{\left(\begin{array}{l}
2 \\
1 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right\}
$$

The dimension of this subspace is also 2 .

Example 3.2.5 Show that

$$
\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right),\left(\begin{array}{l}
1 \\
3 \\
3
\end{array}\right),\left(\begin{array}{l}
0 \\
1 \\
4
\end{array}\right)
$$

is a basis for $\mathbb{R}^{3}$.
There are two things to show, that the set of vectors is independent and that it spans $\mathbb{R}^{3}$. Thus we need to verify that there is exactly one solution to the system of equations

$$
x\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)+y\left(\begin{array}{l}
1 \\
3 \\
3
\end{array}\right)+z\left(\begin{array}{l}
0 \\
1 \\
4
\end{array}\right)=\left(\begin{array}{l}
a \\
b \\
c
\end{array}\right)
$$

for any choice of the right side. Recall how to do this. You set up the augmented matrix and then row reduce it.

$$
\left(\begin{array}{llll}
1 & 1 & 0 & a \\
1 & 3 & 1 & b \\
1 & 3 & 4 & c
\end{array}\right)
$$

After some row operations, this yields

$$
\left(\begin{array}{cccc}
1 & 0 & 0 & \frac{3}{2} a-\frac{2}{3} b+\frac{1}{6} c \\
0 & 1 & 0 & \frac{2}{3} b-\frac{1}{2} a-\frac{1}{6} c \\
0 & 0 & 1 & \frac{1}{3} c-\frac{1}{3} b
\end{array}\right)
$$

Thus there is a unique solution to the system of equations. This shows that the set of vectors is a basis because one solution when the right side of the system equals the zero vector is $x=y=z=0$. Therefore, from what was just done, it is the only solution and so the vectors are linearly independent. As to the span of the vectors equalling $\mathbb{R}^{3}$, this was just shown also.

Example 3.2.6 Show that

$$
\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right),\left(\begin{array}{l}
1 \\
1 \\
3
\end{array}\right),\left(\begin{array}{c}
1 \\
1 \\
-4
\end{array}\right)
$$

is not a basis for $\mathbb{R}^{3}$.
You can do it the same way. It is really a question about whether there exists a unique solution to the system

$$
x\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)+y\left(\begin{array}{l}
1 \\
3 \\
3
\end{array}\right)+z\left(\begin{array}{c}
1 \\
1 \\
-4
\end{array}\right)=\left(\begin{array}{l}
a \\
b \\
c
\end{array}\right)
$$

for any choice of the right side. The augmented matrix is

$$
\left(\begin{array}{cccc}
1 & 1 & 1 & a \\
1 & 1 & 1 & b \\
1 & 3 & -4 & c
\end{array}\right)
$$

After row reduction, this yields

$$
\left(\begin{array}{cccc}
1 & 1 & 1 & a \\
0 & 2 & -5 & c-a \\
0 & 0 & 0 & b-a
\end{array}\right)
$$

Thus there is no solution to the equation unless $b=a$. It follows the span of the given vectors is not all of $\mathbb{R}^{3}$ and so this cannot be a basis.

Example 3.2.7 Show that

$$
\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right),\left(\begin{array}{l}
1 \\
1 \\
3
\end{array}\right)
$$

is not a basis for $\mathbb{R}^{3}$.
If the span of these vectors were all of $\mathbb{R}^{3}$, this would contradict Theorem 3.1.5 because it would be a spanning set which is shorter than a linearly independent set $\left\{e_{1}, e_{2}, e_{3}\right\}$.

Example 3.2.8 Show that

$$
\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right),\left(\begin{array}{l}
1 \\
1 \\
3
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)
$$

is not a basis for $\mathbb{R}^{3}$.
If it were a basis, then it would need to be linearly independent but this cannot happen because it would contradict Theorem 3.1.5 by being an independent set of vectors which is longer than a spanning set.

Theorem 3.2.9 If $V$ is an $n$ dimensional vector space and if $\left\{u_{1}, \cdots, u_{n}\right\}$ is a linearly independent set, then it is a basis. If $m>n$ then $\left\{v_{1}, \cdots, v_{m}\right\}$ is a dependent set. If $V=\operatorname{span}\left(w_{1}, \cdots, w_{m}\right)$, then $m \geq n$ and there is a subset $\left\{u_{1}, \cdots, u_{n}\right\} \subseteq\left\{w_{1}, \cdots, w_{m}\right\}$ such that $\left\{u_{1}, \cdots, u_{n}\right\}$ is a basis. If $\left\{u_{1}, \cdots, u_{k}\right\}$ is linearly independent, then there exists $\left\{u_{1}, \cdots, u_{k}, \cdots, u_{n}\right\}$ which is a basis.

Proof: Say $\left\{u_{1}, \cdots, u_{n}\right\}$ is linearly independent. Is span $\left(u_{1}, \cdots, u_{n}\right)=V$ ? If not, there would be $w \notin \operatorname{span}\left(u_{1}, \cdots, u_{n}\right)$ and then by Lemma 3.1.7 $\left\{u_{1}, \cdots, u_{n}, w\right\}$ would be linearly independent which contradicts Theorem 3.1.5. As to the second claim, $\left\{v_{1}, \cdots, v_{m}\right\}$ cannot be linearly independent because this would contradict Theorem 3.1.5 and so it is dependent.

Now say $V=\operatorname{span}\left(w_{1}, \cdots, w_{m}\right)$. By Theorem 3.1.5 again, you must have $m \geq n$ since spanning sets are at least as long as linearly independent sets, one of which is a basis having
$n$ vectors. If $w_{1}$ is in the span of the other vectors, delete it. Then consider $w_{2}$. If it is in the span of the other vectors, delete it. Continue this way till a shorter list is obtained with the property that no vector is a linear combination of the others, but its span is still $V$. By Proposition 3.1.3, the resulting list of vectors is linearly independent and is therefore, a basis since it spans $V$.

Now suppose for $k<n,\left\{u_{1}, \cdots, u_{k}\right\}$ is linearly independent. Follow the process of Proposition 3.1.8, adding in vectors not in the span and obtaining successively larger linearly independent sets till the process ends. The resulting list must be a basis.

### 3.3 Exercises

1. Show that the following are subspaces of the set of all functions defined on $[a, b]$.
(a) polynomials of degree $\leq n$
(b) polynomials
(c) continuous functions
(d) differentiable functions
2. Show that every subspace of a finite dimensional vector space $V$ is the span of some vectors. It was done above but go over it in your own words.
3. In $\mathbb{R}^{2}$ define a funny addition by $(x, y)+(\hat{x}, \hat{y}) \equiv(3 x+3 \hat{x}, y+\hat{y})$ and let scalar multiplication be the usual thing. Would this be a vector space with these operations?
4. Determine which of the following are subspaces of $\mathbb{R}^{m}$ for some $m . a, b$ are just given numbers in what follows.
(a) $\left\{(x, y) \in \mathbb{R}^{2}: a x+b y=0\right\}$
(b) $\left\{(x, y) \in \mathbb{R}^{2}: a x+b y \geq y\right\}$
(c) $\left\{(x, y) \in \mathbb{R}^{2}: a x+b y=1\right\}$
(d) $\left\{(x, y) \in \mathbb{R}^{2}: x y=0\right\}$
(e) $\left\{(x, y) \in \mathbb{R}^{2}: y \geq 0\right\}$
(f) $\left\{(x, y) \in \mathbb{R}^{2}: x>0\right.$ or $\left.y>0\right\}$
(g) For those who recall the cross product, $\left\{\boldsymbol{x} \in \mathbb{R}^{3}: \boldsymbol{a} \times \boldsymbol{x}=\mathbf{0}\right\}$.
(h) For those who recall the dot product, $\left\{\boldsymbol{x} \in \mathbb{R}^{m}: \boldsymbol{x} \cdot \boldsymbol{a}=\mathbf{0}\right\}$
(i) $\left\{\boldsymbol{x} \in \mathbb{R}^{n}: \boldsymbol{x} \cdot \boldsymbol{a} \geq \mathbf{0}\right\}$
(j) $\left\{\boldsymbol{x} \in \mathbb{R}^{m}: \boldsymbol{x} \cdot \boldsymbol{s}=\mathbf{0}\right.$ for all $\left.s \in S, S \neq \emptyset, S \subseteq \mathbb{R}^{m}\right\}$. This is known as $S^{\perp}$.
5. Show that $\left\{(x, y, z) \in \mathbb{R}^{3}: x+y-z=0\right\}$ is a subspace and find a basis for it.
6. In the subspace of polynomials on $[0,1]$, show that the vectors $\left\{1, x, x^{2}, x^{3}\right\}$ are linearly independent. Show these vectors are a basis for the vector space of polynomials of degree no more than 3 .
7. Determine whether the real valued functions defined on $\mathbb{R}$

$$
\left\{x^{2}+1, x^{3}+2 x^{2}+x, x^{3}+2 x^{2}-1, x^{3}+x^{2}+x\right\}
$$

are linearly independent. Is this a basis for the subspace of polynomials of degree no more than 3 ? Explain why or why not.
8. Determine whether the real valued functions defined on $\mathbb{R}$

$$
\left\{x^{2}+1, x^{3}+2 x^{2}+x, x^{3}+2 x^{2}+x, x^{3}+x^{2}+x\right\}
$$

are linearly independent. Is this a basis for the subspace of polynomials of degree no more than 3 ? Explain why or why not.
9. Show that the following are each a basis for $\mathbb{R}^{3}$.
(a) $\left(\begin{array}{c}3 \\ 2 \\ -1\end{array}\right),\left(\begin{array}{c}2 \\ 2 \\ -1\end{array}\right),\left(\begin{array}{c}-1 \\ -1 \\ 1\end{array}\right)$
(c) $\left(\begin{array}{c}-3 \\ 0 \\ 1\end{array}\right),\left(\begin{array}{c}5 \\ 1 \\ -1\end{array}\right),\left(\begin{array}{c}6 \\ 1 \\ -1\end{array}\right)$
(b) $\left(\begin{array}{c}-2 \\ 0 \\ 2\end{array}\right),\left(\begin{array}{c}3 \\ 1 \\ -2\end{array}\right),\left(\begin{array}{c}4 \\ 1 \\ -2\end{array}\right)$
(d) $\left(\begin{array}{c}1 \\ 2 \\ -1\end{array}\right),\left(\begin{array}{c}2 \\ 2 \\ -1\end{array}\right),\left(\begin{array}{c}-1 \\ -1 \\ 1\end{array}\right)$
10. Show that each of the following is not a basis for $\mathbb{R}^{3}$. Explain why they fail to be a basis.
(a) $\left(\begin{array}{l}1 \\ 1 \\ 1\end{array}\right),\left(\begin{array}{l}0 \\ 1 \\ 1\end{array}\right),\left(\begin{array}{l}3 \\ 5 \\ 5\end{array}\right)$
(d) $\left(\begin{array}{l}1 \\ 0 \\ 1\end{array}\right),\left(\begin{array}{l}1 \\ 1 \\ 0\end{array}\right)$
(b) $\left(\begin{array}{c}1 \\ -1 \\ 1\end{array}\right),\left(\begin{array}{l}0 \\ 1 \\ 1\end{array}\right),\left(\begin{array}{c}3 \\ -1 \\ 5\end{array}\right)$
(e) $\left(\begin{array}{c}1 \\ 2 \\ -1\end{array}\right),\left(\begin{array}{c}2 \\ 2 \\ -1\end{array}\right)$,
(c) $\left(\begin{array}{l}3 \\ 2 \\ 5\end{array}\right),\left(\begin{array}{l}0 \\ 1 \\ 1\end{array}\right),\left(\begin{array}{l}1 \\ 0 \\ 1\end{array}\right)$
$\left(\begin{array}{c}-1 \\ -1 \\ 1\end{array}\right),\left(\begin{array}{l}1 \\ 0 \\ 0\end{array}\right)$
11. Suppose $B$ is a subset of the set of complex valued functions, none equal to 0 and defined on $\Omega$ and it has the property that if $f, g$ are different, then $f g=0$. Show that $B$ must be linearly independent.
12. Suppose $\left\{f_{1}, f_{2}, \cdots, f_{n}\right\}$ are real valued (continuous) functions defined on $[0,1]$, and these satisfy

$$
\int_{0}^{1} f_{i}(x) f_{j}(x) d x=\delta_{i j} \equiv\left\{\begin{array}{l}
1 \text { if } i=j \\
0 \text { if } i \neq j
\end{array}\right.
$$

Show that these functions must be linearly independent.
13. Show that the real valued functions $\cos (2 x), 1, \cos ^{2}(x)$ are linearly dependent.
14. Show that the real valued functions $e^{x} \sin (2 x), e^{x} \cos (2 x)$ are linearly independent.
15. Let the field of scalars be $\mathbb{Q}$ and let the vector space be all vectors (real numbers) of the form $a+b \sqrt{2}$ for $a, b \in \mathbb{Q}$. Show that this really is a vector space and find a basis for $i t$.
16. Consider the two vectors $\binom{2}{1},\binom{1}{2}$ in $\mathbb{R}^{2}$. Show that these are linearly independent. Now consider $\binom{2}{1},\binom{1}{2}$ in $\mathbb{Z}_{3}^{2}$ where the numbers are interpreted as residue classes. Are these vectors linearly independent? If not, give a nontrivial linear combination which is 0 .
17. Is $\mathbb{C}$ a vector space with field of scalars $\mathbb{R}$ ? If so, what is the dimension of this vector space? Give a basis.
18. Is $\mathbb{C}$ a vector space with field of scalars $\mathbb{C}$ ? If so, what is the dimension? Give a basis.
19. The space of real valued continuous functions on $[0,1]$ usually denoted as $C([0,1])$ is a vector space with field of scalars $\mathbb{R}$. Explain why it is not a finite dimensional vector space.
20. Suppose two vector spaces $V, W$ have the same field of scalars $\mathbb{F}$. Show that $V \cap W$ is a subspace of both $V$ and $W$.
21. If $V, W$ are two sub spaces of a vector space $U$, define

$$
V+W \equiv\{v+w: v \in V, w \in W\}
$$

Show that this is a subspace of $U$.
22. If $V, W$ are two sub spaces of a vector space $U$, consider $V \cup W$, the vectors which are in either $V$ or $W$. Will this be a subspace of $U$ ? If so, prove it is the case and if not, give an example which shows that it is not necessarily true.
23. Let $V, W$ be vector spaces. A function $T: V \rightarrow W$ is called a linear transformation if whenever $\alpha, \beta$ are scalars and $u, v$ are vectors in $V$, it follows that

$$
T(\alpha u+\beta v)=\alpha T u+\beta T v
$$

Then $\operatorname{ker}(T) \equiv\{u \in V: T u=0\}, \operatorname{Im}(T) \equiv\{T u: u \in V\}$. Show the first of these is a subspace of $V$ and the second is a subspace of $W$.
24. $\uparrow$ In the situation of the above problem, where $T$ is a linear transformation, suppose $S$ is a linearly independent subset of $W$. Define $T^{-1}(S) \equiv\{u \in V: T u \in S\}$. Show that $T^{-1}(S)$ is linearly independent.
25. $\uparrow$ In the situation of the above problems, $\operatorname{rank}(T)$ is defined as the dimension of $\operatorname{Im}(T)$. Also the nullity of $T$, denoted as null $(T)$ is defined as the dimension of $\operatorname{ker}(T)$. In this problem, you will show that if the dimension of $V$ is $n$, then $\operatorname{rank}(T)+$ $\operatorname{null}(T)=n$.
(a) Let a basis for $\operatorname{ker}(T)$ be $\left\{z_{1}, \cdots, z_{r}\right\}$. Let a basis for $\operatorname{Im}(T)$ be

$$
\left\{T v_{1}, \cdots, T v_{s}\right\}
$$

You need to show that $r+s=n$. Begin with $u \in V$ and consider $T u$. It is a linear combination of $\left\{T v_{1}, \cdots, T v_{s}\right\}$ say $\sum_{i=1}^{s} a_{i} T v_{i}$. Why?
(b) Next explain why $T\left(u-\sum_{i=1}^{s} a_{i} v_{i}\right)=0$. Then explain why there are scalars $b_{j}$ such that $u-\sum_{i=1}^{s} a_{i} v_{i}=\sum_{j=1}^{r} b_{j} z_{j}$.
(c) Observe that $V=\operatorname{span}\left(z_{1}, \cdots, z_{r}, v_{1}, \cdots, v_{s}\right)$. Why?
(d) Finally show that $\left\{z_{1}, \cdots, z_{r}, v_{1}, \cdots, v_{s}\right\}$ is linearly independent. Thus $n=r+s$.

### 3.4 Polynomials and Fields

As an application of the theory of vector spaces, this section considers the problem of field extensions. When you have a polynomial like $x^{2}-3$ which has no rational roots, it turns out you can enlarge the field of rational numbers to obtain a larger field such that this polynomial does have roots in this larger field. I am going to discuss a systematic way to do this. It will turn out that for any polynomial with coefficients in any field, there always exists a possibly larger field such that the polynomial has roots in this larger field. This book mainly features the field of real or complex numbers but this procedure will show how to obtain many other fields. The ideas used in this development are the same as those used later in the material on linear transformations but slightly easier.

Here is an important idea concerning equivalence relations which I hope is familiar. If not, see Section 1.3 on page 6 .

Definition 3.4.1 Let $S$ be a set. The symbol, $\sim$ is called an equivalence relation on $S$ if it satisfies the following axioms.

1. $x \sim x$ for all $x \in S$. (Reflexive)
2. If $x \sim y$ then $y \sim x$. (Symmetric)
3. If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)

Definition 3.4.2 $[x]$ denotes the set of all elements of $S$ which are equivalent to $x$ and $[x]$ is called the equivalence class determined by $x$ or just the equivalence class of $x$.

Also recall the notion of equivalence classes.
Theorem 3.4.3 Let $\sim$ be an equivalence class defined on a set, $S$ and let $\mathscr{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x]=[y]$ or it is not true that $x \sim y$ and $[x] \cap[y]=\emptyset$.

Definition 3.4.4 Let $\mathbb{F}$ be a field, for example the rational numbers, and denote by $\mathbb{F}[x]$ the polynomials having coefficients in $\mathbb{F}$. Suppose $p(x)$ is a polynomial. Let $a(x) \sim b(x)(a(x)$ is similar to $b(x)$ ) when

$$
a(x)-b(x)=k(x) p(x)
$$

for some polynomial $k(x)$. Denote by $(p(x))$ all polynomials of the form $p(x) k(x)$ where $k(x)$ is some polynomial.

Proposition 3.4.5 In the above definition, $\sim$ is an equivalence relation.
Proof: First of all, note that $a(x) \sim a(x)$ because their difference equals $0 p(x)$. If $a(x) \sim b(x)$, then $a(x)-b(x)=k(x) p(x)$ for some $k(x)$. But then

$$
b(x)-a(x)=-k(x) p(x)
$$

and so $b(x) \sim a(x)$. Next suppose $a(x) \sim b(x)$ and $b(x) \sim c(x)$. Then $a(x)-b(x)=$ $k(x) p(x)$ for some polynomial $k(x)$ and also $b(x)-c(x)=l(x) p(x)$ for some polynomial $l(x)$. Then

$$
\begin{aligned}
& a(x)-c(x)=a(x)-b(x)+b(x)-c(x) \\
= & k(x) p(x)+l(x) p(x)=(l(x)+k(x)) p(x)
\end{aligned}
$$

and so $a(x) \sim c(x)$ and this shows the transitive law.
Definition 3.4.6 Let $\mathbb{F}$ be a field and let $p(x) \in \mathbb{F}[x]$ be a nonzero monic polynomial. This means that the coefficient of the highest power is 1 . Also let $p(x)$ have degree at least 1. For the similarity relation of Definition 3.4.4, define the following operations on the equivalence classes. $[a(x)]$ is an equivalence class means that it is the set of all polynomials which are similar to $a(x)$.

$$
\begin{aligned}
{[a(x)]+[b(x)] } & \equiv[a(x)+b(x)] \\
{[a(x)][b(x)] } & \equiv[a(x) b(x)]
\end{aligned}
$$

This collection of equivalence classes is sometimes denoted by $\mathbb{F}[x] /(p(x))$. This is called a quotient space.

The set of equivalence classes just described is a commutative ring. This is like a field except it may fail to have multiplicative inverses. The reason for considering only polynomials of degree at least 1 is that $\mathbb{F}[x] /(1)$ isn't very interesting because $f(x) \sim g(x)$ if and only if their difference is a multiple of 1 . Thus every two polynomials are similar so there is only one similarity class. In particular, [1] $\sim[0]$. It is shown below that this is well defined.

Axiom 3.4.7 Here are the axioms for a commutative ring.

1. $x+y=y+x$, (commutative law for addition)
2. There exists 0 such that $x+0=x$ for all $x$, (additive identity).
3. For each $x \in \mathbb{F}$, there exists $-x \in \mathbb{F}$ such that $x+(-x)=0$, (existence of additive inverse).
4. $(x+y)+z=x+(y+z)$,(associative law for addition).
5. $x y=y x$, (commutative law for multiplication). You could write this as $x \times y=y \times x$.
6. $(x y) z=x(y z)$, (associative law for multiplication).
7. There exists 1 such that $1 x=x$ for all $x$,(multiplicative identity).
8. $x(y+z)=x y+x z .($ distributive law $)$.

Recall that $p(x)$ is irreducible, means the only monic polynomials which divide it are 1 and itself.

Lemma 3.4.8 With the equivalence classes defined in Definition 3.4.6 where $p(x)$ is a monic polynomial of degree at least 1 ,

1. The operations are well defined.
2. $\mathbb{F}[x] /(p(x))$ is a commutative ring
3. If $a, b \in \mathbb{F}$ and $[a]=[b]$, then $a=b$. Thus $\mathbb{F}$ is a subset of $\mathbb{F}[x] /(p(x))$.
4. Also $[q(x)]=0$ if and only if $q(x)=p(x) l(x)$ for some polynomial $l(x)$.
5. $\mathbb{F}[x] /(p(x))$ is a field if and only if $p(x)$ is also irreducible.

Proof: 1.) To show the operations are well defined, suppose

$$
[a(x)]=\left[a^{\prime}(x)\right],[b(x)]=\left[b^{\prime}(x)\right]
$$

It is necessary to show

$$
\begin{aligned}
{[a(x)+b(x)] } & =\left[a^{\prime}(x)+b^{\prime}(x)\right] \\
{[a(x) b(x)] } & =\left[a^{\prime}(x) b^{\prime}(x)\right]
\end{aligned}
$$

Consider the second of the two.

$$
\begin{aligned}
& a^{\prime}(x) b^{\prime}(x)-a(x) b(x) \\
= & a^{\prime}(x) b^{\prime}(x)-a(x) b^{\prime}(x)+a(x) b^{\prime}(x)-a(x) b(x) \\
= & b^{\prime}(x)\left(a^{\prime}(x)-a(x)\right)+a(x)\left(b^{\prime}(x)-b(x)\right)
\end{aligned}
$$

Now by assumption $\left(a^{\prime}(x)-a(x)\right)$ is a multiple of $p(x)$ as is $\left(b^{\prime}(x)-b(x)\right)$, so the above is a multiple of $p(x)$ and by definition this shows $[a(x) b(x)]=\left[a^{\prime}(x) b^{\prime}(x)\right]$. The case for addition is similar.
2.) The various algebraic properties related to these operations are obvious and come directly from the definitions.
3.) Now suppose $[a]=[b]$. This means that $a-b=k(x) p(x)$ for some polynomial $k(x)$. Then $k(x)$ must equal 0 since otherwise the two polynomials $a-b$ and $k(x) p(x)$ could not be equal because they would have different degree. This is where it is important to have the degree of $p(x)$ at least 1 .
4.) $[q(x)]=[0]$ means $q(x) \sim 0$ which means $q(x)=p(x) l(x)$ for some $l(x)$.
5.) Suppose $p(x)$ is irreducible. Let $[q(x)] \in \mathbb{F}[x] /(p(x))$ where $[q(x)] \neq[0]$. Then $q(x)$ is not a multiple of $p(x)$ and so $q(x), p(x)$ are relatively prime. This is because if $\psi(x)$ is a monic polynomial which divides both $q(x)$ and $p(x)$, then since $p(x)$ is irreducible, $\psi(x)$ equals either a multiple of $p(x)$ which is given not to happen since $[q(x)] \neq 0$ or $\psi(x)=1$. Thus there exist $n(x), m(x)$ such that

$$
1=n(x) q(x)+m(x) p(x)
$$

Hence

$$
[1]=[n(x) q(x)]=[n(x)][q(x)]
$$

which shows that $[q(x)]^{-1}=[n(x)]$.
Now suppose $p(x)$ is not irreducible. Then $p(x)=l(x) k(x)$ where $l(x), k(x)$ have smaller degree than $p(x)$. Then $[0]=[l(x)][k(x)]$. Neither $[l(x)]$ nor $[k(x)]$ equals 0 because neither is a multiple of $p(x)$ and this cannot happen in a field. Thus if $p(x)$ is not irreducible, then $\mathbb{F}[x] /(p(x))$ is not a field.

The following proposition is mostly a summary of the above lemma. Recall irreducible means the only monic polynomials which divide $p(x)$ are itself and nonzero scalars.

Proposition 3.4.9 In the situation of Definition 3.4.6 where $p(x)$ is a nonzero monic, irreducible polynomial of degree at least 1, the following are valid.

1. The definitions of addition and multiplication are well defined.
2. If $a, b \in \mathbb{F}$ and $[a]=[b]$, then $a=b$. Thus $\mathbb{F}$ is a subset of $\mathbb{F}[x] /(p(x))$.
3. $\mathbb{F}[x] /(p(x))$ is a field in which the polynomial $p(x)$ has a root.
4. $\mathbb{F}[x] /(p(x))$ is a vector space with field of scalars $\mathbb{F}$ and its dimension is $m$ where $m$ is the degree of the irreducible polynomial $p(x)$.

Proof: 1.) This is shown in Lemma 3.4.8 as is 2.) and 3.) except for the part of 3.) which says $p(x)$ has a root.

The polynomial $p(x)$ has a root in this field because if

$$
\begin{gathered}
p(x)=x^{m}+a_{m-1} x^{m-1}+\cdots+a_{1} x+a_{0} \\
{[0]=[p(x)]=[x]^{m}+\left[a_{m-1}\right][x]^{m-1}+\cdots+\left[a_{1}\right][x]+\left[a_{0}\right]}
\end{gathered}
$$

Thus $[x]$ is a root of this polynomial in the field $\mathbb{F}[x] /(p(x))$.
Consider the last claim. It is clear that $\mathbb{F}[x] /(p(x))$ is a vector space with field of scalars $\mathbb{F}$. Indeed, the operations are defined such that for $\alpha, \beta \in \mathbb{F}$,

$$
\alpha[r(x)]+\beta[b(x)] \equiv[\alpha r(x)+\beta b(x)]
$$

It remains to consider the dimension of this vector space. Let $f(x) \in \mathbb{F}[x] /(p(x))$. Thus $[f(x)]$ is a typical thing in $\mathbb{F}[x] /(p(x))$. Then from the division algorithm,

$$
f(x)=p(x) q(x)+r(x)
$$

where $r(x)$ is either 0 or has degree less than the degree of $p(x)$. Thus

$$
[r(x)]=[f(x)-p(x) q(x)]=[f(x)]
$$

but clearly $[r(x)] \in \operatorname{span}\left([1],[x], \cdots,[x]^{m-1}\right)$ and also it is clear that

$$
\operatorname{span}\left([1], \cdots,[x]^{m-1}\right)=\mathbb{F}[x] /(p(x))
$$

Then $\left\{[1],[x], \cdots,[x]^{m-1}\right\}$ is a basis if these vectors are linearly independent. Suppose then that

$$
\sum_{i=0}^{m-1} c_{i}[x]^{i}=\left[\sum_{i=0}^{m-1} c_{i} x^{i}\right]=0
$$

Then you would need to have $p(x) / \sum_{i=0}^{m-1} c_{i} x^{i}$ which is impossible unless each $c_{i}=0$ because $p(x)$ has degree $m$.

This shows how to enlarge a field to get a new one in which the polynomial has a root. By using a succession of such enlargements, called field extensions, there will exist a field in which the given polynomial can be factored into a product of polynomials having degree one. The field you obtain in this process of enlarging in which the given polynomial factors in terms of linear factors is called a splitting field.

Definition 3.4.10 A commutative ring is just a field in which the assumption that multiplicative inverses for nonzero elements may not exist. An ideal I in a commutative ring $R$ is a subset of $R$ closed with respect to addition and additive inverses such that $r I \subseteq I$ meaning that something in I multiplied by $r \in R$ will yield something in $I$. Then $\mathbb{F}[x]$ is a commutative ring and $(p(x))$ is an example of an ideal. A maximal ideal is an ideal for which the only ideal containing it is itself or the entire ring.
Example 3.4.11 The polynomial $x^{2}-2$ is irreducible in $\mathbb{Q}(x)$. This is because if $x^{2}-$ $2=p(x) q(x)$ where $p(x), q(x)$ both have degree less than 2, then they both have degree 1. Hence you would have $x^{2}-2=(x+a)(x+b)$ which requires that $a+b=0$ so this factorization is of the form $(x-a)(x+a)$ and now you need to have $a=\sqrt{2} \notin \mathbb{Q}$. Now $\mathbb{Q}(x) /\left(x^{2}-2\right)$ is of the form $a+b[x]$ where $a, b \in \mathbb{Q}$ and $[x]^{2}-2=0$. Thus one can regard $[x]$ as $\sqrt{2} . \mathbb{Q}(x) /\left(x^{2}-2\right)$ is of the form $a+b \sqrt{2}$.

In the above example, $\left[x^{2}+x\right]$ is not zero because it is not a multiple of $x^{2}-2$. What is $\left[x^{2}+x\right]^{-1}$ ? You know that the two polynomials are relatively prime and so there exists $n(x), m(x)$ such that

$$
1=n(x)\left(x^{2}-2\right)+m(x)\left(x^{2}+x\right)
$$

Thus $[m(x)]=\left[x^{2}+x\right]^{-1}$. How could you find these polynomials? First of all, it suffices to consider only $n(x)$ and $m(x)$ having degree less than 2 . Otherwise, reiterating the above, $m(x)=p(x) l(x)+r(x)$ where $r(x)$ has degree smaller than the degree of $p(x)$ and you could simply use $r(x)$ in place of $m(x)$.

$$
\begin{gathered}
1=(a x+b)\left(x^{2}-2\right)+(c x+d)\left(x^{2}+x\right) \\
1=a x^{3}-2 b+b x^{2}+c x^{2}+c x^{3}+d x^{2}-2 a x+d x
\end{gathered}
$$

Now you solve the resulting system of equations.

$$
a=\frac{1}{2}, b=-\frac{1}{2}, c=-\frac{1}{2}, d=1
$$

Then the desired inverse is $\left[-\frac{1}{2} x+1\right]$. To check,

$$
\left(-\frac{1}{2} x+1\right)\left(x^{2}+x\right)-1=-\frac{1}{2}(x-1)\left(x^{2}-2\right)
$$

Thus $\left[-\frac{1}{2} x+1\right]\left[x^{2}+x\right]-[1]=[0]$.
The above is an example of something general described in the following definition.
Definition 3.4.12 Let $F \subseteq K$ be two fields. Then clearly $K$ is also a vector space over $F$. Then also, $K$ is called a finite field extension of $F$ if the dimension of this vector space, denoted by $[K: F]$ is finite.

There are some easy things to observe about this.
Proposition 3.4.13 Let $F \subseteq K \subseteq L$ be fields. Then $[L: F]=[L: K][K: F]$.
Proof: Let $\left\{l_{i}\right\}_{i=1}^{n}$ be a basis for $L$ over $K$ and let $\left\{k_{j}\right\}_{j=1}^{m}$ be a basis of $K$ over $F$. Then if $l \in L$, there exist unique scalars $x_{i}$ in $K$ such that $l=\sum_{i=1}^{n} x_{i} l_{i}$. Now $x_{i} \in K$ so there exist $f_{j i}$ such that $x_{i}=\sum_{j=1}^{m} f_{j i} k_{j}$. Then it follows that $l=\sum_{i=1}^{n} \sum_{j=1}^{m} f_{j i} k_{j} l_{i}$. It follows that $\left\{k_{j} l_{i}\right\}$ is a spanning set. If $\sum_{i=1}^{n} \sum_{j=1}^{m} f_{j i} k_{j} l_{i}=0$. Then, since the $l_{i}$ are independent, it follows that $\sum_{j=1}^{m} f_{j i} k_{j}=0$ and since $\left\{k_{j}\right\}$ is independent, each $f_{j i}=0$ for each $j$ for a given arbitrary $i$. Therefore, $\left\{k_{j} l_{i}\right\}$ is a basis.

You will see almost exactly the same argument in exhibiting a basis for $\mathscr{L}(V, W)$ the linear transformations mapping $V$ to $W$.

Note that if $p(x)$ were of degree $n$ and not irreducible, then there still exists an extension $\mathbb{G}$ containing a root of $p(x)$ such that $[\mathbb{G}: \mathbb{F}] \leq n$. You could do this by working with an irreducible factor of $p(x)$.

Theorem 3.4.14 Let $p(x)=x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ be a polynomial with coefficients in a field of scalars $\mathbb{F}$. There exists a larger field $\mathbb{G}$ and $\left\{z_{1}, \cdots, z_{n}\right\}$ contained in $\mathbb{G}$, listed according to multiplicity, such that

$$
p(x)=\prod_{i=1}^{n}\left(x-z_{i}\right)
$$

This larger field is called a splitting field. Furthermore,

$$
[\mathbb{G}: \mathbb{F}] \leq n!
$$

Proof: From Proposition 3.4.9, there exists a field $\mathbb{F}_{1}$ such that $p(x)$ has a root, $z_{1}(=[x])$ Then by the Euclidean algorithm $p(x)=\left(x-z_{1}\right) q_{1}(x)+r$ where $r \in \mathbb{F}_{1}$. Since $p\left(z_{1}\right)=0$, this requires $r=0$. Now do the same for $q_{1}(x)$ that was done for $p(x)$, enlarging the field to $\mathbb{F}_{2}$ if necessary, such that in this new field $q_{1}(x)=\left(x-z_{2}\right) q_{2}(x)$ and so $p(x)=$ $\left(x-z_{1}\right)\left(x-z_{2}\right) q_{2}(x)$. After no more than $n$ such extensions, you will have obtained the necessary field $\mathbb{G}$.

Finally consider the claim about dimension. By Proposition 3.4.9, there is a larger field $\mathbb{G}_{1}$ such that $p(x)$ has a root $a_{1}$ in $\mathbb{G}_{1}$ and $\left[\mathbb{G}_{1}: \mathbb{F}\right] \leq n$. Then $p(x)=\left(x-a_{1}\right) q(x)$. Continue this way until the polynomial equals the product of linear factors. Then by Proposition 3.4.13 applied multiple times, $[\mathbb{G}: \mathbb{F}] \leq n!$.

Example 3.4.15 The polynomial $x^{2}+1$ is irreducible in $\mathbb{R}(x)$, polynomials having real coefficients. To see this is the case, suppose $\psi(x)$ divides $x^{2}+1$. Then $x^{2}+1=\psi(x) q(x)$. If the degree of $\psi(x)$ is less than 2, then it must be either a constant or of the form $a x+b$. In the latter case, $-b / a$ must be a zero of the right side, hence of the left but $x^{2}+1$ has no real zeros. Therefore, the degree of $\psi(x)$ must be two and $q(x)$ must be a constant. Thus the only polynomial which divides $x^{2}+1$ are constants and multiples of $x^{2}+1$. Therefore, this shows $x^{2}+1$ is irreducible. Find the inverse of $\left[x^{2}+x+1\right]$ in the space of equivalence classes, $\mathbb{R}(x) /\left(x^{2}+1\right)$.

You can solve this with partial fractions.

$$
\frac{1}{\left(x^{2}+1\right)\left(x^{2}+x+1\right)}=-\frac{x}{x^{2}+1}+\frac{x+1}{x^{2}+x+1}
$$

and so $1=(-x)\left(x^{2}+x+1\right)+(x+1)\left(x^{2}+1\right)$ which implies $1 \sim(-x)\left(x^{2}+x+1\right)$ and so the inverse is $[-x]$.

The following proposition is interesting. It was essentially proved above but to emphasize it, here it is again.

Proposition 3.4.16 Suppose $p(x) \in \mathbb{F}[x]$ is irreducible and has degree $n$. Then every element of $\mathbb{G}=\mathbb{F}[x] /(p(x))$ is of the form $[0]$ or $[r(x)]$ where the degree of $r(x)$ is less than $n$.

Proof: This follows right away from the Euclidean algorithm for polynomials. If $k(x)$ has degree larger than $n-1$, then $k(x)=q(x) p(x)+r(x)$ where $r(x)$ is either equal to 0 or has degree less than $n$. Hence $[k(x)]=[r(x)]$.

Example 3.4.17 In the situation of the above example where the polynomial is $x^{2}+1$ irreducible in $\mathbb{R}(x)$, find $[a x+b]^{-1}$ assuming $a^{2}+b^{2} \neq 0$. Note this includes all cases of interest thanks to the above proposition.

You can do it with partial fractions as above.

$$
\frac{1}{\left(x^{2}+1\right)(a x+b)}=\frac{b-a x}{\left(a^{2}+b^{2}\right)\left(x^{2}+1\right)}+\frac{a^{2}}{\left(a^{2}+b^{2}\right)(a x+b)}
$$

and so

$$
1=\frac{1}{a^{2}+b^{2}}(b-a x)(a x+b)+\frac{a^{2}}{\left(a^{2}+b^{2}\right)}\left(x^{2}+1\right)
$$

Thus $\frac{1}{a^{2}+b^{2}}(b-a x)(a x+b) \sim 1$ and so

$$
[a x+b]^{-1}=\frac{[(b-a x)]}{a^{2}+b^{2}}=\frac{b-a[x]}{a^{2}+b^{2}}
$$

You might find it interesting to recall that $(a i+b)^{-1}=\frac{b-a i}{a^{2}+b^{2}}$. Didn't this just produce the complex numbers algebraically? If, instead of $\mathbb{R}$ you used $\mathbb{Q}$ this would have just produced a field $\mathbb{Q}+i \mathbb{Q}$.

### 3.4.1 The Algebraic Numbers and Minimum Polynomial

Each polynomial having coefficients in a field $\mathbb{F}$ has a splitting field. Consider the case of all polynomials $p(x)$ having coefficients in a field $\mathbb{F} \subseteq \mathbb{G}$ and consider all roots which are also in $\mathbb{G}$. The theory of vector spaces is very useful in the study of these algebraic numbers. Here is a definition.

Definition 3.4.18 Let $\mathbb{F}$ and $\mathbb{G}$ be two fields, $\mathbb{F} \subseteq \mathbb{G}$. The algebraic numbers $\mathbb{A}$ are those numbers which are in $\mathbb{G}$ and also roots of some polynomial $p(x)$ having coefficients in $\mathbb{F}$. The minimum polynomial ${ }^{1}$ of $a \in \mathbb{A}$ is defined to be the monic polynomial $p(x)$ having smallest degree such that $p(a)=0$. It is also often called the minimum polynomial.

The next theorem is on the uniqueness of the minimum polynomial.

[^1]Theorem 3.4.19 Let $a \in \mathbb{A}$. Then there exists a unique monic irreducible polynomial $p(x)$ having coefficients in $\mathbb{F}$ such that $p(a)=0$. This polynomial is the minimum polynomial.

Proof: Let $p(x)$ be a monic polynomial having smallest degree such that $p(a)=0$. Then $p(x)$ is irreducible because if not, there would exist a polynomial having smaller degree which has $a$ as a root. Now suppose $q(x)$ is monic with smallest degree such that $q(a)=0$. Then $q(x)=p(x) l(x)+r(x)$ where if $r(x) \neq 0$, then it has smaller degree than $p(x)$. But in this case, the equation implies $r(a)=0$ which contradicts the choice of $p(x)$. Hence $r(x)=0$ and so, since $q(x)$ has smallest degree, $l(x)=1$ showing that $p(x)=q(x)$.

Definition 3.4.20 For a an algebraic number, let $\operatorname{deg}(a)$ denote the degree of the minimum polynomial of $a$.

Also, here is another definition.
Definition 3.4.21 Let $a_{1}, \cdots, a_{m}$ be in $\mathbb{A}$. A polynomial in $\left\{a_{1}, \cdots, a_{m}\right\}$ will be an expression of the form

$$
\sum_{k_{1} \cdots k_{n}} a_{k_{1} \cdots k_{n}} a_{1}^{k_{1}} \cdots a_{n}^{k_{n}}
$$

where the $a_{k_{1} \cdots k_{n}}$ are in $\mathbb{F}$, each $k_{j}$ is a nonnegative integer, and all but finitely many of the $a_{k_{1} \cdots k_{n}}$ equal zero. The collection of such polynomials will be denoted by

$$
\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)
$$

The splitting field of $g(x) \in \mathbb{F}[x]$ is $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ where the $\left\{a_{1}, \cdots, a_{m}\right\}$ are the roots of $g(x)$ in $\mathbb{A}$.

Now notice that for $a$ an algebraic number, $\mathbb{F}(a)$ is a finite dimensional vector space with field of scalars $\mathbb{F}$. Similarly, for $\left\{a_{1}, \cdots, a_{m}\right\}$ algebraic numbers, $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is a finite dimensional vector space with field of scalars $\mathbb{F}$. The following fundamental proposition demonstrates this observation. This is a remarkable result.

Proposition 3.4.22 Let $\left\{a_{1}, \cdots, a_{m}\right\}$ be algebraic numbers. Then

$$
\operatorname{dim} \mathbb{F}\left(a_{1}, \cdots, a_{m}\right) \leq \prod_{j=1}^{m} \operatorname{deg}\left(a_{j}\right)
$$

and for an algebraic number a,

$$
\operatorname{dim} \mathbb{F}(a)=\operatorname{deg}(a)
$$

Every element of $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is in $\mathbb{A}$ and $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is a field.
Proof: Let the minimum polynomial of $a$ be

$$
p(x)=x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0} .
$$

If $q(a) \in \mathbb{F}(a)$, then

$$
q(x)=p(x) l(x)+r(x)
$$

where $r(x)$ has degree less than the degree of $p(x)$ if it is not zero. Hence $q(a)=r(a)$. Thus $\mathbb{F}(a)$ is spanned by

$$
\left\{1, a, a^{2}, \cdots, a^{n-1}\right\}
$$

Since $p(x)$ has smallest degree of all polynomials which have $a$ as a root, the above set is also linearly independent. This proves the second claim.

Now consider the first claim. By definition, $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is obtained from all linear combinations of products of $\left\{a_{1}^{k_{1}}, a_{2}^{k_{2}}, \cdots, a_{n}^{k_{n}}\right\}$ where the $k_{i}$ are nonnegative integers. From the first part, it suffices to consider only $k_{j} \leq \operatorname{deg}\left(a_{j}\right)$. This is because $a_{1}^{m}$ can be written as a linear combination of $a_{1}^{k}$ for $k \leq \operatorname{deg}\left(a_{1}\right)$. Therefore, there exists a spanning set for $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ which has $\prod_{i=1}^{m} \operatorname{deg}\left(a_{i}\right)$ entries. By Theorem 3.2.9 a basis has no more vectors than $\prod_{i=1}^{m} \operatorname{deg}\left(a_{i}\right)$. This proves the first claim.

Consider the last claim. Let $g\left(a_{1}, \cdots, a_{m}\right)$ be a polynomial in $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$. Since

$$
\operatorname{dim} \mathbb{F}\left(a_{1}, \cdots, a_{m}\right) \equiv p \leq \prod_{j=1}^{m} \operatorname{deg}\left(a_{j}\right)<\infty
$$

it follows

$$
1, g\left(a_{1}, \cdots, a_{m}\right), g\left(a_{1}, \cdots, a_{m}\right)^{2}, \cdots, g\left(a_{1}, \cdots, a_{m}\right)^{p}
$$

are dependent. It follows $g\left(a_{1}, \cdots, a_{m}\right)$ is the root of some polynomial having coefficients in $\mathbb{F}$. Thus everything in $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is algebraic.

Why is $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ a field? Let $g\left(a_{1}, \cdots, a_{m}\right) \neq 0$ be in $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$. Then it has a minimum polynomial,

$$
p(x)=x^{q}+a_{q-1} x^{q-1}+\cdots+a_{1} x+a_{0}
$$

where the $a_{i} \in \mathbb{F}$. Then $a_{0} \neq 0$ or else the polynomial would not be minimum. You would have

$$
g\left(a_{1}, \cdots, a_{m}\right)\left(g\left(a_{1}, \cdots, a_{m}\right)^{q-1}+a_{q-1} g\left(a_{1}, \cdots, a_{m}\right)^{q-2}+\cdots+a_{1}\right)=0
$$

and so $g\left(a_{1}, \cdots, a_{m}\right)^{q-1}+a_{q-1} g\left(a_{1}, \cdots, a_{m}\right)^{q-2}+\cdots+a_{1}=0$. Therefore, since $a_{0} \neq 0$,

$$
g\left(a_{1}, \cdots, a_{m}\right)\left(g\left(a_{1}, \cdots, a_{m}\right)^{q-1}+a_{q-1} g\left(a_{1}, \cdots, a_{m}\right)^{q-2}+\cdots+a_{1}\right)=-a_{0}
$$

and so the multiplicative inverse for $g\left(a_{1}, \cdots, a_{m}\right)$ is

$$
\frac{g\left(a_{1}, \cdots, a_{m}\right)^{q-1}+a_{q-1} g\left(a_{1}, \cdots, a_{m}\right)^{q-2}+\cdots+a_{1}}{-a_{0}} \in \mathbb{F}\left(a_{1}, \cdots, a_{m}\right)
$$

The other axioms of a field are obvious.
Now from this proposition, it is easy to obtain the following interesting result about the algebraic numbers. Like the above result, it is amazing.

Theorem 3.4.23 The algebraic numbers $\mathbb{A}$, those roots of polynomials in $\mathbb{F}[x]$ which are in $\mathbb{G}$, are a field.

Proof: By definition, each $a \in \mathbb{A}$ has a minimum polynomial. Let $a \neq 0$ be an algebraic number and let $p(x)$ be its minimum polynomial. Then $p(x)$ is of the form

$$
x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}
$$

where $a_{0} \neq 0$. Otherwise $p(x)$ would not have minimum degree. Then plugging in $a$ yields

$$
a \frac{\left(a^{n-1}+a_{n-1} a^{n-2}+\cdots+a_{1}\right)(-1)}{a_{0}}=1 .
$$

and so $a^{-1}=\frac{\left(a^{n-1}+a_{n-1} a^{n-2}+\cdots+a_{1}\right)(-1)}{a_{0}} \in \mathbb{F}(a)$. By Proposition 3.4.22, every element of $\mathbb{F}(a)$ is in $\mathbb{A}$ and this shows that for every nonzero element of $\mathbb{A}$, its inverse is also in $\mathbb{A}$. What about products and sums of things in $\mathbb{A}$ ? Are they still in $\mathbb{A}$ ? Yes. If $a, b \in \mathbb{A}$, then both $a+b$ and $a b \in \mathbb{F}(a, b)$ and from the proposition, each element of $\mathbb{F}(a, b)$ is in $\mathbb{A}$.

A typical example of what is of interest here is when the field $\mathbb{F}$ of scalars is $\mathbb{Q}$, the rational numbers and the field $\mathbb{G}$ is $\mathbb{R}$. However, you can certainly conceive of many other examples by considering the integers mod a prime, for example (See Propositon 1.15.2 on Page 27 for example.) or any of the fields which occur as field extensions in the above.

There is a very interesting thing about $\mathbb{F}\left(a_{1}, \cdots, a_{n}\right)$ in the case where $\mathbb{F}$ is infinite which says that there exists a single algebraic $\gamma$ such that $\mathbb{F}\left(a_{1}, \cdots, a_{n}\right)=\mathbb{F}(\gamma)$. In other words, every field extension of this sort is a simple field extension. I found this fact in an early version of [12].

Proposition 3.4.24 Let $\mathbb{F}$ be infinite. Then $\mathbb{F}\left(a_{1}, \cdots, a_{n}\right)=\mathbb{F}(\gamma)$ for some $\gamma$. Here each $a_{i}$ is algebraic. The $\gamma$ will be of the form

$$
\gamma=a_{1}+\lambda_{1} a_{2}+\cdots+\lambda_{n-1} a_{n}
$$

where the $\lambda_{k}$ are in a splitting field. If the field $\mathbb{F}$ includes $\mathbb{Q}$ so that all polynomials have roots in $\mathbb{C}$, you can conclude that each $\lambda_{i}$ is a positive integer.

Proof: To begin with, consider $\mathbb{F}(\alpha, \beta)$. Let $\gamma=\alpha+\lambda \beta$. Then by Proposition 3.4.22 $\gamma$ is an algebraic number and it is also clear

$$
\mathbb{F}(\gamma) \subseteq \mathbb{F}(\alpha, \beta)
$$

I need to show the other inclusion. This will be done for a suitable choice of $\lambda$. To do this, it suffices to verify that both $\alpha$ and $\beta$ are in $\mathbb{F}(\gamma)$.

Let the minimum polynomials of $\alpha$ and $\beta$ be $f(x)$ and $g(x)$ respectively. Let the distinct roots of $f(x)$ and $g(x)$ be $\left\{\alpha_{1}, \alpha_{2}, \cdots, \alpha_{n}\right\}$ and $\left\{\beta_{1}, \beta_{2}, \cdots, \beta_{m}\right\}$ respectively. These roots are in a field which contains splitting fields of both $f(x)$ and $g(x)$. Let $\alpha=\alpha_{1}$ and $\beta=\beta_{1}$. Now define

$$
h(x) \equiv f(\alpha+\lambda \beta-\lambda x) \equiv f(\gamma-\lambda x)
$$

so that $h(\beta)=f(\alpha)=0$. It follows $(x-\beta)$ divides both $h(x)$ and $g(x)$ since it is given that $g(\beta)=0$. If $(x-\eta)$ is a different linear factor of both $g(x)$ and $h(x)$ then it must be $\left(x-\beta_{j}\right)$ for some $\beta_{j}$ for some $j>1$ because these are the only factors of $g(x)$. Therefore, this would require

$$
0=h\left(\beta_{j}\right)=f\left(\alpha_{1}+\lambda \beta_{1}-\lambda \beta_{j}\right)
$$

and so it would be the case that $\alpha_{1}+\lambda \beta_{1}-\lambda \beta_{j}=\alpha_{k}$ for some $k$. Hence

$$
\lambda=\frac{\alpha_{k}-\alpha_{1}}{\beta_{1}-\beta_{j}}
$$

Now there are finitely many quotients of the above form and if $\lambda$ is chosen to not be any of them, then the above cannot happen and so in this case, the only linear factor of both $g(x)$ and $h(x)$ will be $(x-\beta)$. Choose such a $\lambda$. If all roots are in $\mathbb{C}$ and $\mathbb{F}$ contains $\mathbb{Q}$, then you could pick $\lambda$ a positive integer.

Let $\phi(x)$ be the minimum polynomial of $\beta$ with respect to the field $\mathbb{F}(\gamma)$. Then this minimum polynomial must divide both $h(x)$ and $g(x)$ because $h(\beta)=g(\beta)=0$. However, the only factor these two have in common is $x-\beta$ and so $\phi(x)=x-\beta$ which requires $\beta \in \mathbb{F}(\gamma)$. Now also $\alpha=\gamma-\lambda \beta$ and so $\alpha \in \mathbb{F}(\gamma)$ also. Therefore, both $\alpha, \beta \in \mathbb{F}(\gamma)$ which forces $\mathbb{F}(\alpha, \beta) \subseteq \mathbb{F}(\gamma)$. This proves the proposition in the case that $n=2$. The general result follows right away by observing that

$$
\mathbb{F}\left(a_{1}, \cdots, a_{n}\right)=\mathbb{F}\left(a_{1}, \cdots, a_{n-1}\right)\left(a_{n}\right)
$$

and using induction. $\mathbb{F}\left(a_{1}, \cdots, a_{n-1}\right)\left(a_{n}\right)=\mathbb{F}\left(\gamma_{1}\right)\left(a_{n}\right)=F\left(\gamma_{1}, a_{n}\right)=\mathbb{F}(\gamma)$. Note that $\gamma=\alpha+\lambda \beta$ for two of these. Then

$$
\begin{aligned}
\mathbb{F}\left(a_{1}, a_{2}, a_{3}\right) & =\mathbb{F}\left(a_{1}, a_{2}\right)\left(a_{3}\right)=\mathbb{F}\left(a_{1}+\lambda_{1} a_{2}\right)\left(a_{3}\right) \\
& =\mathbb{F}\left(a_{1}+\lambda_{1} a_{2}, a_{3}\right)=\mathbb{F}\left(a_{1}+\lambda_{1} a_{2}+\lambda_{2} a_{3}\right)
\end{aligned}
$$

continuing this way shows that there are $\lambda_{i}$ in a suitable splitting field such that for $\gamma=$ $a_{1}+\lambda_{1} a_{2}+\cdots+\lambda_{n-1} a_{n}, \mathbb{F}\left(a_{1}, \cdots, a_{n}\right)=\mathbb{F}(\gamma)$. If all the numbers are in $\mathbb{C}$, and your field $\mathbb{F}$ includes $\mathbb{Q}$, you could choose all the $\lambda_{i}$ to be positive integers.

### 3.4.2 Lindermannn Weierstrass Theorem

As another application of the abstract concept of vector spaces, there is an amazing theorem due to Weierstrass and Lindemann.

Theorem 3.4.25 Suppose $a_{1}, \cdots, a_{n}$ are distinct algebraic numbers, roots of a polynomial with rational coefficients. Then it follows that $\sum_{i=1}^{n} a_{i} e^{\alpha_{i}} \neq 0$. In other words, the $\left\{e^{\alpha_{1}}, \cdots, e^{\alpha_{n}}\right\}$ are independent as vectors with field of scalars equal to the algebraic numbers.

There is a proof of this later. It is long and hard but only depends on elementary considerations other than some algebra involving symmetric polynomials. See Theorem 9.2.8. It is presented here to illustrate how the language of linear algebra is useful in describing something which is very exotic, apparently very far removed from something like $\mathbb{R}^{p}$.

A number is transcendental, as opposed to algebraic, if it is not a root of a polynomial which has integer (rational) coefficients. Most numbers are this way but it is hard to verify that specific numbers are transcendental. That $\pi$ is transcendental follows from $e^{0}+e^{i \pi}=0$. By the above theorem, this could not happen if $\pi$ were algebraic because then $i \pi$ would also be algebraic. Recall these algebraic numbers form a field and $i$ is clearly algebraic, being a root of $x^{2}+1$. This fact about $\pi$ was first proved by Lindemann in 1882 and then the general theorem above was proved by Weierstrass in 1885. This fact that $\pi$ is transcendental solved
an old problem called squaring the circle which was to construct a square with the same area as a circle using a straight edge and compass. Such numbers are all algebraic. Thus the fact $\pi$ is transcendental implies this problem is impossible. ${ }^{2}$

### 3.5 Exercises

1. Let $p(x) \in \mathbb{F}[x]$ and suppose that $p(x)$ is the minimum polynomial for $a \in \mathbb{F}$. Consider a field extension of $\mathbb{F}$ called $\mathbb{G}$. Thus $a \in \mathbb{G}$ also. Show that the minimum polynomial of $a$ with coefficients in $\mathbb{G}$ must divide $p(x)$.
2. Here is a polynomial in $\mathbb{Q}[x]: x^{2}+x+3$. Show it is irreducible in $\mathbb{Q}[x]$. Now consider $x^{2}-x+1$. Show that in $\mathbb{Q}[x] /\left(x^{2}+x+3\right)$ it follows that $\left[x^{2}-x+1\right] \neq 0$. Find its inverse in $\mathbb{Q}[x] /\left(x^{2}+x+3\right)$
3. Here is a polynomial in $\mathbb{Q}[x]: x^{2}-x+2$. Show it is irreducible in $\mathbb{Q}[x]$. Now consider $x+2$. Show that in $\mathbb{Q}[x] /\left(x^{2}-x+2\right)$ it follows that $[x+2] \neq 0$. Find its inverse in $\mathbb{Q}[x] /\left(x^{2}-x+2\right)$.
4. Here is a polynomial in $\mathbb{Z}_{3}[x]: x^{2}+x+\overline{2}$. Show it is irreducible in $\mathbb{Z}_{3}[x]$. Show $[x+\overline{2}]$ is not zero in $\mathbb{Z}_{3}[x] /\left(x^{2}+x+\overline{2}\right)$. Now find its inverse in $\mathbb{Z}_{3}[x] /\left(x^{2}+x+\overline{2}\right)$.
5. Suppose the degree of $p(x)$ is $r$ where $p(x)$ is an irreducible monic polynomial with coefficients in a field $\mathbb{F}$. It was shown that the dimension of $\mathbb{F}[x] /(p(x))$ is $r$ and that a basis is $\left\{1,[x],\left[x^{2}\right], \cdots,\left[x^{r-1}\right]\right\}$. Now let $A$ be an $r \times r$ matrix and let $q_{i}(x)=\sum_{k=1}^{r} A_{i j} x^{j-1}$. Show that $\left\{\left[q_{1}(x)\right], \cdots,\left[q_{r}(x)\right]\right\}$ is a basis for $\mathbb{F}[x] /(p(x))$ if and only if the matrix $A$ is invertible.
6. Suppose you have $W$ a subspace of a finite dimensional vector space $V$. Suppose also that $\operatorname{dim}(W)=\operatorname{dim}(V)$. Tell why $W=V$.
7. Suppose $V$ is a vector space with field of scalars $\mathbb{F}$. Let $T \in \mathscr{L}(V, W)$, the space of linear transformations mapping $V$ onto $W$ where $W$ is another vector space (See Problem 23 on Page 62.). Define an equivalence relation on $V$ as follows. $\boldsymbol{v} \sim \boldsymbol{w}$ means $\boldsymbol{v}-\boldsymbol{w} \in \operatorname{ker}(T)$. Recall that $\operatorname{ker}(T) \equiv\{\boldsymbol{v}: T \boldsymbol{v}=\mathbf{0}\}$. Show this is an equivalence relation. Now for $[\boldsymbol{v}]$ an equivalence class define $T^{\prime}[\boldsymbol{v}] \equiv T \boldsymbol{v}$. Show this is well defined. Also show that with the operations

$$
[\boldsymbol{v}]+[\boldsymbol{w}] \equiv[\boldsymbol{v}+\boldsymbol{w}], \alpha[\boldsymbol{v}] \equiv[\alpha \boldsymbol{v}]
$$

this set of equivalence classes, denoted by $V / \operatorname{ker}(T)$ is a vector space. Show next that $T^{\prime}: V / \operatorname{ker}(T) \rightarrow W$ is one to one. This new vector space, $V / \operatorname{ker}(T)$ is called a quotient space. Show its dimension equals the difference between the dimension of $V$ and the dimension of $\operatorname{ker}(T)$.
8. $\uparrow$ Suppose now that $W=T(V)$. Then show that $T^{\prime}$ in the above is one to one and onto. Explain why $\operatorname{dim}(V / \operatorname{ker}(T))=\operatorname{dim}(T(V))$. Now see Problem 25 on Page 62. Show that $\operatorname{rank}(T)+\operatorname{null}(T)=\operatorname{dim}(V)$.

[^2]9. Let $V$ be an $n$ dimensional vector space and let $W$ be a subspace. Generalize the Problem 7 to define and give properties of $V / W$. What is its dimension? What is a basis?
10. A number is transcendental if it is not the root of any nonzero polynomial with rational coefficients. As mentioned, there are many known transcendental numbers. Suppose $\alpha$ is a real transcendental number. Show that $\left\{1, \alpha, \alpha^{2}, \cdots\right\}$ is a linearly independent set of real numbers if the field of scalars is the rational numbers.
11. Suppose $\mathbb{F}$ is a countable field and let $\mathbb{A}$ be the algebraic numbers, those numbers in $\mathbb{G}$ which are roots of a polynomial in $\mathbb{F}[x]$. Show $\mathbb{A}$ is also countable.
12. It was shown in the chapter that $\mathbb{A}$ is a field. Here $\mathbb{A}$ are the numbers in $\mathbb{R}$ which are roots of a rational polynomial. Then it was shown in Problem 11 that it was actually countable. Show that $\mathbb{A}+i \mathbb{A}$ is also an example of a countable field.

## Chapter 4

## Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out that such rectangular arrays of numbers are important from many other different points of view. Numbers are also called scalars. In general, scalars are just elements of some field.

A matrix is a rectangular array of numbers from a field $\mathbb{F}$. For example, here is a matrix.

$$
\left(\begin{array}{cccc}
1 & 2 & 3 & 4 \\
5 & 2 & 8 & 7 \\
6 & -9 & 1 & 2
\end{array}\right)
$$

This matrix is a $3 \times 4$ matrix because there are three rows and four columns. The columns stand upright and are listed in order from left to right. The columns are horizontal and listed in order from top to bottom. The convention in dealing with matrices is to always list the rows first and then the columns. Also, you can remember the columns are like columns in a Greek temple. They stand up right while the rows just lie there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position 2,3 because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase Rowman Catholic. The symbol, $\left(a_{i j}\right)$ refers to a matrix in which the $i$ denotes the row and the $j$ denotes the column. Using this notation on the above matrix, $a_{23}=8, a_{32}=-9, a_{12}=2$, etc.

There are various operations which are done on matrices. They can sometimes be added, multiplied by a scalar and sometimes multiplied.

Definition 4.0.1 Let $A=\left(a_{i j}\right)$ and $B=\left(b_{i j}\right)$ be two $m \times n$ matrices. Then $A+B=C$ where $C=\left(c_{i j}\right)$ for $c_{i j}=a_{i j}+b_{i j}$. Also if $x$ is a scalar, $x A=C$ where the $i j^{\text {th }}$ entry of $C$ is $c_{i j}=x a_{i j}$ where the $i j^{\text {th }}$ entry of $A$ is $a_{i j}$. In short, $c_{i j}=x a_{i j}$. The number $A_{i j}$ will also typically refer to the $i j^{\text {th }}$ entry of the matrix $A$. The zero matrix, denoted by 0 will be the matrix consisting of all zeros.

Do not be upset by the use of the subscripts, $i j$. The expression $c_{i j}=a_{i j}+b_{i j}$ is just saying that you add corresponding entries to get the result of summing two matrices as discussed above.

Note that there are $2 \times 3$ zero matrices, $3 \times 4$ zero matrices, etc. In fact for every size there is a zero matrix.

With this definition, the following properties are all obvious but you should verify all of these properties are valid for $A, B$, and $C, m \times n$ matrices and 0 an $m \times n$ zero matrix.

$$
\begin{equation*}
A+B=B+A \tag{4.1}
\end{equation*}
$$

the commutative law of addition,

$$
\begin{equation*}
(A+B)+C=A+(B+C), \tag{4.2}
\end{equation*}
$$

the associative law for addition,

$$
\begin{equation*}
A+0=A \tag{4.3}
\end{equation*}
$$

the existence of an additive identity,

$$
\begin{equation*}
A+(-A)=0 \tag{4.4}
\end{equation*}
$$

the existence of an additive inverse. Also, for $\alpha, \beta$ scalars, the following also hold.

$$
\begin{gather*}
\alpha(A+B)=\alpha A+\alpha B  \tag{4.5}\\
(\alpha+\beta) A=\alpha A+\beta A  \tag{4.6}\\
\alpha(\beta A)=\alpha \beta(A)  \tag{4.7}\\
1 A=A \tag{4.8}
\end{gather*}
$$

These properties are just the vector space axioms discussed earlier and the fact that the $m \times n$ matrices satisfy these axioms is what is meant by saying this set of matrices with addition and scalar multiplication as defined above forms a vector space.

Definition 4.0.2 Matrices which are $n \times 1$ or $1 \times n$ are especially called vectors and are often denoted by a bold letter. Thus

$$
\boldsymbol{x}=\left(\begin{array}{c}
x_{1} \\
\vdots \\
x_{n}
\end{array}\right)
$$

is an $n \times 1$ matrix also called a column vector while a $1 \times n$ matrix of the form

$$
\left(\begin{array}{lll}
x_{1} & \cdots & x_{n}
\end{array}\right)
$$

is referred to as a row vector.
All the above is fine, but the real reason for considering matrices is that they can be multiplied. This is where things quit being banal. The following is the definition of multiplying an $m \times n$ matrix times a $n \times 1$ vector. Then after this, the product of two matrices is considered.

Definition 4.0.3 First of all, define the product of a $1 \times n$ matrix and a $n \times 1$ matrix.

$$
\left(\begin{array}{lll}
x_{1} & \cdots & x_{n}
\end{array}\right)\left(\begin{array}{c}
y_{1} \\
\vdots \\
y_{n}
\end{array}\right)=\sum_{i} x_{i} y_{i}
$$

If you have $A$ an $m \times n$ matrix and $B$ is an $n \times p$ matrix, then $A B$ will be an $m \times p$ matrix whose $i j^{\text {th }}$ entry is the product of the $i^{\text {th }}$ row of $A$ on the left with the $j^{\text {th }}$ column of $B$ on the right. Thus

$$
(A B)_{i j} \equiv \sum_{k=1}^{n} A_{i k} B_{k j}
$$

and if $B=\left(\begin{array}{lll}\boldsymbol{b}_{1} & \cdots & \boldsymbol{b}_{n}\end{array}\right), A B=\left(\begin{array}{lll}A \boldsymbol{b}_{1} & \cdots & A \boldsymbol{b}_{n}\end{array}\right)$. You can do $(m \times n) \times(n \times p)$ but in order to multiply, you must have the number of columns of the matrix on the left equal to the number of rows of the matrix on the right or else the rule just given makes no sense.

To see the last claim, note that the $j^{\text {th }}$ column of $A B$ involves $\boldsymbol{b}_{j}$ and is of the form

$$
\left(\begin{array}{ccc}
A_{11} & \cdots & A_{1 n} \\
\vdots & & \vdots \\
A_{m 1} & \cdots & A_{m n}
\end{array}\right)\left(\begin{array}{c}
B_{1 j} \\
\vdots \\
B_{n j}
\end{array}\right)=\left(\begin{array}{c}
\sum_{k=1}^{n} A_{1 k} B_{k j} \\
\vdots \\
\sum_{k=1}^{n} A_{m k} B_{k j}
\end{array}\right)=A \boldsymbol{b}_{j}
$$

Here is an example.
Example 4.0.4 Compute the following product in $\mathbb{Z}_{5}$. That is, all the numbers are interpreted as residue classes.

$$
\left(\begin{array}{llll}
1 & 2 & 1 & 3 \\
0 & 2 & 1 & 3 \\
2 & 1 & 4 & 1
\end{array}\right)\left(\begin{array}{ll}
1 & 2 \\
2 & 3 \\
4 & 1 \\
1 & 1
\end{array}\right)
$$

Doing the arithmetic in $\mathbb{Z}_{5}$, you get

$$
\left(\begin{array}{llll}
1 & 2 & 1 & 3 \\
0 & 2 & 1 & 3 \\
2 & 1 & 4 & 1
\end{array}\right)\left(\begin{array}{ll}
1 & 2 \\
2 & 3 \\
4 & 1 \\
1 & 1
\end{array}\right)=\left(\begin{array}{ll}
2 & 2 \\
1 & 0 \\
1 & 2
\end{array}\right)
$$

### 4.1 Properties of Matrix Multiplication

It is sometimes possible to multiply matrices in one order but not in the other order. For example,

$$
\left(\begin{array}{lll}
1 & 2 & 1 \\
2 & 1 & 2
\end{array}\right)\left(\begin{array}{ll}
1 & 2 \\
2 & 1
\end{array}\right) \text { and }\left(\begin{array}{ll}
1 & 2 \\
2 & 1
\end{array}\right)\left(\begin{array}{lll}
1 & 2 & 1 \\
2 & 1 & 2
\end{array}\right)
$$

What if it makes sense to multiply them in either order? Will they be equal then?
Example 4.1.1 Compare $\left(\begin{array}{ll}1 & 2 \\ 3 & 4\end{array}\right)\left(\begin{array}{ll}0 & 1 \\ 1 & 0\end{array}\right)$ and $\left(\begin{array}{ll}0 & 1 \\ 1 & 0\end{array}\right)\left(\begin{array}{ll}1 & 2 \\ 3 & 4\end{array}\right)$.
The first product is

$$
\left(\begin{array}{ll}
1 & 2 \\
3 & 4
\end{array}\right)\left(\begin{array}{ll}
0 & 1 \\
1 & 0
\end{array}\right)=\left(\begin{array}{ll}
2 & 1 \\
4 & 3
\end{array}\right)
$$

the second product is

$$
\left(\begin{array}{ll}
0 & 1 \\
1 & 0
\end{array}\right)\left(\begin{array}{ll}
1 & 2 \\
3 & 4
\end{array}\right)=\left(\begin{array}{ll}
3 & 4 \\
1 & 2
\end{array}\right)
$$

and you see these are not equal. Therefore, you cannot conclude that $A B=B A$ for matrix multiplication. However, there are some properties which do hold.

Proposition 4.1.2 If all multiplications and additions make sense, the following hold for matrices, $A, B, C$ and $a, b$ scalars.

$$
\begin{gather*}
A(a B+b C)=a(A B)+b(A C)  \tag{4.9}\\
(B+C) A=B A+C A  \tag{4.10}\\
A(B C)=(A B) C \tag{4.11}
\end{gather*}
$$

Proof: Using the above definition of matrix multiplication,

$$
\begin{aligned}
(A(a B+b C))_{i j} & =\sum_{k} A_{i k}(a B+b C)_{k j}=\sum_{k} A_{i k}\left(a B_{k j}+b C_{k j}\right) \\
& =a \sum_{k} A_{i k} B_{k j}+b \sum_{k} A_{i k} C_{k j}=a(A B)_{i j}+b(A C)_{i j} \\
& =(a(A B)+b(A C))_{i j}
\end{aligned}
$$

showing that $A(B+C)=A B+A C$ as claimed. Formula 4.10 is entirely similar.
Consider 4.11, the associative law of multiplication. Before reading this, review the definition of matrix multiplication in terms of entries of the matrices.

$$
\begin{aligned}
(A(B C))_{i j} & =\sum_{k} A_{i k}(B C)_{k j}=\sum_{k} A_{i k} \sum_{l} B_{k l} C_{l j} \\
& =\sum_{l}(A B)_{i l} C_{l j}=((A B) C)_{i j}
\end{aligned}
$$

Another important operation on matrices is that of taking the transpose. The following example shows what is meant by this operation, denoted by placing a $T$ as an exponent on the matrix.

$$
\left(\begin{array}{cc}
1 & 1+2 i \\
3 & 1 \\
2 & 6
\end{array}\right)^{T}=\left(\begin{array}{ccc}
1 & 3 & 2 \\
1+2 i & 1 & 6
\end{array}\right)
$$

What happened? The first column became the first row and the second column became the second row. Thus the $3 \times 2$ matrix became a $2 \times 3$ matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. This motivates the following definition of the transpose of a matrix.

Definition 4.1.3 Let $A$ be an $m \times n$ matrix. Then $A^{T}$ denotes the $n \times m$ matrix which is defined as follows.

$$
\left(A^{T}\right)_{i j}=A_{j i}
$$

The transpose of a matrix has the following important property.
Lemma 4.1.4 Let $A$ be an $m \times n$ matrix and let $B$ be a $n \times p$ matrix. Then

$$
\begin{equation*}
(A B)^{T}=B^{T} A^{T} \tag{4.12}
\end{equation*}
$$

and if $\alpha$ and $\beta$ are scalars,

$$
\begin{equation*}
(\alpha A+\beta B)^{T}=\alpha A^{T}+\beta B^{T} \tag{4.13}
\end{equation*}
$$

Proof: From the definition,

$$
\left((A B)^{T}\right)_{i j}=(A B)_{j i}=\sum_{k} A_{j k} B_{k i}=\sum_{k}\left(B^{T}\right)_{i k}\left(A^{T}\right)_{k j}=\left(B^{T} A^{T}\right)_{i j}
$$

4.13 is left as an exercise.

Definition 4.1.5 An $n \times n$ matrix $A$ is said to be symmetric if $A=A^{T}$. It is said to be skew symmetric if $A^{T}=-A$.

Example 4.1.6 Let

$$
A=\left(\begin{array}{ccc}
2 & 1 & 3 \\
1 & 5 & -3 \\
3 & -3 & 7
\end{array}\right)
$$

Then A is symmetric.
Example 4.1.7 Let

$$
A=\left(\begin{array}{ccc}
0 & 1 & 3 \\
-1 & 0 & 2 \\
-3 & -2 & 0
\end{array}\right)
$$

Then A is skew symmetric.
There is a special matrix called $I$ and defined by $I_{i j}=\delta_{i j}$ where $\delta_{i j}$ is the Kronecker symbol defined by

$$
\delta_{i j}=\left\{\begin{array}{l}
1 \text { if } i=j \\
0 \text { if } i \neq j
\end{array}\right.
$$

It is called the identity matrix because it is a multiplicative identity in the following sense.
Lemma 4.1.8 Suppose $A$ is an $m \times n$ matrix and $I_{n}$ is the $n \times n$ identity matrix. Then $A I_{n}=A$. If $I_{m}$ is the $m \times m$ identity matrix, it also follows that $I_{m} A=A$.

Proof: $\left(A I_{n}\right)_{i j}=\sum_{k} A_{i k} \delta_{k j}=A_{i j}$ and so $A I_{n}=A$. The other case is left as an exercise for you.

Definition 4.1.9 An $n \times n$ matrix $A$ has an inverse $A^{-1}$ if and only if there exists a matrix, denoted as $A^{-1}$ such that $A A^{-1}=A^{-1} A=I$ where $I=\left(\delta_{i j}\right)$ for

$$
\delta_{i j} \equiv\left\{\begin{array}{l}
1 \text { if } i=j \\
0 \text { if } i \neq j
\end{array}\right.
$$

Such a matrix is called invertible.
If it acts like an inverse, then it is the inverse. This is the message of the following proposition.
Proposition 4.1.10 Suppose $A B=B A=I$. Then $B=A^{-1}$.
Proof: From the definition, $B$ is an inverse for $A$. Could there be another one $B^{\prime}$ ?

$$
B^{\prime}=B^{\prime} I=B^{\prime}(A B)=\left(B^{\prime} A\right) B=I B=B
$$

Thus, the inverse, if it exists, is unique.

### 4.2 Finding the Inverse of a Matrix

Later a formula is given for the inverse of a matirx. However, it is not a good way to find the inverse for a matrix. There is a much easier way and it is this which is presented here. It is also important to note that not all matrices have inverses.

Example 4.2.1 Let $A=\left(\begin{array}{ll}1 & 1 \\ 1 & 1\end{array}\right)$. Does $A$ have an inverse?
One might think $A$ would have an inverse because it does not equal zero. However,

$$
\left(\begin{array}{ll}
1 & 1 \\
1 & 1
\end{array}\right)\binom{-1}{1}=\binom{0}{0}
$$

and if $A^{-1}$ existed, this could not happen because you could multiply on the left by the inverse $A$ and conclude the vector $(-1,1)^{T}=(0,0)^{T}$. Thus the answer is that $A$ does not have an inverse.

Suppose you want to find $B$ such that $A B=I$. Let

$$
B=\left(\begin{array}{lll}
b_{1} & \cdots & b_{n}
\end{array}\right)
$$

Also the $i^{\text {th }}$ column of $I$ is

$$
\boldsymbol{e}_{i}=\left(\begin{array}{lllllll}
0 & \cdots & 0 & 1 & 0 & \cdots & 0
\end{array}\right)^{T}
$$

Thus, if $A B=I, \boldsymbol{b}_{i}$, the $i^{t h}$ column of $B$ must satisfy the equation $A \boldsymbol{b}_{i}=\boldsymbol{e}_{i}$. The augmented matrix for finding $b_{i}$ is $\left(A \mid e_{i}\right)$. Thus, by doing row operations till $A$ becomes $I$, you end up with $\left(I \mid b_{i}\right)$ where $b_{i}$ is the solution to $A b_{i}=e_{i}$. Now the same sequence of row operations works regardless of the right side of the agumented matrix $\left(A \mid e_{i}\right)$ and so you can save trouble by simply doing the following.

$$
(A \mid I) \xrightarrow{\text { row operations }}(I \mid B)
$$

and the $i^{t h}$ column of $B$ is $\boldsymbol{b}_{i}$, the solution to $A \boldsymbol{b}_{i}=\boldsymbol{e}_{i}$. Thus $A B=I$.
This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the Gauss Jordan procedure. It produces the inverse if the matrix has one. Actually, it produces the right inverse.

Procedure 4.2.2 Suppose $A$ is an $n \times n$ matrix. To find $A^{-1}$ if it exists, form the augmented $n \times 2 n$ matrix, $(A \mid I)$ and then do row operations until you obtain an $n \times 2 n$ matrix of the form

$$
\begin{equation*}
(I \mid B) \tag{4.14}
\end{equation*}
$$

if possible. When this has been done, $B=A^{-1}$. The matrix $A$ has an inverse exactly when it is possible to do row operations and end up with one like 4.14.

Here is a fundamental theorem which describes when a matrix has an inverse.
Theorem 4.2.3 Let $A$ be an $n \times n$ matrix. Then $A^{-1}$ exists if and only if the columns of $A$ are a linearly independent set. Also, if A has a right inverse, then it has an inverse which equals the right inverse.

Proof: $\Rightarrow$ If $A^{-1}$ exists, then $A^{-1} A=I$ and so $A \boldsymbol{x}=0$ if and only if $\boldsymbol{x}=\mathbf{0}$. Why? But this says that the columns of $A$ are linearly independent.
$\Leftarrow$ Say the columns are linearly independent. Then they form a basis for $\mathbb{F}^{n}$. Thus there exists $\boldsymbol{b}_{i} \in \mathbb{F}^{n}$ such that

$$
A b_{i}=e_{i}
$$

where $e_{i}$ is the column vector with 1 in the $i^{t h}$ position and zeros elsewhere. Then from the way we multiply matrices,

$$
A\left(\begin{array}{lll}
b_{1} & \cdots & b_{n}
\end{array}\right)=\left(\begin{array}{lll}
e_{1} & \cdots & e_{n}
\end{array}\right)=I
$$

Thus $A$ has a right inverse. Now letting $B \equiv\left(\begin{array}{lll}\boldsymbol{b}_{1} & \cdots & \boldsymbol{b}_{n}\end{array}\right)$, it follows that $B \boldsymbol{x}=\mathbf{0}$ if and only if $\boldsymbol{x}=\mathbf{0}$. However, this is nothing but a statement that the columns of $B$ are linearly independent. Hence, by what was just shown, $B$ has a right inverse $C, B C=I$. Then from $A B=I$, it follows that

$$
A=A(B C)=(A B) C=I C=C
$$

and so $A B=B C=B A=I$. Thus the inverse exists.
Finally, if $A B=I$, then $B \boldsymbol{x}=\mathbf{0}$ if and only if $\boldsymbol{x}=\mathbf{0}$ and so the columns of $B$ are a linearly independent set in $\mathbb{F}^{n}$. Therefore, it has a right inverse $C$ which by a repeat of the above argument is $A$. Thus $A B=B A=I$.

Similarly, if $A$ has a left inverse then it has an inverse which is the same as the left inverse.

The theorem gives a condition for the existence of the inverse and the above procedure gives a method for finding it.

Example 4.2.4 Let $A=\left(\begin{array}{lll}1 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2\end{array}\right)$. Find $A^{-1}$ in arithmetic of $\mathbb{Z}_{3}$.
Form the augmented matrix

$$
\left(\begin{array}{llllll}
1 & 0 & 1 & 1 & 0 & 0 \\
1 & 2 & 1 & 0 & 1 & 0 \\
1 & 1 & 2 & 0 & 0 & 1
\end{array}\right)
$$

Now do row operations in $\mathbb{Z}_{3}$ until the $n \times n$ matrix on the left becomes the identity matrix. This yields after some computations,

$$
\left(\begin{array}{llllll}
1 & 0 & 0 & 0 & 2 & 2 \\
0 & 1 & 0 & 1 & 2 & 0 \\
0 & 0 & 1 & 1 & 1 & 1
\end{array}\right)
$$

and so the inverse of $A$ is the matrix on the right,

$$
\left(\begin{array}{lll}
0 & 2 & 2 \\
1 & 2 & 0 \\
1 & 1 & 1
\end{array}\right)
$$

Checking the answer is easy. Just multiply the matrices and see if it works.

$$
\left(\begin{array}{lll}
1 & 0 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2
\end{array}\right)\left(\begin{array}{lll}
0 & 2 & 2 \\
1 & 2 & 0 \\
1 & 1 & 1
\end{array}\right)=\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)
$$

All arithmetic is done in $\mathbb{Z}_{3}$. Always check your answer because if you are like some of us, you will usually have made a mistake.

Example 4.2.5 Let $A=\left(\begin{array}{ccc}6 & -1 & 2 \\ -1 & 2 & -1 \\ 2 & -1 & 1\end{array}\right)$. Find $A^{-1}$ in $\mathbb{Q}$.
Set up the augmented matrix $(A \mid I)$

$$
\left(\begin{array}{cccccc}
6 & -1 & 2 & 1 & 0 & 0 \\
-1 & 2 & -1 & 0 & 1 & 0 \\
2 & -1 & 1 & 0 & 0 & 1
\end{array}\right)
$$

Now find row reduced echelon form

$$
\left(\begin{array}{cccccc}
1 & 0 & 0 & 1 & -1 & -3 \\
0 & 1 & 0 & -1 & 2 & 4 \\
0 & 0 & 1 & -3 & 4 & 11
\end{array}\right)
$$

Thus the inverse is

$$
\left(\begin{array}{ccc}
1 & -1 & -3 \\
-1 & 2 & 4 \\
-3 & 4 & 11
\end{array}\right)
$$

Example 4.2.6 Let $A=\left(\begin{array}{lll}1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4\end{array}\right)$. Find $A^{-1}$ in $\mathbb{Q}$.
This time there is no inverse because the columns are not linearly independent. This can be seen by solving the equation

$$
\left(\begin{array}{lll}
1 & 2 & 2 \\
1 & 0 & 2 \\
2 & 2 & 4
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

and finding that there is a nonzero solution which is equivalent to the columns being a dependent set. Thus, by Theorem 4.2.3, there is no inverse.

Example 4.2.7 Consider the matrix

$$
\left(\begin{array}{lll}
1 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 5
\end{array}\right)
$$

Find its inverse in arithmetic of $\mathbb{Q}$ and then find its inverse in $\mathbb{Z}_{5}$.

It has an inverse in $\mathbb{Q}$.

$$
\left(\begin{array}{ccc}
1 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 5
\end{array}\right)^{-1}=\left(\begin{array}{ccc}
1 & -1 & 0 \\
0 & 1 & 0 \\
0 & 0 & \frac{1}{5}
\end{array}\right)
$$

However, in $\mathbb{Z}_{5}$ it has no inverse because $5=0$ in $\mathbb{Z}_{5}$ and so in $\mathbb{Z}_{5}^{3}$, the columns are dependent.
Example 4.2.8 Here is a matrix. $\left(\begin{array}{ll}2 & 1 \\ 1 & 2\end{array}\right)$. Find its inverse in the arithmetic of $\mathbb{Q}$ and then in $\mathbb{Z}_{3}$.

It has an inverse in the arithmetic of $\mathbb{Q} .\left(\begin{array}{ll}2 & 1 \\ 1 & 2\end{array}\right)^{-1}=\left(\begin{array}{cc}\frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3}\end{array}\right)$ However, there is no inverse in the arithmetic of $\mathbb{Z}_{3}$. Indeed, the row reduced echelon form of

$$
\left(\begin{array}{lll}
2 & 1 & 0 \\
1 & 2 & 0
\end{array}\right)
$$

computed in $\mathbb{Z}_{3}$ is $\left(\begin{array}{lll}1 & 2 & 0 \\ 0 & 0 & 0\end{array}\right)$ and so $\binom{1}{1} \in \operatorname{ker}\left(\begin{array}{ll}2 & 1 \\ 1 & 2\end{array}\right)$ which shows that the columns are not independent so there is no inverse in $\mathbb{Z}_{3}^{2}$.

The field of residue classes is not of major importance in this book, but it is included to emphasize that these considerations are completely algebraic in nature, depending only on field axioms. There is no geometry or analysis involved here.

### 4.3 Linear Relations and Row Operations

Suppose you have the following system of equations.

$$
\begin{gathered}
x-5 w-3 z=1 \\
2 w+x+y+z=2 \\
2 w+x+y+z=3
\end{gathered}
$$

You could write it in terms of matrix multiplication as follows.

$$
\left(\begin{array}{cccc}
1 & 0 & -3 & -5 \\
1 & 1 & 1 & 2 \\
1 & 1 & 1 & 2
\end{array}\right)\left(\begin{array}{c}
x \\
y \\
z \\
w
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
3
\end{array}\right)
$$

You could also write it in terms of vector addition as follows.

$$
x\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)+y\left(\begin{array}{l}
0 \\
1 \\
1
\end{array}\right)+z\left(\begin{array}{c}
-3 \\
1 \\
1
\end{array}\right)+w\left(\begin{array}{c}
-5 \\
2 \\
2
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
3
\end{array}\right)
$$

When you find a solution to the system of equations, you are really finding the scalars $x, y, z, w$ such that the vector on the right is the above linear combination of the columns.

We considered writing this system as an augmented matrix

$$
\left(\begin{array}{ccccc}
1 & 0 & -3 & -5 & 1 \\
1 & 1 & 1 & 2 & 2 \\
1 & 1 & 1 & 2 & 3
\end{array}\right)
$$

and then row reducing it to get a matrix in row reduced echelon form from which it was easy to see the solution, finding the last column as a linear combination of the preceding columns. However, this process of row reduction also shows the fourth column as a linear combination of the first three and the third as a linear combination of the first two, so when you reduce to row reduced echelon form, you are really solving many systems of equations at the same time. The important thing was the observation that the row operations did not change the solution set of the system.

However, this could be said differently. The row operations did not change the set of scalars which yield the last column as a linear combination of the first four. Similarly, the row operations did not change the scalars to obtain the fourth column as a linear combination of the first three, and so forth. In other words, if a column is a linear combination of the preceding columns, then after doing row operations, that column will still be the same linear combination of the preceding columns. By permuting the columns, placing a chosen column on the right, the same argument shows that any column after the row operation is the same linear combination of the other columns as it was before the row operation.

Such a relation between a column and other columns will be called a linear relation. Thus we have the following significant observation which is stated here as a theorem.

Theorem 4.3.1 Row operations preserve all linear relations between columns.
Now here is a slightly different description of the row reduced echelon form.
Definition 4.3.2 Let $\boldsymbol{e}_{i}$ denote the column vector which has all zero entries except for the $i^{\text {th }}$ slot which is one. An $m \times n$ matrix is said to be in row reduced echelon form if, in viewing successive columns from left to right, the first nonzero column encountered is $\boldsymbol{e}_{1}$ and if you have encountered $e_{1}, e_{2}, \cdots, e_{k}$, the next column is either $e_{k+1}$ or is a linear combination of the vectors, $e_{1}, e_{2}, \cdots, e_{k}$.

Earlier an algorithm was presented which will produce a matrix in row reduced echelon form. A natural question is whether there is only one row reduced echelon form. In fact, there is only one and this follows easily from the above definition.

Suppose you had two $B, C$ in row reduced echelon form and these came from the same matrix $A$ through row operations. Then they have zero columns in the same positions because row operations preserve all zero columns. Also $B, C$ have $e_{1}$ in the same position because its position is that of the first column of $A$ which is not zero. Similarly $\boldsymbol{e}_{2}, \boldsymbol{e}_{3}$ and so forth must be in the same positions because of the above definition where these positions are defined in terms of a column being the first in $A$ when viewed from the left to the right which is not a linear combination of the columns before it. As to a column after $\boldsymbol{e}_{k}$ and before $e_{k+1}$ if there is such, these columns are an ordered list top to bottom of the scalars which give this column in $A$ as a linear combination of the columns to its left because all linear relations between columns are preserved by doing row operations. Thus
$B, C$ must be exactly the same. This is why there is only one row reduced echelon form for a given matrix and it justifies the use of the definite article when referring to the row reduced echelon form.

This proves the following theorem.
Theorem 4.3.3 The row reduced echelon form is unique.
Now from this theorem, we can obtain the following.
Theorem 4.3.4 Let A be an $n \times n$ matrix. Then it is invertible if and only if there is a sequence of row operations which produces $I$.

Proof: $\Rightarrow$ Since $A$ is invertible, it follows from Theorem 4.2.3 that the columns of $A$ must be independent. Hence, in the row reduced echelon form for $A$, the columns must be $e_{1}, e_{2}, \cdots, e_{n}$ in order from left to right. In other words, there is a sequence of row operations which produces $I$.
$\Leftarrow$ Now suppose such a sequence of row operations produces $I$. Then since row operations preserve linear combinations between columns, it follows that no column is a linear combination of the others and consequently the columns are linearly independent. By Theorem 4.2.3 again, $A$ is invertible.

It would be possible to define things like rank in terms of the row reduced echelon form and this is often done. However, in this book, these things will be defined in terms of vector space language and the row reduced echelon form will be a useful tool to determine the rank.

Definition 4.3.5 Let $A$ be an $m \times n$ matrix, the entries being in $\mathbb{F}$ a field. Then $\operatorname{rank}(A)$ is defined as the dimension of $\operatorname{Im}(A) \equiv A\left(\mathbb{F}^{n}\right)$. Note that, from the way we multiply matrices times a vector, this is just the same as the dimension of $\operatorname{span}($ columns of $A)$, sometimes called the column space.

Now here is a very useful result.
Proposition 4.3.6 Let $A$ be an $m \times n$ matrix. Then $\operatorname{rank}(A)$ equals the number of pivot columns in the row reduced echelon form of $A$. These are the columns of $A$ which are not $a$ linear combination of those columns on the left.

Proof: This is obvious if the matrix is already in row reduced echelon form. In this case, the pivot columns consist of $\boldsymbol{e}_{1}, \boldsymbol{e}_{2}, \cdots, \boldsymbol{e}_{r}$ and every other column is a linear combination of these. Thus the rank of this matrix is $r$ because these vectors are obviously linearly independent. However, the linear relations between a column and its preceeding columns are preserved by row operations and so the columns in $A$ corresponding to the first occurance of $e_{1}$, first occurance of $e_{2}$ and so forth, in the row reduced echelon form, the pivot columns, are also a basis for the span of the columns of $A$ and so there are $r$ of these.

Note that from the description of the row reduced echelon form, the rank is also equal to the number of nonzero rows in the row reduced echelon form.

### 4.4 Block Multiplication of Matrices

Consider the following problem

$$
\left(\begin{array}{ll}
A & B \\
C & D
\end{array}\right)\left(\begin{array}{ll}
E & F \\
G & H
\end{array}\right)
$$

You know how to do this. You get

$$
\left(\begin{array}{cc}
A E+B G & A F+B H \\
C E+D G & C F+D H
\end{array}\right) .
$$

Now what if instead of numbers, the entries, $A, B, C, D, E, F, G$ are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that this is true.

Suppose $A$ is a matrix of the form

$$
A=\left(\begin{array}{ccc}
A_{11} & \cdots & A_{1 m}  \tag{4.15}\\
\vdots & \ddots & \vdots \\
A_{r 1} & \cdots & A_{r m}
\end{array}\right)
$$

where $A_{i j}$ is a $s_{i} \times p_{j}$ matrix where $s_{i}$ is constant for $j=1, \cdots, m$ for each $i=1, \cdots, r$. Such a matrix is called a block matrix, also a partitioned matrix. How do you get the block $A_{i j}$ ? Here is how for $A$ an $m \times n$ matrix:

$$
\overbrace{\left(\begin{array}{lll}
\mathbf{0} & I_{s_{i} \times s_{i}} & \mathbf{0}
\end{array}\right)}^{s_{i} \times m} A \overbrace{\left(\begin{array}{c}
\mathbf{0}  \tag{4.16}\\
I_{p_{j} \times p_{j}} \\
\mathbf{0}
\end{array}\right)}^{n \times p_{j}} .
$$

In the block column matrix on the right, you need to have $c_{j}-1$ rows of zeros above the small $p_{j} \times p_{j}$ identity matrix where the columns of $A$ involved in $A_{i j}$ are $c_{j}, \cdots, c_{j}+p_{j}-1$ and in the block row matrix on the left, you need to have $r_{i}-1$ columns of zeros to the left of the $s_{i} \times s_{i}$ identity matrix where the rows of $A$ involved in $A_{i j}$ are $r_{i}, \cdots, r_{i}+s_{i}$. An important observation to make is that the matrix on the right specifies columns to use in the block and the one on the left specifies the rows used. Thus the block $A_{i j}$ in this case is a matrix of size $s_{i} \times p_{j}$. There is no overlap between the blocks of $A$. Thus the identity $n \times n$ identity matrix corresponding to multiplication on the right of $A$ is of the form

$$
\left(\begin{array}{ccc}
I_{p_{1} \times p_{1}} & & 0 \\
& \ddots & \\
0 & & I_{p_{m} \times p_{m}}
\end{array}\right)
$$

these little identity matrices don't overlap. A similar conclusion follows from consideration of the matrices $I_{s_{i} \times s_{i}}$.

Next consider the question of multiplication of two block matrices. Let $B$ be a block matrix of the form

$$
\left(\begin{array}{ccc}
B_{11} & \cdots & B_{1 p}  \tag{4.17}\\
\vdots & \ddots & \vdots \\
B_{r 1} & \cdots & B_{r p}
\end{array}\right)
$$

and $A$ is a block matrix of the form

$$
\left(\begin{array}{ccc}
A_{11} & \cdots & A_{1 m}  \tag{4.18}\\
\vdots & \ddots & \vdots \\
A_{p 1} & \cdots & A_{p m}
\end{array}\right)
$$

and that for all $i, j$, it makes sense to multiply $B_{i s} A_{s j}$ for all $s \in\{1, \cdots, p\}$. (That is the two matrices, $B_{i s}$ and $A_{s j}$ are conformable.) and that for fixed $i j$, it follows $B_{i s} A_{s j}$ is the same size for each $s$ so that it makes sense to write $\sum_{s} B_{i s} A_{s j}$.

The following theorem says essentially that when you take the product of two matrices, you can do it two ways. One way is to simply multiply them forming $B A$. The other way is to partition both matrices, formally multiply the blocks to get another block matrix and this one will be $B A$ partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

Lemma 4.4.1 Consider the following product.

$$
\left(\begin{array}{l}
0 \\
I \\
0
\end{array}\right)\left(\begin{array}{lll}
0 & I & 0
\end{array}\right)
$$

where the first is $n \times r$ and the second is $r \times n$. The small identity matrix I is an $r \times r$ matrix and there are $l$ zero rows above $I$ and $l$ zero columns to the left of $I$ in the right matrix. Then the product of these matrices is a block matrix of the form

$$
\left(\begin{array}{lll}
0 & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{array}\right)
$$

Proof: From the definition of the way you multiply matrices, the product is

$$
\left(\left(\begin{array}{l}
0 \\
I \\
0
\end{array}\right) 0 \cdots\left(\begin{array}{l}
0 \\
I \\
0
\end{array}\right) 0\left(\begin{array}{c}
0 \\
I \\
0
\end{array}\right) e_{1} \cdots\left(\begin{array}{l}
0 \\
I \\
0
\end{array}\right) e_{r}\left(\begin{array}{l}
0 \\
I \\
0
\end{array}\right) 0 \cdots\left(\begin{array}{c}
0 \\
I \\
0
\end{array}\right) 0\right)
$$

which yields the claimed result. In the formula $\boldsymbol{e}_{j}$ refers to the column vector of length $r$ which has a 1 in the $j^{t h}$ position.
Theorem 4.4.2 Let $B$ be a $q \times p$ block matrix as in 4.17 and let $A$ be a $p \times n$ block matrix as in 4.18 such that $B_{i s}$ is conformable with $A_{s j}$ and each product, $B_{i s} A_{s j}$ for $s=1, \cdots, p$ is of the same size so they can be added. Then BA can be obtained as a block matrix such that the $i j^{\text {th }}$ block is of the form

$$
\begin{equation*}
\sum_{s} B_{i s} A_{s j} \tag{4.19}
\end{equation*}
$$

Proof: From 4.16

$$
B_{i s} A_{s j}=\left(\begin{array}{lll}
\mathbf{0} & I_{r_{i} \times r_{i}} & \mathbf{0}
\end{array}\right) B\left(\begin{array}{c}
\mathbf{0} \\
I_{p_{s} \times p_{s}} \\
\mathbf{0}
\end{array}\right)\left(\begin{array}{lll}
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0}
\end{array}\right) A\left(\begin{array}{c}
\mathbf{0} \\
I_{q_{j} \times q_{j}} \\
\mathbf{0}
\end{array}\right)
$$

where here it is assumed $B_{i s}$ is $r_{i} \times p_{s}$ and $A_{s j}$ is $p_{s} \times q_{j}$. The product involves the $s^{t h}$ block in the $i^{\text {th }}$ row of blocks for $B$ and the $s^{\text {th }}$ block in the $j^{\text {th }}$ column of $A$. Thus there are the same number of rows above the $I_{p_{s} \times p_{s}}$ as there are columns to the left of $I_{p_{s} \times p_{s}}$ in those two inside matrices. Then from Lemma 4.4.1

$$
\left(\begin{array}{c}
\mathbf{0} \\
I_{p_{s} \times p_{s}} \\
\mathbf{0}
\end{array}\right)\left(\begin{array}{lll}
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0}
\end{array}\right)=\left(\begin{array}{ccc}
\mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0}
\end{array}\right)
$$

Since the blocks of small identity matrices do not overlap,

$$
\sum_{s}\left(\begin{array}{ccc}
\mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0}
\end{array}\right)=\left(\begin{array}{ccc}
I_{p_{1} \times p_{1}} & & 0 \\
& \ddots & \\
0 & & I_{p_{p} \times p_{p}}
\end{array}\right)=I
$$

and so $\sum_{s} B_{i s} A_{s j}=$

$$
\begin{aligned}
& \sum_{s}\left(\begin{array}{lll}
\mathbf{0} & I_{r_{i} \times r_{i}} & \mathbf{0}
\end{array}\right) B\left(\begin{array}{c}
\mathbf{0} \\
I_{p_{s} \times p_{s}} \\
\mathbf{0}
\end{array}\right)\left(\begin{array}{lll}
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0}
\end{array}\right) A\left(\begin{array}{c}
\mathbf{0} \\
I_{q_{j} \times q_{j}} \\
\mathbf{0}
\end{array}\right) \\
= & \left(\begin{array}{lll}
\mathbf{0} & I_{r_{i} \times r_{i}} & \mathbf{0}
\end{array}\right) B \sum_{s}\left(\begin{array}{c}
\mathbf{0} \\
I_{p_{s} \times p_{s}} \\
\mathbf{0}
\end{array}\right)\left(\begin{array}{lll}
\mathbf{0} & I_{p_{s} \times p_{s}} & \mathbf{0}
\end{array}\right) A\left(\begin{array}{c}
\mathbf{0} \\
I_{q_{j} \times q_{j}} \\
\mathbf{0}
\end{array}\right) \\
= & \left(\begin{array}{lll}
\mathbf{0} & I_{r_{i} \times r_{i}} & \mathbf{0}
\end{array}\right) B I A\left(\begin{array}{c}
\mathbf{0} \\
I_{q_{j} \times q_{j}} \\
\mathbf{0}
\end{array}\right)=\left(\begin{array}{lll}
\mathbf{0} & I_{r_{i} \times r_{i}} & \mathbf{0}
\end{array}\right) B A\left(\begin{array}{c}
\mathbf{0} \\
I_{q_{j} \times q_{j}} \\
\mathbf{0}
\end{array}\right)
\end{aligned}
$$

which equals the $i j^{\text {th }}$ block of $B A$. Hence the $i j^{\text {th }}$ block of $B A$ equals the formal multiplication according to matrix multiplication, $\sum_{s} B_{i s} A_{s j}$.

Example 4.4.3 Let an $n \times n$ matrix have the form

$$
A=\left(\begin{array}{ll}
a & b \\
c & P
\end{array}\right)
$$

where $P$ is $n-1 \times n-1$. Multiply it by

$$
B=\left(\begin{array}{ll}
p & \boldsymbol{q} \\
\boldsymbol{r} & Q
\end{array}\right)
$$

where $B$ is also an $n \times n$ matrix and $Q$ is $n-1 \times n-1$.

You use block multiplication

$$
\left(\begin{array}{ll}
a & \boldsymbol{b} \\
\boldsymbol{c} & P
\end{array}\right)\left(\begin{array}{ll}
p & \boldsymbol{q} \\
r & Q
\end{array}\right)=\left(\begin{array}{ll}
a p+\boldsymbol{b r} & a \boldsymbol{q}+\boldsymbol{b} Q \\
p \boldsymbol{c}+P r & c q+P Q
\end{array}\right)
$$

Note that this all makes sense. For example, $\boldsymbol{b}=1 \times n-1$ and $\boldsymbol{r}=n-1 \times 1$ so $\boldsymbol{b} \boldsymbol{r}$ is a $1 \times 1$. Similar considerations apply to the other blocks.

Here is a very significant application. A matrix is called block diagonal if it has all zeros except for square blocks down the diagonal. That is, it is of the form

$$
A=\left(\begin{array}{cccc}
A_{1} & & & 0 \\
& A_{2} & & \\
& & \ddots & \\
0 & & & A_{m}
\end{array}\right)
$$

where $A_{j}$ is a $r_{j} \times r_{j}$ matrix whose main diagonal lies on the main diagonal of $A$. Then by block multiplication, if $p \in \mathbb{N}$ the positive integers,

$$
A^{p}=\left(\begin{array}{cccc}
A_{1}^{p} & & & 0  \tag{4.20}\\
& A_{2}^{p} & & \\
& & \ddots & \\
0 & & & A_{m}^{p}
\end{array}\right)
$$

Also, $A^{-1}$ exists if and only if each block is invertible and in fact, $A^{-1}$ is given by the above when $p=-1$.

### 4.5 Elementary Matrices

The elementary matrices result from doing a row operation to the identity matrix. Recall the following definition.

## Definition 4.5.1 The row operations consist of the following

1. Switch two rows.
2. Multiply a row by a nonzero number.
3. Replace a row by a multiple of another row added to it.

The elementary matrices are given in the following definition.
Definition 4.5.2 The elementary matrices consist of those matrices which result by applying a row operation to an identity matrix. Those which involve switching rows of the identity are called permutation matrices ${ }^{1}$.

[^3]The importance of elementary matrices is that when you multiply on the left by one, it does the row operation which was used to produce the elementary matrix.

Now consider what these elementary matrices look like. First consider the one which involves switching row $i$ and row $j$ where $i<j$. This matrix is of the form

$$
\left(\begin{array}{lllll}
\ddots & & & & \\
& 0 & & 1 & \\
& & \ddots & & \\
& 1 & & 0 & \\
& & & & \ddots
\end{array}\right)
$$

Note how the $i^{\text {th }}$ and $j^{\text {th }}$ rows are switched in the identity matrix and there are thus all ones on the main diagonal except for those two positions indicated. The two exceptional rows are shown. The $i^{t h}$ row was the $j^{t h}$ and the $j^{t h}$ row was the $i^{t h}$ in the identity matrix. Now consider what this does to a column vector.

$$
\left(\begin{array}{ccccc}
\ddots & & & & \\
& 0 & & 1 & \\
& & \ddots & & \\
& 1 & & 0 & \\
& & & & \ddots
\end{array}\right)\left(\begin{array}{c}
\vdots \\
x_{i} \\
\vdots \\
x_{j} \\
\vdots
\end{array}\right)=\left(\begin{array}{c}
\vdots \\
x_{j} \\
\vdots \\
x_{i} \\
\vdots
\end{array}\right)
$$

Now denote by $P^{i j}$ the elementary matrix which comes from the identity from switching rows $i$ and $j$. From what was just explained,

$$
P^{i j}\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{i 1} & a_{i 2} & \cdots & a_{i p} \\
\vdots & \vdots & & \vdots \\
a_{j 1} & a_{j 2} & \cdots & a_{j p} \\
\vdots & \vdots & & \vdots
\end{array}\right)=\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{j 1} & a_{j 2} & \cdots & a_{j p} \\
\vdots & \vdots & & \vdots \\
a_{i 1} & a_{i 2} & \cdots & a_{i p} \\
\vdots & \vdots & & \vdots
\end{array}\right)
$$

This has established the following lemma.
Lemma 4.5.3 Let $P^{i j}$ denote the elementary matrix which involves switching the $i^{\text {th }}$ and the $j^{\text {th }}$ rows. Then

$$
P^{i j} A=B
$$

where $B$ is obtained from $A$ by switching the $i^{\text {th }}$ and the $j^{\text {th }}$ rows.
Example 4.5.4 Consider the following.

$$
\left(\begin{array}{lll}
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{array}\right)\left(\begin{array}{ll}
a & b \\
g & d \\
e & f
\end{array}\right)=\left(\begin{array}{ll}
g & d \\
a & b \\
e & f
\end{array}\right)
$$

Next consider the row operation which involves multiplying the $i^{t h}$ row by a nonzero constant, $c$. The elementary matrix which results from applying this operation to the $i^{\text {th }}$ row of the identity matrix is of the form

$$
\left(\begin{array}{ccccc}
\ddots & & & & 0 \\
& 1 & & & \\
& & c & & \\
& & & 1 & \\
0 & & & & \ddots
\end{array}\right)
$$

Now consider what this does to a column vector.

$$
\left(\begin{array}{ccccc}
\ddots & & & & 0 \\
& 1 & & & \\
& & c & & \\
& & & 1 & \\
0 & & & & \ddots
\end{array}\right)\left(\begin{array}{c}
\vdots \\
v_{i-1} \\
v_{i} \\
v_{i+1} \\
\vdots
\end{array}\right)=\left(\begin{array}{c}
\vdots \\
v_{i-1} \\
c v_{i} \\
v_{i+1} \\
\vdots
\end{array}\right)
$$

Denote by $E(c, i)$ this elementary matrix which multiplies the $i^{t h}$ row of the identity by the nonzero constant, $c$. Then from what was just discussed,

$$
E(c, i)\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{(i-1) 1} & a_{(i-1) 2} & \cdots & a_{(i-1) p} \\
a_{i 1} & a_{i 2} & \cdots & a_{i p} \\
a_{(i+1) 1} & a_{(i+1) 2} & \cdots & a_{(i+1) p} \\
\vdots & \vdots & & \vdots
\end{array}\right)=\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{(i-1) 1} & a_{(i-1) 2} & \cdots & a_{(i-1) p} \\
c a_{i 1} & c a_{i 2} & \cdots & c a_{i p} \\
a_{(i+1) 1} & a_{(i+1) 2} & \cdots & a_{(i+1) p} \\
\vdots & \vdots & & \vdots
\end{array}\right)
$$

This proves the following lemma.
Lemma 4.5.5 Let $E(c, i)$ denote the elementary matrix corresponding to the row operation in which the $i^{\text {th }}$ row is multiplied by the nonzero constant, $c$. Thus $E(c, i)$ involves multiplying the $i^{\text {th }}$ row of the identity matrix by $c$. Then

$$
E(c, i) A=B
$$

where $B$ is obtained from $A$ by multiplying the $i^{\text {th }}$ row of $A$ by $c$.
Example 4.5.6 Consider this.

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 5 & 0 \\
0 & 0 & 1
\end{array}\right)\left(\begin{array}{ll}
a & b \\
c & d \\
e & f
\end{array}\right)=\left(\begin{array}{cc}
a & b \\
5 c & 5 d \\
e & f
\end{array}\right)
$$

Finally consider the third of these row operations. Denote by $E(c \times i+j)$ the elementary matrix which replaces the $j^{t h}$ row with the $j^{\text {th }}$ row added to $c$ times the $i^{\text {th }}$ row. In case
$i<j$ this will be of the form

$$
\left(\begin{array}{ccccc}
\ddots & & & & 0 \\
& 1 & & & \\
& & \ddots & & \\
& c & & 1 & \\
0 & & & & \ddots
\end{array}\right)
$$

Now consider what this does to a column vector.

$$
\left(\begin{array}{ccccc}
\ddots & & & & 0 \\
& 1 & & & \\
& & \ddots & & \\
& c & & 1 & \\
& & & & \ddots
\end{array}\right)\left(\begin{array}{c}
\vdots \\
v_{i} \\
\vdots \\
v_{j} \\
\vdots
\end{array}\right)=\left(\begin{array}{c}
\vdots \\
v_{i} \\
\vdots \\
c v_{i}+v_{j} \\
\vdots
\end{array}\right)
$$

Now from this,

$$
\begin{aligned}
& E(c \times i+j)\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{i 1} & a_{i 2} & \cdots & a_{i p} \\
\vdots & \vdots & & \vdots \\
a_{j 1} & a_{j 2} & \cdots & a_{j p} \\
\vdots & \vdots & & \vdots
\end{array}\right) \\
& =\left(\begin{array}{cccc}
\vdots & \vdots & & \vdots \\
a_{i 1} & a_{i 2} & \cdots & a_{i p} \\
\vdots & \vdots & & \vdots \\
c a_{i 1}+a_{j 1} & c a_{i 2}+a_{j 2} & \cdots & c a_{i p}+a_{j p} \\
\vdots & \vdots & & \vdots
\end{array}\right)
\end{aligned}
$$

The case where $i>j$ is handled similarly. This proves the following lemma.
Lemma 4.5.7 Let $E(c \times i+j)$ denote the elementary matrix obtained from I by replacing the $j^{\text {th }}$ row with $c$ times the $i^{\text {th }}$ row added to it. Then

$$
E(c \times i+j) A=B
$$

where $B$ is obtained from $A$ by replacing the $j^{\text {th }}$ row of $A$ with itself added to $c$ times the $i^{\text {th }}$ row of $A$.

Example 4.5.8 Consider the third row operation.

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
2 & 0 & 1
\end{array}\right)\left(\begin{array}{ll}
a & b \\
c & d \\
e & f
\end{array}\right)=\left(\begin{array}{cc}
a & b \\
c & d \\
2 a+e & 2 b+f
\end{array}\right)
$$

The next theorem is the main result.
Theorem 4.5.9 To perform any of the three row operations on a matrix $A$, it suffices to do the row operation on the identity matrix obtaining an elementary matrix $E$ and then take the product, $E A$. Furthermore, if $E$ is an elementary matrix, then there is another elementary matrix $\hat{E}$ such that $E \hat{E}=\hat{E} E=I$.

Proof: The first part of this theorem has been proved in Lemmas 4.5.3-4.5.7. It only remains to verify the claim about the matrix $\hat{E}$. Consider first the elementary matrices corresponding to row operation of type three.

$$
E(-c \times i+j) E(c \times i+j)=I
$$

This follows because the first matrix takes $c$ times row $i$ in the identity and adds it to row $j$. When multiplied on the left by $E(-c \times i+j)$ it follows from the first part of this theorem that you take the $i^{\text {th }}$ row of $E(c \times i+j)$ which coincides with the $i^{\text {th }}$ row of $I$ since that row was not changed, multiply it by $-c$ and add to the $j^{t h}$ row of $E(c \times i+j)$ which was the $j^{t h}$ row of $I$ added to $c$ times the $i^{\text {th }}$ row of $I$. Thus $E(-c \times i+j)$ multiplied on the left, undoes the row operation which resulted in $E(c \times i+j)$. The same argument applied to the product $E(c \times i+j) E(-c \times i+j)$ replacing $c$ with $-c$ in the argument yields that this product is also equal to $I$. Therefore, there is an elementary matrix of the same sort which when multiplied by $E$ on either side gives the identity.

Similar reasoning shows that for $E(c, i)$ the elementary matrix which comes from multiplying the $i^{t h}$ row by the nonzero constant $c$, you can take $\hat{E}=E((1 / c), i)$.

Finally, consider $P^{i j}$ which involves switching the $i^{t h}$ and the $j^{\text {th }}$ rows $P^{i j} P^{i j}=I$ because by the first part of this theorem, multiplying on the left by $P^{i j}$ switches the $i^{t h}$ and $j^{t h}$ rows of $P^{i j}$ which was obtained from switching the $i^{t h}$ and $j^{t h}$ rows of the identity. First you switch them to get $P^{i j}$ and then you multiply on the left by $P^{i j}$ which switches these rows again and restores the identity matrix.

Using Theorem 4.3.4, this shows the following result.
Theorem 4.5.10 Let $A$ be an $n \times n$ matrix. Then if $R$ is its row reduced echelon form, there is a sequence of elementary matrices $E_{i}$ such that

$$
E_{1} E_{2} \cdots E_{m} A=R
$$

In particular, $A$ is invertible if and only if there is a sequence of elementary matrices as above such that $E_{1} E_{2} \cdots E_{m} A=I$. Inverting these, $A=E_{m}^{-1} \cdots E_{2}^{-1} E_{1}^{-1}$ a product of elementary matrices.

### 4.6 Exercises

1. In 4.1-4.8 describe $-A$ and 0 .
2. Let $A$ be an $n \times n$ matrix. Show $A$ equals the sum of a symmetric and a skew symmetric matrix.
3. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form $a_{i i}$. It runs from the upper left down to the lower right.
4. We used the fact that the columns of a matrix $A$ are independent if and only if $A \boldsymbol{x}=\mathbf{0}$ has only the zero solution for $\boldsymbol{x}$. Why is this so?
5. If $A$ is $m \times n$ where $n>m$, explain why there exists $\boldsymbol{x} \in \mathbb{F}^{n}$ such that $A \boldsymbol{x}=\mathbf{0}$ but $\boldsymbol{x} \neq \mathbf{0}$.
6. Using only the properties $4.1-4.8$ show $-A$ is unique.
7. Using only the properties $4.1-4.8$ show 0 is unique.
8. Using only the properties $4.1-4.8$ show $0 A=0$. Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for $m \times n$ matrices.
9. Using only the properties 4.1-4.8 and previous problems show $(-1) A=-A$.
10. Prove that $I_{m} A=A$ where $A$ is an $m \times n$ matrix.
11. Let $A$ and be a real $m \times n$ matrix and let $\boldsymbol{x} \in \mathbb{R}^{n}$ and $\boldsymbol{y} \in \mathbb{R}^{m}$. Show $(A \boldsymbol{x}, \boldsymbol{y})_{\mathbb{R}^{m}}=$ $\left(\boldsymbol{x}, A^{T} \boldsymbol{y}\right)_{\mathbb{R}^{n}}$ where $(\cdot, \cdot)_{\mathbb{R}^{k}}$ denotes the dot product in $\mathbb{R}^{k}$. You need to know about the dot product. It will be discussed later but hopefully it has been seen in physics or calculus.
12. Use the result of Problem 11 to verify directly that $(A B)^{T}=B^{T} A^{T}$ without making any reference to subscripts. However, note that the treatment in the chapter did not depend on a dot product.
13. Let $\boldsymbol{x}=(-1,-1,1)$ and $\boldsymbol{y}=(0,1,2)$. Find $\boldsymbol{x}^{T} \boldsymbol{y}$ and $\boldsymbol{x} \boldsymbol{y}^{T}$ if possible.
14. Give an example of matrices, $A, B, C$ such that $B \neq C, A \neq 0$, and yet $A B=A C$.
15. Let $A=\left(\begin{array}{cc}1 & 1 \\ -2 & -1 \\ 1 & 2\end{array}\right), B=\left(\begin{array}{ccc}1 & -1 & -2 \\ 2 & 1 & -2\end{array}\right), C=\left(\begin{array}{ccc}1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0\end{array}\right)$. Find if possible the following products. $A B, B A, A C, C A, C B, B C$.
16. Show $(A B)^{-1}=B^{-1} A^{-1}$.
17. Show that if $A$ is an invertible $n \times n$ matrix, then so is $A^{T}$ and $\left(A^{T}\right)^{-1}=\left(A^{-1}\right)^{T}$.
18. Show that if $A$ is an $n \times n$ invertible matrix and $\boldsymbol{x}$ is a $n \times 1$ matrix such that $A \boldsymbol{x}=\boldsymbol{b}$ for $\boldsymbol{b}$ an $n \times 1$ matrix, then $\boldsymbol{x}=A^{-1} \boldsymbol{b}$.
19. Give an example of a matrix $A$ such that $A^{2}=I$ and yet $A \neq I$ and $A \neq-I$.
20. Give an example of matrices, $A, B$ such that neither $A$ nor $B$ equals zero and yet $A B=0$.
21. Give another example other than the one given in this section of two square matrices, $A$ and $B$ such that $A B \neq B A$.
22. Suppose $A$ and $B$ are square matrices of the same size. Which of the following are correct?
(a) $(A-B)^{2}=A^{2}-2 A B+B^{2}$
(b) $(A B)^{2}=A^{2} B^{2}$
(c) $(A+B)^{2}=A^{2}+2 A B+B^{2}$
(d) $(A+B)^{2}=A^{2}+A B+B A+B^{2}$
(e) $A^{2} B^{2}=A(A B) B$
(f) $(A+B)^{3}=A^{3}+3 A^{2} B+3 A B^{2}+B^{3}$
(g) $(A+B)(A-B)=A^{2}-B^{2}$
(h) None of the above. They are all wrong.
(i) All of the above. They are all right.
23. Let $A=\left(\begin{array}{cc}-1 & -1 \\ 3 & 3\end{array}\right)$. Find all $2 \times 2$ matrices, $B$ such that $A B=0$.
24. Prove that if $A^{-1}$ exists and $A \boldsymbol{x}=\mathbf{0}$ then $\boldsymbol{x}=\mathbf{0}$.
25. Let

$$
A=\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 1 & 4 \\
1 & 0 & 2
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.
26. Let

$$
A=\left(\begin{array}{lll}
1 & 0 & 3 \\
2 & 3 & 4 \\
1 & 0 & 2
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.
27. Let

$$
A=\left(\begin{array}{ccc}
1 & 2 & 3 \\
2 & 1 & 4 \\
4 & 5 & 10
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.
28. Let

$$
A=\left(\begin{array}{cccc}
1 & 2 & 0 & 2 \\
1 & 1 & 2 & 0 \\
2 & 1 & -3 & 2 \\
1 & 2 & 1 & 2
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why.
29. Let

$$
A=\left(\begin{array}{ll}
2 & 1 \\
1 & 3
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why. Do this in $\mathbb{Q}^{2}$ and in $\mathbb{Z}_{5}^{2}$.
30. Let

$$
A=\left(\begin{array}{ll}
2 & 1 \\
1 & 2
\end{array}\right)
$$

Find $A^{-1}$ if possible. If $A^{-1}$ does not exist, determine why. Do this in $\mathbb{Q}^{2}$ and in $\mathbb{Z}_{3}^{2}$.
31. If you have any system of equations $A \boldsymbol{x}=\boldsymbol{b}$, let $\operatorname{ker}(A) \equiv\{\boldsymbol{x}: A \boldsymbol{x}=\mathbf{0}\}$. Show that all solutions of the system $A \boldsymbol{x}=\boldsymbol{b}$ are in $\operatorname{ker}(A)+\boldsymbol{y}_{p}$ where $A \boldsymbol{y}_{p}=\boldsymbol{b}$. This means that every solution of this last equation is of the form $\boldsymbol{y}_{p}+\boldsymbol{z}$ where $A \boldsymbol{z}=\mathbf{0}$.
32. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$
\left(\begin{array}{lll}
1 & -1 & 2 \\
1 & -2 & 1 \\
3 & -4 & 5
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

33. Using Problem 32 find the general solution to the following linear system.

$$
\left(\begin{array}{lll}
1 & -1 & 2 \\
1 & -2 & 1 \\
3 & -4 & 5
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
4
\end{array}\right)
$$

34. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$
\left(\begin{array}{lll}
0 & -1 & 2 \\
1 & -2 & 1 \\
1 & -4 & 5
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

35. Using Problem 34 find the general solution to the following linear system.

$$
\left(\begin{array}{lll}
0 & -1 & 2 \\
1 & -2 & 1 \\
1 & -4 & 5
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{c}
1 \\
-1 \\
1
\end{array}\right)
$$

36. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$
\left(\begin{array}{lll}
1 & -1 & 2 \\
1 & -2 & 0 \\
3 & -4 & 4
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

37. Using Problem 36 find the general solution to the following linear system.

$$
\left(\begin{array}{lll}
1 & -1 & 2 \\
1 & -2 & 0 \\
3 & -4 & 4
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
4
\end{array}\right) .
$$

38. Show that 4.20 is valid for $p=-1$ if and only if each block has an inverse and that this condition holds if and only if $A$ is invertible.
39. Let $A$ be an $n \times n$ matrix and let $P^{i j}$ be the permutation matrix which switches the $i^{t h}$ and $j^{t h}$ rows of the identity. Show that $P^{i j} A P^{i j}$ produces a matrix which is similar to $A$ which switches the $i^{\text {th }}$ and $j^{\text {th }}$ entries on the main diagonal.
40. You could define column operations by analogy to row operations. That is, you switch two columns, multiply a column by a nonzero scalar, or add a scalar multiple of a column to another column. Let $E$ be one of these column operations applied to the identity matrix. Show that $A E$ produces the column operation on $A$ which was used to define $E$.
41. Consider the symmetric $3 \times 3$ matrices, those for which $A=A^{T}$. Show that with respect to the usual notions of addition and scalar multiplication this is a vector space of dimension 6 . What is the dimension of the set of skew symmetric matrices?
42. You have an $m \times n$ matrix of rank $r$. Explain why if you delete a column, the resulting matrix has rank $r$ or rank $r-1$.
43. Using the fact that multiplication on the left by an elementary matrix accomplishes a row operation, show easily that row operations produce no change in linear relations between columns.

## Chapter 5

## Linear Transformations

This chapter is on functions which map a vector space to another one which are also linear. The description of these is in the following definition. Linear algebra is all about understanding these kinds of mappings.

Definition 5.0.1 Let $V$ and $W$ be two finite dimensional vector spaces. A function, $L$ which maps $V$ to $W$ is called a linear transformation and written $L \in \mathscr{L}(V, W)$ if for all scalars $\alpha$ and $\beta$, and vectors $v, w, L(\alpha v+\beta w)=\alpha L(v)+\beta L(w)$.

Example 5.0.2 Let $V=\mathbb{R}^{3}, W=\mathbb{R}$, and let $\boldsymbol{a} \in \mathbb{R}^{3}$ be given vector in $V$. Define $T: V \rightarrow W$ by $T \boldsymbol{v} \equiv \sum_{i=1}^{3} a_{i} v_{i}$

It is left as an exercise to verify that this is indeed linear. Here is an interesting observation.

Proposition 5.0.3 Let $L: \mathbb{F}^{n} \rightarrow \mathbb{F}^{m}$ be linear. Then there exists a unique $m \times n$ matrix $A$ such that $L \boldsymbol{x}=A \boldsymbol{x}$ for all $\boldsymbol{x}$. Also, matrix multiplication yields a linear transformation.

Proof: Note that $\boldsymbol{x}=\sum_{i=1}^{n} x_{i} \boldsymbol{e}_{i}$ and so

$$
\begin{aligned}
& L \boldsymbol{x}=L\left(\sum_{i=1}^{n} x_{i} e_{i}\right)=\sum_{i=1}^{n} x_{i} L e_{i}=\left(\begin{array}{ccc}
L e_{1} & \cdots & L e_{n}
\end{array}\right)\left(\begin{array}{c}
x_{1} \\
\vdots \\
x_{n}
\end{array}\right) \\
& =\left(\begin{array}{lll}
L e_{1} & \cdots & L e_{n}
\end{array}\right) x
\end{aligned}
$$

The matrix is $A$. The last claim follows from the properties of matrix multiplication.
I will abuse terminology slightly and say that a $m \times n$ matrix is one to one if the linear transformation it determines is one to one, similarly for the term onto.

## 5.1 $\mathscr{L}(V, W)$ as a Vector Space

The linear transformations can be considered as a vector space as described next.
Definition 5.1.1 Given $L, M \in \mathscr{L}(V, W)$ define a new element of $\mathscr{L}(V, W)$, denoted by $L+M$ according to the rule ${ }^{1}$

$$
(L+M) v \equiv L v+M v
$$

For $\alpha$ a scalar and $L \in \mathscr{L}(V, W)$, define $\alpha L \in \mathscr{L}(V, W)$ by

$$
\alpha L(v) \equiv \alpha(L v)
$$

You should verify that all the axioms of a vector space hold for $\mathscr{L}(V, W)$ with the above definitions of vector addition and scalar multiplication. In fact, is just a subspace of the set of functions mapping $V$ to $W$ which is a vector space thanks to Example 3.0.3. What about the dimension of $\mathscr{L}(V, W)$ ? What about a basis for $\mathscr{L}(V, W)$ ?

Before answering this question, here is a useful lemma. It gives a way to define linear transformations and a way to tell when two of them are equal.

[^4]Lemma 5.1.2 Let $V$ and $W$ be vector spaces and suppose $\left\{v_{1}, \cdots, v_{n}\right\}$ is a basis for $V$. Then if $L: V \rightarrow W$ is given by $L v_{k}=w_{k} \in W$ and

$$
L\left(\sum_{k=1}^{n} a_{k} v_{k}\right) \equiv \sum_{k=1}^{n} a_{k} L v_{k}=\sum_{k=1}^{n} a_{k} w_{k}
$$

then $L$ is well defined and is in $\mathscr{L}(V, W)$. Also, if $L, M$ are two linear transformations such that $L v_{k}=M v_{k}$ for all $k$, then $M=L$.

Proof: $L$ is well defined on $V$ because, since $\left\{v_{1}, \cdots, v_{n}\right\}$ is a basis, there is exactly one way to write a given vector of $V$ as a linear combination. Next, observe that $L$ is obviously linear from the definition. If $L, M$ are equal on the basis, then if $\sum_{k=1}^{n} a_{k} v_{k}$ is an arbitrary vector of $V$,

$$
L\left(\sum_{k=1}^{n} a_{k} v_{k}\right)=\sum_{k=1}^{n} a_{k} L v_{k}=\sum_{k=1}^{n} a_{k} M v_{k}=M\left(\sum_{k=1}^{n} a_{k} v_{k}\right)
$$

and so $L=M$ because they give the same result for every vector in $V$.
The message is that when you define a linear transformation, it suffices to tell what it does to a basis.

Example 5.1.3 A basis for $\mathbb{R}^{2}$ is

$$
\binom{1}{1},\binom{1}{0}
$$

Suppose $T$ is a linear transformation which satisfies

$$
T\binom{1}{1}=\binom{2}{1}, T\binom{1}{0}=\binom{-1}{1}
$$

Find $T\binom{3}{2}$.

$$
\begin{aligned}
T\binom{3}{2} & =T\left(2\binom{1}{1}+\binom{1}{0}\right) \\
& =2 T\binom{1}{1}+T\binom{1}{0} \\
& =2\binom{2}{1}+\binom{-1}{1}=\binom{3}{3}
\end{aligned}
$$

Theorem 5.1.4 Let $V$ and $W$ be finite dimensional linear spaces of dimension $n$ and $m$ respectively. Then $\operatorname{dim}(\mathscr{L}(V, W))=m n$.

Proof: Let two sets of bases be $\left\{v_{1}, \cdots, v_{n}\right\}$ and $\left\{w_{1}, \cdots, w_{m}\right\}$ for $V$ and $W$ respectively. Using Lemma 5.1.2, let $w_{i} v_{j} \in \mathscr{L}(V, W)$ be the linear transformation defined on the basis, $\left\{v_{1}, \cdots, v_{n}\right\}$, by

$$
w_{i} v_{k}\left(v_{j}\right) \equiv w_{i} \delta_{j k}
$$

where $\delta_{i k}=1$ if $i=k$ and 0 if $i \neq k$. I will show that $L \in \mathscr{L}(V, W)$ is a linear combination of these special linear transformations called dyadics, also rank one transformations.

Then let $L \in \mathscr{L}(V, W)$. Since $\left\{w_{1}, \cdots, w_{m}\right\}$ is a basis, there exist constants, $d_{j k}$ such that

$$
L v_{r}=\sum_{j=1}^{m} d_{j r} w_{j}
$$

Now consider the following sum of dyadics. $\sum_{j=1}^{m} \sum_{i=1}^{n} d_{j i} w_{j} v_{i}$. Apply this to $v_{r}$. This yields

$$
\begin{equation*}
\sum_{j=1}^{m} \sum_{i=1}^{n} d_{j i} w_{j} v_{i}\left(v_{r}\right)=\sum_{j=1}^{m} \sum_{i=1}^{n} d_{j i} w_{j} \delta_{i r}=\sum_{j=1}^{m} d_{j r} w_{j}=L v_{r} \tag{5.1}
\end{equation*}
$$

Therefore, $L=\sum_{j=1}^{m} \sum_{i=1}^{n} d_{j i} w_{j} v_{i}$ showing the span of the dyadics is all of $\mathscr{L}(V, W)$.
Now consider whether these special linear transformations are a linearly independent set. Suppose

$$
\sum_{i, k} d_{i k} w_{i} v_{k}=\mathbf{0} .
$$

Are all the scalars $d_{i k}$ equal to 0 ?

$$
\mathbf{0}=\sum_{i, k} d_{i k} w_{i} v_{k}\left(v_{l}\right)=\sum_{i=1}^{m} d_{i l} w_{i}
$$

and so, since $\left\{w_{1}, \cdots, w_{m}\right\}$ is a basis, $d_{i l}=0$ for each $i=1, \cdots, m$. Since $l$ is arbitrary, this shows $d_{i l}=0$ for all $i$ and $l$. Thus these linear transformations form a basis and this shows that the dimension of $\mathscr{L}(V, W)$ is $m n$ as claimed because there are $m$ choices for the $w_{i}$ and $n$ choices for the $v_{j}$.

Note that from 5.1, these coefficients which obtain $L$ as a linear combination of the diadics are given by the equation

$$
\begin{equation*}
\sum_{j=1}^{m} d_{j r} w_{j}=L v_{r} \tag{5.2}
\end{equation*}
$$

Thus $L v_{r}$ is in the span of the $w_{j}$.

### 5.2 The Matrix of a Linear Transformation

In order to do computations based on a linear transformation, we usually work with its matrix. This is what is described here.

Theorem 5.1.4 says that the rank one transformations defined there in terms of two bases, one for $V$ and the other for $W$ are a basis for $\mathscr{L}(V, W)$. Thus if $A \in \mathscr{L}(V, W)$, there are scalars $A_{i j}$ such that

$$
A=\sum_{i=1}^{n} \sum_{j=1}^{m} A_{i j} w_{i} v_{j}
$$

Here we have $1 \leq i \leq n$ and $1 \leq j \leq m$. We can arrange these scalars in a rectangular shape as follows.

$$
\left(\begin{array}{ccccc}
A_{11} & A_{12} & \cdots & A_{1(n-1)} & A_{1 n} \\
A_{21} & A_{22} & \cdots & A_{2(n-1)} & A_{2 n} \\
\vdots & \vdots & & \vdots & \vdots \\
A_{m 1} & A_{m 2} & \cdots & A_{m(n-1)} & A_{m n}
\end{array}\right)
$$

Here this is an $m \times n$ matrix because it has $m$ rows and $n$ columns. It is called the matrix of the linear transformation $A$ with respect to the two bases $\left\{v_{1}, \cdots, v_{n}\right\}$ for $V$ and $\left\{w_{1}, \cdots, w_{m}\right\}$ for $W$. Now, as noted earlier, if $v=\sum_{r=1}^{n} x_{r} v_{r}$,

$$
\begin{gathered}
A v=\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i j} w_{i} v_{j}\left(\sum_{r=1}^{n} x_{r} v_{r}\right) \\
=\sum_{r=1}^{n} x_{r} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i j} w_{i} v_{j}\left(v_{r}\right)=\sum_{r=1}^{n} x_{r} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i j} w_{i} \delta_{j r} \\
=\sum_{r=1}^{n} x_{r} \sum_{i=1}^{m} A_{i r} w_{i}=\sum_{i=1}^{m}\left(\sum_{r=1}^{n} A_{i r} x_{r}\right) w_{i}
\end{gathered}
$$

What does this show? It shows that if the component vector of $v$ is

$$
\boldsymbol{x}=\left(\begin{array}{c}
x_{1} \\
\vdots \\
x_{n}
\end{array}\right)
$$

meaning that $v=\sum_{i} x_{i} v_{i}$, then the component vector of $w$ has $i^{t h}$ component equal to

$$
\sum_{r=1}^{n} A_{i r} x_{r}=(A \boldsymbol{x})_{i}
$$

The idea is that acting on a vector $v$ with a linear transformation $T$ yields a new vector $w$ whose component vector is obtained as the matrix of the linear transformation times the component vector of $v$. It is helpful for some of us to think of this in terms of diagrams. On the other hand, some people hate such diagrams. Use them if it helps. Otherwise ignore them and go right to the algebraic definition 5.2.

Let $\beta=\left\{v_{1}, \cdots, v_{n}\right\}$ be a basis for $V$ and let $\left\{w_{1}, \cdots, w_{m}\right\}=\gamma$ be a basis for $W$. Then let $q_{\beta}: \mathbb{F}^{n} \rightarrow V, q_{\gamma}: \mathbb{F}^{m} \rightarrow W$ be defined as

$$
q_{\beta} \boldsymbol{x} \equiv \sum_{i=1}^{n} x_{i} v_{i}, \quad q_{\gamma} \boldsymbol{y} \equiv \sum_{j=1}^{m} y_{j} w_{j}
$$

Thus these mappings are linear and take the component vector to the vector determined by the component vector.

Then the diagram which describes the matrix of the linear transformation $L$ is in the following picture.

$$
\begin{array}{lrcll}
\beta=\left\{v_{1}, \cdots, v_{n}\right\} & & L & & \\
& q_{\beta} \uparrow & \circ & \uparrow q_{\gamma} \\
& \mathbb{F}^{n} & \rightarrow & \mathbb{F}^{m}  \tag{5.3}\\
& & & \\
& {[L]_{\gamma \beta}} &
\end{array}
$$

In terms of this diagram, the matrix $[L]_{\gamma \beta}$ is the matrix chosen to make the diagram "commute". It is the matrix of the linear transformation because it takes the component
vector of $v$ to the component vector for $L v$. As implied by the diagram and as shown above, for $A=[L]_{\gamma \beta}$,

$$
L v_{i}=\sum_{j=1}^{m} A_{j i} w_{j}
$$

## Gimmick for finding matrix of a linear transformation

It may be useful to write this in the form

$$
\left(\begin{array}{lll}
L v_{1} & \cdots & L v_{n}
\end{array}\right)=\left(\begin{array}{lll}
w_{1} & \cdots & w_{m} \tag{5.4}
\end{array}\right) A, A \text { is } m \times n
$$

and multiply formally as if the $L v_{i}, w_{j}$ were numbers.
Example 5.2.1 Let $L \in \mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{m}\right)$ and let the two bases be

$$
\left\{\begin{array}{lll}
e_{1} & \cdots & e_{n}
\end{array}\right\},\left\{\begin{array}{lll}
e_{1} & \cdots & e_{m}
\end{array}\right\}
$$

$e_{i}$ denoting the column vector of zeros except for $a l$ in the $i^{\text {th }}$ position. Then from the above, you need to have

$$
L \boldsymbol{e}_{i}=\sum_{j=1}^{m} A_{j i} \boldsymbol{e}_{j}
$$

which says that

$$
\left(\begin{array}{lll}
L e_{1} & \cdots & L e_{n}
\end{array}\right)_{m \times n}=\left(\begin{array}{lll}
e_{1} & \cdots & e_{m}
\end{array}\right)_{m \times m} A_{m \times n}
$$

and so $L e_{i}$ equals the $i^{\text {th }}$ column of $A$. In other words,

$$
A=\left(\begin{array}{lll}
L e_{1} & \cdots & L e_{n}
\end{array}\right)
$$

Then for $\boldsymbol{x}=\left(\begin{array}{lll}x_{1} & \cdots & x_{n}\end{array}\right)^{T}$

$$
\begin{aligned}
A \boldsymbol{x} & =A\left(\sum_{i=1}^{n} x_{i} e_{i}\right)=\sum_{i=1}^{n} x_{i} A \boldsymbol{e}_{i} \\
& =\sum_{i=1}^{n} x_{i} L e_{i}=L\left(\sum_{i=1}^{n} x_{i} \boldsymbol{e}_{i}\right)=L \boldsymbol{x}
\end{aligned}
$$

Thus, doing $L$ to a vector $\boldsymbol{x}$ is the same as multiplying on the left by the matrix $A$.
Example 5.2.2 Let

$$
\begin{aligned}
V & \equiv\{\text { polynomials of degree } 3 \text { or less }\} \\
W & \equiv\{\text { polynomials of degree } 2 \text { or less }\}
\end{aligned}
$$

and $L \equiv D$ where $D$ is the differentiation operator. $A$ basis for $V$ is $\beta=\left\{1, x, x^{2}, x^{3}\right\}$ and a basis for $W$ is $\gamma=\left\{1, x, x^{2}\right\}$.

What is the matrix of this linear transformation with respect to this basis? Using 5.4,

$$
\left(\begin{array}{cccc}
0 & 1 & 2 x & 3 x^{2}
\end{array}\right)=\left(\begin{array}{ccc}
1 & x & x^{2}
\end{array}\right)[D]_{\gamma \beta}
$$

It follows from this that the first column of $[D]_{\gamma \beta}$ is

$$
\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
$$

The next three columns of $[D]_{\gamma \beta}$ are

$$
\left(\begin{array}{l}
1 \\
0 \\
0
\end{array}\right),\left(\begin{array}{l}
0 \\
2 \\
0
\end{array}\right),\left(\begin{array}{l}
0 \\
0 \\
3
\end{array}\right)
$$

and so

$$
[D]_{\gamma \beta}=\left(\begin{array}{cccc}
0 & 1 & 0 & 0 \\
0 & 0 & 2 & 0 \\
0 & 0 & 0 & 3
\end{array}\right)
$$

Say you have $a+b x+c x^{2}+d x^{3}$. Then doing $D$ to it gives $b+2 c x+3 d x^{2}$. The component vector of the function is

$$
\left(\begin{array}{llll}
a & b & c & d
\end{array}\right)^{T}
$$

and after doing $D$ to the function, you get for the component vector

$$
\left(\begin{array}{lll}
b & 2 c & 3 d
\end{array}\right)^{T}
$$

This is the same result you get when you multiply by $[D]$.

$$
\left(\begin{array}{llll}
0 & 1 & 0 & 0 \\
0 & 0 & 2 & 0 \\
0 & 0 & 0 & 3
\end{array}\right)\left(\begin{array}{l}
a \\
b \\
c \\
d
\end{array}\right)=\left(\begin{array}{c}
b \\
2 c \\
3 d
\end{array}\right)
$$

Of course, this is what it means to be the matrix of the transformation.
Now consider the important case where $V=\mathbb{F}^{n}, W=\mathbb{F}^{m}$, and the basis chosen is the standard basis of vectors $\boldsymbol{e}_{i}$ described above.

$$
\beta=\left\{\boldsymbol{e}_{1}, \cdots, e_{n}\right\}, \gamma=\left\{\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{m}\right\}
$$

Let $L$ be a linear transformation from $\mathbb{F}^{n}$ to $\mathbb{F}^{m}$ and let $A$ be the matrix of the transformation with respect to these bases. In this case the coordinate maps $q_{\beta}$ and $q_{\gamma}$ are simply the identity maps on $\mathbb{F}^{n}$ and $\mathbb{F}^{m}$ respectively, and can be accomplished by simply multiplying by the appropriate sized identity matrix. The requirement that $A$ is the matrix of the transformation amounts to

$$
L b=A b
$$

What about the situation where different pairs of bases are chosen for $V$ and $W$ ? How are the two matrices with respect to these choices related? Consider the following diagram which illustrates the situation.


In this diagram $q_{\beta_{i}}$ and $q_{\gamma_{i}}$ are coordinate maps as described above. From the diagram,

$$
q_{\gamma_{1}}^{-1} q_{\gamma_{2}} A_{2} q_{\beta_{2}}^{-1} q_{\beta_{1}}=A_{1}
$$

where $q_{\beta_{2}}^{-1} q_{\beta_{1}}$ and $q_{\gamma_{1}}^{-1} q_{\gamma_{2}}$ are one to one, onto, and linear maps which may be accomplished by multiplication by a square matrix. Thus there exist matrices $P, Q$ such that $P: \mathbb{F}^{n} \rightarrow \mathbb{F}^{n}$ and $Q: \mathbb{F}^{m} \rightarrow \mathbb{F}^{m}$ are invertible and

$$
P A_{2} Q=A_{1} .
$$

Example 5.2.3 Let $\beta \equiv\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ and $\gamma \equiv\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{n}\right\}$ be two bases for $V$. Let $L$ be the linear transformation which maps $\boldsymbol{v}_{i}$ to $\boldsymbol{w}_{i}$. Find $[L]_{\gamma \beta}$.

Letting $\delta_{i j}$ be the symbol which equals 1 if $i=j$ and 0 if $i \neq j$, it follows that $L=$ $\sum_{i, j} \delta_{i j} \boldsymbol{w}_{i} \boldsymbol{v}_{j}$ and so $[L]_{\gamma \beta}=I$ the identity matrix.

Definition 5.2.4 In the special case where $V=W$ and only one basis is used for $V=W$, this becomes

$$
q_{\beta_{1}}^{-1} q_{\beta_{2}} A_{2} q_{\beta_{2}}^{-1} q_{\beta_{1}}=A_{1}
$$

Letting $S$ be the matrix of the linear transformation $q_{\beta_{2}}^{-1} q_{\beta_{1}}$ with respect to the standard basis vectors in $\mathbb{F}^{n}$,

$$
\begin{equation*}
S^{-1} A_{2} S=A_{1} \tag{5.5}
\end{equation*}
$$

When this occurs, $A_{1}$ is said to be similar to $A_{2}$ and $A \rightarrow S^{-1} A S$ is called a similarity transformation.

Recall the following.
Definition 5.2.5 Let $S$ be a set. The symbol $\sim$ is called an equivalence relation on $S$ if it satisfies the following axioms.

1. $x \sim x$ for all $x \in S$. (Reflexive)
2. If $x \sim y$ then $y \sim x$. (Symmetric)
3. If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)

Definition 5.2.6 $[x]$ denotes the set of all elements of $S$ which are equivalent to $x$ and $[x]$ is called the equivalence class determined by $x$ or just the equivalence class of $x$.

Also recall the notion of equivalence classes.
Theorem 5.2.7 Let $\sim$ be an equivalence class defined on a set $S$ and let $\mathscr{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x]=[y]$ or it is not true that $x \sim y$ and $[x] \cap[y]=\emptyset$.

Theorem 5.2.8 In the vector space of $n \times n$ matrices, define

$$
A \sim B
$$

if there exists an invertible matrix $S$ such that

$$
A=S^{-1} B S
$$

Then $\sim$ is an equivalence relation and $A \sim B$ if and only if whenever $V$ is an dimensional vector space, there exists $L \in \mathscr{L}(V, V)$ and bases $\left\{v_{1}, \cdots, v_{n}\right\}$ and $\left\{w_{1}, \cdots, w_{n}\right\}$ such that $A$ is the matrix of $L$ with respect to $\left\{v_{1}, \cdots, v_{n}\right\}$ and $B$ is the matrix of $L$ with respect to $\left\{w_{1}, \cdots, w_{n}\right\}$.

Proof: $A \sim A$ because $S=I$ works in the definition. If $A \sim B$, then $B \sim A$, because

$$
A=S^{-1} B S
$$

implies $B=S A S^{-1}$. If $A \sim B$ and $B \sim C$, then $A=S^{-1} B S, B=T^{-1} C T$ and so

$$
A=S^{-1} T^{-1} C T S=(T S)^{-1} C T S
$$

which implies $A \sim C$. This verifies the first part of the conclusion.
Now let $V$ be an $n$ dimensional vector space, $A \sim B$ so $A=S^{-1} B S$ and pick a basis for V,

$$
\beta \equiv\left\{v_{1}, \cdots, v_{n}\right\}
$$

Define $L \in \mathscr{L}(V, V)$ by $L v_{i} \equiv \sum_{j} a_{j i} v_{j}$ where $A=\left(a_{i j}\right)$. Thus $A$ is the matrix of the linear transformation $L$. Consider the diagram

| $\mathbb{F}^{n}$ | $\underline{B}$ | $\mathbb{F}^{n}$ |
| ---: | ---: | ---: |
| $q_{\gamma} \downarrow$ | $\circ$ | $q_{\gamma} \downarrow$ |
| $V$ | $\underline{L}$ | $V$ |
| $q_{\beta} \uparrow$ | $\circ$ | $q_{\beta} \uparrow$ |
| $\mathbb{F}^{n}$ | $\xrightarrow{A}$ | $\mathbb{F}^{n}$ |

where $q_{\gamma}$ is chosen to make the diagram commute. Thus we need $S=q_{\gamma}^{-1} q_{\beta}$ which requires $q_{\gamma}=q_{\beta} S^{-1}$. Then it follows that $B$ is the matrix of $L$ with respect to the basis

$$
\left\{q_{\gamma} \boldsymbol{e}_{1}, \cdots, q_{\gamma} \boldsymbol{e}_{n}\right\} \equiv\left\{w_{1}, \cdots, w_{n}\right\} .
$$

That is, $A$ and $B$ are matrices of the same linear transformation $L$. Conversely, suppose whenever $V$ is an $n$ dimensional vector space, there exists $L \in \mathscr{L}(V, V)$ and bases $\left\{v_{1}, \cdots, v_{n}\right\}$ and $\left\{w_{1}, \cdots, w_{n}\right\}$ such that $A$ is the matrix of $L$ with respect to $\left\{v_{1}, \cdots, v_{n}\right\}$ and $B$ is the matrix of $L$ with respect to $\left\{w_{1}, \cdots, w_{n}\right\}$. Then it was shown above that $A \sim B$.

What if the linear transformation consists of multiplication by a matrix $A$ and you want to find the matrix of this linear transformation with respect to another basis? Is there an easy way to do it? The next proposition considers this.

Proposition 5.2.9 Let $A$ be an $m \times n$ matrix and consider it as a linear transformation by multiplication on the left by $A$. Then the matrix $M$ of this linear transformation with respect to the bases $\beta=\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{n}\right\}$ for $\mathbb{F}^{n}$ and $\gamma=\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{m}\right\}$ for $\mathbb{F}^{m}$ is given by

$$
M=\left(\begin{array}{lll}
\boldsymbol{w}_{1} & \cdots & \boldsymbol{w}_{m}
\end{array}\right)^{-1} A\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right)
$$

where $\left(\begin{array}{lll}\boldsymbol{w}_{1} & \cdots & \boldsymbol{w}_{m}\end{array}\right)$ is the $m \times m$ matrix which has $\boldsymbol{w}_{j}$ as its $j^{\text {th }}$ column. Note that also

$$
\left(\begin{array}{lll}
\boldsymbol{w}_{1} & \cdots & \boldsymbol{w}_{m}
\end{array}\right) M\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right)^{-1}=A
$$

Proof: Consider the following diagram.

$$
\begin{array}{rll} 
& A & \\
\mathbb{F}^{n} & \rightarrow & \mathbb{F}^{m} \\
q_{\beta} \uparrow & \circ & \uparrow q_{\gamma} \\
\mathbb{F}^{n} & \rightarrow & \mathbb{F}^{m} \\
& M &
\end{array}
$$

Here the coordinate maps are defined in the usual way. Thus

$$
q_{\beta}(\boldsymbol{x}) \equiv \sum_{i=1}^{n} x_{i} \boldsymbol{u}_{i}=\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right) \boldsymbol{x}
$$

Therefore, $q_{\beta}$ can be considered the same as multiplication of a vector in $\mathbb{F}^{n}$ on the left by the matrix $\left(\begin{array}{lll}\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}\end{array}\right)$. Similar considerations apply to $q_{\gamma}$. Thus it is desired to have the following for an arbitrary $\boldsymbol{x} \in \mathbb{F}^{n}$.

$$
A\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right) \boldsymbol{x}=\left(\begin{array}{lll}
\boldsymbol{w}_{1} & \cdots & \boldsymbol{w}_{n}
\end{array}\right) M \boldsymbol{x}
$$

Therefore, the conclusion of the proposition follows.
The second formula in the above is pretty useful. You might know the matrix $M$ of a linear transformation with respect to a funny basis and this formula gives the matrix of the linear transformation in terms of the usual basis which is really what you want.

Definition 5.2.10 Let $A \in \mathscr{L}(X, Y)$ where $X$ and $Y$ are finite dimensional vector spaces. Define $\operatorname{rank}(A)$ to equal the dimension of $A(X)$.

Lemma 5.2.11 Let $M$ be an $m \times n$ matrix. Then $M$ can be considered as a linear transformation as follows.

$$
M(\boldsymbol{x}) \equiv M \boldsymbol{x}
$$

That is, you multiply on the left by $M$.
Proof: This follows from the properties of matrix multiplication. In particular,

$$
M(a \boldsymbol{x}+b \boldsymbol{y})=a M \boldsymbol{x}+b M \boldsymbol{y}
$$

Note also that, as explained earlier, the image of this transformation is just the span of the columns, known as the column space.

The following theorem explains how the rank of $A$ is related to the rank of the matrix of $A$.

Theorem 5.2.12 Let $A \in \mathscr{L}(X, Y)$. Then $\operatorname{rank}(A)=\operatorname{rank}(M)$ where $M$ is the matrix of $A$ taken with respect to a pair of bases for the vector spaces $X$, and $Y$. Here $M$ is considered as a linear transformation by matrix multiplication.

Proof: Recall the diagram which describes what is meant by the matrix of $A$. Here the two bases are as indicated.

$$
\begin{array}{ccccc}
\beta=\left\{v_{1}, \cdots, v_{n}\right\} & X & \xrightarrow{A} & Y & \left\{w_{1}, \cdots, w_{m}\right\}=\gamma \\
& q_{\beta} \uparrow & \circ & \uparrow q_{\gamma} \\
\mathbb{F}^{n} & \xrightarrow{M} & \mathbb{F}^{m}
\end{array}
$$

Let $\left\{A x_{1}, \cdots, A x_{r}\right\}$ be a basis for $A X$. Thus

$$
\left\{q_{\gamma} M q_{\beta}^{-1} x_{1}, \cdots, q_{\gamma} M q_{\beta}^{-1} x_{r}\right\}
$$

is a basis for $A X$. It follows that

$$
\left\{M q_{X}^{-1} x_{1}, \cdots, M q_{X}^{-1} x_{r}\right\}
$$

is linearly independent and so $\operatorname{rank}(A) \leq \operatorname{rank}(M)$. However, one could interchange the roles of $M$ and $A$ in the above argument and thereby turn the inequality around.

The following result is a summary of many concepts.
Theorem 5.2.13 Let $L \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space. Then the following are equivalent.

1. L is one to one.
2. L maps a basis to a basis.
3. $L$ is onto.
4. If $L v=0$ then $v=0$.

Proof: Suppose first $L$ is one to one and let $\beta=\left\{v_{i}\right\}_{i=1}^{n}$ be a basis. Then if $\sum_{i=1}^{n} c_{i} L v_{i}=$ 0 it follows $L\left(\sum_{i=1}^{n} c_{i} v_{i}\right)=0$ which means that since $L(0)=0$, and $L$ is one to one, it must be the case that $\sum_{i=1}^{n} c_{i} v_{i}=0$. Since $\left\{v_{i}\right\}$ is a basis, each $c_{i}=0$ which shows $\left\{L v_{i}\right\}$ is a linearly independent set. Since there are $n$ of these, it must be that this is a basis.

Now suppose 2.). Then letting $\left\{v_{i}\right\}$ be a basis, and $y \in V$, it follows from part 2.) that there are constants, $\left\{c_{i}\right\}$ such that $y=\sum_{i=1}^{n} c_{i} L v_{i}=L\left(\sum_{i=1}^{n} c_{i} v_{i}\right)$. Thus $L$ is onto. It has been shown that 2.) implies 3.).

Now suppose 3.). Then $L(V)=V$. If $\left\{v_{1}, \cdots, v_{n}\right\}$ is a basis of $V$, then

$$
V=\operatorname{span}\left(L v_{1}, \cdots, L v_{n}\right) .
$$

It follows that $\left\{L v_{1}, \cdots, L v_{n}\right\}$ must be linearly independent because if not, one of the vectors could be deleted and you would then have a spanning set with fewer vectors than $\operatorname{dim}(V)$. If $L v=0$,

$$
v=\sum_{i} x_{i} v_{i}
$$

then doing $L$ to both sides, $0=\sum_{i} x_{i} L v_{i}$ which imiplies each $x_{i}=0$ and consequently $v=0$. Thus 4. follows.

Now suppose 4.) and suppose $L v=L w$. Then $L(v-w)=0$ and so by 4.$), v-w=0$ showing that $L$ is one to one.

Also it is important to note that composition of linear transformations corresponds to multiplication of the matrices. Consider the following diagram in which $[A]_{\gamma \beta}$ denotes the matrix of $A$ relative to the bases $\gamma$ on $Y$ and $\beta$ on $X,[B]_{\delta \gamma}$ defined similarly.

| $X$ | $\xrightarrow{A}$ | $Y$ | $\xrightarrow{B}$ | $Z$ |
| :---: | :---: | :---: | :---: | :---: |
| $q_{\beta} \uparrow$ | $\circ$ | $\uparrow q_{\gamma}$ | $\circ$ | $\uparrow q_{\delta}$ |
| $\mathbb{F}^{n}$ | $\xrightarrow{[A]_{\gamma \beta}}$ | $\mathbb{F}^{m}$ | $\xrightarrow{[B]_{\delta \gamma}}$ | $\mathbb{F}^{p}$ |

where $A$ and $B$ are two linear transformations, $A \in \mathscr{L}(X, Y)$ and $B \in \mathscr{L}(Y, Z)$. Then $B \circ A \in$ $\mathscr{L}(X, Z)$ and so it has a matrix with respect to bases given on $X$ and $Z$, the coordinate maps for these bases being $q_{\beta}$ and $q_{\delta}$ respectively. Then

$$
B \circ A=q_{\delta}[B]_{\delta \gamma} q_{\gamma}^{-1} q_{\gamma}[A]_{\gamma \beta} q_{\beta}^{-1}=q_{\delta}[B]_{\delta \gamma}[A]_{\gamma \beta} q_{\beta}^{-1}
$$

But this shows that $[B]_{\delta \gamma}[A]_{\gamma \beta}$ plays the role of $[B \circ A]_{\delta \beta}$, the matrix of $B \circ A$. Hence the matrix of $B \circ A$ equals the product of the two matrices $[A]_{\gamma \beta}$ and $[B]_{\delta \gamma}$. Of course it is interesting to note that although $[B \circ A]_{\delta \beta}$ must be unique, the matrices, $[A]_{\gamma \beta}$ and $[B]_{\delta \gamma}$ are not unique because they depend on $\gamma$, the basis chosen for $Y$.

Theorem 5.2.14 The matrix of the composition of linear transformations equals the product of the matrices of these linear transformations.

### 5.3 Rotations About a Given Vector*

As an application, consider the problem of rotating counter clockwise about a given unit vector which is possibly not one of the unit vectors in coordinate directions. First consider a pair of perpendicular unit vectors, $\boldsymbol{u}_{1}$ and $\boldsymbol{u}_{2}$ and the problem of rotating in the counterclockwise direction about $\boldsymbol{u}_{3}$ where $\boldsymbol{u}_{3}=\boldsymbol{u}_{1} \times \boldsymbol{u}_{2}$ so that $\boldsymbol{u}_{1}, \boldsymbol{u}_{2}, \boldsymbol{u}_{3}$ forms a right handed orthogonal coordinate system. See the appendix on the cross product if this is not familiar. Thus the vector $\boldsymbol{u}_{\mathbf{3}}$ is coming out of the page.


Let $T$ denote the desired rotation. Then

$$
\begin{gathered}
T\left(a \boldsymbol{u}_{1}+b \boldsymbol{u}_{2}+c \boldsymbol{u}_{3}\right)=a T \boldsymbol{u}_{1}+b T \boldsymbol{u}_{2}+c \boldsymbol{T} \boldsymbol{u}_{3} \\
=(a \cos \theta-b \sin \theta) \boldsymbol{u}_{1}+(a \sin \theta+b \cos \theta) \boldsymbol{u}_{2}+c \boldsymbol{u}_{3} .
\end{gathered}
$$

Thus in terms of the basis $\gamma \equiv\left\{\boldsymbol{u}_{1}, \boldsymbol{u}_{2}, \boldsymbol{u}_{3}\right\}$, the matrix of this transformation is

$$
[T]_{\gamma} \equiv\left(\begin{array}{ccc}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{array}\right)
$$

This is not desirable because it involves a funny basis. I want to obtain the matrix of the transformation in terms of the usual basis $\beta \equiv\left\{\boldsymbol{e}_{1}, e_{2}, e_{3}\right\}$ because it is in terms of this basis that we usually deal with vectors in $\mathbb{R}^{3}$. From Proposition 5.2.9, if $[T]_{\beta}$ is this matrix,

$$
\begin{aligned}
& \left(\begin{array}{ccc}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{array}\right) \\
= & \left(\begin{array}{lll}
\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \boldsymbol{u}_{3}
\end{array}\right)^{-1}\left[\begin{array}{l}
{[]_{\beta}}
\end{array}\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \boldsymbol{u}_{3}
\end{array}\right)\right.
\end{aligned}
$$

and so you can solve for $[T]_{\beta}$ if you know the $\boldsymbol{u}_{i}$.
Recall why this is so.

$$
\begin{array}{rcr}
\mathbb{R}^{3} & \xrightarrow{[T]_{\gamma}} & \mathbb{R}^{3} \\
q_{\gamma} \downarrow & \circ & q_{\gamma} \downarrow \\
\mathbb{R}^{3} & \xrightarrow{T} & \mathbb{R}^{3} \\
I \uparrow & \circ & I \uparrow \\
\mathbb{R}^{3} & \xrightarrow{[T]_{\beta}} & \mathbb{R}^{3}
\end{array}
$$

The map $q_{\gamma}$ is accomplished by a multiplication on the left by $\left(\begin{array}{lll}\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \boldsymbol{u}_{3}\end{array}\right)$. Thus

$$
[T]_{\beta}=q_{\gamma}[T]_{\gamma} q_{\gamma}^{-1}=\left(\begin{array}{ccc}
\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \boldsymbol{u}_{3}
\end{array}\right)[T]_{\gamma}\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \boldsymbol{u}_{3}
\end{array}\right)^{-1}
$$

Suppose the unit vector $\boldsymbol{u}_{3}$ about which the counterclockwise rotation takes place is $(a, b, c)$. Then I obtain vectors, $\boldsymbol{u}_{1}$ and $\boldsymbol{u}_{2}$ such that $\left\{\boldsymbol{u}_{1}, \boldsymbol{u}_{2}, \boldsymbol{u}_{3}\right\}$ is a right handed orthonormal system with $\boldsymbol{u}_{3}=(a, b, c)$ and then use the above result. It is of course somewhat arbitrary how this is accomplished. I will assume however, that $|c| \neq 1$ since otherwise you are looking at either clockwise or counter clockwise rotation about the positive $z$ axis and this is a problem which is fairly easy. Indeed, the matrix of such a rotation in terms of the usual basis is just

$$
\left(\begin{array}{ccc}
\cos \theta & -\sin \theta & 0  \tag{5.6}\\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{array}\right)
$$

Then let $\boldsymbol{u}_{3}=(a, b, c)$ and $\boldsymbol{u}_{2} \equiv \frac{1}{\sqrt{a^{2}+b^{2}}}(b,-a, 0)$. This one is perpendicular to $\boldsymbol{u}_{3}$. If $\left\{\boldsymbol{u}_{1}, \boldsymbol{u}_{2}, \boldsymbol{u}_{3}\right\}$ is to be a right hand system it is necessary to have

$$
\boldsymbol{u}_{1}=\boldsymbol{u}_{2} \times \boldsymbol{u}_{3}=\frac{1}{\sqrt{\left(a^{2}+b^{2}\right)\left(a^{2}+b^{2}+c^{2}\right)}}\left(-a c,-b c, a^{2}+b^{2}\right)
$$

Now recall that $\boldsymbol{u}_{3}$ is a unit vector and so the above equals

$$
\frac{1}{\sqrt{\left(a^{2}+b^{2}\right)}}\left(-a c,-b c, a^{2}+b^{2}\right)
$$

Then from the above, $A$ is given by

$$
\left(\begin{array}{ccc}
\frac{-a c}{\sqrt{\left(a^{2}+b^{2}\right)}} & \frac{b}{\sqrt{a^{2}+b^{2}}} & a \\
\frac{-b c}{\sqrt{\left(a^{2}+b^{2}\right)}} & \frac{-a}{\sqrt{a^{2}+b^{2}}} & b \\
\sqrt{a^{2}+b^{2}} & 0 & c
\end{array}\right)\left(\begin{array}{ccc}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{array}\right)\left(\begin{array}{ccc}
\frac{-a c}{\sqrt{\left(a^{2}+b^{2}\right)}} & \frac{b}{\sqrt{a^{2}+b^{2}}} & a \\
\frac{-b c}{\sqrt{\left(a^{2}+b^{2}\right)}} & \frac{-a}{\sqrt{a^{2}+b^{2}}} & b \\
\sqrt{a^{2}+b^{2}} & 0 & c
\end{array}\right)^{-1}
$$

It is easy to take the inverse of this matrix on the left. You can check right away that its inverse is nothing but its transpose. Then doing the computation and then some simplification yields

$$
=\left(\begin{array}{ccc}
a^{2}+\left(1-a^{2}\right) \cos \theta & a b(1-\cos \theta)-c \sin \theta & a c(1-\cos \theta)+b \sin \theta  \tag{5.7}\\
a b(1-\cos \theta)+c \sin \theta & b^{2}+\left(1-b^{2}\right) \cos \theta & b c(1-\cos \theta)-a \sin \theta \\
a c(1-\cos \theta)-b \sin \theta & b c(1-\cos \theta)+a \sin \theta & c^{2}+\left(1-c^{2}\right) \cos \theta
\end{array}\right) .
$$

With this, it is clear how to rotate clockwise about the unit vector, $(a, b, c)$. Just rotate counter clockwise through an angle of $-\theta$. Thus the matrix for this clockwise rotation is just

$$
=\left(\begin{array}{ccc}
a^{2}+\left(1-a^{2}\right) \cos \theta & a b(1-\cos \theta)+c \sin \theta & a c(1-\cos \theta)-b \sin \theta \\
a b(1-\cos \theta)-c \sin \theta & b^{2}+\left(1-b^{2}\right) \cos \theta & b c(1-\cos \theta)+a \sin \theta \\
a c(1-\cos \theta)+b \sin \theta & b c(1-\cos \theta)-a \sin \theta & c^{2}+\left(1-c^{2}\right) \cos \theta
\end{array}\right) .
$$

In deriving 5.7 it was assumed that $c \neq \pm 1$ but even in this case, it gives the correct answer. Suppose for example that $c=1$ so you are rotating in the counter clockwise direction about the positive $z$ axis. Then $a, b$ are both equal to zero and 5.7 reduces to 5.6.

### 5.4 Exercises

1. If $A, B$, and $C$ are each $n \times n$ matrices and $A B C$ is invertible, why are each of $A, B$, and $C$ invertible?
2. Give an example of a $3 \times 2$ matrix with the property that the linear transformation determined by this matrix is one to one but not onto.
3. Explain why $A \boldsymbol{x}=\mathbf{0}$ always has a solution whenever $A$ is a linear transformation.
4. Recall that a line in $\mathbb{R}^{n}$ is of the form $\boldsymbol{x}+t \boldsymbol{v}$ where $t \in \mathbb{R}$. Recall that $\boldsymbol{v}$ is a "direction vector". Show that if $T: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ is linear, then the image of $T$ is either a line or a point.
5. In the following examples, a linear transformation, $T$ is given by specifying its action on a basis $\beta$. Find its matrix with respect to this basis.
(a) $T\binom{1}{2}=2\binom{1}{2}+1\binom{-1}{1}, T\binom{-1}{1}=\binom{-1}{1}$
(b) $T\binom{0}{1}=2\binom{0}{1}+1\binom{-1}{1}, T\binom{-1}{1}=\binom{0}{1}$
(c) $T\binom{1}{0}=2\binom{1}{2}+1\binom{1}{0}, T\binom{1}{2}=1\binom{1}{0}-\binom{1}{2}$
6. $\uparrow$ In each example above, find a matrix $A$ such that for every $\boldsymbol{x} \in \mathbb{R}^{2}, T \boldsymbol{x}=A \boldsymbol{x}$.
7. Consider the linear transformation $T_{\theta}$ which rotates every vector in $\mathbb{R}^{2}$ through the angle of $\theta$. Find the matrix $A_{\theta}$ such that $T_{\theta} \boldsymbol{x}=A_{\theta} \boldsymbol{x}$. Hint: You need to have the columns of $A_{\theta}$ be $T e_{1}$ and $T e_{2}$. Review why this is before using this. Then simply find these vectors from trigonometry.
8. $\uparrow$ If you did the above problem right, you got

$$
A_{\theta}=\left(\begin{array}{cc}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{array}\right)
$$

Derive the famous trig. identities for the sum of two angles by using the fact that $A_{\theta+\phi}=A_{\theta} A_{\phi}$ and the above description.
9. Let $\beta=\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{n}\right\}$ be a basis for $\mathbb{F}^{n}$ and let $T: \mathbb{F}^{n} \rightarrow \mathbb{F}^{n}$ be defined as follows.

$$
T\left(\sum_{k=1}^{n} a_{k} \boldsymbol{u}_{k}\right)=\sum_{k=1}^{n} a_{k} b_{k} \boldsymbol{u}_{k}
$$

First show that $T$ is a linear transformation. Next show that the matrix of $T$ with respect to this basis is $[T]_{\beta}=$

$$
\left(\begin{array}{ccc}
b_{1} & & \\
& \ddots & \\
& & b_{n}
\end{array}\right)
$$

Show that the above definition is equivalent to simply specifying $T$ on the basis vectors of $\beta$ by

$$
T\left(\boldsymbol{u}_{k}\right)=b_{k} \boldsymbol{u}_{k}
$$

10. Let $T$ be given by specifying its action on the vectors of a basis

$$
\beta=\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{n}\right\}
$$

as follows.

$$
T \boldsymbol{u}_{k}=\sum_{j=1}^{n} a_{j k} \boldsymbol{u}_{j}
$$

Letting $A=\left(a_{i j}\right)$, verify that $[T]_{\beta}=A$. It is done in the chapter, but go over it yourself. Show that $[T]_{\gamma}=$

$$
\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right)[T]_{\beta}\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n} \tag{5.8}
\end{array}\right)^{-1}
$$

11. Let $\boldsymbol{a}$ be a fixed vector. The function $T_{\boldsymbol{a}}$ defined by $T_{\boldsymbol{a}} \boldsymbol{v}=\boldsymbol{a}+\boldsymbol{v}$ has the effect of translating all vectors by adding $\boldsymbol{a}$. Show this is not a linear transformation. Explain why it is not possible to realize $T_{a}$ in $\mathbb{R}^{3}$ by multiplying by a $3 \times 3$ matrix.
12. $\uparrow$ In spite of Problem 11 we can represent both translations and linear transformations by matrix multiplication at the expense of using higher dimensions. This is done by the homogeneous coordinates. I will illustrate in $\mathbb{R}^{3}$ where most interest in this is found. For each vector $\boldsymbol{v}=\left(v_{1}, v_{2}, v_{3}\right)^{T}$, consider the vector in $\mathbb{R}^{4}\left(v_{1}, v_{2}, v_{3}, 1\right)^{T}$. What happens when you do

$$
\left(\begin{array}{cccc}
1 & 0 & 0 & a_{1} \\
0 & 1 & 0 & a_{2} \\
0 & 0 & 1 & a_{3} \\
0 & 0 & 0 & 1
\end{array}\right)\left(\begin{array}{c}
v_{1} \\
v_{2} \\
v_{3} \\
1
\end{array}\right) ?
$$

Describe how to consider both linear transformations and translations all at once by forming appropriate $4 \times 4$ matrices.
13. You want to add $(1,2,3)$ to every point in $\mathbb{R}^{3}$ and then rotate about the $z$ axis counter clockwise through an angle of $30^{\circ}$. Find what happens to the point $(1,1,1)$.
14. Let $P_{3}$ denote the set of real polynomials of degree no more than 3, defined on an interval $[a, b]$. Show that $P_{3}$ is a subspace of the vector space of all functions defined on this interval. Show that a basis for $P_{3}$ is $\left\{1, x, x^{2}, x^{3}\right\}$. Now let $D$ denote the differentiation operator which sends a function to its derivative. Show $D$ is a linear transformation which sends $P_{3}$ to $P_{3}$. Find the matrix of this linear transformation with respect to the given basis.
15. Generalize the above problem to $P_{n}$, the space of polynomials of degree no more than $n$ with basis $\left\{1, x, \cdots, x^{n}\right\}$.
16. If $A$ is an $n \times n$ invertible matrix, show that $A^{T}$ is also and that in fact, $\left(A^{T}\right)^{-1}=$ $\left(A^{-1}\right)^{T}$.
17. Suppose you have an invertible $n \times n$ matrix $A$. Consider the polynomials

$$
\left(\begin{array}{c}
p_{1}(x) \\
\vdots \\
p_{n}(x)
\end{array}\right)=A\left(\begin{array}{c}
1 \\
\vdots \\
x^{n-1}
\end{array}\right)
$$

Show that these polynomials $p_{1}(x), \cdots, p_{n}(x)$ are a linearly independent set of functions.
18. Let the linear transformation be $T=D^{2}+1$, defined as $T f=f^{\prime \prime}+f$. Find the matrix of this linear transformation with respect to the given basis $\left\{1, x, x^{2}, x^{3}\right\}$.
19. Let $L$ be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by

$$
D^{2}+2 D+1
$$

where $D$ is the differentiation operator. Find the matrix of this linear transformation relative to the basis $\left\{1, x, x^{2}, x^{3}\right\}$. Find the matrix directly and then find the matrix with respect to the differential operator $D+1$ and multiply this matrix by itself. You should get the same thing. Why?
20. Let $L$ be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by $D^{2}+5 D+4$ where $D$ is the differentiation operator. Find the matrix of this linear transformation relative to the bases $\left\{1, x, x^{2}, x^{3}\right\}$. Find the matrix directly and then find the matrices with respect to the differential operators $D+1, D+4$ and multiply these two matrices. You should get the same thing. Why?
21. Suppose $A \in \mathscr{L}(V, W)$ where $\operatorname{dim}(V)>\operatorname{dim}(W)$. Show $\operatorname{ker}(A) \neq\{0\}$. That is, show there exist nonzero vectors $\boldsymbol{v} \in V$ such that $A \boldsymbol{v}=\mathbf{0}$.
22. A vector $\boldsymbol{v}$ is in the convex hull of a nonempty set if there are finitely many vectors of $S,\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{m}\right\}$ and nonnegative scalars $\left\{t_{1}, \cdots, t_{m}\right\}$ such that

$$
\boldsymbol{v}=\sum_{k=1}^{m} t_{k} \boldsymbol{v}_{k}, \sum_{k=1}^{m} t_{k}=1
$$

Such a linear combination is called a convex combination. Suppose now that $S \subseteq V$, a vector space of dimension $n$. Show that if $\boldsymbol{v}=\sum_{k=1}^{m} t_{k} \boldsymbol{v}_{k}$ is a vector in the convex hull for $m>n+1$, then there exist other scalars $\left\{t_{k}^{\prime}\right\}$ such that

$$
\boldsymbol{v}=\sum_{k=1}^{m-1} t_{k}^{\prime} \boldsymbol{v}_{k}
$$

Thus every vector in the convex hull of $S$ can be obtained as a convex combination of at most $n+1$ points of $S$. This incredible result is in Rudin [37]. Hint: Consider $L: \mathbb{R}^{m} \rightarrow V \times \mathbb{R}$ defined by

$$
L(\boldsymbol{a}) \equiv\left(\sum_{k=1}^{m} a_{k} \boldsymbol{v}_{k}, \sum_{k=1}^{m} a_{k}\right)
$$

Explain why $\operatorname{ker}(L) \neq\{\mathbf{0}\}$. Next, letting $\boldsymbol{a} \in \operatorname{ker}(L) \backslash\{\mathbf{0}\}$ and $\lambda \in \mathbb{R}$, note that $\lambda \boldsymbol{a} \in \operatorname{ker}(L)$. Thus for all $\lambda \in \mathbb{R}$,

$$
\boldsymbol{v}=\sum_{k=1}^{m}\left(t_{k}+\lambda a_{k}\right) \boldsymbol{v}_{k}
$$

Now vary $\lambda$ till some $t_{k}+\lambda a_{k}=0$ for some $a_{k} \neq 0$.
23. For those who know about compactness, use Problem 22 to show that if $S \subseteq \mathbb{R}^{n}$ and $S$ is compact, then so is its convex hull.
24. Show that if $L \in \mathscr{L}(V, W)$ (linear transformation) where $V$ and $W$ are vector spaces, then if $L \boldsymbol{y}_{p}=\boldsymbol{f}$ for some $\boldsymbol{y}_{p} \in V$, then the general solution of $L \boldsymbol{y}=\boldsymbol{f}$ is of the form $\operatorname{ker}(L)+\boldsymbol{y}_{p}$.
25. Suppose $A \boldsymbol{x}=\boldsymbol{b}$ has a solution. Explain why the solution is unique precisely when $A \boldsymbol{x}=\mathbf{0}$ has only the trivial (zero) solution.
26. Let $L: \mathbb{R}^{n} \rightarrow \mathbb{R}$ be linear. Show that there exists a vector $\boldsymbol{a} \in \mathbb{R}^{n}$ such that $L \boldsymbol{y}=\boldsymbol{a}^{T} \boldsymbol{y}$.
27. Let the linear transformation $T$ be determined by

$$
T \boldsymbol{x}=\left(\begin{array}{cccc}
1 & 0 & -5 & -7 \\
0 & 1 & -3 & -9 \\
1 & 1 & -8 & -16
\end{array}\right) \boldsymbol{x}
$$

Find the rank of this transformation.
28. Let $T f=\left(D^{2}+5 D+4\right) f$ for $f$ in the vector space of polynomials of degree no more than 3 where we consider $T$ to map into the same vector space. Find the rank of $T$. You might want to use Proposition 4.3.6.
29. (Extra important) Let $A$ be an $n \times n$ matrix. The trace of $A$, $\operatorname{trace}(A)$ is defined as $\sum_{i} A_{i i}$. It is just the sum of the entries on the main diagonal. Show trace $(A)=$ trace $\left(A^{T}\right)$. Suppose $A$ is $m \times n$ and $B$ is $n \times m$. Show that trace $(A B)=\operatorname{trace}(B A)$. Now show that if $A$ and $B$ are similar $n \times n$ matrices, then $\operatorname{trace}(A)=\operatorname{trace}(B)$. Recall that $A$ is similar to $B$ means $A=S^{-1} B S$ for some matrix $S$.
30. Suppose you have a monic polynomial $\phi(\lambda)$ which is irreducible over $\mathbb{F}$ the field of scalars. Remember that this means that no polynomial divides it except scalar multiples of $\phi(\lambda)$ and scalars. Say

$$
\phi(\lambda)=a_{0}+a_{1} \lambda+\cdots+a_{d-1} \lambda^{d-1}+\lambda^{d}
$$

Now consider $A \in \mathscr{L}(V, V)$ where $V$ is a vector space. Consider $\operatorname{ker}(\phi(A))$ and suppose this is not 0 . For $x \in \operatorname{ker}(\phi(A)), x \neq 0$, let $\beta_{x}=\left\{x, A x, \cdots, A^{d-1} x\right\}$. Show that $\beta_{x}$ is an independent set of vectors if $x \neq 0$.
31. $\uparrow$ Let $V$ be a finite dimensional vector space and let $A \in \mathscr{L}(V, V)$. Also let $W$ be a subspace of $V$ such that $A(W) \subseteq W$. We call such a subspace an $A$ invariant subspace. Say $\left\{w_{1}, \cdots, w_{s}\right\}$ is a basis for $W$. Also let $x \in U \backslash W$ where $U$ is an $A$ invariant subspace which is contained in $\operatorname{ker}(\phi(A))$. Then you know that $\left\{w_{1}, \cdots, w_{s}, x\right\}$ is linearly independent. Show that in fact $\left\{w_{1}, \cdots, w_{s}, \beta_{x}\right\}$ is linearly independent where $\beta_{x}$ is given in the above problem. Hint: Suppose you have

$$
\begin{equation*}
\sum_{k=1}^{s} a_{k} w_{k}+\sum_{j=1}^{d} b_{j} A^{j-1} x=0 \tag{*}
\end{equation*}
$$

You need to verify that the second sum is 0 . From this it will follow that each $b_{j}$ is 0 and then each $a_{k}=0$. Let $S=\sum_{j=1}^{d} b_{j} A^{j-1} x$. Observe that $\beta_{S} \subseteq \beta_{x}$ and if $S \neq 0$, then $\beta_{S}$ is independent from the above problem and both $\beta_{x}$ and $\beta_{S}$ have the same dimension. You will argue that $\operatorname{span}\left(\beta_{S}\right) \subseteq W \cap \operatorname{span}\left(\beta_{x}\right) \subseteq \operatorname{span}\left(\beta_{x}\right)$ and then use Problem 6 on Page 74..
32. $\uparrow$ In the situation of the above problem, show that there exist finitely many vectors in $U,\left\{x_{1}, \cdots, x_{m}\right\}$ such that $\left\{w_{1}, \cdots, w_{s}, \beta_{x_{1}}, \cdots, \beta_{x_{m}}\right\}$ is a basis for $U+W$. This last vector space is defined as the set of all $y+w$ where $y \in U$ and $w \in W$.
33. $\uparrow$ In the situation of the above where $\phi(\lambda)$ is irreducible. Let $U$ be defined as

$$
U=\phi(A)\left(\operatorname{ker}\left(\phi(A)^{m}\right)\right)
$$

Explain why $U \subseteq \operatorname{ker}\left(\phi(A)^{m-1}\right)$. Suppose you have a linearly independent set in $U$ which is $\left\{\beta_{x_{1}}, \cdots, \beta_{x_{r}}\right\}$. Here the notation means

$$
\beta_{x} \equiv\left\{x, A x, \cdots, A^{m-1} x\right\}
$$

where these vectors are independent but $A^{m} x$ is in the span of these. Such exists any time you have $x \in \operatorname{ker}(g(A))$ for $g(\lambda)$ a polynomial. Letting $\phi(A) y_{i}=x_{i}$, explain why $\left\{\beta_{y_{1}}, \cdots, \beta_{y_{r}}\right\}$ is also linearly independent. This is like the theorem presented earlier that the inverse image of a linearly independent set is linearly independent but it is more complicated here because instead of single vectors, we are considering sets $\beta_{x}$.

## Chapter 6

## Direct Sums and Block Diagonal Matrices

This is a convenient place to put a very interesting result about direct sums and block diagonal matrices. First is the notion of a direct sum. In all of this, $V$ will be a finite dimensional vector space of dimension $n$ and field of scalars $\mathbb{F}$.

Definition 6.0.1 Let $\left\{V_{i}\right\}_{i=1}^{r}$ be subspaces of $V$. Then $\sum_{i=1}^{r} V_{i} \equiv V_{1}+\cdots+V_{r}$ denotes all sums of the form $\sum_{i=1}^{r} v_{i}$ where $v_{i} \in V_{i}$. If whenever $\sum_{i=1}^{r} v_{i}=0, v_{i} \in V_{i}$, it follows that $v_{i}=0$ for each $i$, then a special notation is used to denote $\sum_{i=1}^{r} V_{i}$. This notation is $V_{1} \oplus \cdots \oplus V_{r}$, or sometimes to save space $\bigoplus_{i=1}^{r} V_{i}$ and it is called a direct sum of subspaces. A subspace $W$ of $V$ is called $A$ invariant for $A \in \mathscr{L}(V, V)$ if $A W \subseteq W$.

The next lemma tells how to recognize a direct sum.
Lemma 6.0.2 For the $V_{i}$ subspaces as above, $V_{1}+\cdots+V_{r}=V_{1} \oplus \cdots \oplus V_{r}$ if and only if

$$
0=\left(V_{1}+\cdots+V_{i-1}+V_{i+1}+\cdots+V_{r}\right) \cap V_{i}=0
$$

for each $i$.
Proof: Suppose the sum is a direct sum. Then if $m \in M_{i} \cap \sum_{j \neq i} M_{j}$ it follows that $m=m_{i}=\sum_{j \neq i} m_{j}$ where $m_{j} \in M_{j}$ for all $j$ and so $0=-m_{i}+\sum_{j \neq i} m_{j}$ so all the $m_{j}=0$ and $m_{i}=0$. Next suppose the condition about the intersection.

Then if $\sum_{i} m_{i}=0$ it follows that $-m_{i}=\sum_{j \neq i} m_{j}$ and so $m=-m_{i}=\sum_{j \neq i} m_{j} \in M_{i} \cap$ $\sum_{j \neq i} M_{j}$ and so $m=0$. Since $i$ was arbitrary, each $m_{i}=0$.

The important idea is that you seek to understand $A$ by looking at what it does on each $V_{i}$. It is a lot like knowing $A$ by knowing what it does to a basis, an idea used earlier.

Lemma 6.0.3 If $V=V_{1} \oplus \cdots \oplus V_{r}$ and if $\beta_{i}=\left\{v_{1}^{i}, \cdots, v_{m_{i}}^{i}\right\}$ is a basis for $V_{i}$, then a basis for $V$ is $\left\{\beta_{1}, \cdots, \beta_{r}\right\}$. Thus

$$
\operatorname{dim}(V)=\sum_{i=1}^{r} \operatorname{dim}\left(V_{i}\right)=\sum_{i=1}^{r}\left|\beta_{i}\right|
$$

where $\left|\beta_{i}\right|$ denotes the number of vectors in $\beta_{i}$. Conversely, if $\beta_{i}$ linearly independent and if a basis for $V$ is $\left\{\beta_{1}, \cdots, \beta_{r}\right\}$, then $V=\operatorname{span}\left(\beta_{1}\right) \oplus \cdots \oplus \operatorname{span}\left(\beta_{r}\right)$

Proof: Suppose $\sum_{i=1}^{r} \sum_{j=1}^{m_{i}} c_{i j} v_{j}^{i}=0$. Since a direct sum, for each $i, \sum_{j=1}^{m_{i}} c_{i j} v_{j}^{i}=0$ and now, since $\left\{v_{1}^{i}, \cdots, v_{m_{i}}^{i}\right\}$ is a basis, each $c_{i j}=0$ for each $j$, this for each $i$.

Suppose now that each $\beta_{i}$ is independent and a basis is $\left\{\beta_{1}, \cdots, \beta_{r}\right\}$. Then clearly

$$
V=\operatorname{span}\left(\beta_{1}\right)+\cdots+\operatorname{span}\left(\beta_{r}\right)
$$

Suppose then that $0=\sum_{i=1}^{r} \sum_{j=1}^{m_{i}} c_{i j} v_{j}^{i}$, the inside sum being something in span $\left(\beta_{i}\right)$. Since $\left\{\beta_{1}, \cdots, \beta_{r}\right\}$ is a basis, each $c_{i j}=0$. Thus each $\sum_{j=1}^{m_{i}} c_{i j} v_{j}^{i}=0$ and so $V=\operatorname{span}\left(\beta_{1}\right) \oplus$ $\cdots \oplus \operatorname{span}\left(\beta_{r}\right)$.

Thus, from this lemma, we can produce a basis for $V$ of the form $\left\{\beta_{1}, \cdots, \beta_{r}\right\}$, so what is the matrix of a linear transformation $A$ such that each $V_{i}$ is $A$ invariant?

Theorem 6.0.4 Suppose $V$ is a vector space with field of scalars $\mathbb{F}$ and $A \in \mathscr{L}(V, V)$. Suppose also $V=V_{1} \oplus \cdots \oplus V_{q}$ where each $V_{k}$ is A invariant. ( $A V_{k} \subseteq V_{k}$ ) Also let $\beta_{k}$ be an ordered basis for $V_{k}$ and let $A_{k}$ denote the restriction of $A$ to $V_{k}$. Letting $M^{k}$ denote the matrix of $A_{k}$ with respect to this basis, it follows the matrix of $A$ with respect to the basis $\left\{\beta_{1}, \cdots, \beta_{q}\right\}$ is

$$
\left(\begin{array}{ccc}
M^{1} & & 0 \\
& \ddots & \\
0 & & M^{q}
\end{array}\right)
$$

Proof: Let $\beta$ denote the ordered basis $\left\{\beta_{1}, \cdots, \beta_{q}\right\},\left|\beta_{k}\right|$ being the number of vectors in $\beta_{k}$. Let $q_{k}: \mathbb{F}^{\left|\beta_{k}\right|} \rightarrow V_{k}$ be the usual map such that the following diagram commutes.

$$
\begin{array}{rcl} 
& A_{k} & \\
V_{k} & \rightarrow & V_{k} \\
q_{k} \uparrow & \circ & \uparrow q_{k} \\
\mathbb{F}^{\left|\beta_{k}\right|} & \rightarrow & \mathbb{F}^{\left|\beta_{k}\right|} \\
& M^{k} &
\end{array}
$$

Thus $A_{k} q_{k}=q_{k} M^{k}$. Then if $q$ is the map from $\mathbb{F}^{n}$ to $V$ corresponding to the ordered basis $\beta$ just described,

$$
q\left(\begin{array}{lllll}
\mathbf{0} & \cdots & \boldsymbol{x} & \cdots & \mathbf{0}
\end{array}\right)^{T}=q_{k} \boldsymbol{x}
$$

where $\boldsymbol{x}$ occupies the positions between $\sum_{i=1}^{k-1}\left|\beta_{i}\right|+1$ and $\sum_{i=1}^{k}\left|\beta_{i}\right|$. Then $M$ will be the matrix of $A$ with respect to $\beta$ if and only if a similar diagram to the above commutes. Thus it is required that $A q=q M$. However, from the description of $q$ just made, and the invariance of each $V_{k}$,

$$
A q\left(\begin{array}{c}
\mathbf{0} \\
\vdots \\
\boldsymbol{x} \\
\vdots \\
\mathbf{0}
\end{array}\right)=A_{k} q_{k} \boldsymbol{x}=q_{k} M^{k} \boldsymbol{x}=q\left(\begin{array}{ccccc}
M^{1} & & & & 0 \\
& \ddots & & & \\
& & M^{k} & & \\
& & & \ddots & \\
0 & & & & M^{q}
\end{array}\right)\left(\begin{array}{c}
\mathbf{0} \\
\vdots \\
\boldsymbol{x} \\
\vdots \\
\mathbf{0}
\end{array}\right)
$$

It follows that the above block diagonal matrix is the matrix of $A$ with respect to the given ordered basis.

The matrix of $A$ with respect to the ordered basis $\beta$ which is described above is called a block diagonal matrix. Sometimes the blocks consist of a single number.

Example 6.0.5 Consider the following matrix.

$$
A \equiv\left(\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & -1 \\
-2 & 2 & 3
\end{array}\right)
$$

Let $V_{1} \equiv \operatorname{span}\left(\left(\begin{array}{l}1 \\ 1 \\ 0\end{array}\right),\left(\begin{array}{l}1 \\ 0 \\ 1\end{array}\right)\right), V_{2} \equiv \operatorname{span}\left(\left(\begin{array}{c}0 \\ -1 \\ 2\end{array}\right)\right)$. Show that $\mathbb{R}^{3}=V_{1} \oplus V_{2}$ and
and that $V_{i}$ is $A$ invariant. Find the matrix of $A$ with respect to the ordered basis

$$
\left\{\left(\begin{array}{l}
1  \tag{*}\\
1 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right),\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)\right\}
$$

First note that

$$
\left(\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & -1 \\
-2 & 2 & 3
\end{array}\right)\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right)=\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right),\left(\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & -1 \\
-2 & 2 & 3
\end{array}\right)\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)=\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)
$$

Therefore, $A\left(V_{1}\right) \subseteq V_{1}$. Similarly,

$$
\left(\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & -1 \\
-2 & 2 & 3
\end{array}\right)\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)=\left(\begin{array}{c}
0 \\
-2 \\
4
\end{array}\right)=2\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)
$$

and so $A\left(V_{2}\right) \subseteq V_{2}$. The vectors in $*$ clearly are a basis for $\mathbb{R}^{3}$. You can verify this by observing that there is a unique solution $x, y, z$ to the system of equations

$$
x\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right)+y\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)+z\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)=\left(\begin{array}{l}
a \\
b \\
c
\end{array}\right)
$$

for any choice of the right side. Therefore, by Lemma 6.0.3, $\mathbb{R}^{3}=V_{1} \oplus V_{2}$.
If you look at the restriction of $A$ to $V_{1}$, what is the matrix of this restriction? It satisfies

$$
\left(A\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right), A\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right)=\left(\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right)\left(\begin{array}{ll}
a & b \\
c & d
\end{array}\right)
$$

Thus, from what was observed above, you need the matrix on the right to satisfy

$$
\left(\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right)=\left(\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right)\left(\begin{array}{ll}
a & b \\
c & d
\end{array}\right)
$$

and so the matix on the right is just $\left(\begin{array}{ll}1 & 0 \\ 0 & 1\end{array}\right)$. As to the matrix of $A$ restricted to $V_{2}$, we need

$$
A\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)=2\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)=a\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)
$$

where $a$ is a $1 \times 1$ matrix. Thus $a=2$ and so the matrix of $A$ with respect to the ordered basis given above is

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 2
\end{array}\right)
$$

What if you changed the order of the vectors in the basis? Suppose you had them ordered as

$$
\left\{\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right),\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right),\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right\}
$$

Then you would have three invariant subspaces whose direct sum is $\mathbb{R}^{3}$,

$$
\operatorname{span}\left(\left(\begin{array}{l}
1 \\
1 \\
0
\end{array}\right)\right), \operatorname{span}\left(\left(\begin{array}{c}
0 \\
-1 \\
2
\end{array}\right)\right), \text { and } \operatorname{span}\left(\left(\begin{array}{l}
1 \\
0 \\
1
\end{array}\right)\right)
$$

Then the matrix of $A$ with respect to this ordered basis is

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 1
\end{array}\right)
$$

Example 6.0.6 Consider the following matrix.

$$
A=\left(\begin{array}{ccc}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{array}\right)
$$

Let

$$
V_{1} \equiv \operatorname{span}\left(\left(\begin{array}{l}
0 \\
0 \\
1
\end{array}\right)\right), V_{2} \equiv \operatorname{span}\left(\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right),\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right)\right)
$$

Show that these are $A$ invariant subspaces and find the matrix of $A$ with respect to the ordered basis

$$
\left\{\left(\begin{array}{l}
0 \\
0 \\
1
\end{array}\right),\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right),\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)\right\}
$$

First note that

$$
\left(\left(\begin{array}{ccc}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{array}\right)-2\left(\begin{array}{ccc}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)\right)\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)=\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right)
$$

$$
\begin{aligned}
& \text { and so } A\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right) \text { is in the span of }\left\{\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right),\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)\right\} . \text { Also } \\
& \left(\begin{array}{ccc}
3 & 1 & 0 \\
-1 & 1 & 0 \\
-1 & -1 & 1
\end{array}\right)\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right)=\left(\begin{array}{c}
2 \\
-2 \\
0
\end{array}\right) \in \operatorname{span}\left(\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right),\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)\right)
\end{aligned}
$$

Thus $V_{2}$ is $A$ invariant. What is the matrix of $A$ restricted to $V_{2}$ ? We need

$$
\left(A\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right), A\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)\right)=\left(\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right),\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)\right)\left(\begin{array}{ll}
a & b \\
c & d
\end{array}\right)
$$

Now it was shown above that

$$
A\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)=2\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right)+\left(\begin{array}{c}
1 \\
-1 \\
0
\end{array}\right)
$$

and so the matrix is of the form $\left(\begin{array}{cc}a & 1 \\ c & 2\end{array}\right)$. Then it was also shown that $A\left(\begin{array}{c}1 \\ -1 \\ 0\end{array}\right)=$ $2\left(\begin{array}{c}1 \\ -1 \\ 0\end{array}\right)$ and so the matrix is of the form $\left(\begin{array}{cc}2 & 1 \\ 0 & 2\end{array}\right)$. As to $V_{1}, A\left(\begin{array}{l}0 \\ 0 \\ 1\end{array}\right)=\left(\begin{array}{l}0 \\ 0 \\ 1\end{array}\right)$ and the matrix of $A$ restricted to $V_{1}$ is just the $1 \times 1$ matrix consisting of the number 1 . Thus the matrix of $A$ with respect to this basis is

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 2 & 1 \\
0 & 0 & 2
\end{array}\right)
$$

How can you find $V$ as a direct sum of invariant subspaces? In the next section, I will give a systematic way based on a profound theorem of Sylvester. However, there is also a very easy way to come up with an invariant subspace. Let $v \in V$ an $n$ dimensional vector space and let $A \in \mathscr{L}(V, V)$. Let $W \equiv \operatorname{span}\left(v, A v, A^{2} v, \cdots\right)$. It is left as an exercise to verify that $W$ is a finite dimensional subspace of $V$. Recall that the span is the set of all finite linear combinations. Of course $W$ might be all of $V$ or it might be a proper subset of $V$. The method of Sylvester will end up typically giving proper invariant subspaces whose direct sum is the whole space. An outline of the following presentation is as follows.

1. Sylvester's theorem $\operatorname{dim}\left(\operatorname{ker}\left(\prod_{i=1}^{m} L_{i}\right)\right) \leq \sum_{i=1}^{m} \operatorname{dim}\left(\operatorname{ker}\left(L_{i}\right)\right)$
2. If $L_{i} L_{j}=L_{j} L_{i}, L_{i}$ one to one on $\operatorname{ker}\left(L_{i}\right)$ then $\operatorname{ker}\left(\prod_{i=1}^{m} L_{i}\right)=\bigoplus_{i=1}^{m} \operatorname{ker}\left(L_{i}\right)$
3. $L \in \mathscr{L}(V, V)$ having minimum polyinomial $\prod_{i=1}^{m} \phi_{i}(\lambda)^{r_{i}}$. Then

$$
V \equiv \operatorname{ker}\left(\prod_{i=1}^{m} \phi_{i}(L)^{r_{i}}\right)=\bigoplus_{i=1}^{m} \operatorname{ker}\left(\phi_{i}(L)\right)
$$

### 6.1 A Theorem of Sylvester, Direct Sums

The notation is defined as follows. First recall the definition of ker in Problem 23 on Page 62.

Definition 6.1.1 Let $L \in \mathscr{L}(V, W)$. Then $\operatorname{ker}(L) \equiv\{v \in V: L v=0\}$.
Lemma 6.1.2 Whenever $L \in \mathscr{L}(V, W), \operatorname{ker}(L)$ is a subspace. Also, if $V$ is an $n$ dimensional vector space and $W$ is a subspace of $V$, then $W=V$ if and only if $\operatorname{dim}(W)=n$.

Proof: If $a, b$ are scalars and $v, w$ are in $\operatorname{ker}(L)$, then

$$
L(a v+b w)=a L(v)+b L(w)=0+0=0
$$

As to the last claim, it is clear that $\operatorname{dim}(W) \leq n$. If $\operatorname{dim}(W)=n$, then, letting $\left\{w_{1}, \cdots, w_{n}\right\}$ be a basis for $W$, there can be no $v \in V \backslash W$ because then $v \notin \operatorname{span}\left(w_{1}, \cdots, w_{n}\right)$ and so by Lemma 3.1.7 $\left\{w_{1}, \cdots, w_{n}, v\right\}$ would be independent which is impossible by Theorem 3.1.5. You have an independent set which is longer than a spanning set.

Suppose now that $A \in \mathscr{L}(V, W)$ and $B \in \mathscr{L}(W, U)$ where $V, W, U$ are all finite dimensional vector spaces. Then it is interesting to consider $\operatorname{ker}(B A)$. The following theorem of Sylvester is a very useful and important result.

Theorem 6.1.3 Let $A \in \mathscr{L}(V, W)$ and $B \in \mathscr{L}(W, U)$ where $V, W, U$ are all vector spaces over a field $\mathbb{F}$. Suppose also that $\operatorname{ker}(A)$ and $A(\operatorname{ker}(B A))$ are finite dimensional subspaces. Then

$$
\operatorname{dim}(\operatorname{ker}(B A)) \leq \operatorname{dim}(\operatorname{ker}(B))+\operatorname{dim}(\operatorname{ker}(A))
$$

Equality holds if and only if $A(\operatorname{ker}(B A))=\operatorname{ker}(B)$.
Proof: If $\boldsymbol{x} \in \operatorname{ker}(B A)$, then $A \boldsymbol{x} \in \operatorname{ker}(B)$ and so $A(\operatorname{ker}(B A)) \subseteq \operatorname{ker}(B)$. The following picture may help.


$$
\begin{array}{cc} 
& \text { Basis } \\
A(\operatorname{ker}(B A)) & \left\{A y_{1}, \cdots, A y_{m}\right\} \\
\operatorname{ker}(A) & \left\{x_{1}, \cdots, x_{n}\right\} \\
\operatorname{ker}(B) & \left\{A y_{1}, \cdots, A y_{m}, w_{1}, \cdots, w_{s}\right\}
\end{array}
$$

Now let $\left\{x_{1}, \cdots, x_{n}\right\}$ be a basis of $\operatorname{ker}(A)$ and let $\left\{A y_{1}, \cdots, A y_{m}\right\}$ one for $A(\operatorname{ker}(B A))$, each $y_{i} \in \operatorname{ker}(B A)$. Take any $z \in \operatorname{ker}(B A)$. Then $A z=\sum_{i=1}^{m} a_{i} A y_{i}$ and so

$$
A\left(z-\sum_{i=1}^{m} a_{i} y_{i}\right)=\mathbf{0}
$$

which means $z-\sum_{i=1}^{m} a_{i} y_{i} \in \operatorname{ker}(A)$ and so there are scalars $b_{i}$ such that

$$
z-\sum_{i=1}^{m} a_{i} y_{i}=\sum_{j=1}^{n} b_{i} x_{i}
$$

It follows span $\left(x_{1}, \cdots, x_{n}, y_{1}, \cdots, y_{m}\right)=\operatorname{ker}(B A)$ and so by the first part, (See the picture.)

$$
\begin{aligned}
\operatorname{dim}(\operatorname{ker}(B A)) & \leq n+m \leq \operatorname{dim}(\operatorname{ker}(A))+\operatorname{dim}(A(\operatorname{ker}(B A))) \\
& \leq \operatorname{dim}(\operatorname{ker}(A))+\operatorname{dim}(\operatorname{ker}(B))
\end{aligned}
$$

Now $\left\{x_{1}, \cdots, x_{n}, y_{1}, \cdots, y_{m}\right\}$ is linearly independent because if

$$
\sum_{i} a_{i} x_{i}+\sum_{j} b_{j} y_{j}=0
$$

then you could do $A$ to both sides and conclude that $\sum_{j} b_{j} A y_{j}=0$ which requires that each $b_{j}=0$. Then it follows that each $a_{i}=0$ also because it implies $\sum_{i} a_{i} x_{i}=0$. Thus the first inequality in the above list is an equal sign and $\left\{x_{1}, \cdots, x_{n}, y_{1}, \cdots, y_{m}\right\}$ is a basis for $\operatorname{ker}(B A)$. Each vector is in $\operatorname{ker}(B A)$, they are linearly independent, and their span is $\operatorname{ker}(B A)$. Then by Lemma 6.1.2, $A(\operatorname{ker}(B A))=\operatorname{ker}(B)$ if and only if $m=\operatorname{dim}(\operatorname{ker}(B))$ if and only if

$$
\operatorname{dim}(\operatorname{ker}(B A))=m+n=\operatorname{dim}(\operatorname{ker}(B))+\operatorname{dim}(\operatorname{ker}(A))
$$

Of course this result holds for any finite product of linear transformations by induction. One way this is quite useful is in the case where you have a finite product of linear transformations $\prod_{i=1}^{l} L_{i}$ all in $\mathscr{L}(V, V)$. Then $\operatorname{dim}\left(\operatorname{ker} \prod_{i=1}^{l} L_{i}\right) \leq \sum_{i=1}^{l} \operatorname{dim}\left(\operatorname{ker} L_{i}\right)$.

Now here is a useful lemma which is likely already understood.
Lemma 6.1.4 Let $L \in \mathscr{L}(V, W)$ where $V, W$ are $n$ dimensional vector spaces. Then $L$ is one to one, if and only if $L$ is also onto. In fact, if $\left\{v_{1}, \cdots, v_{n}\right\}$ is a basis, then so is $\left\{L v_{1}, \cdots, L v_{n}\right\}$.

Proof: Let $\left\{v_{1}, \cdots, v_{n}\right\}$ be a basis for $V$. Then I claim that $\left\{L v_{1}, \cdots, L v_{n}\right\}$ is a basis for $W$. First of all, I show $\left\{L v_{1}, \cdots, L v_{n}\right\}$ is linearly independent. Suppose $\sum_{k=1}^{n} c_{k} L v_{k}=0$. Then $L\left(\sum_{k=1}^{n} c_{k} v_{k}\right)=0$ and since $L$ is one to one, it follows $\sum_{k=1}^{n} c_{k} v_{k}=0$ which implies each $c_{k}=0$. Therefore, $\left\{L v_{1}, \cdots, L v_{n}\right\}$ is linearly independent. If there exists $w$ not in the span of these vectors, then by Lemma 3.1.7, $\left\{L v_{1}, \cdots, L v_{n}, w\right\}$ would be independent and this contradicts the exchange theorem, Theorem 3.1.5 because it would be a linearly independent set having more vectors than the spanning set $\left\{v_{1}, \cdots, v_{n}\right\}$.

Conversely, suppose $L$ is onto. Then there exists a basis for $W$ which is of the form $\left\{L v_{1}, \cdots, L v_{n}\right\}$. It follows that $\left\{v_{1}, \cdots, v_{n}\right\}$ is linearly independent. Hence it is a basis for $V$ by similar reasoning to the above. Then if $L x=0$, it follows that there are scalars $c_{i}$ such that $x=\sum_{i} c_{i} v_{i}$ and consequently $0=L x=\sum_{i} c_{i} L v_{i}$. Therefore, each $c_{i}=0$ and so $x=0$ also. Thus $L$ is one to one.

Here is a fundamental lemma which gives a typical situation where a vector space is the direct sum of subspaces.

Lemma 6.1.5 Let $L_{i}$ be in $\mathscr{L}(V, V)$ and suppose for $i \neq j, L_{i} L_{j}=L_{j} L_{i}$ and also $L_{i}$ is one to one on $\operatorname{ker}\left(L_{j}\right)$ whenever $i \neq j$. Then

$$
\operatorname{ker}\left(\prod_{i=1}^{p} L_{i}\right)=\operatorname{ker}\left(L_{1}\right) \oplus+\cdots+\oplus \operatorname{ker}\left(L_{p}\right)
$$

Here $\prod_{i=1}^{p} L_{i}$ is the product of all the linear transformations. It signifies

$$
L_{p} \circ L_{p-1} \circ \cdots \circ L_{1}
$$

or the product in any other order since the transformations commute.
Proof : Note that since the operators commute, $L_{j}: \operatorname{ker}\left(L_{i}\right) \rightarrow \operatorname{ker}\left(L_{i}\right)$. Here is why. If $L_{i} y=0$ so that $y \in \operatorname{ker}\left(L_{i}\right)$, then $L_{i} L_{j} y=L_{j} L_{i} y=L_{j} 0=0$ and so $L_{j}: \operatorname{ker}\left(L_{i}\right) \mapsto$ $\operatorname{ker}\left(L_{i}\right)$. Next observe that it is obvious that, since the operators commute, $\sum_{i=1}^{p} \operatorname{ker}\left(L_{p}\right) \subseteq$ $\operatorname{ker}\left(\prod_{i=1}^{p} L_{i}\right)$.

Next, why is $\sum_{i} \operatorname{ker}\left(L_{p}\right)=\operatorname{ker}\left(L_{1}\right) \oplus \cdots \oplus \operatorname{ker}\left(L_{p}\right)$ ? Suppose $\sum_{i=1}^{p} v_{i}=0, v_{i} \in \operatorname{ker}\left(L_{i}\right)$, but some $v_{i} \neq 0$. Then do $\prod_{j \neq i} L_{j}$ to both sides. Since the linear transformations commute, this results in

$$
\left(\prod_{j \neq i} L_{j}\right)\left(v_{i}\right)=0
$$

which contradicts the assumption that these $L_{j}$ are one to one on $\operatorname{ker}\left(L_{i}\right)$ and the observation that they map $\operatorname{ker}\left(L_{i}\right)$ to $\operatorname{ker}\left(L_{i}\right)$. Thus if $\sum_{i} v_{i}=0, v_{i} \in \operatorname{ker}\left(L_{i}\right)$ then each $v_{i}=0$. It follows that

$$
\begin{equation*}
\operatorname{ker}\left(L_{1}\right) \oplus+\cdots+\oplus \operatorname{ker}\left(L_{p}\right) \subseteq \operatorname{ker}\left(\prod_{i=1}^{p} L_{i}\right) \tag{*}
\end{equation*}
$$

From Sylvester's theorem and the observation about direct sums in Lemma 6.0.3,

$$
\begin{aligned}
\sum_{i=1}^{p} \operatorname{dim}\left(\operatorname{ker}\left(L_{i}\right)\right) & =\operatorname{dim}\left(\operatorname{ker}\left(L_{1}\right) \oplus+\cdots+\oplus \operatorname{ker}\left(L_{p}\right)\right) \\
& \leq \operatorname{dim}\left(\operatorname{ker}\left(\prod_{i=1}^{p} L_{i}\right)\right) \leq \sum_{i=1}^{p} \operatorname{dim}\left(\operatorname{ker}\left(L_{i}\right)\right)
\end{aligned}
$$

which implies all these are equal. Now in general, if $W$ is a subspace of $V$, a finite dimensional vector space and the two have the same dimension, then $W=V$, Lemma 6.1.2. It follows from * that

$$
\operatorname{ker}\left(L_{1}\right) \oplus+\cdots+\oplus \operatorname{ker}\left(L_{p}\right)=\operatorname{ker}\left(\prod_{i=1}^{p} L_{i}\right)
$$

So how does the above situation occur? First recall the following theorem and corollary about polynomials. It was Theorem 6.1.6 and Corollary 6.1.7 proved earlier.

Theorem 6.1.6 Let $f(\lambda)$ be a nonconstant polynomial with coefficients in $\mathbb{F}$. Then there is some $a \in \mathbb{F}$ such that $f(\lambda)=a \prod_{i=1}^{n} \phi_{i}(\lambda)$ where $\phi_{i}(\lambda)$ is an irreducible nonconstant monic polynomial and repeats are allowed. Furthermore, this factorization is unique in the sense that any two of these factorizations have the same nonconstant factors in the product, possibly in different order and the same constant $a$.

Corollary 6.1.7 Let $q(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}$ where the $k_{i}$ are positive integers and the $\phi_{i}(\lambda)$ are irreducible monic polynomials. Suppose also that $p(\lambda)$ is a monic polynomial which divides $q(\lambda)$. Then $p(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{r_{i}}$ where $r_{i}$ is a nonnegative integer no larger than $k_{i}$.

Now I will show how to use these basic theorems about polynomials to produce $L_{i}$ such that the above major result follows. This is going to have a striking similarity to the notion of a minimum polynomial in the context of algebraic numbers.

Definition 6.1.8 Let $V$ be an $n$ dimensional vector space, $n \geq 1$, and let $L \in \mathscr{L}(V, V)$ which is a vector space of dimension $n^{2}$ by Theorem 5.1.4. Then $p(\lambda)$ will be the non constant monic polynomial such that $p(L)=0$ and out of all polynomials $q(\lambda)$ such that $q(L)=0$, the degree of $p(\lambda)$ is the smallest. This is called the minimum polynomial. It is always understood that $L \neq 0$. It is not interesting to fuss with this case of the zero linear transformation.

In the following, we always define $L^{0} \equiv I$.
Theorem 6.1.9 The above definition is well defined. Also, if $q(L)=0$, then $p(\lambda)$ divides $q(\lambda)$.

Proof: The dimension of $\mathscr{L}(V, V)$ is $n^{2}$. Therefore, $I, L, \cdots, L^{n^{2}}$ are linearly dependent and so there is some polynomial $q(\lambda)$ such that $q(L)=0$. Let $m$ be the smallest degree of any polynomial with this property. Such a smallest number exists by well ordering of $\mathbb{N}$. To obtain a monic polynomial $p(\lambda)$ with degree $m$, divide such a polynomial with degree $m$, having the property that $p(L)=0$ by the leading coefficient. Now suppose $q(\lambda)$ is any polynomial such that $q(L)=0$. Then by the Euclidean algorithm, there is $r(\lambda)$ either zero or having degree less than the degree of $p(\boldsymbol{\lambda})$ such that $q(\boldsymbol{\lambda})=p(\boldsymbol{\lambda}) k(\boldsymbol{\lambda})+r(\boldsymbol{\lambda})$ for some polynomial $k(\lambda)$. But then

$$
0=q(L)=k(L) p(L)+r(L)=r(L)
$$

If $r(\lambda) \neq 0$, then this is a contradiction to $p(\lambda)$ having the smallest degree. Therefore, $p(\lambda)$ divides $q(\lambda)$. Now suppose $\hat{p}(\lambda)$ and $p(\lambda)$ are two monic polynomials of degree $m$. Then from what was just shown $\hat{p}(\lambda)$ divides $p(\lambda)$ and $p(\lambda)$ divides $\hat{p}(\boldsymbol{\lambda})$. Since they are both monic polynomials, they must be equal. Thus the minimum polynomial is unique and this shows the above definition is well defined.

Now here is the major result which comes from Sylvester's theorem given above.
Theorem 6.1.10 Let $L \in \mathscr{L}(V, V)$ where $V$ is an $n$ dimensional vector space with field of scalars $\mathbb{F}$. Letting $p(\lambda)$ be the minimum polynomial for $L$,

$$
p(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}
$$

where the $k_{i}$ are positive integers and the $\phi_{i}(\lambda)$ are distinct irreducible monic polynomials. Also the linear maps $\phi_{i}(L)^{k_{i}}$ commute and $\phi_{i}(L)^{k_{i}}$ is one to one on $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$ for all $j \neq i$ as is $\phi_{i}(L)$. Also

$$
V=\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\phi_{p}(L)^{k_{p}}\right)
$$

and each $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ is invariant with respect to L. Letting $L_{j}$ be the restriction of $L$ to

$$
\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)
$$

it follows that the minimum polynomial of $L_{j}$ equals $\phi_{j}(\lambda)^{k_{j}}$. Also $p \leq n$.
Proof: By Theorem 6.1.6, the minimum polynomial $p(\lambda)$ is of the form $a \prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}$ where $\phi_{i}(\lambda)$ is monic and irreducible with $\phi_{i}(\lambda) \neq \phi_{j}(\lambda)$ if $i \neq j$. Since $p(\lambda)$ is monic, it follows that $a=1$. Since $L$ commutes with itself, all of these $\phi_{i}(L)^{k_{i}}$ commute. Also

$$
\phi_{i}(L): \operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right) \rightarrow \operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)
$$

because all of these operators commute.
Now consider $\phi_{i}(L)$. Is it one to one on $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$ ? Suppose not. Suppose that for some $j \neq i, \phi_{i}(L)$ is not one to one on $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$. We know that $\phi_{i}(\lambda), \phi_{j}(\lambda)^{k_{j}}$ are relatively prime meaning the monic polynomial of greatest degree which divides them both is 1 . Why is this? If some polynomial divided both, then it would need to be $\phi_{i}(\lambda)$ or 1 because $\phi_{i}(\lambda)$ is irreducible. But $\phi_{i}(\lambda)$ cannot divide $\phi_{j}(\lambda)^{k_{j}}$ unless it equals $\phi_{j}(\lambda)$, this by Corollary 6.1.7 and they are assumed unequal. Hence there are polynomials $l(\lambda), m(\lambda)$ such that $1=l(\lambda) \phi_{i}(\lambda)+m(\lambda) \phi_{j}(\lambda)^{k_{j}}$. By what we mean by equality of polynomials, that coefficients of equal powers of $\lambda$ are equal, it follows that for $I$ the identity transformation,

$$
I=l(L) \phi_{i}(L)+m(L) \phi_{j}(L)^{k_{j}}
$$

Say $v \in \operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$ and $v \neq 0$ while $\phi_{i}(L) v=0$. Then from the above equation,

$$
v=l(L) \phi_{i}(L) v+m(L) \phi_{j}(L)^{k_{j}} v=0+0=0
$$

a contradiction. Thus $\phi_{i}(L)$ and hence $\phi_{i}(L)^{k_{i}}$ is one to one on $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$. (Recall that, since these commute, $\phi_{i}(L)$ maps $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ to $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$.) On $V_{j} \equiv$ $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right), \phi_{i}(L)$ actually has an inverse. In fact, the above equation says that for $v \in V_{j}, v=l(L) \phi_{i}(L) v$. hence an inverse for $\phi_{i}(L)^{m}$ is $l(L)^{m}$.

Thus, from Lemma 6.1.5,

$$
V=\operatorname{ker}\left(\prod_{i=1}^{p} \phi_{i}(L)^{k_{i}}\right)=\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\phi_{p}(L)^{k_{p}}\right)
$$

Next consider the claim about the minimum polynomial of $L_{j}$. Denote this minimum polynomial as $p_{j}(\lambda)$. Then since $\phi_{j}(L)^{k_{j}}=\phi_{j}\left(L_{j}\right)^{k_{j}}=0$ on $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$, it must be the case that $p_{j}(\boldsymbol{\lambda})$ must divide $\phi_{j}(\boldsymbol{\lambda})^{k_{j}}$ and so by Corollary 6.1 .7 this means $p_{j}(\boldsymbol{\lambda})=\phi_{j}(\boldsymbol{\lambda})^{r_{j}}$ where $r_{j} \leq k_{j}$. If $r_{j}<k_{j}$, consider the polynomial

$$
\prod_{i=1, i \neq j}^{p} \phi_{i}(\lambda)^{k_{i}} \phi_{j}(\lambda)^{r_{j}} \equiv r(\lambda)
$$

Then since these operators $\phi_{i}(L)^{k_{i}}$ commute with each other, $r(L)=0$ because $r(L) v=$ 0 for every $v \in \operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ and also $r(L) v=0$ for $v \in \operatorname{ker}\left(\phi_{j}(L)^{r_{j}}\right)$. However, this violates the definition of the minimum polynomial for $L, p(\lambda)$ because here is a polynomial $r(\lambda)$ such that $r(L)=0$ but $r(\lambda)$ has smaller degree than $p(\lambda)$. Thus $r_{j}=k_{j}$.

Consider the claim that $p \leq n$ the dimension of $V$. Let $v_{i} \in \operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right), v_{i} \neq 0$. Then it must be the case that $\left\{v_{1}, \cdots, v_{p}\right\}$ is a linearly independent set because $\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus$ $\cdots \oplus \operatorname{ker}\left(\phi_{p}(L)^{k_{p}}\right)$ is a direct sum. Hence $p \leq n$ because a linearly independent set is never longer than a spanning set one of which has $n$ elements.

Letting $\beta_{i}$ be an ordered basis for $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ and letting $\beta \equiv\left(\beta_{1}, \beta_{2}, \cdots, \beta_{p}\right)$, it follows from Theorem 6.0.4, that if $M_{j}$ is the matrix for $L_{j}$, the restriction of $L$ to $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$, then the matrix of $L$ with respect to the basis $\beta$ is a block diagonal matrix of the form

$$
\left(\begin{array}{ccc}
M_{1} & & 0 \\
& \ddots & \\
0 & & M_{p}
\end{array}\right)
$$

The study of cannonical forms has to do with choosing the bases $\beta_{i}$ in an auspicious manner. This topic will be discussed more later.

### 6.2 Finding the Minimum Polynomial

All of this depends on the minimum polynomial. It was shown above that this polynomial exists, but how can you find it? In fact, it is not all that hard to find. Recall that if $L \in$ $\mathscr{L}(V, V)$ where the dimension of $V$ is $n$, then $I, L^{2}, \cdots, L^{n^{2}}$ is linearly independent. Thus some linear combination equals zero. The minimum polynomial was the polynomial $p(\lambda)$ of smallest degree which is monic and which has $p(L)=0$. At this point, we only know that this degree is no more than $n^{2}$. However, it will be shown later in the proof of the Cayley Hamilton theorem that there exists a polynomial $q(\lambda)$ of degree $n$ such that $q(L)=0$. Then from Theorem 6.1.9 it follows that $p(\lambda)$ divides $q(\lambda)$ and so the degree of $p(\lambda)$ will always be no more than $n$.

Another observation to make is that it suffices to find the minimum polynomial for the matrix of the linear transformation taken with respect to any basis. Recall the relation of this matrix and $L$.

$$
\begin{array}{rll} 
& L & \\
V & \rightarrow & V \\
q \uparrow & \circ & \uparrow q \\
\mathbb{F}^{n} & \rightarrow & \mathbb{F}^{n} \\
& A &
\end{array}
$$

where $q$ is a one to one and onto linear map from $\mathbb{F}^{n}$ to $V$. Thus if $p(L)$ is a polynomial in $L$,

$$
p(L)=p\left(q^{-1} A q\right)
$$

A typical term on the right is of the form

$$
c_{k}(\overbrace{\left(q^{-1} A q\right)\left(q^{-1} A q\right)\left(q^{-1} A q\right) \cdots q^{-1} A q}^{k \text { times }})=q^{-1}\left(c_{k} A^{k}\right) q
$$

Thus, applying this to each term and factoring out $q^{-1}$ and $q, p(L)=q^{-1} p(A) q$. Recall the convention that $A^{0}=I$ the identity matrix and $L^{0}=I$, the identity linear transformation. Thus $p(L)=0$ if and only if $p(A)=0$ and so the minimum polynomial for $A$ is exactly the same as the minimum polynomial for $L$. However, in case of $A$, the multiplication is just matrix multiplication so we can compute with it easily.

This shows that it suffices to learn how to find the minimum polynomial for an $n \times n$ matrix. I will show how to do this with some examples. The process can be made much more systematic, but I will try to keep it pretty short because it is often the case that it is easy to find it without going through a long computation.

Example 6.2.1 Find the minimum polynomial of

$$
\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)
$$

Go right to the definition and use the fact that you only need to have three powers of this matrix in order to get things to work, which will be shown later. Thus the minimum polynomial involves finding $a, b, c, d$ scalars such that

$$
\begin{gathered}
a\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)+b\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)+ \\
c\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)^{2}+d\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)=0
\end{gathered}
$$

You could include all nine powers if you want, but there is no point in doing so from what will be presented later. You will be able to find a polynomial of degree no larger than 3 which will work.

There is such a solution from the above theory and it is only a matter of finding it. Thus you need to find scalars such that

$$
\begin{array}{r}
a\left(\begin{array}{ccc}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)+b\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right) \\
+c\left(\begin{array}{ccc}
-5 & 0 & 18 \\
3 & 1 & -9 \\
-3 & 0 & 10
\end{array}\right)+d\left(\begin{array}{ccc}
-13 & 0 & 42 \\
7 & 1 & -21 \\
-7 & 0 & 22
\end{array}\right)=0
\end{array}
$$

Lets try the diagonal entries first and then lets pick the bottom left corner.

$$
\begin{gathered}
a-b-5 c-13 d=0 \\
a+b+c+d=0 \\
a+4 b+10 c+22 d=0 \\
-b+-3 c+-7 d=0
\end{gathered}
$$

Thus we row reduce the matrix

$$
\left(\begin{array}{ccccc}
1 & -1 & -5 & -13 & 0 \\
1 & 1 & 1 & 1 & 0 \\
1 & 4 & 10 & 22 & 0 \\
0 & -1 & -3 & -7 & 0
\end{array}\right)
$$

which yields after some computations

$$
\left(\begin{array}{ccccc}
1 & 0 & -2 & -6 & 0 \\
0 & 1 & 3 & 7 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}\right)
$$

We can take $d=0$ and $c=1$ and find that $a=2, b=-3$. A candidate for minimum polynomial is

$$
\lambda^{2}-3 \lambda+2
$$

Could you have a smaller degree polynomial? No you could not because if you took both $c$ and $d$ equal to 0 , then you would be forced to have $a, b$ both be zero as well. Hence this must be the minimum polynomial provided the matrix satisfies this equation. You verify this by plugging the matrix in to the polynomial and checking to see if you get 0 . If it didn't work, you would simply include another equation in the above computation for $a, b, c, d$.

$$
\begin{gathered}
\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)^{2}-3\left(\begin{array}{ccc}
-1 & 0 & 6 \\
1 & 1 & -3 \\
-1 & 0 & 4
\end{array}\right)+2\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right) \\
\\
=\left(\begin{array}{lll}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{array}\right)
\end{gathered}
$$

It is a little tedious, but completely routine to find this minimum polynomial. To be more systematic, you would take the powers of the matrix and string each of them out into a long $n^{2} \times 1$ vector and make these the columns of a matrix which would then be row reduced. However, as shown above, you can get away with less as in the above example, but you need to be sure to check that the matrix satisfies the equation you come up with.

Now here is an example where $\mathbb{F}=\mathbb{Z}_{5}$ and the arithmetic is in $\mathbb{F}$ so $A$ is the matrix of a linear transformation which maps $\mathbb{F}^{3}$ to $\mathbb{F}^{3}$.

Example 6.2.2 The matrix is

$$
A=\left(\begin{array}{lll}
1 & 2 & 3 \\
0 & 3 & 1 \\
4 & 1 & 1
\end{array}\right)
$$

Find the minimum polynomial.
Powers of the matrix are

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right),\left(\begin{array}{lll}
1 & 2 & 3 \\
0 & 3 & 1 \\
4 & 1 & 1
\end{array}\right),\left(\begin{array}{lll}
3 & 1 & 3 \\
4 & 0 & 4 \\
3 & 2 & 4
\end{array}\right),\left(\begin{array}{lll}
0 & 2 & 3 \\
0 & 2 & 1 \\
4 & 1 & 0
\end{array}\right)
$$

If we pick the top left corners, the middle entry, the bottom right corner and the entries in the middle of the bottom row, an appropriate augmented matrix is

$$
\left(\begin{array}{lllll}
1 & 1 & 3 & 0 & 0 \\
1 & 3 & 0 & 2 & 0 \\
1 & 1 & 4 & 0 & 0 \\
0 & 1 & 2 & 1 & 0
\end{array}\right)
$$

Then row reduced echelon form in $\mathbb{Z}_{5}$ is

$$
\left(\begin{array}{lllll}
1 & 0 & 0 & 4 & 0 \\
0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}\right)
$$

so it would seem a possible minimum polynomial is obtained by $a=1, b=-1=4, c=$ $0, d=1$. Thus it has degree 3 . There cannot be any polynomial of smaller degree because of the first three columns so it would seem that this should be the minimum polynomial,

$$
1+4 \lambda+\lambda^{3}
$$

Does it send the matrix to 0 ? This just involves checking whether it does and in fact, this is the case using the arithmetic in the residue class.

In summary, it is not all that hard to find the minimum polynomial.

### 6.3 Eigenvalues, Eigenvectors

We begin with the following fundamental definition.
Definition 6.3.1 Let $L \in \mathscr{L}(V, V)$ where $V$ is a vector space of dimension $n$ with field of scalars $\mathbb{F}$. An eigen-pair consists of a scalar $\lambda \in \mathbb{F}$ called an eigenvalue and a NON-ZERO $v \in V$ such that

$$
(\lambda I-L) v=\mathbf{0}
$$

Do eigen-pairs exist? Recall that from Theorem 6.1.10 the minimum polynomial can be factored in a unique way as $p(\lambda)=\prod_{i=1}^{p} \phi_{i}(\lambda)^{k_{i}}$ where each $\phi_{i}(\lambda)$ is irreducible and monic. Then the following theorem is obtained.

Theorem 6.3.2 Let $L \in \mathscr{L}(V, V)$ and let its minimum polynomial $p(\lambda)$ have a root $\mu$ in the field of scalars. Then $\mu$ is an eigenvalue of $L$.

Proof: Since $p(\lambda)$ has a root, we know $p(\lambda)=(\lambda-\mu) q(\lambda)$ where the degree of $q(\lambda)$ is less than the degree of $p(\lambda)$. Therefore, there is a vector $u$ such that $q(L) u \equiv v \neq 0$. Otherwise, $p(\lambda)$ is not really the minimum polynomial because $q(\lambda)$ would work better. Then $(L-\mu I) q(L) u=(L-\mu I) v=0$ and so $\mu$ is indeed an eigenvalue.

Theorem 6.3.3 Suppose the minimum polynomial $p(\lambda)$ of $L \in \mathscr{L}(V, V)$ factors completely into linear factors (splits) so that

$$
p(\lambda)=\prod_{i=1}^{p}\left(\lambda-\mu_{i}\right)^{k_{i}}
$$

Then the $\mu_{i}$ are distinct eigenvalues and corresponding to each of these eigenvalues, there is an eigenvector $w_{i} \neq 0$ such that $L w_{i}=\mu_{i} w_{i}$. Also, there are no other eigenvalues than these $\mu_{i}$. Also

$$
V=\operatorname{ker}\left(L-\mu_{1} I\right)^{k_{1}} \oplus \cdots \oplus \operatorname{ker}\left(L-\mu_{p} I\right)^{k_{p}}
$$

and if $L_{i}$ is the restriction of $L$ to $\operatorname{ker}\left(A-\mu_{i} I\right)^{k_{i}}$, then $L_{i}$ has exactly one eigenvalue and it is $\mu_{i}$.

Proof: By Theorem 6.3.2, each $\mu_{i}$ is an eigenvalue and we can let $w_{i}$ be a corresponding eigenvector. By Theorem 6.1.10,

$$
V=\operatorname{ker}\left(L-\mu_{1} I\right)^{k_{1}} \oplus \cdots \oplus \operatorname{ker}\left(L-\mu_{p} I\right)^{k_{p}}
$$

Also by this theorem, the minimum polynomial of $L_{i}$ is $\left(\lambda-\mu_{i}\right)^{k_{i}}$ and so it has an eigenvalue $\mu_{i}$. Could $L_{i}$ have any other eigenvalue $v \neq \mu_{i}$ ? To save notation, denote by $m$ the exponent $k_{i}$ and by $\mu$ the eigenvalue $\mu_{i}$. Also let $w$ denote an eigenvector of $L_{i}$ with respect to $v$. Then since the minimum polynomial for $L_{i}$ is $(\lambda-\mu)^{m}$,

$$
\begin{aligned}
0 & =(L-\mu I)^{m} w=(L-v I+(v-\mu) I)^{m} w \\
& =\sum_{k=0}^{m}\binom{m}{k}(L-v I)^{m-k}(v-\mu)^{k} w=(v-\mu)^{m} w
\end{aligned}
$$

which is impossible because $w \neq 0$. Thus there can be no other eigenvalue for $L_{i}$.
Consider the claim about $L$ having no other eigenvalues than the $\mu_{i}$. Say $\mu$ is another eigenvalue with eigenvector $w$. Then let $w=\sum_{i} z_{i}, z_{i} \in \operatorname{ker}\left(L-\mu_{i} I\right)^{k_{i}}$. Then not every $z_{i}=0$ and

$$
0=(L-\mu I) \sum_{i} z_{i}=\sum_{i}\left(L z_{i}-\mu z_{i}\right)=\sum_{i} L_{i} z_{i}-\mu z_{i}
$$

Since this is a direct sum and each $\operatorname{ker}\left(L-\mu_{i} I\right)^{k_{i}}$ is invariant with respect to $L$, we must have each $L_{i} z_{i}-\mu z_{i}=0$. This is impossible unless $\mu$ equals some $\mu_{i}$ because not every $z_{i}$ is 0 .

Example 6.3.4 The minimum polynomial for the matrix

$$
A=\left(\begin{array}{ccc}
4 & 0 & -6 \\
-1 & 2 & 3 \\
1 & 0 & -1
\end{array}\right)
$$

is $\lambda^{2}-3 \lambda+2$. This factors as $(\lambda-2)(\lambda-1)$ and so the eigenvalues are 1,2 . Find the eigen-pairs. Then determine the matrix with respect to a basis of these eigenvectors if possible.

First consider the eigenvalue 2 . There exists a nonzero vector $v$ such that $(A-2 I) v=0$. This follows from the above theory. However, it is best to just find it directly rather than try to get it by using the proof of the above theorem. The augmented matrix to consider is then

$$
\left(\begin{array}{cccc}
4-2 & 0 & -6 & 0 \\
-1 & 2-2 & 3 & 0 \\
1 & 0 & -1-2 & 0
\end{array}\right)
$$

Row reducing this yields

$$
\left(\begin{array}{cccc}
1 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Thus the solution is any vector of the form

$$
\left(\begin{array}{c}
3 z \\
y \\
z
\end{array}\right)=z\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right)+y\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right), z, y \text { not both } 0
$$

Now consider the eigenvalue 1 . This time you row reduce

$$
\left(\begin{array}{cccc}
4-1 & 0 & -6 & 0 \\
-1 & 2-1 & 3 & 0 \\
1 & 0 & -1-1 & 0
\end{array}\right)
$$

which yields for the row reduced echelon form

$$
\left(\begin{array}{cccc}
1 & 0 & -2 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Thus an eigenvector is of the form

$$
\left(\begin{array}{c}
2 z \\
-z \\
z
\end{array}\right), z \neq 0
$$

Consider a basis for $\mathbb{R}^{n}$ of the form

$$
\left\{\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right),\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right),\left(\begin{array}{c}
2 \\
-1 \\
1
\end{array}\right)\right\}
$$

You might want to consider Problem 9 on Page 114 at this point. This problem shows that the matrix with respect to this basis is diagonal.

When the matrix of a linear transformation can be chosen to be a diagonal matrix, the transformation is said to be nondefective. Also, note that the term applies to the matrix of a linear transformation and so I will specialize to the consideration of matrices in what follows. As shown above, this is equivalent to saying that any matrix of the linear transformation is similar to one which is diagonal. That is, the matrix of a linear transformation, or more generally just a square matrix $A$ has the property that there exists $S$ such that $S^{-1} A S=D$ where $D$ is a diagonal matrix.

Here is a definition which also introduces one of the most horrible adjectives in all of mathematics.

### 6.4 Diagonalizability

Diagonalizability is a term intended to be descriptive of whether a given matrix is similar to a diagonal matrix. More precisely one has the following definition.

Definition 6.4.1 Let $A$ be an $n \times n$ matrix. Then $A$ is diagonalizable if there exists an invertible matrix $S$ such that $S^{-1} A S=D$ where $D$ is a diagonal matrix. This means $D$ has a zero as every entry except for the main diagonal. More precisely, $D_{i j}=0$ unless $i=j$. Such matrices look like the following.

$$
\left(\begin{array}{ccc}
* & & 0 \\
& \ddots & \\
0 & & *
\end{array}\right)
$$

where $*$ might not be zero.
The most important theorem about diagonalizability ${ }^{1}$ is the following major result. First here is a simple observation.

Observation 6.4.2 Let $S=\left(\begin{array}{lll}s_{1} & \cdots & s_{n}\end{array}\right)$ where $S$ is $n \times n$. Then here is the result of multiplying on the right by a diagonal matrix.

$$
\left(\begin{array}{lll}
s_{1} & \cdots & s_{n}
\end{array}\right)\left(\begin{array}{ccc}
\lambda_{1} & & \\
& \ddots & \\
& & \lambda_{n}
\end{array}\right)=\left(\begin{array}{lll}
\lambda_{1} s_{1} & \cdots & \lambda_{n} s_{n}
\end{array}\right)
$$

This follows from the way we multiply matrices. The diagonal matrix has $i j j^{\text {th }}$ entry equal to $\delta_{i j} \lambda_{j}$ and the $i j^{\text {th }}$ entry of the matrix on the far left is $s^{j i}$ where

$$
s_{i}=\left(\begin{array}{llll}
s^{1 i} & s^{2 i} & \cdots & s^{n i}
\end{array}\right)^{T}
$$

Thus the $i j^{\text {th }}$ entry of the product on the left is $\sum_{k} s^{i k} \delta_{k j} \lambda_{j}=s^{i j} \lambda_{j}$. It follows that the $j^{\text {th }}$ column is

$$
\left(\begin{array}{cccc}
s^{1 j} \lambda_{j} & s^{2 j} \lambda_{j} & \cdots & s^{n j} \lambda_{j}
\end{array}\right)^{T}=\lambda_{j} s_{j}
$$

[^5]Theorem 6.4.3 An $n \times n$ matrix is diagonalizable if and only if $\mathbb{F}^{n}$ has a basis of eigenvectors of $A$. Furthermore, you can take the matrix $S$ described above, to be given as

$$
S=\left(\begin{array}{llll}
s_{1} & s_{2} & \cdots & s_{n}
\end{array}\right)
$$

where here the $s_{k}$ are the eigenvectors in the basis for $\mathbb{F}^{n}$. If $A$ is diagonalizable, the eigenvalues of $A$ are the diagonal entries of the diagonal matrix.

Proof: To say that $A$ is diagonalizable, is to say that for some $S$,

$$
S^{-1} A S=\left(\begin{array}{ccc}
\lambda_{1} & & \\
& \ddots & \\
& & \lambda_{n}
\end{array}\right)
$$

the $\lambda_{i}$ being elements of $\mathbb{F}$. This is to say that for $S=\left(\begin{array}{lll}s_{1} & \cdots & s_{n}\end{array}\right), s_{k}$ being the $k^{\text {th }}$ column,

$$
A\left(\begin{array}{lll}
s_{1} & \cdots & s_{n}
\end{array}\right)=\left(\begin{array}{lll}
s_{1} & \cdots & s_{n}
\end{array}\right)\left(\begin{array}{ccc}
\lambda_{1} & & \\
& \ddots & \\
& & \lambda_{n}
\end{array}\right)
$$

which is equivalent, from the way we multiply matrices and the above observation, that

$$
\left(\begin{array}{lll}
A s_{1} & \cdots & A s_{n}
\end{array}\right)=\left(\begin{array}{lll}
\lambda_{1} s_{1} & \cdots & \lambda_{n} s_{n}
\end{array}\right)
$$

which is equivalent to saying that the columns of $S$ are eigenvectors and the diagonal matrix has the eigenvectors down the main diagonal. Since $S^{-1}$ is invertible, these eigenvectors are a basis. Similarly, if there is a basis of eigenvectors, one can take them as the columns of $S$ and reverse the above steps, finally concluding that $A$ is diagonalizable.
Corollary 6.4.4 Let A be an $n \times n$ matrix with minimum polynomial

$$
p(\lambda)=\prod_{i=1}^{p}\left(\lambda-\mu_{i}\right)^{k_{i}}, \text { the } \mu_{i} \text { being distinct } .
$$

Then $A$ is diagonalizable if and only if each $k_{i}=1$.
Proof: Suppose first that it $A$ is diagonalizable with a basis of eigenvectors $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ with $A \boldsymbol{v}_{i}=\mu_{i} \boldsymbol{v}_{i}$. Since $n \geq p$, there may be some repeats here, a $\mu_{i}$ going with more than one $\boldsymbol{v}_{i}$. Say $k_{i}>1$. Now consider $\hat{p}(\boldsymbol{\lambda}) \equiv \prod_{j=1, j \neq i}^{p}\left(\lambda-\mu_{j}\right)^{k_{j}}\left(\lambda-\mu_{i}\right)$. Thus this is a monic polynomial which has smaller degree than $p(\lambda)$. If you have $\boldsymbol{v} \in \mathbb{F}^{n}$, since this is a basis, there are scalars $c_{i}$ such that $\boldsymbol{v}=\sum_{j} c_{j} \boldsymbol{v}_{j}$. Then $\hat{p}(A) \boldsymbol{v}=\mathbf{0}$. Since $\boldsymbol{v}$ is arbitrary, this shows that $\hat{p}(A)=0$ contrary to the definition of the minimum polynomial being $p(\lambda)$. Thus each $k_{i}$ must be 1 .

Conversely, if each $k_{i}=1$, then

$$
\mathbb{F}^{n}=\operatorname{ker}\left(A-\mu_{1} I\right) \oplus \cdots \oplus \operatorname{ker}\left(A-\mu_{p} I\right)
$$

and you simply let $\beta_{i}$ be a basis for $\operatorname{ker}\left(A-\mu_{i} I\right)$ which consists entirely of eigenvectors by definition of what you mean by $\operatorname{ker}\left(A-\mu_{i} I\right)$. Then a basis of eigenvectors consists of $\left\{\beta_{1}, \beta_{2}, \cdots, \beta_{p}\right\}$ and so the matrix $A$ is diagonalizable.

Example 6.4.5 The minimum polynomial for the matrix

$$
A=\left(\begin{array}{ccc}
10 & 12 & -6 \\
-4 & -4 & 3 \\
3 & 4 & -1
\end{array}\right)
$$

is $\lambda^{3}-5 \lambda^{2}+8 \lambda-4$. This factors as $(\lambda-2)^{2}(\lambda-1)$ and so the eigenvalues are 1,2 . Find the eigen-pairs. Then determine the matrix with respect to a basis of these eigenvectors if possible. If it is not possible to find a basis of eigenvectors, find a block diagonal matrix similar to the matrix. Note that from the above theorem, it is not possible to diagonalize this matrix.

First find the eigenvectors for 2 . You need to row reduce

$$
\left(\begin{array}{cccc}
10-2 & 12 & -6 & 0 \\
-4 & -4-2 & 3 & 0 \\
3 & 4 & -1-2 & 0
\end{array}\right)
$$

This yields

$$
\left(\begin{array}{cccc}
1 & 0 & -3 & 0 \\
0 & 1 & \frac{3}{2} & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Thus the eigenvectors which go with 2 are

$$
\left(\begin{array}{ccc}
6 z & -3 z & 2 z
\end{array}\right)^{T}, z \in \mathbb{R}, z \neq 0
$$

The eigenvectors which go with 1 are

$$
z\left(\begin{array}{lll}
2 & -1 & 1
\end{array}\right)^{T}, z \in \mathbb{R}, z \neq 0
$$

By Theorem 6.3.3, there are no other eigenvectors than those which correspond to eigenvalues 1,2 . Thus there is no basis of eigenvectors because the span of the eigenvectors has dimension two.

However, we can consider

$$
\mathbb{R}^{3}=\operatorname{ker}\left((A-2 I)^{2}\right) \oplus \operatorname{ker}(A-I)
$$

The second of these is just span $\left.\left(\begin{array}{ccc}2 & -1 & 1\end{array}\right)^{T}\right)$. What is the first? We find it by row reducing the following matrix which is the square of $A-2 I$ augmented with a column of zeros.

$$
\left(\begin{array}{cccc}
-2 & 0 & 6 & 0 \\
1 & 0 & -3 & 0 \\
-1 & 0 & 3 & 0
\end{array}\right)
$$

Row reducing this yields

$$
\left(\begin{array}{cccc}
1 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

which says that solutions are of the form

$$
\left(\begin{array}{c}
3 z \\
y \\
z
\end{array}\right), y, z \in \mathbb{R} \text { not both } 0
$$

This is the nonzero vectors of.

$$
\operatorname{span}\left(\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right),\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right)\right)
$$

Note these are not eigenvectors. They are called generalized eigenvectors because they pertain to $\operatorname{ker}\left((A-2 I)^{2}\right)$ rather than $\operatorname{ker}((A-2 I))$. What is the matrix of the restriction of $A$ to this subspace having ordered basis

$$
\begin{aligned}
& \left\{\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right),\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right),\left(\begin{array}{c}
2 \\
-1 \\
1
\end{array}\right)\right\} \\
& A\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right)=\left(\begin{array}{ccc}
10 & 12 & -6 \\
-4 & -4 & 3 \\
3 & 4 & -1
\end{array}\right)\left(\begin{array}{l}
3 \\
0 \\
1
\end{array}\right)=\left(\begin{array}{c}
24 \\
-9 \\
8
\end{array}\right) \\
& A\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right)=\left(\begin{array}{ccc}
10 & 12 & -6 \\
-4 & -4 & 3 \\
3 & 4 & -1
\end{array}\right)\left(\begin{array}{l}
0 \\
1 \\
0
\end{array}\right)=\left(\begin{array}{c}
12 \\
-4 \\
4
\end{array}\right)
\end{aligned}
$$

Then

$$
\left(\begin{array}{cc}
24 & 12  \tag{6.1}\\
-9 & -4 \\
8 & 4
\end{array}\right)=\left(\begin{array}{cc}
3 & 0 \\
0 & 1 \\
1 & 0
\end{array}\right) M
$$

and so some computations yield

$$
M=\left(\begin{array}{cc}
8 & 4 \\
-9 & -4
\end{array}\right)
$$

Indeed this works

$$
\left(\begin{array}{ll}
3 & 0 \\
0 & 1 \\
1 & 0
\end{array}\right)\left(\begin{array}{cc}
8 & 4 \\
-9 & -4
\end{array}\right)=\left(\begin{array}{cc}
24 & 12 \\
-9 & -4 \\
8 & 4
\end{array}\right)
$$

Then the matrix associated with the other eigenvector is just 1 . Hence the matrix with respect to the above ordered basis is

$$
\left(\begin{array}{ccc}
8 & 4 & 0 \\
-9 & -4 & 0 \\
0 & 0 & 1
\end{array}\right)
$$

So what are some convenient computations which will allow you to find $M$ easily? Take the transpose of both sides of 6.1. Then you would have

$$
\left(\begin{array}{ccc}
24 & -9 & 8 \\
12 & -4 & 4
\end{array}\right)=M^{T}\left(\begin{array}{ccc}
3 & 0 & 1 \\
0 & 1 & 0
\end{array}\right)
$$

Thus

$$
M^{T}\binom{0}{1}=\binom{-9}{-4}, M^{T}\binom{1}{0}=\binom{8}{4}
$$

and so $M^{T}=\left(\begin{array}{cc}8 & -9 \\ 4 & -4\end{array}\right)$ so $M=\left(\begin{array}{cc}8 & 4 \\ -9 & -4\end{array}\right)$.
The eigenvalue problem is one of the hardest problems in algebra because of our inability to exactly solve polynomial equations. Therefore, estimating the eigenvalues becomes very significant. In the case of the complex field of scalars, there is a very elementary result due to Gerschgorin. It can at least give an upper bound for the size of the eigenvalues.

Theorem 6.4.6 Let A be an $n \times n$ matrix. Consider the $n$ Gerschgorin discs defined as

$$
D_{i} \equiv\left\{\lambda \in \mathbb{C}:\left|\lambda-a_{i i}\right| \leq \sum_{j \neq i}\left|a_{i j}\right|\right\} .
$$

Then every eigenvalue is contained in some Gerschgorin disc.
This theorem says to add up the absolute values of the entries of the $i^{\text {th }}$ row which are off the main diagonal and form the disc centered at $a_{i i}$ having this radius. The union of these discs contains $\sigma(A)$.

Proof: Suppose $A \boldsymbol{x}=\boldsymbol{\lambda} \boldsymbol{x}$ where $\boldsymbol{x} \neq \mathbf{0}$. Then for $A=\left(a_{i j}\right)$, let $\left|x_{k}\right| \geq\left|x_{j}\right|$ for all $x_{j}$. Thus $\left|x_{k}\right| \neq 0$.

$$
\sum_{j \neq k} a_{k j} x_{j}=\left(\lambda-a_{k k}\right) x_{k} .
$$

Then

$$
\left|x_{k}\right| \sum_{j \neq k}\left|a_{k j}\right| \geq \sum_{j \neq k}\left|a_{k j}\right|\left|x_{j}\right| \geq\left|\sum_{j \neq k} a_{k j} x_{j}\right|=\left|\lambda-a_{i i}\right|\left|x_{k}\right| .
$$

Now dividing by $\left|x_{k}\right|$, it follows $\lambda$ is contained in the $k^{t h}$ Gerschgorin disc.
In these examples given above, it was possible to factor the minimum polynomial and explicitly determine eigenvalues and eigenvectors and obtain information about whether the matrix was diagonalizable by explicit computations. Well, what if you can't factor the minimum polynomial? What then? This is the typical situation, not what was presented in the above examples. Just write down a $3 \times 3$ matrix and see if you can find the eigenvalues explicitly using algebra. Is there a way to determine whether a given matrix is diagonalizable in the case that the minimum polynomial factors although you might have trouble finding the factors? Amazingly, the answer is yes. One can answer this question completely using only methods from algebra.

### 6.5 A Formal Derivative and Diagonalizability

For $p(\lambda)=a_{n} \lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}$ where $n$ is a positive integer, define

$$
p^{\prime}(\lambda) \equiv n a_{n} \lambda^{n-1}+(n-1) a_{n-1} \lambda^{n-2}+\cdots+a_{1}
$$

In other words, you use the usual rules of differentiation in calculus to write down this formal derivative. It has absolutely no physical significance in this context because the coefficients are just elements of some field, possibly $\mathbb{Z}_{p}$. It is a purely algebraic manipulation. A term like $k a$ where $k \in \mathbb{N}$ and $a \in \mathbb{F}$ means to add $a$ to itself $k$ times. There are no limits or anything else. However, this has certain properties. In particular, the "derivative" of a sum equals the sum of the derivatives. This is fairly clear from the above definition. You just need to always be considering polynomials. Also

$$
\begin{aligned}
& \left(b \lambda^{m}\left(a_{n} \lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}\right)\right)^{\prime} \\
= & \left(a_{n} b \lambda^{n+m}+b a_{n-1} \lambda^{m+(n-1)}+\cdots+b a_{1} \lambda^{1+m}+a_{0} b \lambda^{m}\right)^{\prime} \\
\equiv & a_{n} b(n+m) \lambda^{n+m-1}+b a_{n-1}(m+n-1) \lambda^{m+n-2}+ \\
& \cdots+b a_{1}(m+1) \lambda^{m}+a_{0} b m \lambda^{m-1}
\end{aligned}
$$

Will the product rule give the same thing? Is it true that the above equals

$$
\begin{aligned}
& \left(b \lambda^{m}\right)^{\prime}\left(a_{n} \lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}\right) \\
& +b \lambda^{m}\left(a_{n} \lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}\right)^{\prime} ?
\end{aligned}
$$

A short computation shows that this is indeed the case. Then by induction one can conclude that

$$
\left(\prod_{i=1}^{p} p_{i}(\boldsymbol{\lambda})\right)=\sum_{j=1}^{p} p_{j}^{\prime}(\boldsymbol{\lambda}) \prod_{i \neq j} p_{i}(\boldsymbol{\lambda})
$$

In particular, if

$$
p(\lambda)=\prod_{i=1}^{p}\left(\lambda-\mu_{i}\right)^{k_{i}}
$$

then

$$
p^{\prime}(\lambda)=\sum_{j=1}^{p} k_{j}\left(\lambda-\mu_{j}\right)^{k_{j}-1} \prod_{i \neq j}\left(\lambda-\mu_{i}\right)^{k_{i}}
$$

I want to emphasize that this is an arbitrary field of scalars, but if one is only interested in the real or complex numbers, then all of this follows from standard calculus theorems.

Proposition 6.5.1 Suppose the minimum polynomial $p(\lambda)$ of an $n \times n$ matrix $A$ completely factors into linear factors. Then $A$ is diagonalizable if and only if $p(\lambda), p^{\prime}(\lambda)$ are relatively prime.

Proof: Suppose $p(\lambda), p^{\prime}(\lambda)$ are relatively prime. Say

$$
p(\lambda)=\prod_{i=1}^{n}\left(\lambda-\mu_{i}\right)^{k_{i}}, \mu_{i} \text { are distinct }
$$

From the above discussion,

$$
p^{\prime}(\lambda)=\sum_{j=1}^{p} k_{j}\left(\lambda-\mu_{j}\right)^{k_{j}-1} \prod_{i \neq j}\left(\lambda-\mu_{i}\right)^{k_{i}}
$$

and $p^{\prime}(\lambda), p(\lambda)$ are relatively prime if and only if each $k_{i}=1$. Then by Corollary 6.4.4 this is true if and only if $A$ is diagonalizable.

Example 6.5.2 Find whether the matrix

$$
A=\left(\begin{array}{ccc}
1 & -1 & 2 \\
0 & 1 & 2 \\
1 & -1 & 1
\end{array}\right)
$$

is diagonalizable. Assume the field of scalars is $\mathbb{C}$ because in this field, the minimum polynomial will factor thanks to the fundamental theorem of algebra.

Successive powers of the matrix are

$$
\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right),\left(\begin{array}{ccc}
1 & -1 & 2 \\
0 & 1 & 2 \\
1 & -1 & 1
\end{array}\right),\left(\begin{array}{ccc}
3 & -4 & 2 \\
2 & -1 & 4 \\
2 & -3 & 1
\end{array}\right),\left(\begin{array}{ccc}
5 & -9 & 0 \\
6 & -7 & 6 \\
3 & -6 & -1
\end{array}\right)
$$

Then we need to have for a linear combination involving $a, b, c, d$ as scalars

$$
\begin{gathered}
a+b+3 c+5 d=0 \\
2 c+6 d=0 \\
b+2 c+3 d=0 \\
-b-3 c-6 d=0
\end{gathered}
$$

Then letting $d=1$, this gives only one solution, $a=1, b=3, c=-3$ and so the candidate for the minimum polynomial is $\lambda^{3}-3 \lambda^{2}+3 \lambda+1$. In fact, this does work as is seen by substituting $A$ for $\lambda$. So is this polynomial and its derivative relatively prime? The derivative is $3 \lambda^{2}-6 \lambda+3$. Dividing, one obtains

$$
\lambda^{3}-3 \lambda^{2}+3 \lambda+1=\frac{1}{3}(\lambda-1)\left(3 \lambda^{2}-6 \lambda+3\right)+2
$$

and clearly $\left(3 \lambda^{2}-6 \lambda+3\right)$ and 2 are relatively prime. Hence this matrix is diagonalizable. Of course, finding its diagonalization is another matter. For an algorithm for determining whether two polynomials are relatively prime, see Problem 34 on Page 35 or the process described in Section 2.3 on Page 44.

Of course this was an easy example thanks to Problem 12 on Page 146. because there are three distinct eigenvalues, one real and two complex which must be complex conjugates. This problem says that eigenvectors corresponding to distinct eigenvalues are an independent set. Be sure to do this problem.

Consider the following example in which the eigenvalues are not distinct, consisting of $a, a$.

## Example 6.5.3 Find whether the matrix

$$
A=\left(\begin{array}{cc}
a+1 & 1 \\
-1 & a-1
\end{array}\right)
$$

is diagonalizable.
Listing the powers of the matrix,

$$
\left(\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right),\left(\begin{array}{cc}
a+1 & 1 \\
-1 & a-1
\end{array}\right),\left(\begin{array}{cc}
a^{2}+2 a & 2 a \\
-2 a & a^{2}-2 a
\end{array}\right)
$$

Then we need to have for a linear combination involving scalars $x, y, z$

$$
\begin{aligned}
x+(a+1) y+\left(a^{2}+2 a\right) z & =0 \\
y+2 a z & =0 \\
x+(a-1) y+\left(a^{2}-2 a\right) z & =0
\end{aligned}
$$

Then some routine row operations yield $x=a^{2} z, y=-2 a z$ and $z$ is arbitrary. For the minimum polynomial, we take $z=1$ because this is a monic polynomial. Thus the minimum polynomial is $a^{2}-2 a \lambda+\lambda^{2}=(\lambda-a)^{2}$ and clearly this and its derivative are not relatively prime. Thus this matrix is not diagonalizable for any choice of $a$.

### 6.6 Exercises

1. For the linear transformation determined by multiplication by the following matrices, find the minimum polynomial.
(a) $\left(\begin{array}{cc}3 & 1 \\ -4 & -1\end{array}\right)$
(e) $\left(\begin{array}{ccc}1 & 0 & 0 \\ -2 & -1 & 0 \\ 3 & 6 & 2\end{array}\right)$
(b) $\left(\begin{array}{cc}0 & -2 \\ 1 & 3\end{array}\right)$
(f) $\left(\begin{array}{ccc}2 & 1 & 0 \\ -2 & 2 & 1 \\ 5 & -1 & -1\end{array}\right)$
(c) $\left(\begin{array}{ccc}2 & 1 & 0 \\ -1 & 0 & 0 \\ 2 & 5 & 2\end{array}\right)$
(g) $\left(\begin{array}{ccc}5 & 0 & 12 \\ -2 & 1 & -6 \\ -2 & 0 & -5\end{array}\right)$
2. Here is a matrix:

$$
\left(\begin{array}{ccc}
-1 & -4 & -2 \\
2 & 4 & 2 \\
-1 & -1 & 0
\end{array}\right)
$$

Its minimum polynomial is $\lambda^{3}-3 \lambda^{2}+4 \lambda-2=(\lambda-1)\left(-2 \lambda+\lambda^{2}+2\right)$. Obtain a block diagonal matrix similar to this one.
3. Suppose $A \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space and suppose $p(\lambda)$ is the minimum polynomial. Say $p(\lambda)=\lambda^{m}+a_{m-1} \lambda^{m-1}+\cdots+a_{1} \lambda+a_{0}$. If $A$ is one to one, show that it is onto and also that $A^{-1} \in \mathscr{L}(V, V)$. In this case, explain why $a_{0} \neq 0$. In this case, give a formula for $A^{-1}$ as a polynomial in $A$.
4. Let $A=\left(\begin{array}{cc}0 & -2 \\ 1 & 3\end{array}\right)$. Its minimum polynomial is $\lambda^{2}-3 \lambda+2$. Find $A^{10}$ exactly. Hint: You can do long division and get $\lambda^{10}=l(\lambda)\left(\lambda^{2}-3 \lambda+2\right)+1023 \lambda-1022$.
5. Suppose $A \in \mathscr{L}(V, V)$ and it has minimum polynomial $p(\lambda)$ which has degree $m$. It is desired to compute $A^{n}$ for $n$ large. Show that it is possible to obtain $A^{n}$ in terms of a polynomial in $A$ of degree less than $m$.
6. Determine whether the following matrices are diagonalizable. Assume the field of scalars is $\mathbb{C}$.
(a) $\left(\begin{array}{ccc}1 & 1 & 1 \\ -1 & 2 & 1 \\ 0 & 1 & 1\end{array}\right)$
(d) $\left(\begin{array}{ccc}1 & 1 & -1 \\ 2 & 1 & -1 \\ 0 & 1 & 2\end{array}\right)$
(b) $\left(\begin{array}{cc}\sqrt{2}+1 & 1 \\ -1 & \sqrt{2}-1\end{array}\right)$
(e) $\left(\begin{array}{ccc}2 & 1 & 0 \\ -1 & 0 & 0 \\ 2 & 2 & 1\end{array}\right)$
(c) $\left(\begin{array}{cc}a+1 & 1 \\ -1 & a-1\end{array}\right)$ where $a \in \mathbb{R}$
7. The situation for diagonalizability was presented for the situation in which the minimum polynomial factors completely as a product of linear factors since this is certainly the case of most interest, including $\mathbb{C}$. What if the minimum polynomial does not split? Is there a theorem available that will allow one to conclude that the matrix is diagonalizable in a splitting field, possibly larger than the given field? It is a reasonable question because the assumption that $p(\boldsymbol{\lambda}), p^{\prime}(\boldsymbol{\lambda})$ are relatively prime may be determined without factoring the polynomials and involves only computations involving the given field $\mathbb{F}$. If you enlarge the field, what happens to the minimum polynomial? Does it stay the same or does it change? Remember, the matrix has entries all in the smaller field $\mathbb{F}$ while a splitting field is $\mathbb{G}$ larger than $\mathbb{F}$, but you can determine the minimum polynomial using row operations on vectors in $\mathbb{F}^{n^{2}}$.
8. Suppose $V$ is a finite dimensional vector space and suppose $N \in \mathscr{L}(V, V)$ satisfies $N^{m}=0$ for some $m \geq 1$. Show that the only eigenvalue is 0 .
9. Suppose $V$ is an $n$ dimensional vector space and suppose $\beta$ is a basis for $V$. Consider the map $\mu I: V \rightarrow V$ given by $\mu I v=\mu v$. What is the matrix of this map with respect to the basis $\beta$ ? Hint: You should find that it is $\mu$ times the identity matrix whose $i j^{\text {th }}$ entry is $\delta_{i j}$ which is 1 if $i=j$ and 0 if $i \neq j$. Thus the $i j^{t h}$ entry of this matrix will be $\mu \delta_{i j}$.
10. In the case that the minimum polynomial factors, which was discussed above, we had

$$
V=\operatorname{ker}\left(L-\mu_{1} I\right)^{k_{1}} \oplus \cdots \oplus \operatorname{ker}\left(L-\mu_{p} I\right)^{k_{p}}
$$

If $V_{i}=\operatorname{ker}\left(L-\mu_{i} I\right)^{k_{i}}$, then by definition, $\left(L_{i}-\mu_{i} I\right)^{k_{i}}=0$ where here $L_{i}$ is the restriction of $L$ to $V_{i}$. If $N=L_{i}-\mu_{i} I$, then $N: V_{i} \rightarrow V_{i}$ and $N^{k_{i}}=0$. This is the definition of a nilpotent transformation, one which has a high enough power equal to 0 . Suppose then that $N: V \rightarrow V$ where $V$ is an $m$ dimensional vector space. We will show that there is a basis for $V$ such that with respect to this basis, the matrix of $N$ is block diagonal and of the form

$$
\left(\begin{array}{ccc}
N_{1} & & 0 \\
& \ddots & \\
0 & & N_{s}
\end{array}\right)
$$

where $N_{i}$ is an $r_{i} \times r_{i}$ matrix of the form

$$
\left(\begin{array}{llll}
0 & 1 & & 0 \\
& 0 & \ddots & \\
& & \ddots & 1 \\
0 & & & 0
\end{array}\right)
$$

That is, there are ones down the superdiagonal and zeros everywhere else. Now consider the case where $N_{i}=L_{i}-\mu_{i} I$ on one of the $V_{i}$ as just described. Use the preceding problem and the special basis $\beta_{i}$ just described for $N_{i}$ to show that the matrix of $L_{i}$ with respect to this basis is of the form

$$
J\left(\mu_{i}\right) \equiv\left(\begin{array}{ccc}
J_{1}\left(\mu_{i}\right) & & 0 \\
& \ddots & \\
0 & & J_{s}\left(\mu_{i}\right)
\end{array}\right)
$$

where $J_{r}\left(\mu_{i}\right)$ is of the form

$$
\left(\begin{array}{cccc}
\mu_{i} & 1 & & 0 \\
& \mu_{i} & \ddots & \\
& & \ddots & 1 \\
0 & & & \mu_{i}
\end{array}\right)
$$

This is called a Jordan block. Now let $\beta=\left(\beta_{1}, \cdots, \beta_{p}\right)$. Explain why the matrix of $L$ with respect to this basis is of the form

$$
\left(\begin{array}{ccc}
J\left(\mu_{1}\right) & & 0 \\
& \ddots & \\
0 & & J\left(\mu_{p}\right)
\end{array}\right)
$$

This special matrix is called the Jordan canonical form. This problem shows that it reduces to the study of the matrix of a nilpotent matrix. You see that it is a block diagonal matrix such that each block is a block diagonal matrix which is also an upper triangular matrix having the eigenvalues down the main diagonal and strings of ones on the super diagonal.
11. Now in this problem, the method for finding the special basis for a nilpotent transformation is given. Let $V$ be a vector space and let $N \in \mathscr{L}(V, V)$ be nilpotent. First note the only eigenvalue of $N$ is 0 . Why? (See Problem 8.) Let $v_{1}$ be an eigenvector. Then $\left\{v_{1}, v_{2}, \cdots, v_{r}\right\}$ is called a chain based on $v_{1}$ if $N v_{k+1}=v_{k}$ for all $k=1,2, \cdots, r$ and $v_{1}$ is an eigenvector so $N v_{1}=0$. It will be called a maximal chain if there is no solution $v$, to the equation, $N v=v_{r}$. Now there will be a sequence of steps leading to the desired basis.
(a) Show the vectors in any chain are linearly independent and for $\left\{v_{1}, v_{2}, \cdots, v_{r}\right\}$ a chain based on $v_{1}$,

$$
\begin{equation*}
N: \operatorname{span}\left(v_{1}, v_{2}, \cdots, v_{r}\right) \mapsto \operatorname{span}\left(v_{1}, v_{2}, \cdots, v_{r}\right) \tag{6.2}
\end{equation*}
$$

Also if $\left\{v_{1}, v_{2}, \cdots, v_{r}\right\}$ is a chain, then $r \leq n$. Hint: If $0=\sum_{i=1}^{r} c_{i} v_{i}$, and the last nonzero scalar occurs at $l$, do $N^{l-1}$ to the sum and see what happens to $c_{l}$.
(b) Consider the set of all chains based on eigenvectors. Since all have total length no larger than $n$ it follows there exists one of maximal length, $\left\{v_{1}^{1}, \cdots, v_{r_{1}}^{1}\right\} \equiv$ $B_{1}$. If span $\left(B_{1}\right)$ contains all eigenvectors of $N$, then stop. Otherwise, consider all chains based on eigenvectors not in span $\left(B_{1}\right)$ and pick one, $B_{2} \equiv$ $\left\{v_{1}^{2}, \cdots, v_{r_{2}}^{2}\right\}$ which is as long as possible. Thus $r_{2} \leq r_{1}$. If $\operatorname{span}\left(B_{1}, B_{2}\right)$ contains all eigenvectors of $N$, stop. Otherwise, consider all chains based on eigenvectors not in $\operatorname{span}\left(B_{1}, B_{2}\right)$ and pick one, $B_{3} \equiv\left\{v_{1}^{3}, \cdots, v_{r_{3}}^{3}\right\}$ such that $r_{3}$ is as large as possible. Continue this way. Thus $r_{k} \geq r_{k+1}$. Then show that the above process terminates with a finite list of chains $\left\{B_{1}, \cdots, B_{s}\right\}$ because for any $k,\left\{B_{1}, \cdots, B_{k}\right\}$ is linearly independent. Hint: From part a. you know this is true if $k=1$. Suppose true for $k-1$ and letting $L\left(B_{i}\right)$ denote a linear combination of vectors of $B_{i}$, suppose $\sum_{i=1}^{k} L\left(B_{i}\right)=0$. Then we can assume $L\left(B_{k}\right) \neq 0$ by induction. Let $v_{i}^{k}$ be the last term in $L\left(B_{k}\right)$ which has nonzero scalar. Now act on the whole thing with $N^{i-1}$ to find $v_{1}^{k}$ as a linear combination of vectors in $\left\{B_{1}, \cdots, B_{k-1}\right\}$, a contradiction to the construction. You fill in the details.
(c) Suppose $N w=0$. ( $w$ is an eigenvector). Show that there exist scalars, $c_{i}$ such that $w=\sum_{i=1}^{s} c_{i} v_{1}^{i}$. Recall that $v_{1}^{i}$ is the eigenvector in the $i^{\text {th }}$ chain on which this chain is based. You know that $w$ is a linear combination of the vectors in $\left\{B_{1}, \cdots, B_{s}\right\}$. This says that in fact it is a linear combination of the bottom vectors in the $B_{i}$. Hint: You know that $w=\sum_{i=1}^{s} L\left(B_{i}\right)$. Let $v_{i}^{s}$ be the last in $L\left(B_{s}\right)$ which has nonzero scalar. Suppose that $i>1$. Now do $N^{i-1}$ to both sides and obtain that $v_{1}^{s}$ is in the span of $\left\{B_{1}, \cdots, B_{s-1}\right\}$ which is a contradiction. Hence $i=1$ and so the only term of $L\left(B_{s}\right)$ is one involving an eigenvector. Now do something similar to $L\left(B_{s-1}\right), L\left(B_{s-2}\right)$ etc. You fill in details.
(d) If $N w=0$, then $w \in \operatorname{span}\left(B_{1}, \cdots, B_{s}\right)$. This was what was just shown. In fact, it was a particular linear combination involving the bases of the chains. What if $N^{k} w=0$ ? Does it follow that $w \in \operatorname{span}\left(B_{1}, \cdots, B_{s}\right)$ ? Show that if $N^{k} w=0$, then $w \in \operatorname{span}\left(B_{1}, \cdots, B_{s}\right)$. Hint: Say $k$ is as small as possible such that $N^{k} w=0$. Then you have $N^{k-1} w$ is an eigenvector and so $N^{k-1} w=\sum_{i=1}^{s} c_{i} v_{1}^{i}$ If $N^{k-1} w$ is the base of some chain $B_{i}$, then there is nothing to show. Otherwise, consider the chain $N^{k-1} w, N^{k-2} w, \cdots, w$. It cannot be any longer than any of the chains $B_{1}, B_{2}, \cdots, B_{s}$ why? Therefore, $v_{1}^{i}=N^{k-1} v_{k}^{i}$. Why is $v_{k}^{i} \in B_{i}$ ? This is where you
use that this is no longer than any of the $B_{i}$. Thus

$$
N^{k-1}\left(w-\sum_{i=1}^{s} c_{i} v_{k}^{i}\right)=0
$$

By induction, (details) $w-\sum_{i=1}^{s} c_{i} v_{k}^{i} \in \operatorname{span}\left(B_{1}, \cdots, B_{s}\right)$.
(e) Since $N$ is nilpotent, $\operatorname{ker}\left(N^{m}\right)=V$ for some $m$ so $V=\operatorname{span}\left(B_{1}, \cdots, B_{s}\right)$.
(f) Explain why the matrix with respect to the ordered basis $\left(B_{1}, \cdots, B_{s}\right)$ is the kind of thing desired and described in the above problem. Also explain why the size of the blocks decreases from upper left to lower right. To see why the matrix is like the above, consider

$$
\left(\begin{array}{llll}
0 & v_{1}^{i} & \cdots & v_{r_{i}-1}^{i}
\end{array}\right)=\left(\begin{array}{llll}
v_{1}^{i} & v_{2}^{i} & \cdots & v_{r_{i}}^{i}
\end{array}\right) M_{i}
$$

where $M_{i}$ is the $i^{t h}$ block and $r_{i}$ is the length of the $i^{t h}$ chain.
If you have gotten through this, then along with the previous problem, you have proved the existence of the Jordan canonical form, one of the greatest results in linear algebra. It will be considered a different way later. Specifically, you have shown that if the minimum polynomial splits, then the linear transformation has a matrix of the following form:

$$
\left(\begin{array}{ccc}
J\left(\mu_{1}\right) & & 0 \\
& \ddots & \\
0 & & J\left(\mu_{p}\right)
\end{array}\right)
$$

where without loss of generality, you can arrange these blocks to be decreasing in size from the upper left to the lower right and $J\left(\mu_{i}\right)$ is of the form

$$
\left(\begin{array}{ccc}
J_{r_{1}}\left(\mu_{i}\right) & & 0 \\
& \ddots & \\
0 & & J_{r_{s}}\left(\mu_{i}\right)
\end{array}\right)
$$

Where $J_{r}\left(\mu_{i}\right)$ is the $r \times r$ matrix which is of the following form

$$
J_{r}\left(\mu_{i}\right)=\left(\begin{array}{cccc}
\mu_{i} & 1 & & 0 \\
& \mu_{i} & \ddots & \\
& & \ddots & 1 \\
0 & & & \mu_{i}
\end{array}\right)
$$

and the blocks $J_{r}\left(\mu_{i}\right)$ can also be arranged to have their size decreasing from the upper left to lower right.
12. (Extra important) The following theorem gives an easy condition for which the Jordan canonical form will be a diagonal matrix.

Theorem 6.6.1 Let $A \in \mathscr{L}(V, V)$ and suppose $\left(u_{i}, \lambda_{i}\right), i=1,2, \cdots, m$ are eigenpairs such that if $i \neq j$, then $\lambda_{i} \neq \lambda_{j}$. Then $\left\{u_{1}, \cdots, u_{m}\right\}$ is linearly independent. In words, eigenvectors from distinct eigenvalues are linearly independent.

Hint: Suppose $\sum_{i=1}^{k} c_{i} u_{i}=0$ where $k$ is as small as possible such that not all of the $c_{i}=0$. Then $c_{k} \neq 0$. Explain why $k>1$ and

$$
\sum_{i=1}^{k} c_{i} \lambda_{k} u_{i}=\sum_{i=1}^{k} c_{i} \lambda_{i} u_{i}
$$

Now

$$
\sum_{i=1}^{k} c_{i}\left(\lambda_{k}-\lambda_{i}\right) u_{i}=0
$$

Obtain a contradiction of some sort at this point. Thus if the $n \times n$ matrix has $n$ distinct eigenvalues, then the corresponding eigenvectors will be a linearly independent set and so the matrix will be diagonal and all the Jordan blocks will be single numbers.
13. This and the next few problems will give another presentation of the Jordan canonical form. Let $A \in \mathscr{L}(V, V)$ be a nonzero linear transformation where $V$ has finite dimensions. Consider

$$
\left\{x, A x, A^{2} x, \cdots, A^{m-1} x\right\}
$$

where for $k \leq m-1, A^{k} x$ is not in,

$$
A^{k} x \notin \operatorname{span}\left(x, A x, A^{2} x, \cdots, A^{k-1} x\right)
$$

show that then $\left\{x, A x, A^{2} x, \cdots, A^{m-1} x\right\}$ must be linearly independent. Hint: Let $\eta(\lambda)$ be the minimum polynomial for $A$. Then let $\phi(\lambda)$ be the monic polynomial of smallest degree such that $\phi(A) x=0$. Explain why $\phi(\lambda)$ divides $\eta(\lambda)$. Then show that if the degree of $\phi(\lambda)$ is $d$, then $\left\{x, A x, A^{2} x, \cdots, A^{d-1} x\right\}$ is linearly independent and if $k$ is as described above, then $k \leq d$. Note: linear dependence implies the existence of a polynomial $\psi(\boldsymbol{\lambda})$ such that $\psi(A) x=0$. An ordered set of vectors of the form $x, A x, A^{2} x, \cdots, A^{m-1} x$ where $A^{m} x \in \operatorname{span}\left(x, A x, A^{2} x, \cdots, A^{m-1} x\right)$ with $m$ as small as possible is called a cyclic set.
14. $\uparrow$ Suppose now that $N \in \mathscr{L}(V, V)$ for $V$ a finite dimensional vector space and the minimum polynomial for $N$ is $\lambda^{p}$. In other words, $N$ is nilpotent, $N^{p}=0$ and $p$ as small as possible. For $x \neq 0$, let $\beta_{x}=\left\{x, N x, N^{2} x, \cdots, N^{m-1} x\right\}$ where we keep the order of these vectors in $\beta_{x}$ and here $m$ is such that $N^{m} x \in \operatorname{span}\left(x, N x, N^{2} x, \cdots, N^{m-1} x\right)$ with $m$ as small as possible.
(a) Show that $N^{m} x=0$. Hint: You know from the assumption that

$$
N^{m} x \in \operatorname{span}\left(x, N x, N^{2} x, \cdots, N^{m-1} x\right)
$$

that there is a monic polynomial $\eta(\lambda)$ of degree $m$ such that $\eta(N) x=0$. Explain why $\eta(\lambda)$ divides the minimum polynomial $\lambda^{p}$. Then $\eta(\lambda)=\lambda^{m}$. Thus $N^{m} x=0$.
(b) For each $x \neq 0$, there is such a $\beta_{x}$ and let $V_{1} \equiv \operatorname{span}\left(\beta_{x_{1}}\right)$. Explain why $N$ : $V_{1} \rightarrow V_{1}$.
(c) Let $N_{1}$ be the restriction of $N$ to $V_{1}$. Find the matrix of $N_{1}$ with respect to the ordered basis $\left\{N^{m-1} x_{1}, \cdots, N x_{1}, x_{1}\right\}$. Note that we reverse the order of these vectors. This is just the traditional way of doing it. Show this matrix is of the form

$$
B \equiv\left(\begin{array}{cccc}
0 & 1 & & 0  \tag{6.3}\\
& 0 & \ddots & \\
& & \ddots & 1 \\
0 & & & 0
\end{array}\right)
$$

15. $\uparrow$ In the context of the above problems where

$$
N^{p}=0, N \in \mathscr{L}(V, V)
$$

and $\beta_{x}$ is defined as above, show that for each $k \leq p$, if $W$ is a subspace of $\operatorname{ker}\left(N^{k}\right)$ which is invariant with respect to $N$ meaning $N(W) \subseteq W$, then there are finitely many $y_{i} \in W$ such that

$$
W=\operatorname{span}\left(\beta_{y_{1}}, \beta_{y_{2}}, \cdots, \beta_{y_{s}}\right), \text { some } s
$$

and $\left\{\beta_{y_{1}}, \beta_{y_{2}}, \cdots, \beta_{y_{s}}\right\}$ is linearly independent. This is called a cyclic basis. Hint: If $W \subseteq \operatorname{ker}(N)$, this is obviously true because in this case, $\beta_{x}=x$ for $x \in \operatorname{ker}(N)$. Now suppose the assertion is true for $k<p$ and consider invariant $W \subseteq \operatorname{ker}\left(N^{k+1}\right)$. Argue as follows:
(a) Explain why $N(W)$ is an invariant subspace of $\operatorname{ker}\left(N^{k}\right)$. Thus, by induction, $N(W)=\operatorname{span}\left(\beta_{x_{1}}, \boldsymbol{\beta}_{x_{2}}, \cdots, \boldsymbol{\beta}_{x_{s}}\right)$ where that in $(\cdot)$ is a basis.
(b) Let $z \in W$ so $N z=\sum_{i=1}^{s} \sum_{j=0}^{r_{i}-1} a_{i j} N^{j} x_{j}$. Let $y_{j} \in W$ such that $N y_{j}=x_{j}$. Explain why

$$
N\left(z-\sum_{i=1}^{s} \sum_{j=0}^{r_{i}-1} a_{i j} N^{j} y_{i}\right)=0
$$

where the length of $\beta_{x_{i}}$ is $r_{i}$. Explain why there is an eigenvector $y_{0}$ such that

$$
z=\sum_{i=1}^{s} \sum_{j=0}^{r_{i}-1} a_{i j} N^{j} y_{i}+y_{0}
$$

(c) Note that $\beta_{y_{0}}=y_{0}$. Explain why

$$
\operatorname{span}\left(\beta_{y_{0}}, \beta_{y_{1}}, \cdots, \beta_{y_{s}}\right) \supseteq W
$$

Then explain why $\left\{\beta_{y_{0}}, \beta_{y_{1}}, \cdots, \beta_{y_{s}}\right\}$ is linearly independent. Hint: If

$$
\sum_{i=1}^{s} \sum_{j=0}^{r_{i}-1} a_{i j} N^{j} y_{i}+b y_{0}=0
$$

Do $N$ to both sides and use induction to conclude all $a_{i j}=0$.
16. $\uparrow$ Now in the above situation show that there is a basis for $V=\operatorname{ker}\left(N^{p}\right)$ such that with respect to this basis, the matrix of $N$ is block diagonal of the form

$$
\left(\begin{array}{cccc}
B_{1} & & &  \tag{6.4}\\
& B_{2} & & \\
& & \ddots & \\
& & & B_{r}
\end{array}\right)
$$

where the size of the blocks is decreasing from upper left to lower right and each block is of the form given in 6.3. Hint: Repeat the argument leading to this equation for each $\beta_{y_{i}}$ where the ordered basis for $\operatorname{ker}\left(N^{p}\right)$ is of the form

$$
\left\{\beta_{y_{1}}, \beta_{y_{2}}, \cdots, \beta_{y_{r}}\right\}
$$

arranged so that the length of $\beta_{y_{i}}$ is at least as long as the length of $\beta_{y_{i+1}}$.
17. $\uparrow$ Now suppose the minimum polynomial for $A \in \mathscr{L}(V, V)$ is

$$
p(\boldsymbol{\lambda})=\prod_{i=1}^{r}\left(\boldsymbol{\lambda}-\mu_{i}\right)^{m_{i}}
$$

Thus from what was shown above,

$$
V=\bigoplus_{i=1}^{r} \operatorname{ker}\left(\left(A-\mu_{i} I\right)^{m_{i}}\right) \equiv \bigoplus_{i=1}^{r} \operatorname{ker}\left(N_{i}^{m_{i}}\right)
$$

where $N_{i}$ is the restriction of $\left(A-\mu_{i} I\right)$ to $V_{i} \equiv \operatorname{ker}\left(\left(A-\mu_{i} I\right)^{m_{i}}\right)$. Explain why there are ordered bases $\beta_{1}, \cdots, \beta_{r}, \beta_{j}$ being a basis for $V_{j}$ such that with respect to this basis, the matrix of $N_{i}$ has the form

$$
\left(\begin{array}{cccc}
B_{1} & & & \\
& B_{2} & & \\
& & \ddots & \\
& & & B_{s_{i}}
\end{array}\right)
$$

each $B_{k}$ having ones down the super diagonal and zeros elsewhere. Now explain why each $V_{i}$ is $A$ invariant and the basis just described yields a matrix for $A$ which is of the form

$$
\left(\begin{array}{ccc}
J_{1} & & \\
& \ddots & \\
& & J_{r}
\end{array}\right)
$$

where

$$
J_{k}=\left(\begin{array}{lll}
J_{1}\left(\mu_{k}\right) & & \\
& \ddots & \\
& & J_{s_{k}}\left(\mu_{k}\right)
\end{array}\right)
$$

with the size of the diagonal blocks decreasing and $J_{m}\left(\mu_{k}\right)$ having ones down the super diagonal and $\mu_{k}$ down the diagonal. Hint: Explain why, for $I$ the identity on $V_{k}$ the matrix of $\mu_{k} I$ with respect to any basis is just the diagonal matrix having $\mu_{k}$ down the diagonal. Thus the matrix of $A$ restricted to $V_{k}$ relative to the basis $\beta_{k}$ will be of the desired form. Note that on $V_{k}, A=N_{k}+\mu_{k} I$. This yields the Jordan canonical form. Another argument based on rings and modules will be introduced in the following chapter to obtain both the rational and Jordan form.

## Chapter 7

## Canonical Forms

Linear algebra is really all about linear transformations and the fundamental question is whether a matrix comes from some linear transformation with respect to some basis. In other words, are two matrices really from the same linear transformation? As proved above, this happens if and only if the two are similar. Canonical forms allow one to answer this question. There are two main kinds of canonical form, the Jordan canonical form for the case where the minimum polynomial splits and the rational canonical form in the other case. Of the two, the Jordan canonical form is the one which is used the most in applied math. However, the other one is also pretty interesting. In what follows $V, W$ will denote vector spaces over the field of scalars $\mathbb{F}$.

### 7.1 Reduction to Diagonal Matrix

This is a really interesting result on diagonalization. It is an approach used in Jacobsen [26]. It concerns matrices whose entries are polynomials having coefficients in $\mathbb{F}$ and is an application of row operations and division of polynomials.

Recall the elementary matrices which involved doing a row operation to the identity matrix. The elementary matrices which involve switching two rows or adding a multiple of one row to another result in elementary matrices which are invertible. Similarly, these two column operations may be accomplished by multiplying on the right by an elementary matrix which involves adding a multiple of a column to another column or switching two columns. See Problem 40 on Page 99. It all works just as well if the multiple is an element of a commutative ring. In the theorem which follows, the entries will be polynomials $\delta(q)$ will denote the degree of the nonzero polynomial $q(x)$ which is undefined if $q=0$. Also $A_{i j}$ will be a polynomial. When we write $A B$ we mean the matrix whose $i j^{\text {th }}$ entry is just $\sum_{k} A_{i k} B_{k j}$ which may be a polynomial. The identity matrix is the same as usual. An inverse is also the same as before, $P P^{-1}=I$. Recall that if $\alpha, \beta \in \mathbb{F}[x]$, the polynomials with coefficients in $\mathbb{F}$, there exists $\kappa$ such that

$$
\alpha=\kappa \beta+\rho, \delta(\rho)<\delta(\beta) \text { or else } \rho=0
$$

Theorem 7.1.1 Let A be an $m \times n$ matrix whose entries are polynomials. Then there are invertible matrices $P, Q$ of the right size such that $P A Q=B$ where $B$ is a diagonal matrix.

Proof: If $A=0$ there is nothing to show. Just let $P, Q$ be appropriate identity matrices. Assume then that $A \neq 0$. Begin with $P$ and $Q$ appropriate sized identity matrices. Let $\delta\left(A_{i j}\right)$ be the smallest of the degrees of all entries of $A$ which are not zero. Now choosing a switch of columns and rows, we can modify $P, Q$ such that $B_{11}=A_{i j}$. Consider $B_{i 1}$, the first entry in the $i^{\text {th }}$ row. By the Euclidean algorithm,

$$
B_{i 1}=B_{11} q+r_{i 1}, \delta\left(r_{i 1}\right)<\delta\left(B_{11}\right)
$$

or else $r_{i 1}=0$. Take $-q$ times the first row of $B$ and add to the $i^{t h}$ row to place a $r_{i 1}$ in the $i 1$ position in place of $B_{i 1}$. This involves adjusting $P$ to get this new $B$. It is desired to get a 0 in the $i 1$ position which might have occurred if the $r_{i 1}$ had been 0 . Otherwise, out of all entries of the new matrix $B$ the $B_{r s}$ which has $\delta\left(B_{r s}\right)$ the smallest is in the $i^{t h}$ row and $\delta\left(B_{i s}\right) \leq \delta\left(r_{i 1}\right)$. Switch rows and columns till $B_{i s}$ is in the 11 position. Now repeat the argument just given, replacing the first entry of the $i^{\text {th }}$ row with a remainder $r^{\prime}$ where it
is either zero or $\boldsymbol{\delta}\left(r^{\prime}\right)<\boldsymbol{\delta}\left(B_{i s}\right)$. Continuing in this way, eventually the remainder $r$ must be zero because the process yields a strictly decreasing sequence of nonnegative integers, yielding a 0 in the first column, $i^{t h}$ row. Now do a similar process to the other rows of the resulting matrix. When this is done, do the same thing using column operations to eventually obtain

$$
P A Q=\left(\begin{array}{cc}
B_{11} & 0^{T}  \tag{*}\\
\mathbf{0} & \hat{B}
\end{array}\right)
$$

If $\hat{B}=0$ we are done. If not, do the same thing working with the rows of $\hat{B}$ and then the columns of $\hat{B}$ adjusting $P$ and $Q$ as the process continues, to obtain

$$
P A Q=\left(\begin{array}{ccc}
B_{11} & & 0 \\
& B_{22} & \\
0 & & \hat{B}
\end{array}\right)
$$

where $\hat{B}$ is now $(m-2) \times(n-2)$. Eventually, the result is a diagonal matrix.
Since adustments are constantly made to make the degree smaller, this matrix $B$ will consist of numbers from $\mathbb{F}$.

Definition 7.1.2 For $L \in \mathscr{L}(V, V)$, if $p(x) \in \mathbb{F}[x], p(L)$ will simply be the linear transformation which involves replacing $x$ with $L$ and the constant term a with aI. Denote by $D$ the polynomials $\mathbb{F}[x]$. For $p \in D, p v \equiv p(L) v$. For $v \in V, D v$ will consist of all vectors of the form $p v$. Thus $V$ is an Abelian group and if $p, q \in D$ and $v, w \in V$ it follows that

$$
\begin{aligned}
& p v \in V,(p+q) v=p v+q v \\
& 1 v=v, p(v+w)=p v+p w
\end{aligned}
$$

Since the "scalars" are coming from $D=\mathbb{F}[x]$ rather than a field, $V$ is called a module. As in the case of vector spaces, $W$ is called a submodule if it satisfies the above conditions and is a subset of $V$.

Definition 7.1.3 Suppose you have two modules, $V, W$ over $D=\mathbb{F}[x]$. A mapping $h: V \rightarrow$ $W$ is called a morphism if it does the following. For $\alpha, \beta \in D$ and $v, w$ in $V, W$ respectively,

$$
h(\alpha v+\beta w)=\alpha h(v)+\beta h(w)
$$

When a morphism $h$ is one to one, it is called a monomorphism and if the morphism $h$ is onto in addition to this then it is called an isomorphism.

We say that $W$ is a submodule of $V$ if it is a subset of $V$ and is a module over $D$. Thus, just as in the case of a subspace, $W$ is a submodule if and only if $0 \in W$ and whenever $\alpha, \beta \in D$ and $w, v \in W, \alpha w+\beta v \in W$.

We would say $h$ is linear if the coefficients were from the field $\mathbb{F}$.
Example 7.1.4 Let $h: V \rightarrow W$ be a morphism. Then $\operatorname{ker}(h)$ is a submodule of $V$.
This is clear because if $v, w \in \operatorname{ker}(h), h(\alpha v+\beta w)=\alpha h(v)+\beta h(w)=\alpha 0+\beta 0=0$. Thus $\alpha v+\beta w \in \operatorname{ker}(h)$.

### 7.2 Quotients

One can consider quotients of modules. This involves a set of equivalence classes as described below.

Definition 7.2.1 Let $A$ be a module over $D=\mathbb{F}[x]$ and let $B$ be a submodule. Then $A / B$ denotes sets of the form $a+B$ defined by $\{a+b: b \in B\}$, with the operations defined by

$$
\begin{aligned}
a+B+(\hat{a}+B) & \equiv a+\hat{a}+B \\
\lambda(a+B) & \equiv \lambda a+B
\end{aligned}
$$

To make the notation shorter, we can write $[a]$ instead of $a+B$. Thus the above definition says that $[a]+[\hat{a}]=[a+\hat{a}]$ and $\lambda[a] \equiv[\lambda a]$. We also have $[a]=[\hat{a}]$ if and only if $a-\hat{a} \in B$.

Lemma 7.2.2 $[a]=[\hat{a}]$ if and only if $a-\hat{a} \in B$.
Proof: From the definition, $[a]=[\hat{a}]$ if and only if whenever $b \in B$, it follows that there exists $\hat{b}$ such that $a+b=\hat{a}+\hat{b}$ if and only if $a-\hat{a}=b-\hat{b} \in B$.

The main result about quotients is in the following. It will be reminiscent of what was done with the field $\mathbb{Z}_{p}$ for a $p$ a prime.

Proposition 7.2.3 $A / B$ is a module over $D$.
Proof: I need to verify that the operations are well defined. If $[a]=[\hat{a}]$, and if $[b]=[\hat{b}]$, is $[a+\hat{a}]=[b+\hat{b}]$ ? Yes because $(b+\hat{b})-(a+\hat{a})=(b-a)+(\hat{b}-\hat{a}) \in B$. Similarly if $[a]=[\hat{a}]$ then $\lambda a-\lambda \hat{a}=\lambda(a-\hat{a}) \in B$. Thus $\lambda[a]=\lambda[\hat{a}]$ and so the operations are well defined. As to their algebraic properties, these follow directly from the fact that $A$ is a module.

Definition 7.2.4 Let $h$ be a morphism, $h: V \rightarrow W$ for $V, W$ modules. Let $\hat{h}: V / \operatorname{ker}(h) \rightarrow W$ be defined by $\hat{h}([v]) \equiv h v$. Then $\hat{h}$ is one to one and a morphism.

Lemma 7.2.5 In Definition 7.2.4 $\hat{h}$ is well defined, a morphism and maps onto $h(V)$ which is a submodule of $W$.

Proof: It is clear that $\hat{h}$ is a morphism if it is well defined. It suffices to show that if $[v]=[\hat{v}]$ so $v-\hat{v}=0$, then $h v=h \hat{v}$. But $h(v)-h(\hat{v})=h(v-\hat{v})=0$, so these are equal. Thus $\hat{h}$ is well defined. $\hat{h}$ is one to one because $\hat{h}[v]=0$ means $h(v)=0$ and so $v \in \operatorname{ker}(h)$ so $[v]=0 . \hat{h}$ is obviously onto $h(V)$.

Example 7.2.6 Let $v \in V$ a module. Then $D v$ defined as $\{\alpha v: \alpha \in D\}$ is a module. This is called a cyclical module. It is a submodule of $V$.

Definition 7.2.7 A module $V$ is said to have the ascending chain condition and is called a Noetherian module if whenever there is a chain of submodules $V_{1} \subseteq V_{2} \subseteq V_{3} \cdots$, these are eventually constant. That is, for large enough $n, V_{n}=V_{m}$ whenever $m \geq n$.

Lemma 7.2.8 Suppose $A, B$ are modules and $\tau: A \rightarrow B$ is a morphism. Then if $C$ is $a$ sub module of $B$, it follows that $\tau^{-1}(C)$ is a submodule of $A$ which contains $\operatorname{ker}(\tau)$.

Proof: Let $a, b \in \tau^{-1}(C)$ and let $\alpha, \beta \in D$. Then $\tau(\alpha a+\beta b)=\alpha \tau(a)+\beta \tau(b) \in C$ and so $\tau^{-1}(C)$ is indeed a submodule of $A$. If $v \in \operatorname{ker}(\tau)$, then $\tau(v)=0 \in C$ so $v \in \tau^{-1}(C)$.

When is a module Noetherian?
Lemma 7.2.9 Suppose $A, C$ are Noetherian modules and $A \xrightarrow{\theta} B \xrightarrow{\eta} C$ where $\theta, \eta$ are morphisms. Suppose also that $\theta$ is one to one, and $\operatorname{ker}(\eta)=\theta(A)$. Then if $A, C$ are Noetherian, so is B.

Proof: Suppose $B_{n}$ is an ascending chain of sub modules in $B$. Then $\eta\left(B_{n}\right)$ is eventually constant because these are submodules of $C$. Also $\theta^{-1}\left(B_{n}\right)$ is eventually constant for the same reason. Let these be constant for all $n \geq m$. If $n>m$, let $b \in B_{n}$. Then there is $\hat{b} \in B_{m}$ such that $\eta(\hat{b})=\eta(b)$ and so $b-\hat{b} \in \operatorname{ker}(\eta)=\operatorname{Im}(\theta)$ so there is $a \in \theta^{-1}\left(B_{n}\right)=\theta^{-1}\left(B_{m}\right)$ with $b-\hat{b}=\theta a \in B_{m}$ showing that $b \in B_{m}$ also.

This sequence in which $\operatorname{ker}(\eta)=\operatorname{Im}(\theta)$ is called a short exact sequence.
Proposition 7.2.10 For $n \in \mathbb{N}, D^{n}$ is a Noetherian module over $D$. Here the usual conventions are being followed. $\gamma\left(\alpha_{1}, \ldots, \alpha_{n}\right)=\left(\gamma \alpha_{1}, \ldots, \gamma \alpha_{n}\right)$. If $K$ is a submodule of $D^{n}$ then there are vectors $\boldsymbol{z}_{k}$ such that $K=D \boldsymbol{z}_{1}+D \boldsymbol{z}_{2}+\cdots+D \boldsymbol{z}_{n}$.

Proof: First note that $D$ is a Noetherian module for $D$. The submodules of $D$ are just the ideals. If you have an increasing chain $I_{1} \subseteq I_{2} \subseteq \cdots$ of submodules, then you could consider $I \equiv \cup_{k} I_{k}$ and this would also be a submodule because if $\alpha_{1}, \beta_{1} \in I$ and $\alpha, \beta \in D$, then for large enough $k$, both $\alpha_{1}, \beta_{1} \in I_{k}$ and so $\alpha \alpha_{1}+\beta \beta_{1} \in I_{k} \subseteq I$. Let $\sigma \in I$ have smallest degree. Then for $\alpha \in I$, it follows that for all $k$ large enough, $\alpha, \sigma \in I_{k}$ and $\alpha=\beta \sigma+\rho$ where either $\rho$ has smaller degree than $\sigma$ which is impossible because $\rho=\alpha-\beta \sigma \in I_{k}$ or else $\rho=0$. Thus for all $k$ large enough, every $\alpha \in I_{k}$ is a multiple of $\sigma$ and so $I_{k}$ equals $D \sigma$ for all $k$ large enough.

Consider $D^{n-1} \xrightarrow{\theta} D^{n} \xrightarrow{\eta} D$ where $\theta(\boldsymbol{b}) \equiv(\boldsymbol{b}, 0), \eta(\boldsymbol{c}) \equiv c_{n}$ where

$$
\boldsymbol{c}=\left(c_{1}, \cdots, c_{n}\right)
$$

It is clear that $\theta$ is one to one. Also $\operatorname{ker}(\eta)=\left\{(\boldsymbol{b}, 0): \boldsymbol{b} \in D^{n-1}\right\}=\theta\left(D^{n-1}\right)$. Now use Lemma 7.2.9 for $n=2$ to find $D^{2}$ is Noetherian and then for $n=3$ to find that $D^{3}$ is Noetherian and so forth. Thus $D^{n}$ is Noetherian.

Now let $K$ be a submodule of $D^{n}$. I need to show $K=D \boldsymbol{z}_{1}+D \boldsymbol{z}_{2}+\cdots+D \boldsymbol{z}_{n}$ for suitable $\boldsymbol{z}_{k}$. Pick $\boldsymbol{z}_{1} \in D^{n}$. If $K=D \boldsymbol{z}_{1}$ stop. Otherwise consider $\boldsymbol{z}_{2} \notin D \boldsymbol{z}_{1}$. If $D \boldsymbol{z}_{1}+$ $D \boldsymbol{z}_{2}=K$, stop otherwise continue. Now these $D \boldsymbol{z}_{1}+D \boldsymbol{z}_{2}+\cdots+D \boldsymbol{z}_{k}, k=1,2, \ldots$ form an increasing chain of submodules of $K$ and so it must eventually be constant at which point you have what is desired.

Definition 7.2.11 Also recall the concept of direct sums of subspaces. $V=\bigoplus_{k=1}^{n} V_{k}$ means $V=\sum_{k=1}^{n} V_{k}$ and if $0=\sum_{k=1}^{n} v_{k}$, then each $v_{k}=0$. The direct sum of modules has the same definition. Also, if $K=D \boldsymbol{z}_{1}+D \boldsymbol{z}_{2}+\cdots+D \boldsymbol{z}_{n}$ we say that $K=\operatorname{span}\left(\boldsymbol{z}_{1}, \ldots, \boldsymbol{z}_{n}\right)$.

Let $e_{k} \in D^{n}$ denote the usual thing, a column of entries of $D$ with a 1 in the $k^{t h}$ slot down from the top. Then $\left\{e_{1}, \ldots, e_{n}\right\}$ is linearly independent in the usual way meaning that if $\sum_{k=1}^{n} \alpha_{k} \boldsymbol{e}_{k}=\mathbf{0}$, then each $\alpha_{k}=0 \in D$. Also, if

$$
\left(\begin{array}{lll}
e_{1}^{\prime} & \cdots & e_{n}^{\prime}
\end{array}\right)=\left(\begin{array}{lll}
e_{1} & \cdots & e_{n}
\end{array}\right) P \text {, meaning } e_{k}^{\prime}=\sum_{j} e_{j} P_{k j},
$$

where $P$ is an invertible $n \times n$ matrix of entries of $D$, then $\left\{\boldsymbol{e}_{1}^{\prime}, \ldots, \boldsymbol{e}_{n}^{\prime}\right\}$ is linearly independent in the same way. Indeed, if $\mathbf{0}=\sum_{k} \sigma_{k} e_{k}^{\prime}$ then

$$
\mathbf{0}=\sum_{k} \sigma_{k} e_{k}^{\prime}=\sum_{k} \sigma_{k} \sum_{j} e_{j} P_{k j}=\sum_{j}\left(\sum_{k} \sigma_{k} P_{k j}\right) e_{j}
$$

and so $\sum_{k} \sigma_{k} P_{k j}=0$ for each $j$. Since $P$ is invertible, this gives each $\sigma_{k}=0$. This observation is useful in the following proof.

### 7.3 Cyclic Decomposition

Theorem 7.3.1 Let $V$ be a finite dimensional vector space and let $L \in \mathscr{L}(V, V)$. Then there are vectors $m_{1}, \ldots, m_{p}$ such that $V=D m_{1} \oplus \cdots \oplus D m_{p}$ where $D=\mathbb{F}[x]$ and for $p \in$ $D, v \in V, p v \equiv p(L)(v)$.

Proof: Let $V=D b_{1}+\cdots+D b_{n}$ for $b_{k} \in V$. This is possible because $V$ is finite dimensional. In fact, we could pick a basis and let this be a direct sum of $\mathbb{F} b_{k}$. However, the point here is that there are more vectors in $D b_{k}$ than in $\mathbb{F} b_{k}$. Thus $p$ will likely be smaller than $n$ if the dimension of $V$ is $n$. Now define $\eta: D^{n} \rightarrow V$ by

$$
\eta\left(\sum_{i=1}^{n} \sigma_{i} e_{i}\right) \equiv \eta(\boldsymbol{\sigma}) \equiv \sum_{i=1}^{n} \sigma_{i} b_{i}, \sigma_{i} b_{i} \equiv \sigma_{i}(L) b_{i}
$$

Here $e_{i}$ has 1 in the $i^{t h}$ position and 0 elsewhere and $\sigma_{i}$ will denote a polynomial. Then it follows that $\eta$ is a morphism. For $\boldsymbol{v}, \hat{\boldsymbol{v}} \in D^{n}$ and $\sigma \in D$,

$$
\begin{equation*}
\eta(\boldsymbol{v}+\hat{\boldsymbol{v}})=\eta(\boldsymbol{v})+\eta(\hat{\boldsymbol{v}}), \eta(\sigma \boldsymbol{v})=\sigma \eta(\boldsymbol{v}) \tag{7.1}
\end{equation*}
$$

Let $K \equiv \operatorname{ker}(\eta)$.
Now use Lemma 7.2.5 to define the one to one and onto morphism $\hat{\eta}: D^{n} / K \rightarrow V$ as $\hat{\eta}([\boldsymbol{v}]) \equiv \eta(\boldsymbol{v})$.

It was shown above in Proposition 7.2.10 that $D^{n}$ is Noetherian and so the submodule $K=\operatorname{ker}(\eta)$ is $\operatorname{span}\left(\boldsymbol{z}_{1}, \cdots, \boldsymbol{z}_{m}\right)$ for some $m$. Out of all such spans, let $m$ be as small as possible. Let the matrix $A$ be defined by $\boldsymbol{z}_{k}=\sum_{j} A_{j k} \boldsymbol{e}_{j}$ written as

$$
\left(\begin{array}{lll}
z_{1} & \cdots & z_{m}
\end{array}\right)=\left(\begin{array}{lll}
e_{1} & \cdots & e_{n}
\end{array}\right) A_{n \times m}
$$

By Theorem 7.1.1, there are invertible $P, Q$ such that $P A Q=B$ for $B_{n \times m}$ a matrix 0 off the main diagonal, and so $A=P^{-1} B Q^{-1}$ where $B$ is $n \times m$. Therefore,

$$
\left(\begin{array}{ccc}
z_{1}^{\prime} & \cdots & z_{m}^{\prime}
\end{array}\right) \equiv\left(\begin{array}{lll}
z_{1} & \cdots & z_{m}
\end{array}\right) Q=\left(\begin{array}{lll}
e_{1} & \cdots & e_{n} \tag{7.2}
\end{array}\right) P^{-1} B
$$

Thus span $\left(\boldsymbol{z}_{1}^{\prime}, \ldots, \boldsymbol{z}_{m}^{\prime}\right)$ is the same as the span of the $\boldsymbol{z}_{k}$ which is $K$. Then if $m \leq n$ the above gives

$$
\equiv\left(\begin{array}{lll}
e_{1}^{\prime} & \cdots & e_{n}^{\prime}
\end{array}\right) B=\left(\begin{array}{lll}
\delta_{1} e_{1}^{\prime} & \cdots & \delta_{m} e_{m}^{\prime} \tag{7.3}
\end{array}\right)
$$

where $\delta_{k}=\hat{B}_{k k}, k \leq m$. If $m>n$ then some of the $\boldsymbol{z}_{k}^{\prime}$ equals $\mathbf{0}$ and so $m$ would not be as small as possible.

It follows that $\boldsymbol{z}_{k}^{\prime}=\boldsymbol{\delta}_{k} \boldsymbol{e}_{k}^{\prime}, k \leq m$ and no $\boldsymbol{\delta}_{k} \boldsymbol{e}_{k}^{\prime}$ is $\mathbf{0}$ because you could then delete some $\boldsymbol{z}_{k}^{\prime}$ which are $\mathbf{0}$ and still have a spanning set for $K$, but $m$ was as small as possible. Thus $\delta_{k} e_{k}^{\prime} \neq 0$.

If $\boldsymbol{e}_{k}^{\prime}=\sum_{i=1}^{m} \alpha_{i} z_{i}^{\prime} \in K$, then $\boldsymbol{e}_{k}^{\prime}=\sum_{i=1}^{m} \alpha_{i} \delta_{i} e_{i}^{\prime}$. This cannot happen if $k>m$ (see 7.2). Thanks to linear independence of the $e_{k}^{\prime}$, if $k \leq m$, then, each $\alpha_{i} \delta_{i}=0$ for $i \neq k$ and $\alpha_{k} \delta_{k}=1$ so $\delta_{k}$ and $\alpha_{k}$ are in $\mathbb{F}$, are nonzero and $D e_{k}^{\prime}=D z_{k}^{\prime}$ since $\delta_{k} e_{k}^{\prime}=\boldsymbol{z}_{k}^{\prime}$ and $\delta_{k}$ is invertible. The $e_{i}^{\prime}$ are divided into two classes, those which are an invertible multiple of some $z_{i}^{\prime}$ and those which are not. Those $e_{k}^{\prime}$ in the latter class are not in $K$. Let $S$ be those $i$ for which $e_{i}^{\prime}$ is NOT in $K$ so $S^{C}$ are those $i$ for which $e_{i}^{\prime}$ are in $K$.

Thus $D^{n} / K=\sum_{k \in S}\left(D\left(e_{k}^{\prime}+K\right)\right)$. In fact this is a direct sum. If $\sum_{k \in S}\left(\alpha_{k} e_{k}^{\prime}+c_{k}\right) \in K$, then $\sum_{k \in S} \alpha_{k} e_{k}^{\prime}=w=\sum_{k \notin S} \beta_{k} e_{k}^{\prime} \in K$ so $\alpha_{k}, \beta_{k}$ are all 0 by linear independence of the $e_{k}^{\prime}$. Write the direct sum in the form $\bigoplus_{k \in S}\left(D e_{k}^{\prime}+K\right)$. Since $\hat{\eta}$ is an isomorphism, it follows $M=\bigoplus_{k \in S} D \eta\left(e_{k}^{\prime}\right)$. Let $m_{k}=\eta\left(e_{k}^{\prime}\right)$.

Now note that $D m$ is a submodule of $V$ which implies that $\alpha D m \subseteq D m$. In terms of linear transformations, this says that $p(L) m \in D m$ for every $p(x) \in \mathbb{F}[x]$. Also $D m$ is a subspace of $V$. Thus it is an "invariant subspace". $p(L): D m \rightarrow D m$ for any $p(x) \in \mathbb{F}[x]$. In particular, this holds for $p(x)=x$ and $D m$ is an invariant subspace with respect to $L$. This has exhibited $V$ as a direct sum of invariant subspaces, one for each $m_{k}$.

Corollary 7.3.2 In the context of Theorem 7.3.1 where $V=D m_{1} \oplus \cdots \oplus D m_{p}$, there exists an integer, $l_{k}$ such that

$$
D m_{k}=\operatorname{span}\left(m_{k}, L m_{k}, L^{2} m_{k}, \cdots, L^{l_{k}-1} m_{k}\right)
$$

Also $\left\{m_{k}, L m_{k}, L^{2} m_{k}, \cdots, L^{l_{k}-1} m_{k}\right\}$ is a linearly independent set. Recall that $p \in D, v \in$ $V, p v \equiv p(L)(v)$ for $L$ given in $\mathscr{L}(V, V)$.

Proof: As just noted, $D m_{k}$ is a subspace of a finite dimensional vector space $V$ and $L: D m_{k} \rightarrow D m_{k}$ so there is a minimum polynomial for this restricted $L$ to $D m_{k}, \sigma_{k}(x)$ for which $\sigma_{k}(L): D m_{k} \rightarrow 0$. Say the degree of this minimum polynomial is $l_{k}$. Then from the division algorithm for polynomials, a generic element of $D m_{k}$ is of the form $\alpha(L) m_{k}$ where $\alpha(x)$ has degree less than $l_{k}$. However, such examples are all elements of $\operatorname{span}\left(m_{k}, L m_{k}, L^{2} m_{k}, \cdots, L^{l_{k}-1} m_{k}\right)$. Why is $\left\{m_{k}, L m_{k}, L^{2} m_{k}, \cdots, L^{l_{k}-1} m_{k}\right\}$ independent? If not, there exist $a_{j} \in \mathbb{F}$ such that

$$
\sum_{j=0}^{l_{k}-1} a_{j} L^{j} m_{k}=\alpha(L) m_{k}=0
$$

where the degree of $\alpha(x)$ is smaller than $l_{k}$. However, if $\beta \in \mathbb{F}[x]$, then $\alpha\left(\beta m_{k}\right)=$ $\beta\left(\alpha m_{k}\right)=\beta 0=0$ and so $\sigma$ was not the minimum polynomial after all. A scalar multiple of $\alpha(x)$ having smaller degree would be the minimum polynomial.
$\left\{m_{k}, L m_{k}, L^{2} m_{k}, \cdots, L^{l_{k}-1} m_{k}\right\}$ is called a cyclic set.

### 7.4 A Direct Sum Decomposition

This approach is in [26]. The end result is essentially Theorem 6.1 .10 presented more quickly. For $p$ an irreducible polynomial and $L \in \mathscr{L}(V, V)$ for $V$ a finite dimensional
vector space over $\mathbb{F}$. As usual, for $\alpha \in \mathbb{F}[x] \equiv D, \alpha m \equiv \alpha(L) m$. For $p$ a monic irreducible polynomial,

$$
V_{p} \equiv\left\{m \in V: p^{k} m=0 \text { for some } k \in \mathbb{N}\right\}
$$

That is, eventually $p^{k} m=0$. It might be possible that $k$ could change for different $m \in V$. Note that if $p$ is invertible, then $V_{p}=0$ because $x^{p} m=0$ if and only if $m=\left(x^{-1}\right)^{p} 0=0$, so nothing is lost from considering only irreducible non constant polynomials. It is obvious that $V_{p}$ is a subgroup of the module $V$ and is itself a module. Indeed, if $m \in V_{p}$ so that $p^{k} m=$ 0 for some $k$, then $0=\alpha p^{k} m=p^{k} \alpha m=0$ also and so $\alpha m \in V$. If $m, \hat{m} \in V_{p}$, then letting $k_{m}, k_{\hat{m}}$ be the exponents for $m, \hat{m}$, let $k \geq \max \left(k_{m}, k_{\hat{m}}\right)$ and $p^{k}(m+\hat{m})=p^{k} m+p^{k} \hat{m}=0$ so the sum $m+\hat{m}$ is in $V_{p}$ if $m, \hat{m}$ are.

Proposition 7.4.1 Let $p_{1}, \cdots, p_{n}$ be monic irreducible nonconstant polynomials. Let $V$ be a finite dimensional vector space. Then

$$
\left(V_{p_{1}}+\cdots+V_{p_{j-1}}+V_{p_{j+1}}+\cdots+V_{p_{n}}\right) \cap V_{p_{j}}=0
$$

and so $\sum_{i} V_{p_{i}}=\bigoplus_{i} V_{p_{i}}$.
Proof: This follows from the observation that $\prod_{i \neq j} p_{i}^{k_{i}}$ and $p_{j}^{k_{j}}$ are relatively prime. If $q$ is monic and divides the second, then it is of the form $p_{j}^{m_{j}}, m_{j} \leq k_{j}$. If $q$ divides the first, then $q$ is $\prod_{i \neq j} p_{i}^{m_{i}}, m_{i} \leq k_{i}$. Thus $\prod_{i \neq j} p_{i}^{m_{i}}=p_{j}^{m_{j}}$ contradicting Theorem 1.13.9 about uniqueness of factorization. Since the irreducible polynomials are distinct, we must have all $m_{j}, m_{i}$ equal to 0 and $q=1$ so these two, $\prod_{i \neq j} p_{i}^{k_{i}}$ and $p_{j}^{k_{j}}$ are relatively prime as claimed. If $m \in\left(V_{p_{1}}+\cdots+V_{p_{j-1}}+V_{p_{j+1}}+\cdots+V_{p_{n}}\right) \cap V_{p_{j}}$, then $m=\sum_{i \neq j} m_{i}$ and so there exist $k_{i}, k_{j}$ such that $p_{i}^{k_{i}} m_{i}=0$ and $p_{j}^{k_{j}} m=0$. Since $\prod_{i \neq j} p_{i}^{k_{i}}$ and $p_{j}^{k_{j}}$ are relatively prime, there exist $\sigma, \tau$ such that

$$
\begin{equation*}
1=\sigma \prod_{i \neq j} p_{i}^{k_{i}}+\tau p_{j}^{k_{j}} \tag{*}
\end{equation*}
$$

Then do both sides of $*$ to $m$.

$$
m=\left(\sigma \prod_{i \neq j} p_{i}^{k_{i}}\right)\left(\sum_{i \neq j}^{=m} m_{i}\right)+\tau p_{j}^{k_{j}} m=0
$$

This yields $m=0$ and verifies the conclusion of the proposition.
It follows from Lemma 6.0.2 that if $m_{i} \in V_{p_{i}}$, and if $\sum_{i} m_{i}=0$, then each $m_{i}=0$ so $\sum_{i} V_{p_{i}}=\bigoplus_{i} V_{p_{i}}$.

Lemma 7.4.2 Let $V$ be a vector space over a field $\mathbb{F}$ and let $L \in \mathscr{L}(V, V)$ have minimum polynomial $\alpha(x)=\prod_{i=1}^{n} p_{i}^{k_{i}}(x)$ in which the $p_{i}$ are distinct irreducible monic polynomials. Let $D \equiv \mathbb{F}[x]$ and for $\alpha \in D, \alpha(m) \equiv \alpha(L)(m)$. Let

$$
\begin{equation*}
\hat{V}_{p_{i}} \equiv\left\{m \in V: p_{i}^{k_{i}} m=0\right\} \subseteq V_{p_{i}} \tag{7.4}
\end{equation*}
$$

(Note that here $k_{i}$ is fixed. ) Then

$$
\begin{equation*}
D m \subseteq \hat{V}_{p_{1}} \oplus \cdots \oplus \hat{V}_{p_{n}} \subseteq V_{p_{1}} \oplus \cdots \oplus V_{p_{n}} \tag{7.5}
\end{equation*}
$$

Proof: First note that from Proposition 7.4.1 $V_{p_{1}}+\cdots+V_{p_{n}}=V_{p_{1}} \oplus \cdots \oplus V_{p_{n}}$. It follows from 7.4 that $\hat{V}_{p_{1}}+\cdots+\hat{V}_{p_{n}}=\hat{V}_{p_{1}} \oplus \cdots \oplus \hat{V}_{p_{n}}$ and that the second subset in 7.5 holds.

Consider the first inclusion. This will be shown by establishing the following claim by induction.

Claim: For $m \in V$, if $\prod_{i=1}^{s} p_{i}^{k_{i}} m=0$ then $D m \subseteq \hat{V}_{p_{1}} \oplus \cdots \oplus \hat{V}_{p_{s}}$.
Proof: First suppose $s=1$. Then for $\beta \in D, \beta m \subseteq \hat{V}_{p_{1}}$ because $p^{k_{1}} \beta m=\beta p^{k_{1}} m=$ $\beta(0)=0$.

Suppose the claim is true for some $s-1 \geq 1$ and let $\prod_{i=1}^{s} p_{i}^{k_{i}}(m)=0$. Since $p_{n}^{k_{s}}(x)$ and $\prod_{i=1}^{s-1} p_{i}^{k_{i}}$ are relatively prime, there exist polynomials $\sigma, \tau$ such that $1=\sigma p_{s}^{k_{s}}+\tau \prod_{i=1}^{s-1} p_{i}^{k_{i}}$.

$$
\begin{equation*}
m=\sigma p_{s}^{k_{s}} m+\overbrace{\tau \prod_{i=1}^{s-1} p_{i}^{k_{i} m}}^{\in \hat{V}_{p_{s}}} \tag{7.6}
\end{equation*}
$$

Then by assumption, $\prod_{i=1}^{s-1} p_{i}^{k_{i}}\left(\sigma p_{s}^{k_{s}} m\right)=\sigma \prod_{i=1}^{s} p_{i}^{k_{i}}(m)=0$ and so, by induction,

$$
\begin{equation*}
D \sigma p_{s}^{k_{s}} m \in \hat{V}_{p_{1}} \oplus \cdots \oplus \hat{V}_{p_{s-1}} \tag{7.7}
\end{equation*}
$$

and since $\prod_{i=1}^{s-1} p_{i}^{k_{i}} m \in \hat{V}_{p_{s}}$ a repeat of the first part of the argument in which there is only on space in the direct sum shows that $D \prod_{i=1}^{s-1} p_{i}^{k_{i}} m \in \hat{V}_{p_{s}}$ so from 7.6,

$$
\begin{equation*}
D m \subseteq D \sigma p_{s}^{k_{s}} m+D \prod_{i=1}^{s-1} p_{i}^{k_{i}} m \subseteq \hat{V}_{p_{1}} \oplus \cdots \oplus \hat{V}_{p_{s-1}}+\hat{V}_{p_{s}}=\bigoplus_{i=1}^{s} \hat{V}_{p_{i}} \tag{7.8}
\end{equation*}
$$

Now the result follows from letting $s=n$ and the observation that $\prod_{i=1}^{n} p_{i}^{k_{i}}(x)$ is the minimum polynomial and so one can apply the above claim to any $m \in V$.

The following is the main result.
Theorem 7.4.3 Let $V$ be a finite dimensional vector space over the field $\mathbb{F}$ and let $\alpha(x)=$ $\prod_{i=1}^{n} p_{i}^{k_{i}}(x)$ be the minimum polynomial of $L \in \mathscr{L}(V, V)$ where each $p_{i}$ is irreducible and monic. Let $D \equiv \mathbb{F}[x]$ and $\alpha m \equiv \alpha(L)(m)$. Then

$$
\hat{V}_{p_{i}} \equiv\left\{m \in V: p_{i}^{k_{i}} m=0\right\}=\left\{m \in V: p_{i}^{k} m=0 \text { for some } k\right\} \equiv V_{p_{i}}
$$

and $V=V_{p_{1}} \oplus \cdots \oplus V_{p_{n}}$.
Proof: Let a basis be $\left(m_{1}, \ldots, m_{p}\right)$. Then from Lemma 7.4.2,

$$
\begin{equation*}
V=\mathbb{F} m_{1} \oplus \cdots \oplus \mathbb{F} m_{p} \subseteq D m_{1}+\cdots+D m_{p} \subseteq V_{p_{1}} \oplus \cdots \oplus V_{p_{n}} \subseteq V \tag{7.9}
\end{equation*}
$$

### 7.5 Uniqueness

The following discussion follows [26]. From Theorem 7.3.1 if $V$ is a finite dimensional vector space there exist vectors $m_{k}$ such that $V=D m_{1} \oplus \cdots \oplus D m_{p}$. From Theorem 6.1.10

$$
V=\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\phi_{p}(L)^{k_{p}}\right)
$$

where the minimum polynomial for $L \in \mathscr{L}(V, V)$ is $\prod_{j=1}^{p} \phi_{j}(x)^{k_{j}}$ where each $\phi_{j}(x)$ is monic and irreducible. I want to understand the cyclic decomposition for $\operatorname{ker}\left(\phi(L)^{q}\right)$ where $\phi(x)$ is monic and irreducible. First note that $L: \operatorname{ker}\left(\phi(L)^{q}\right) \rightarrow \operatorname{ker}\left(\phi(L)^{q}\right)$ so $\operatorname{ker}\left(\phi(L)^{q}\right)$ is a finite dimensional vector space with respect to the restriction of $L$ to $\operatorname{ker}\left(\phi(L)^{q}\right)$. Also, there is a minimum polynomial $\sigma$ for the restriction of $L$ to $\operatorname{ker}\left(\phi(L)^{q}\right)$. Since $\sigma$ divides $\phi(L)^{q}, \sigma$ must be of the form $\phi(L)^{l}, l \leq q$.

Letting $M=\operatorname{ker}\left(\phi(L)^{q}\right)$, where $\phi(x)^{q}$ is the minimum polynomial for $L$ restricted to $M$, Theorem 7.3.1 implies

$$
\begin{equation*}
M=D v_{1} \oplus \cdots \oplus D v_{s}, \text { no } v_{j}=0 \tag{7.10}
\end{equation*}
$$

First consider whether $L: D v_{k} \rightarrow D v_{k}$. If $p \in D, L\left(p v_{k}\right) \equiv L p(L) v_{k} \in D v_{k}$ so this is clearly true. Let the minimum polynomial for $L$ restricted to $D v_{k}$ be $\phi^{l_{k}}$. Then it follows $l_{k} \leq q$. Note that if $\phi^{l} v_{k}=0$, then so does $\phi^{l+1} v_{k}=0$.

I want to show that these $l_{k}$ are unique where $\phi^{l_{k}} v_{k}=0$ and $\phi^{l_{k}}(x)$ is the minimum polynomial of $L$ restricted to $D v_{k}$. Let these $D v_{k}$ be numbered such that $l_{1} \leq l_{2} \leq \cdots \leq l_{s} \leq$ $q$. Then $l_{s}=q$ since if not, the minimum polynomial would equal to $\phi^{l_{s}}$ instead of $\phi^{q}$. By Lemma 3.4.8 on Page $65, D / D \phi \equiv \hat{D}$ is a field. In that lemma, $(p(x))$ was written instead of $D p$.

In addition, we have $M \supseteq \phi M \supseteq \phi^{2} M \supseteq \cdots$. These are each submodules of $M$.
Lemma 7.5.1 $\phi^{k} M / \phi^{k+1} M$ is a vector space over $\hat{D}$ if $[\alpha]\left(m+\phi^{k+1} M\right) \equiv \alpha m+\phi^{k+1} M$.
Proof: This is clear if it is shown that the scalar multiplication by the elements of the field is well defined so let $[\alpha]=[\hat{\alpha}]$. I need to verify that $\alpha m+\phi^{k+1} M=\hat{\alpha} \hat{m}+\phi^{k+1} M$ This is so exactly when $\alpha m-\hat{\alpha} \hat{m} \in \phi^{k+1} M$. However, $\alpha-\hat{\alpha} \in D \phi$

$$
\begin{aligned}
\alpha m-\hat{\alpha} \hat{m} & =\alpha m-\alpha \hat{m}+\alpha \hat{m}-\hat{\alpha} \hat{m} \\
& =\alpha(m-\hat{m})+(\alpha-\hat{\alpha}) \hat{m} \in \phi^{k+1} M+D \phi \hat{m}
\end{aligned}
$$

Now $\hat{m} \in \phi^{k} M$ and so the last term is also in $\phi^{k+1} M$. The rest follows from the description of quotient spaces given earlier.

If $k \geq l_{s}=q$, then $\phi^{k} M=0$ and so for such $k, \phi^{k} M / \phi^{k+1} M=0$. The other case is that $k<l_{s}=q$. Say $k \in\left[l_{j}, l_{j+1}\right)$. Then

$$
\phi^{k} M=D \phi^{k} v_{j+1}+D \phi^{k} v_{j+2}+\cdots+D \phi^{k} v_{s}
$$

This is because if $k \geq l_{j}$, then for $i \leq j, \phi^{k} v_{i}=0$ and so all that survives on multiplication by $\phi^{k}$ is the above sum. Is

$$
\left\{\phi^{k} v_{j+1}+\phi^{k+1} M, \phi^{k} v_{j+2}+\phi^{k+1} M, \cdots, \phi^{k} v_{s}+\phi^{k+1} M\right\}
$$

linearly independent over $\hat{D}$ ? Suppose

$$
\sum_{r=j+1}^{s}\left[\alpha_{r}\right]\left(\phi^{k} v_{r}+\phi^{k+1} M\right)=0
$$

Then $\sum_{r=j+1}^{s} \alpha_{r} \phi^{k} v_{r}+\phi^{k+1} M=0$. This requires $\sum_{r=j+1}^{s} \alpha_{r} \phi^{k} v_{r}=\phi^{k+1} m$ for some $m \in M$. However, since $M$ is a sum, as in 7.10 , there is $\beta_{r}$ such that

$$
\sum_{r=j+1}^{s} \alpha_{r} \phi^{k} v_{r}=\phi^{k+1} m=\sum_{r=j+1}^{s} \phi^{k+1} \beta_{r} v_{r}
$$

Thus, since $M$ is a direct sum, for each $r, \phi^{k} \alpha_{r}=\phi^{k+1} \beta_{r}$ and so $\alpha_{r}$ is a multiple of $\phi$. Hence $\left[\alpha_{r}\right]=0$ for each $r$ and this is indeed a basis for $\phi^{k} M / \phi^{k+1} M$ over $\hat{D}$. It follows that the dimension of $\phi^{k} M / \phi^{k+1} M$ over $\hat{D}$ is $s-j$ where $k \in\left[l_{j}, l_{j+1}\right)$, the number of $v_{j+1}$ for $l_{j+1}>k$.

Suppose that $M=D w_{1} \oplus \cdots \oplus D w_{t}$ such that the minimum polynomial of $L$ restricted to $D w_{j}$ is $\phi^{m_{j}}(x)$. The question is whether $s=t$ and $m_{j}=l_{j}$. It was just shown that for $k$ a positive integer, the dimension of $\phi^{k} M / \phi^{k+1} M$ is the number of $v_{j}$ for $l_{j}>k$. Similarly it is the number of $w_{j}$ for $m_{j}>k$ and this must be the same number because $\phi^{k} M / \phi^{k+1} M$ does not depend on the $v_{j}$ or the $w_{r}$. Any two bases have the same number of vectors. In other words, for each $k$ there are the same number of $m_{j}$ larger than $k$ as there are $l_{j}$ larger than $k$. Hence $s=t$. Also the $l_{j}$ coincide with the $m_{j}$ in addition to having the same number of them. To see the last claim, suppose not. Then consider the first $i$ such that $l_{i} \neq m_{i}$. Let $k$ be the smaller of the two to contradict that there are the same number of $l_{j}$ and $m_{j}$ larger than $k$. Say $k=l_{i}$. There are $r+1$ of the $l_{j}$ larger than $k-1$ and $r$ of the $m_{j}$ larger than $k-1$.

Theorem 7.5.2 Suppose $V$ is a finite dimensional vector space and $L \in \mathscr{L}(V, V)$ with minimum polynomial $\phi(x)^{q}$ where $\phi$ is irreducible. Then

$$
V=D v_{1} \oplus \cdots \oplus D v_{s}, \text { no } v_{j}=0
$$

It follows that the restriction of $L$ to $D v_{j}$ is $\phi^{l_{j}}$ for some $l_{j} \leq q$. If the direct summands are listed in the order that the $l_{i}$ are increasing (or decreasing), then $s$ is independent of the choice of the $v_{j}$ and any other such cyclic direct sum for $V$ will have the same sequence of $l_{j}$.

### 7.6 Canonical Forms

Let $L \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space over $\mathbb{F}$. By Theorem 6.1.10,

$$
V=\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\phi_{n}(L)^{k_{n}}\right)
$$

where the minimum polynomial is $\prod_{j=1}^{n} \phi_{j}(x)^{k_{j}}$ with each $\phi_{j}(x)$ irreducible. For $L$ restricted to the invariant subspace $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$, the minimum polynomial is just $\phi_{j}(L)^{k_{j}}$ because if not, then the minimum polynomial would be $\phi_{j}(x)^{l_{j}}$ for $l_{j}<k_{j}$ and then the claimed minimum polynomial $\prod_{j=1}^{n} \phi_{j}(x)^{k_{j}}$ would not be the minimum polynomial after all. This is, a direct sum and so none of the $\phi_{i}(x)^{k_{i}}$ for $i \neq j$ can send to 0 any nonzero vector of $\operatorname{ker}\left(\phi_{j}(L)^{k_{j}}\right)$. Thus, the exponent $k_{j}$ could be replaced with $l_{j}$.

Then we have the following theorem.
Theorem 7.6.1 Let $V$ be a finite dimensional vector space over a field of scalars $\mathbb{F}$. Also suppose the minimum polynomial is $\prod_{i=1}^{n}\left(\phi_{i}(x)\right)^{k_{i}}$ where $k_{i}$ is a positive integer and the degree of $\phi_{i}(x)$ is $d_{i}$, these $\phi_{i}(x)$ being monic and irreducible (prime in $\mathbb{F}[x]$ ). Then

$$
V=\operatorname{ker}\left(\phi_{1}(L)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\phi_{n}(L)^{k_{n}}\right)
$$

Furthermore, for each $i$, in $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$, there are vectors $v_{1}, \cdots, v_{s_{i}}$ and positive integers $l_{1}, \cdots, l_{s_{i}}$ each no larger than $k_{i}$ such that a basis for $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ is given by

$$
\left\{\beta_{v_{1}}^{l_{1} d_{i}-1}, \cdots, \beta_{v_{s_{i}}}^{l_{s_{i}} d_{i}-1}\right\}
$$

where the symbol $\beta_{v_{j}}^{l_{j} d_{i}-1}$ signifies the ordered basis

$$
\left(v_{j}, L v_{j}, L^{2} v_{j}, \cdots, L^{l_{j} d_{i}-2} v_{j}, L^{l_{j} d_{i}-1} v_{j}\right)
$$

Its length is the degree of $\phi_{j}(x)^{k_{j}}$ and is therefore, determined completely by the $l_{j}$. Thus the lengths of the $\beta_{v_{j}}^{l_{j} d_{i}-1}$ are uniquely determined if they are listed in order of increasing or decreasing length.

The last claim of this theorem will mean that the various canonical forms are uniquely determined.

It is clear that the span of $\beta_{v_{j}}^{l_{j} d_{i}-1}$ is invariant with respect to $L$ because, as discussed above, this span is $D v_{j}$ where $D=\mathbb{F}[x]$ and $\phi_{i}(L)^{l_{j}}$ is the minimum polynomial of $L$ restricted to span $\left(\beta_{v_{j}}^{l_{j} d_{i}-1}\right)$ Let

$$
\phi_{i}(x)^{l_{j}}=x^{l_{j} d_{i}}+a_{l_{j} d_{i}-1} x^{l_{j} d_{i}-1}+\cdots+a_{1} x+a_{0}
$$

Recall that the minimum polynomial has leading coefficient equal to 1 . Of course this makes no difference in the above presentation because $a_{n}$ is invertible but it is convenient to let this happen since otherwise, the blocks for the rational canonical form will not be standard. Then what is the matrix of $L$ restricted to $D v_{j}$ ?

$$
\left(\begin{array}{llll}
L v_{j} & \cdots & L^{l_{j} d_{i}-1} v_{j} & L^{l_{j} d_{i}} v_{j}
\end{array}\right)=\left(\begin{array}{llll}
v_{j} & \cdots & L^{l_{j} d_{i}-2} v_{j} & L^{l_{j} d_{i}-1} v_{j}
\end{array}\right) M
$$

where $M$ is the desired matrix. Now ann $\left(D v_{j}\right)=\left(\phi_{i}(x)^{l_{j}}\right)$ and so

$$
L^{l_{j} d_{i}} v_{j}=(-1)\left(a_{n-1} L^{l_{j} d_{i}-1} v_{j}+\cdots+a_{1} L v_{j}+a_{0} v_{j}\right)
$$

Thus the matrix $M$ must be of the form

$$
\left(\begin{array}{cccc}
0 & & & -a_{0}  \tag{7.11}\\
1 & & & -a_{1} \\
& \ddots & & \vdots \\
0 & & 1 & -a_{n-1}
\end{array}\right)
$$

It follows that the matrix of $L$ with respect to the basis obtained as above will be a block diagonal with blocks like the above. This is the rational canonical form.

Of course, those blocks corresponding to $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ can be arranged in any order by just listing the $\beta_{v_{1}}^{l_{1} d_{i}-1}, \cdots, \beta_{v_{s_{i}}}^{l_{s_{i}} d_{i}-1}$ in various orders. If we want the blocks to be larger
in the top left and get smaller towards the lower right, we just re-number it to have $l_{i}$ be a decreasing sequence.

What about uniqueness of the rational canonical form given an order of the spaces $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ and under the convention the blocks associated with $\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)$ should be increasing or degreasing in size from upper left toward lower right? In other words, suppose you have

$$
\operatorname{ker}\left(\phi_{i}(L)^{k_{i}}\right)=D v_{1} \oplus \cdots \oplus D v_{s}=D w_{1} \oplus \cdots \oplus D w_{t}
$$

and $\phi_{i}(L)^{l_{i}}$ is the minimum polynomial for $D v_{i}$ and we choose the order such that the $l_{i}$ are increasing (decreasing). Then from Theorem 7.5.2 the canonical form will be unique.

In the case that the minimum polynomial splits the following is also obtained.
Corollary 7.6.2 Let $V$ be a finite dimensional vector space over a field of scalars $\mathbb{F}$. Also let the minimum polynomial be $\prod_{i=1}^{n}\left(x-\mu_{i}\right)^{k_{i}}$ where $k_{i}$ is a positive integer. Then

$$
V=\operatorname{ker}\left(\left(L-\mu_{1} I\right)^{k_{1}}\right) \oplus \cdots \oplus \operatorname{ker}\left(\left(L-\mu_{n} I\right)^{k_{n}}\right)
$$

Furthermore, for each $i$, in $\operatorname{ker}\left(\left(L-\mu_{i} I\right)^{k_{i}}\right)$, there are vectors $v_{1}, \cdots, v_{s_{i}}$ and positive integers $l_{1}, \cdots, l_{s_{i}}$ each no larger than $k_{i}$ such that a basis for $\operatorname{ker}\left(\left(L-\mu_{i} I\right)^{k_{i}}\right)$ is given by

$$
\left\{\beta_{v_{1}}^{l_{1}-1}, \cdots, \beta_{v_{s_{i}}}^{l_{s_{i}}-1}\right\}
$$

where the symbol $\beta_{v_{j}}^{l_{j}-1}$ signifies the ordered basis

$$
\left(\left(L-\mu_{i} I\right)^{l_{j}-1} v_{j},\left(L-\mu_{i} I\right)^{l_{j}-2} v_{j}, \cdots,\left(L-\mu_{i} I\right)^{2} v_{j},\left(L-\mu_{i}\right) v_{j}, v_{j}\right)
$$

(Note how this is the reverse order to the above. This is to follow the usual convention in the Jordan form in which the string of ones is on the super diagonal.)

Proof: The proof is essentially the same.

$$
\operatorname{ker}\left(\left(L-\mu_{i}\right)^{k_{i}}\right)=D v_{1} \oplus \cdots \oplus D v_{s_{i}}
$$

ann $\left(D v_{j}\right)$ for $v_{j} \in \operatorname{ker}\left(\left(L-\mu_{i}\right)^{k_{i}}\right)$ is $D \sigma$ where $\sigma(x) /\left(x-\mu_{i}\right)^{k_{i}}$ and so $\sigma(x)$ is of the form $\left(x-\mu_{i}\right)^{l_{j}}$ where $0 \leq l_{j} \leq k_{i}$. Then as before,

$$
v_{j},\left(L-\mu_{i} I\right) v_{j},\left(L-\mu_{i} I\right)^{2} v_{j}, \cdots,\left(L-\mu_{i} I\right)^{l_{j}-1} v_{j}
$$

is a basis for $D v_{j}$.
This gives the Jordan form right away. In this case, $\left(L-\mu_{i} I\right)^{l_{j}} v_{j}=0$ and so the matrix of the transformation $L-\mu_{i} I$ with respect to this basis on $D v_{j}$ obtained in the usual way.

$$
\left(\begin{array}{llll}
0 & \left(L-\mu_{i} I\right)^{l_{j}-1} v_{j} & \cdots & \left(L-\mu_{i} I\right) v_{j}
\end{array}\right)=
$$

$$
\left(\begin{array}{lllll}
\left(L-\mu_{i} I\right)^{l_{j}-1} v_{j} & \left(L-\mu_{i} I\right)^{l_{j}-2} v_{j} & \cdots & v_{j}
\end{array}\right)\left(\begin{array}{cccc}
0 & 1 & & 0 \\
& 0 & \ddots & \\
& & \ddots & 1 \\
0 & & & 0
\end{array}\right)
$$

a Jordan block for the nilpotent matrix $\left(L-\mu_{i} I\right)$ (a power of the matrix equals 0 ). Thus, with respect to this basis, the block associated with $L=\mu_{i} I+\left(L-\mu_{i} I\right)$ is

$$
\left(\begin{array}{cccc}
\mu_{i} & 1 & & 0 \\
& \mu_{i} & \ddots & \\
& & \ddots & 1 \\
0 & & & \mu_{i}
\end{array}\right)
$$

This has proved the existence of the Jordan form. It is a block diagonal matrix consisting of strings of blocks of the above form for each eigenvalue $\mu_{i}$. Of course, these can be arranged so that the size of the blocks is decreasing from upper left to lower right. As with the rational canonical form, once it is decided to have the blocks be decreasing (increasing) in size from upper left to lower right, the Jordan form is unique.

The main item of interest concerning the Jordan canonical form is that it exists. However, it can always be found if you know the eigenvalues.

Example 7.6.3 Find the Jordan form for $A=\left(\begin{array}{cccc}1 & 2 & 1 & -1 \\ 1 & 2 & 1 & 0 \\ -1 & -3 & -1 & 1 \\ 1 & -2 & -1 & 3\end{array}\right)$.
The minimum polynomial is $(\lambda-2)(\lambda-1)^{3}$. Now find a eigenvector for $\lambda=1$. There is only one eigenvector and it is $\boldsymbol{v}_{0}=\left(\begin{array}{cccc}-1 & 0 & 1 & 1\end{array}\right)^{T}$. Therefore, you look for "generalized" eigenvectors. One of these is of the form $\boldsymbol{v}_{1}$ such that $(A-\lambda I) \boldsymbol{v}_{1}=\boldsymbol{v}_{0}$. A solution to this is $\boldsymbol{v}_{1}=\left(\begin{array}{cccc}0 & -1 & 1 & 0\end{array}\right)^{T}$. This is still not enough because the algebraic multiplicity of $\lambda=1$ is 3 so you need to find a solution $\boldsymbol{v}_{2}$ to $\boldsymbol{v}_{1}=(A-\lambda I) \boldsymbol{v}_{2}$. Then a solution to this is $\left(\begin{array}{cccc}0 & 1 & -2 & 0\end{array}\right)^{T}$. Finally, an eigenvector for $\lambda=2$ is $\boldsymbol{v}_{3}=$ $\left(\begin{array}{cccc}-1 & 0 & 1 & 2\end{array}\right)^{T}$. Lets consider what has just been obtained. $\boldsymbol{v}_{0}=A \boldsymbol{v}_{0}, \boldsymbol{v}_{0}+\boldsymbol{v}_{1}=$ $A \boldsymbol{v}_{1}, \boldsymbol{v}_{1}+\boldsymbol{v}_{2}=A \boldsymbol{v}_{2}, 2 \boldsymbol{v}_{3}=A \boldsymbol{v}_{3}$. To find the matrix with respect to the ordered basis $\left(\boldsymbol{v}_{0}, \boldsymbol{v}_{1}, \boldsymbol{v}_{2}, \boldsymbol{v}_{3}\right)$ recall that formally we find $J$ such that

$$
\left(\begin{array}{cccc}
A \boldsymbol{v}_{0} & A \boldsymbol{v}_{1} & A \boldsymbol{v}_{2} & A \boldsymbol{v}_{3}
\end{array}\right)=\left(\begin{array}{llll}
\boldsymbol{v}_{0} & \boldsymbol{v}_{1} & \boldsymbol{v}_{2} & \boldsymbol{v}_{3}
\end{array}\right) J
$$

and so $\left(\begin{array}{ccc}\boldsymbol{v}_{0} & \boldsymbol{v}_{0}+\boldsymbol{v}_{1} & \boldsymbol{v}_{1}+\boldsymbol{v}_{2} \\ 2 \boldsymbol{v}_{3}\end{array}\right)=\left(\begin{array}{llll}\boldsymbol{v}_{0} & \boldsymbol{v}_{1} & \boldsymbol{v}_{2} & \boldsymbol{v}_{3}\end{array}\right) J$
so $J=\left(\begin{array}{cccc}1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2\end{array}\right)$. We know there is a unique Jordan form from the above theory
and so this must be it. This is the process used in beginning differential equations to find solutions to a system when there is a repeated eigenvalue and the matrix is defective. I think it is a fairly easy way to remember things. Other cases are similar.
Example 7.6.4 Find the Jordan form for $A=\left(\begin{array}{ccccc}3 & 3 & 0 & 0 & -1 \\ -1 & 1 & -3 & -4 & -3 \\ 2 & 2 & 7 & 7 & 5 \\ 0 & -2 & 4 & 7 & 5 \\ -2 & 0 & -10 & -13 & -9\end{array}\right)$.
The eigenvalues are 2,1 . This was of course cooked up. You can't find eigenvalues in general. The minimum polynomial is $\lambda^{3}-5 \lambda^{2}+8 \lambda-4$. This can be found through the methods described earlier. It equals $(\lambda-1)(\lambda-2)^{2}$. The eigenvectors for $\lambda=2$ are $\left(\begin{array}{lllll}-\frac{3}{2} t_{4}-\frac{1}{2} t_{5} & \frac{1}{2} t_{4}+\frac{1}{2} t_{5} & -t_{4}-t_{5} & t_{4} & t_{5}\end{array}\right)^{T}$. I can get two independent eigenvectors,

$$
\left(\begin{array}{ccccc}
-3 & 1 & -2 & 2 & 0
\end{array}\right)^{T},\left(\begin{array}{ccccc}
-1 & 1 & -2 & 0 & 2
\end{array}\right)^{T}
$$

However, this will not be enough so I look for generalized eigenvectors which yield these eigenvectors. Two of these are respectively

$$
\left(\begin{array}{ccccc}
-3 & 0 & -2 & 2 & 0
\end{array}\right)^{T},\left(\begin{array}{ccccc}
-1 & 0 & 0 & 0 & 0
\end{array}\right)^{T}
$$

An eigenvector for $\lambda=1$ is $\left(\begin{array}{ccccc}1 & 2 & -3 & -4 & 8\end{array}\right)^{T}$. Then using this ordered basis ordered as eigenvector followed by generalized eigenvector, we obtain the matrix of $A$ with respect to this basis as follows

$$
\begin{gathered}
\left(\begin{array}{ccccc}
-3 & -3 & -1 & -1 & 1 \\
1 & 0 & 1 & 0 & 2 \\
-2 & -2 & -2 & 0 & -3 \\
2 & 2 & 0 & 0 & -4 \\
0 & 0 & 2 & 0 & 8
\end{array}\right)\left(\begin{array}{ccccc}
3 & 3 & 0 & 0 & -1 \\
-1 & 1 & -3 & -4 & -3 \\
2 & 2 & 7 & 7 & 5 \\
0 & -2 & 4 & 7 & 5 \\
-2 & 0 & -10 & -13 & -9
\end{array}\right) \\
\left(\begin{array}{ccccc}
-3 & -3 & -1 & -1 & 1 \\
1 & 0 & 1 & 0 & 2 \\
-2 & -2 & -2 & 0 & -3 \\
2 & 2 & 0 & 0 & -4 \\
0 & 0 & 2 & 0 & 8
\end{array}\right)=\left(\begin{array}{ccccc}
2 & 1 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 \\
0 & 0 & 2 & 1 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 1
\end{array}\right)
\end{gathered}
$$

which is the Jordan form.

### 7.7 Exercises

1. In the discussion of Nilpotent transformations, it was asserted that if two $n \times n$ matrices $A, B$ are similar, then $A^{k}$ is also similar to $B^{k}$. Why is this so? If two matrices are similar, why must they have the same rank?
2. If $A, B$ are both invertible, then they are both row equivalent to the identity matrix. Are they necessarily similar? Explain.
3. Suppose you have two nilpotent matrices $A, B$ and $A^{k}$ and $B^{k}$ both have the same rank for all $k \geq 1$. Does it follow that $A, B$ are similar? What if it is not known that $A, B$ are nilpotent? Does it follow then?
4. (Review problem.) When we say a polynomial equals zero, we mean that all the coefficients equal 0 . If we assign a different meaning to it which says that a polynomial $p(\lambda)$ equals zero when it is the zero function, $(p(\lambda)=0$ for every $\lambda \in \mathbb{F}$.) does this amount to the same thing? Is there any difference in the two definitions for ordinary fields like $\mathbb{Q}$ ? Hint: Consider for the field of scalars $\mathbb{Z}_{2}$, the integers mod 2 and consider $p(\lambda)=\lambda^{2}+\lambda$.
5. Let $A \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space with field of scalars $\mathbb{F}$. Let $p(\lambda)$ be the minimum polynomial and suppose $\phi(\lambda)$ is any nonzero polynomial such that $\phi(A)$ is not one to one and $\phi(\lambda)$ has smallest possible degree such that $\phi(A)$ is nonzero and not one to one. Show $\phi(\lambda)$ must divide $p(\lambda)$.
6. Let $A \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space with field of scalars $\mathbb{F}$. Let $p(\lambda)$ be the minimum polynomial and suppose $\phi(\lambda)$ is an irreducible polynomial with the property that $\phi(A) x=0$ for some specific $x \neq 0$. Show that $\phi(\lambda)$ must divide $p(\lambda)$. Hint: First write $p(\lambda)=\phi(\lambda) g(\lambda)+r(\lambda)$ where $r(\lambda)$ is either 0 or has degree smaller than the degree of $\phi(\lambda)$. If $r(\lambda)=0$ you are done. Suppose it is not 0 . Let $\eta(\lambda)$ be the monic polynomial of smallest degree with the property that $\eta(A) x=0$. Now use the Euclidean algorithm to divide $\phi(\lambda)$ by $\eta(\lambda)$. Contradict the irreducibility of $\phi(\lambda)$.
7. Let $A=\left(\begin{array}{ccc}1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0\end{array}\right)$ Find the minimum polynomial for $A$.
8. Suppose $A$ is an $n \times n$ matrix and let $\boldsymbol{v}$ be a vector. Consider the $A$ cyclic set of vectors $\left\{\boldsymbol{v}, A \boldsymbol{v}, \cdots, A^{m-1} \boldsymbol{v}\right\}$ where this is an independent set of vectors but $A^{m} \boldsymbol{v}$ is a linear combination of the preceding vectors in the list. Show how to obtain a monic polynomial of smallest degree, $m, \phi_{\boldsymbol{v}}(\lambda)$ such that $\phi_{\boldsymbol{v}}(A) \boldsymbol{v}=\mathbf{0}$. Now let $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{n}\right\}$ be a basis and let $\phi(\lambda)$ be the least common multiple of the $\phi_{\boldsymbol{w}_{k}}(\lambda)$. Explain why this must be the minimum polynomial of $A$. Give a reasonably easy algorithm for computing $\phi_{v}(\lambda)$.
9. Here is a matrix.

$$
\left(\begin{array}{ccc}
-7 & -1 & -1 \\
-21 & -3 & -3 \\
70 & 10 & 10
\end{array}\right)
$$

Using the process of Problem 8 find the minimum polynomial of this matrix. Determine whether it can be diagonalized from its minimum polynomial.
10. Let $A$ be an $n \times n$ matrix with field of scalars $\mathbb{C}$ or more generally, the minimum polynomial splits. Letting $\lambda$ be an eigenvalue, show the dimension of the eigenspace
equals the number of Jordan blocks in the Jordan canonical form which are associated with $\lambda$. Recall the eigenspace is $\operatorname{ker}(\lambda I-A)$.
11. For any $n \times n$ matrix, why is the dimension of the eigenspace always less than or equal to the algebraic multiplicity of the eigenvalue as a root of the characteristic equation? Hint: Note the algebraic multiplicity is the size of the appropriate block in the Jordan form.
12. Give an example of two nilpotent matrices which are not similar but have the same minimum polynomial if possible.
13. Here is a matrix. Find its Jordan canonical form by directly finding the eigenvectors and generalized eigenvectors based on these to find a basis which will yield the Jordan form. The eigenvalues are 1 and 2.

$$
\left(\begin{array}{cccc}
-3 & -2 & 5 & 3 \\
-1 & 0 & 1 & 2 \\
-4 & -3 & 6 & 4 \\
-1 & -1 & 1 & 3
\end{array}\right)
$$

Why is it typically impossible to find the Jordan canonical form?
14. Let $A$ be an $n \times n$ matrix and let $J$ be its Jordan canonical form. Here $\mathbb{F}=\mathbb{R}$ or $\mathbb{C}$. Recall $J$ is a block diagonal matrix having blocks $J_{k}(\lambda)$ down the diagonal. Each of these blocks is of the form

$$
J_{k}(\lambda)=\left(\begin{array}{cccc}
\lambda & 1 & & 0 \\
& \lambda & \ddots & \\
& & \ddots & 1 \\
0 & & & \lambda
\end{array}\right)
$$

Now for $\varepsilon>0$ given, let the diagonal matrix $D_{\varepsilon}$ be given by

$$
D_{\varepsilon}=\left(\begin{array}{cccc}
1 & & & 0 \\
& \varepsilon & & \\
& & \ddots & \\
0 & & & \varepsilon^{k-1}
\end{array}\right)
$$

Show that $D_{\varepsilon}^{-1} J_{k}(\lambda) D_{\varepsilon}$ has the same form as $J_{k}(\lambda)$ but instead of ones down the super diagonal, there is $\varepsilon$ down the super diagonal. That is $J_{k}(\lambda)$ is replaced with

$$
\left(\begin{array}{llll}
\lambda & \varepsilon & & 0 \\
& \lambda & \ddots & \\
& & \ddots & \varepsilon \\
0 & & & \lambda
\end{array}\right)
$$

Now show that for $A$ an $n \times n$ matrix, it is similar to one which is just like the Jordan canonical form except instead of the blocks having 1 down the super diagonal, it has $\varepsilon$.
15. Let $A$ be in $\mathscr{L}(V, V)$ and suppose that $A^{p} x \neq 0$ for some $x \neq 0$. Show that $A^{p} e_{k} \neq 0$ for some $e_{k} \in\left\{e_{1}, \cdots, e_{n}\right\}$, a basis for $V$. If you have a matrix which is nilpotent, ( $A^{m}=0$ for some $m$ ) will it always be possible to find its Jordan form? Describe how to do it if this is the case. Hint: First explain why all the eigenvalues are 0. Then consider the way the Jordan form for nilpotent transformations was constructed in the above.
16. Show that if two $n \times n$ matrices $A, B$ are similar, then they have the same minimum polynomial and also that if this minimum polynomial is of the form $p(\lambda)=$ $\prod_{i=1}^{s} \phi_{i}(\lambda)^{r_{i}}$ where the $\phi_{i}(\lambda)$ are irreducible and monic, then $\operatorname{ker}\left(\phi_{i}(A)^{r_{i}}\right)$ and $\operatorname{ker}\left(\phi_{i}(B)^{r_{i}}\right)$ have the same dimension. Why is this so? This was what was responsible for the blocks corresponding to an eigenvalue being of the same size.
17. Show that each cyclic set $\beta_{x}$ is associated with a monic polynomial $\eta_{x}(\lambda)$ such that $\eta_{x}(A)(x)=0$ and this polynomial has smallest possible degree such that this happens. Show that the cyclic sets $\beta_{x_{i}}$ can be arranged such that $\eta_{x_{i+1}}(\lambda) / \eta_{x_{i}}(\lambda)$.
18. Show that if $A$ is a complex $n \times n$ matrix, then $A$ and $A^{T}$ are similar. Hint: Consider a Jordan block. Note that

$$
\left(\begin{array}{lll}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{array}\right)\left(\begin{array}{lll}
\lambda & 1 & 0 \\
0 & \lambda & 1 \\
0 & 0 & \lambda
\end{array}\right)\left(\begin{array}{lll}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{array}\right)=\left(\begin{array}{ccc}
\lambda & 0 & 0 \\
1 & \lambda & 0 \\
0 & 1 & \lambda
\end{array}\right)
$$

19. (Extra important) Let $A$ be an $n \times n$ matrix. The trace of $A$, trace $(A)$ is defined as $\sum_{i} A_{i i}$. It is just the sum of the entries on the main diagonal. Show trace $(A)=$ trace $\left(A^{T}\right)$. Suppose $A$ is $m \times n$ and $B$ is $n \times m$. Show that trace $(A B)=\operatorname{trace}(B A)$. Now show that if $A$ and $B$ are similar $n \times n$ matrices, then $\operatorname{trace}(A)=\operatorname{trace}(B)$. Recall that $A$ is similar to $B$ means $A=S^{-1} B S$ for some matrix $S$.
20. (Extra important) If $A$ is an $n \times n$ matrix and the minimum polynomial splits in $\mathbb{F}$ the field of scalars, show that trace $(A)$ equals the sum of the eigenvalues listed according to multiplicity according to number of times they occur in the Jordan form.
21. Let $A$ be a linear transformation defined on a finite dimensional vector space $V$. Let the minimum polynomial be $\prod_{i=1}^{q} \phi_{i}(\lambda)^{m_{i}}$ and let $\left(\beta_{v_{1}^{i}}^{i}, \cdots, \beta_{v_{r_{i}}}^{i}\right)$ be the cyclic sets such that $\left\{\beta_{v_{1}^{i}}^{i}, \cdots, \beta_{v_{r_{i}}}^{i}\right\}$ is a basis for $\operatorname{ker}\left(\phi_{i}(A)^{m_{i}}\right)$. Let $v=\sum_{i} \sum_{j} v_{j}^{i}$. Now let $q(\boldsymbol{\lambda})$ be any polynomial and suppose that $q(A) v=0$. Show that it follows $q(A)=0$. Hint: First consider the special case where a basis for $V$ is $\left\{x, A x, \cdots, A^{n-1} x\right\}$ and $q(A) x=0$.
22. Find the minimum polynomial for $A=\left(\begin{array}{ccc}1 & 2 & 3 \\ 2 & 1 & 4 \\ -3 & 2 & 1\end{array}\right)$ assuming the field of scalars is the rational numbers.
23. Show, using the rational root theorem, the minimum polynomial for $A$ in the above problem is irreducible with respect to $\mathbb{Q}$. Letting the field of scalars be $\mathbb{Q}$ find the rational canonical form and a similarity transformation which will produce it.
24. Letting the field of scalars be $\mathbb{Q}$, find the rational canonical form for the matrix

$$
\left(\begin{array}{cccc}
1 & 2 & 1 & -1 \\
2 & 3 & 0 & 2 \\
1 & 3 & 2 & 4 \\
1 & 2 & 1 & 2
\end{array}\right)
$$

25. Let $A: \mathbb{Q}^{3} \rightarrow \mathbb{Q}^{3}$ be linear. Suppose the minimum polynomial is

$$
(\lambda-2)\left(\lambda^{2}+2 \lambda+7\right)
$$

Find the rational canonical form. Can you give generalizations of this rather simple problem to other situations?
26. Find the rational canonical form with respect to the field of scalars equal to $\mathbb{Q}$ for the matrix

$$
A=\left(\begin{array}{ccc}
0 & 0 & 1 \\
1 & 0 & -1 \\
0 & 1 & 1
\end{array}\right)
$$

Observe that this particular matrix is already a companion matrix of $\lambda^{3}-\lambda^{2}+\lambda-1$. Then find the rational canonical form if the field of scalars equals $\mathbb{C}$ or $\mathbb{Q}+i \mathbb{Q}$.
27. Consider $\left(\begin{array}{llll}1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1\end{array}\right),\left(\begin{array}{llll}1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1\end{array}\right)$. They are not similar because they are different Jordan forms. However, show that they have the same minimum polynomial. If two matrices are similar, show that then they do have the same minimum polynomial.
28. Let $L \in \mathscr{L}(V, V)$ where $V$ is a finite dimensional vector space with field of scalars $\mathbb{F}$ and consider for $v \in V, v \neq 0$ the cycle $\left\{v, L v, \cdots, L^{m} v\right\}$. Show that if $L^{m} v \neq 0$ then the resulting set of vectors is linearly independent.
29. Suppose you have two $n \times n$ matrices $A, B$ whose entries are in a field $\mathbb{F}$ and suppose $\mathbb{G}$ is an extension of $\mathbb{F}$. For example, you could have $\mathbb{F}=\mathbb{Q}$ and $\mathbb{G}=\mathbb{C}$. Suppose $A$ and $B$ are similar with respect to the field $\mathbb{G}$. Can it be concluded that they are similar with respect to the field $\mathbb{F}$ ? Hint: Let the minimum polynomial of $A$ with respect to $\mathbb{F}$ be $\prod_{i=1}^{q} \phi_{i}(\lambda)^{p_{i}}$. Say $\beta_{v_{j}}$ is the cyclical set associated with the polynomial $\phi(\lambda)^{l}$ as described in the proof for the rational canonical form. Here $\phi(\lambda)$ is one of the $\phi_{i}(\lambda)$ having degree $d$. Thus, as described above the length of $\left|\beta_{v_{j}}\right|$ equals $l d$ and the term in the cyclical decomposition of $\operatorname{ker}\left(\phi(A)^{p}\right)$ satisfies $\phi(A)^{l} v_{j}=0$ and $l$ is as small as possible for that $v_{j}$. Here $v_{j} \in \mathbb{F}^{n}$. A basis for a block in the rational form corresponding to the field $\mathbb{F}$ would be.

$$
\begin{equation*}
\left\{v_{j}, A v_{j}, \cdots, A^{d-1} v_{j}, \cdots, A^{l d-1} v_{j}\right\}, v_{j} \in \mathbb{F}^{n} \tag{7.12}
\end{equation*}
$$

This corresponds to

$$
\begin{aligned}
0 & =\phi(A)^{l} v_{j}=\left(a_{0}+a_{1} A+\ldots+a_{d-1} A^{d-1}+A^{d}\right)^{l} v_{j} \\
& =\phi(A)^{l-1}\left(a_{0}+a_{1} A+\ldots+a_{d-1} A^{d-1}+A^{d}\right) v_{j}
\end{aligned}
$$

Could you have $\phi(A)^{l-1} A^{r} v_{j}=0$ ? No because the minimum polynomial for $D \phi^{l}$ is $\phi^{l}$. Thus the columns of the following table give bases for blocks for the Jordan form for $\phi(A)$ corresponding to $D v_{j}$ provided the vectors in the columns taken together are linearly independent. From the above problem, each column is linearly independent. So assume this for now.

| $\alpha_{0}^{j}$ | $\alpha_{1}^{j}$ | $\alpha_{2}^{j}$ | $\cdots$ | $\alpha_{d-1}^{j}$ |
| :---: | :---: | :---: | :---: | :---: |
| $v_{j}$ | $A v_{j}$ | $A^{2} v_{j}$ | $\cdots$ | $A^{d-1} v_{j}$ |
| $\phi(A) v_{j}$ | $\phi(A) A v_{j}$ | $\phi(A) A^{2} v_{j}$ | $\cdots$ | $\phi(A) A^{d-1} v_{j}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |  | $\vdots$ |
| $\phi(A)^{l-1} v_{j}$ | $\phi(A)^{l-1} A v_{j}$ | $\phi(A)^{l-1} A^{2} v_{j}$ | $\cdots$ | $\phi(A)^{l-1} A^{d-1} v_{j}$ |

These columns would yield $d-1$ blocks in the Jordan form for $\phi(A)$ in $\mathbb{C}^{n^{2}}$. If all such cyclical bases like 7.12 fail to include 7.12 with $v_{j}$ replaced with some $w_{j} \in \mathbb{F}^{n}$ and $A$ with $B$, then the Jordan form for $\phi(B)$ in $\mathbb{C}^{n^{2}}$ will be different than the Jordan form for $\phi(A)$ contradicting the similarity of $\phi(A), \phi(B)$. Thus it remains to verify linear independence as indicated above.
30. This entire presentation is based on modules over the commutative ring $\mathbb{F}[x]$. This ring is an integral domain because if $\alpha(x) \beta(x)=0$ then one of $\alpha(x)$ or $\beta(x)$ is 0 . This is shown in Lemma 1.13.2 on Page 21. For a commutative ring $R$, an ideal $I$ is a subset of $R$ with the property that $\alpha I \subseteq I$ for all $\alpha$ and whenever $\alpha, \beta \in I$, so is $\alpha+\beta$ and if $\alpha \in I$, so is $-\alpha$. See Definition 3.4.10. A principle ideal domain $D$ called a p.i.d. is an integral domain in which the only ideals are of the form $D \alpha$ for some $\alpha \in D$. These are called principle ideals. Earlier these were written as $(\alpha)$ indicating all multiples of $\alpha$. Show $\mathbb{F}[x]$ is a principle ideal domain. A commutative ring $R$ is called a Noetherian ring if every increasing sequence of ideals is eventually constant. Show that a principle ideal domain must be Noetherian.
31. You can define $R / I$ similar to what was done above for modules. Here $I$ is an ideal in the commutative ring $R$. That is, let $r \sim \hat{r}$ if $r-\hat{r} \in I$ and let $[r]$ be the equivalence class determined by $r$. Show this is the same as considering $[r]=r+I$. Define $[r][\hat{r}] \equiv[r \hat{r}]$ and $[r]+[\hat{r}] \equiv[r+\hat{r}]$. Verify these operations make $R / I$ into a ring and are well defined. Next show that if $I$ is a maximal ideal (Definition 3.4.10) then $R / I$ is also a field. For $R=\mathbb{F}[x]$, show that $(p(x))$ is a maximal ideal if $p(x)$ is irreducible over $\mathbb{F}$.
32. Say you have a commutative ring $R$ and $a \in R \backslash\{1\}, a \neq 0$ which is not invertible. Explain why $(a)$ is an ideal which does not contain 1 . Show there exists a maximal ideal. A maximal ideal $I$ is one which is strictly smaller than $R$ and that there is
no ideal $J$ other than $R$ which contains $I$. Hint: You could let $\mathscr{F}$ denote all ideals which do not contain 1 . It is nonempty by assumption. Now partially order this by set inclusion. Consider a maximal chain. This uses the Hausdorff maximal theorem in the appendix.
33. It is always assumed that the rings used here are commutative rings and that they have a multiplicative identity 1 . However, sometimes people have considered things which they have called rings $R$ which have all the same axioms except that there is no multiplicative identity 1 . However, all such things can be considered to be in a sense embedded in a real ring which has a multiplicative identity. You consider $\mathbb{Z} \times R$ and define addition in the obvious way $(k, r)+(\hat{k}, \hat{r}) \equiv(k+\hat{k}, r+\hat{r})$ and multiplication as follows. $(k, r)(\hat{k}, \hat{r}) \equiv(k \hat{k}, k \hat{r}+\hat{k} r+r \hat{r})$ Then the multiplicative identity is just $(1,0)$. You have $(1,0)(k, r) \equiv(k, r)$. You just have to verify the other axioms like the distributive laws and that multiplication is associative.
34. Let $R$ be a p.i.d. Then $p \in R$ is prime if it is divisible only by invertible elements of $R$ and $x p$ where $x$ is invertible. Show that the ideal $R p$ for $p$ a noninvertible prime is a maximal ideal. Thus $R / R p$ is a field from the above problems.
35. Let $R$ be the ring of continuous functions defined on $[0,1]$. Here it is understood that $f=g$ means the usual thing, that $f(x)=g(x)$. Multiplication and addition are defined in the usual way. Pick $x_{0} \in[0,1]$ and let $I_{x_{0}} \equiv\left\{f \in R: f\left(x_{0}\right)=0\right\}$. Show that this is a maximal ideal of $R$. Then show that there are no other maximal ideals. Hint: For the second part, let $I$ be a maximal ideal. Show using a compactness argument and continuity of the functions that unless there exists some $x_{0}$ for which all $f \in I$ are zero, then there exists a function in $I$ which is never 0 . Then since this is an ideal, you can show that it contains 1 . Explain why this ring cannot be an integral domain.

## Chapter 8

## Determinants

The determinant is a number which comes from an $n \times n$ matrix of elements of a field $\mathbb{F}$. It is easiest to give a definition of the determinant which is clearly well defined and then prove the one which involves Laplace expansion which the reader might have seen already. Let $\left(i_{1}, \cdots, i_{n}\right)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1,2,3)$ and $(2,1,3)$ are different. Two books which give a good introduction to determinants are Apostol [1] and Rudin [36]. Some recent books which also have a good introduction are Baker [4], and Baker and Kuttler [6]. The approach here is less elegant than in these other books but it amounts to the same thing. I have just tried to avoid the language of permutations in the presentation. The function sgn presented in what follows is really the sign of a permutation however.

The determinant is an alternating multilinear form $d\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right)$, meaning that it is linear in each entry and when two are switched, it changes sign. It is normalized by requiring $d\left(e_{1}, \cdots, e_{n}\right)=1$. These two conditions are sufficient to define the determinant algebraically as will be seen. Geometrically, it gives the signed volume of a $n$ dimensional parallelepiped, the sign giving a precise meaning for orientation. This will become clear in what follows. I do not wish this presentation to depend on geometric intuition which some pretend to have more of than others, so I am presenting this algebraically. Yes, I have much more confidence in algebra than my geometric intuition, even if I agree that the main interest in the determinant has to do with understanding geometry. It is the same thing which often happens in math. The most transparent explanations are in one place and the applications and most significant meaning are somewhere else. I take the point of view that algebra should be a tool for understanding geometry, not the other way around, because for me, this approach involves less pretense that I understand something which I really don't understand, although it may seem plausible based on smaller dimensional situations. However, plausibility is not proof. What exactly is the "volume" of an $n$ dimensional parallelepiped anyway? I know about the volume of a Cartesian product of intervals, but slanty things are not so clear to me. What is the orientation? In the case of general fields of scalars, what do such geometric concepts even mean? Even in my favorite case of the real field of scalars, my right hand is not sufficient to consider orientation all that well in $\mathbb{R}^{8}$. However, this will be discussed later, including the manner in which the signed volume delivered by the determinant is the only geometrically reasonable way to generalize to higher dimensions the Euclidean length of a line segment, but the proofs of such things depend on algebraic considerations. This also leads to proofs of some very interesting approximation results like the Muntz theorem, an amazing generalization of the Weierstrass approximation theorem, which has very little obvious connection to most people's favorite applications of determinants. Another thing to note is that historically, determinants came before the elegant material presented earlier and they were an algebraic entity right from the beginning. This is why we have Cramer's rule, Laplace expansion, and the amazing Cauchy theorem presented in this chapter. Each of these men lived before the lovely vector space material presented earlier, and also before modern ideas of geometry.

### 8.1 The Function sgn

The following Lemma will be essential in the definition of the determinant.
Lemma 8.1.1 There exists a function, $\operatorname{sgn}_{n}$ which maps each ordered list of numbers from
$\{1, \cdots, n\}$ to one of the three numbers, 0,1 , or -1 which also has the following properties.

$$
\begin{gather*}
\operatorname{sgn}_{n}(1, \cdots, n)=1  \tag{8.1}\\
\operatorname{sgn}_{n}\left(i_{1}, \cdots, p, \cdots, q, \cdots, i_{n}\right)=-\operatorname{sgn}_{n}\left(i_{1}, \cdots, q, \cdots, p, \cdots, i_{n}\right) \tag{8.2}
\end{gather*}
$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1 . Also, in the case where $n>1$ and $\left\{i_{1}, \cdots, i_{n}\right\}=\{1, \cdots, n\}$ so that every number from $\{1, \cdots, n\}$ appears in the ordered list, $\left(i_{1}, \cdots, i_{n}\right)$,

$$
\begin{gather*}
\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_{n}\right) \equiv \\
(-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(i_{1}, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n}\right) \tag{8.3}
\end{gather*}
$$

where $n=i_{\theta}$ in the ordered list, $\left(i_{1}, \cdots, i_{n}\right)$.
Proof: Define $\operatorname{sign}(x)=1$ if $x>0,-1$ if $x<0$ and 0 if $x=0$. If $n=1$, there is only one list and it is just the number 1 . Thus one can define $\operatorname{sgn}_{1}(1) \equiv 1$. For the general case where $n>1$, simply define

$$
\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{n}\right) \equiv \operatorname{sign}\left(\prod_{r<s}\left(i_{s}-i_{r}\right)\right)
$$

This delivers either $-1,1$, or 0 by definition. What about the other claims? Suppose you switch $i_{p}$ with $i_{q}$ where $p<q$ so two numbers in the ordered list $\left(i_{1}, \cdots, i_{n}\right)$ are switched. Denote the new ordered list of numbers as $\left(j_{1}, \cdots, j_{n}\right)$. Thus $j_{p}=i_{q}$ and $j_{q}=i_{p}$ and if $r \notin\{p, q\}, j_{r}=i_{r}$. See the following illustration

| $i_{1}$ | $i_{2}$ | $\cdots$ | $i_{p}$ | $\cdots$ | $i_{q}$ | $\cdots$ | $i_{n}$ |
| :---: | :---: | :--- | :---: | :---: | :---: | :--- | :---: |
| 1 | 2 |  | $p$ |  | $q$ |  | $n$ |
| $i_{1}$ | $i_{2}$ | $\cdots$ | $i_{q}$ | $\cdots$ | $i_{p}$ | $\cdots$ | $i_{n}$ |
| 1 | 2 |  | $p$ |  | $q$ |  | $n$ |
| $j_{1}$ | $j_{2}$ | $\cdots$ | $j_{p}$ | $\cdots$ | $j_{q}$ | $\cdots$ | $j_{n}$ |
| 1 | 2 |  | $p$ |  | $q$ |  | $n$ |

Then

$$
\begin{gathered}
\operatorname{sgn}_{n}\left(j_{1}, \cdots, j_{n}\right) \equiv \operatorname{sign}\left(\prod_{r<s}\left(j_{s}-j_{r}\right)\right) \\
=\operatorname{sign}\left(\left(\begin{array}{c}
\text { both } p, q \\
\left.i_{p}-i_{q}\right)
\end{array} \prod_{p<j<q}\left(i_{j}-i_{q}\right) \prod_{p<j<q}^{\text {one of } p, q}\left(i_{p}-i_{j}\right) \prod_{r<s, r, s \notin\{p, q\}}^{\text {neither } p \text { nor } q}\left(i_{s}-i_{r}\right)\right)\right.
\end{gathered}
$$

The last product consists of the product of terms which were in $\prod_{r<s}\left(i_{s}-i_{r}\right)$ while the two products in the middle both introduce $q-p-1$ minus signs. Thus their product is positive. The first factor is of opposite sign to the $i_{q}-i_{p}$ which occured in $\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{n}\right)$. Therefore, this switch introduced a minus sign and

$$
\operatorname{sgn}_{n}\left(j_{1}, \cdots, j_{n}\right)=-\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{n}\right)
$$

Now consider the last claim. In computing $\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_{n}\right)$ there will be the product of $n-\theta$ negative terms

$$
\left(i_{\theta+1}-n\right) \cdots\left(i_{n}-n\right)
$$

and the other terms in the product for computing $\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_{n}\right)$ are those which are required to compute $\operatorname{sgn}_{n-1}\left(i_{1}, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n}\right)$ multiplied by terms of the form $\left(n-i_{j}\right)$ which are nonnegative. It follows that

$$
\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_{n}\right)=(-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(i_{1}, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n}\right)
$$

It is obvious that if there are repeats in the list the function gives 0 .
Lemma 8.1.2 Every ordered list of distinct numbers from $\{1,2, \cdots, n\}$ can be obtained from every other ordered list of distinct numbers by a finite number of switches. Also, $\operatorname{sgn}_{n}$ is unique.

Proof: This is obvious if $n=1$ or 2 . Suppose then that it is true for sets of $n-1$ elements. Take two ordered lists of numbers, $P_{1}, P_{2}$. Make one switch in both if necessary to place $n$ at the end. Call the result $P_{1}^{n}$ and $P_{2}^{n}$. Then using induction, there are finitely many switches in $P_{1}^{n}$ so that it will coincide with $P_{2}^{n}$. Now switch the $n$ in what results to where it was in $P_{2}$.

To see $\operatorname{sgn}_{n}$ is unique, if there exist two functions, $f$ and $g$ both satisfying 8.1 and 8.2 , you could start with $f(1, \cdots, n)=g(1, \cdots, n)=1$ and applying the same sequence of switches, eventually arrive at $f\left(i_{1}, \cdots, i_{n}\right)=g\left(i_{1}, \cdots, i_{n}\right)$. If any numbers are repeated, then 8.2 gives both functions are equal to zero for that ordered list.

Definition 8.1.3 An ordered list of distinct numbers from $\{1,2, \cdots, n\}$, say $\left(i_{1}, \cdots, i_{n}\right)$, is called a permutation. The symbol for all such permutations is $S_{n}$. The number defined above $\operatorname{sgn}_{n}\left(i_{1}, \cdots, i_{n}\right)$ is called the sign of the permutation.

A permutation can also be considered as a function from the set

$$
\{1,2, \cdots, n\} \text { to }\{1,2, \cdots, n\}
$$

as follows. Let $f(k)=i_{k}$. Permutations are of fundamental importance in certain areas of math. For example, it was by considering permutations that Galois was able to give a criterion for solution of polynomial equations by radicals, but this is a different direction than what is being attempted here.

In what follows sgn will often be used rather than $\operatorname{sgn}_{n}$ because the context supplies the appropriate $n$.

### 8.2 The Definition of the Determinant

Definition 8.2.1 Let $f$ be a real valued function which has the set of ordered lists of numbers from $\{1, \cdots, n\}$ as its domain. Define $\sum_{\left(k_{1}, \cdots, k_{n}\right)} f\left(k_{1} \cdots k_{n}\right)$ to be the sum of all the $f\left(k_{1} \cdots k_{n}\right)$ for all possible choices of ordered lists $\left(k_{1}, \cdots, k_{n}\right)$ of numbers of $\{1, \cdots, n\}$. For example,

$$
\sum_{\left(k_{1}, k_{2}\right)} f\left(k_{1}, k_{2}\right)=f(1,2)+f(2,1)+f(1,1)+f(2,2) .
$$

Definition 8.2.2 Let $\left(a_{i j}\right)=A$ denote an $n \times n$ matrix. The determinant of $A$, denoted by $\operatorname{det}(A)$ is defined by $\operatorname{det}(A) \equiv \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots a_{n k_{n}}$ where the sum is taken over all ordered lists of numbers from $\{1, \cdots, n\}$. Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are, $\operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right)=0$ and so that term contributes 0 to the sum.

Let $A$ be an $n \times n$ matrix $A=\left(a_{i j}\right)$ and let $\left(r_{1}, \cdots, r_{n}\right)$ denote an ordered list of $n$ numbers from $\{1, \cdots, n\}$. Let $A\left(r_{1}, \cdots, r_{n}\right)$ denote the matrix whose $k^{\text {th }}$ row is the $r_{k}$ row of the matrix $A$. Thus

$$
\begin{equation*}
\operatorname{det}\left(A\left(r_{1}, \cdots, r_{n}\right)\right)=\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{r_{1} k_{1}} \cdots a_{r_{n} k_{n}} \tag{8.4}
\end{equation*}
$$

and $A(1, \cdots, n)=A$.
Proposition 8.2.3 Let $\left(r_{1}, \cdots, r_{n}\right)$ be an ordered list of numbers from $\{1, \cdots, n\}$. Then

$$
\begin{align*}
\operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right) \operatorname{det}(A) & =\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{r_{1} k_{1}} \cdots a_{r_{n} k_{n}}  \tag{8.5}\\
& =\operatorname{det}\left(A\left(r_{1}, \cdots, r_{n}\right)\right) \tag{8.6}
\end{align*}
$$

Proof: Let $(1, \cdots, n)=(1, \cdots, r, \cdots s, \cdots, n)$ so $r<s$.

$$
\begin{equation*}
\operatorname{det}(A(1, \cdots, r, \cdots, s, \cdots, n))= \tag{8.7}
\end{equation*}
$$

$$
\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{r}, \cdots, k_{s}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots a_{r k_{r}} \cdots a_{s k_{s}} \cdots a_{n k_{n}}
$$

and renaming the variables, calling $k_{s}, k_{r}$ and $k_{r}, k_{s}$, this equals

$$
\begin{gather*}
=\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{s}, \cdots, k_{r}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots a_{r k_{s}} \cdots a_{s k_{r}} \cdots a_{n k_{n}} \\
=\sum_{\left(k_{1}, \cdots, k_{n}\right)}-\operatorname{sgn}\left(\begin{array}{c}
k_{1}, \cdots, \overbrace{k_{r}, \cdots, k_{s}}^{\text {These got switched }}, \cdots, k_{n}) a_{1 k_{1}} \cdots a_{s k_{r}} \cdots a_{r k_{s}} \cdots a_{n k_{n}} \\
=-\operatorname{det}(A(1, \cdots, s, \cdots, r, \cdots, n))
\end{array} .\right.
\end{gather*}
$$

Consequently,

$$
\operatorname{det}(A(1, \cdots, s, \cdots, r, \cdots, n))=-\operatorname{det}(A(1, \cdots, r, \cdots, s, \cdots, n))=-\operatorname{det}(A)
$$

Now letting $A(1, \cdots, s, \cdots, r, \cdots, n)$ play the role of $A$, and continuing in this way, switching pairs of numbers,

$$
\operatorname{det}\left(A\left(r_{1}, \cdots, r_{n}\right)\right)=(-1)^{p} \operatorname{det}(A)
$$

where it took $p$ switches to obtain $\left(r_{1}, \cdots, r_{n}\right)$ from $(1, \cdots, n)$. By Lemma 8.1.1, this implies

$$
\operatorname{det}\left(A\left(r_{1}, \cdots, r_{n}\right)\right)=(-1)^{p} \operatorname{det}(A)=\operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right) \operatorname{det}(A)
$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $\left(r_{1}, \cdots, r_{n}\right)$.However, if there is a repeat, say the $r^{t h}$ row equals the $s^{\text {th }}$ row, then the reasoning of $8.7-8.8$ shows that $\operatorname{det}\left(A\left(r_{1}, \cdots, r_{n}\right)\right)=0$ and also $\operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right)=0$ so the formula holds in this case also.

Observation 8.2.4 There are $n$ ! ordered lists of distinct numbers from $\{1, \cdots, n\}$.
To see this, consider $n$ slots placed in order. There are $n$ choices for the first slot. For each of these choices, there are $n-1$ choices for the second. Thus there are $n(n-1)$ ways to fill the first two slots. Then for each of these ways there are $n-2$ choices left for the third slot. Continuing this way, there are $n$ ! ordered lists of distinct numbers from $\{1, \cdots, n\}$ as stated in the observation.

### 8.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\operatorname{det}(A)=\operatorname{det}\left(A^{T}\right)$.

Corollary 8.3.1 The following formula for $\operatorname{det}(A)$ is valid.

$$
\begin{equation*}
\operatorname{det}(A)=\frac{1}{n!} \cdot \sum_{\left(r_{1}, \cdots, r_{n}\right)} \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right) \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{r_{1} k_{1}} \cdots a_{r_{n} k_{n}} \tag{8.9}
\end{equation*}
$$

And also $\operatorname{det}\left(A^{T}\right)=\operatorname{det}(A)$ where $A^{T}$ is the transpose of $A$. (Recall that for $A^{T}=\left(a_{i j}^{T}\right)$, $a_{i j}^{T}=a_{j i}$.)

Proof: From Proposition 8.2.3, if the $r_{i}$ are distinct,

$$
\operatorname{det}(A)=\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right) \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{r_{1} k_{1}} \cdots a_{r_{n} k_{n}}
$$

Summing over all ordered lists, $\left(r_{1}, \cdots, r_{n}\right)$ where the $r_{i}$ are distinct, (If the $r_{i}$ are not distinct, $\operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right)=0$ and so there is no contribution to the sum.)

$$
n!\operatorname{det}(A)=\sum_{\left(r_{1}, \cdots, r_{n}\right)} \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(r_{1}, \cdots, r_{n}\right) \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{r_{1} k_{1}} \cdots a_{r_{n} k_{n}}
$$

This proves the corollary since the formula gives the same number for $A$ as it does for $A^{T}$.

Corollary 8.3.2 If two rows or two columns in an $n \times n$ matrix $A$, are switched, the determinant of the resulting matrix equals $(-1)$ times the determinant of the original matrix. If $A$ is an $n \times n$ matrix in which two rows are equal or two columns are equal then $\operatorname{det}(A)=0$. Suppose the $i^{\text {th }}$ row of $A$ equals $\left(x a_{1}+y b_{1}, \cdots, x a_{n}+y b_{n}\right)$. Then

$$
\operatorname{det}(A)=x \operatorname{det}\left(A_{1}\right)+y \operatorname{det}\left(A_{2}\right)
$$

where the $i^{\text {th }}$ row of $A_{1}$ is $\left(a_{1}, \cdots, a_{n}\right)$ and the $i^{t h}$ row of $A_{2}$ is $\left(b_{1}, \cdots, b_{n}\right)$, all other rows of $A_{1}$ and $A_{2}$ coinciding with those of $A$. In other words, det is a linear function of each row A. The same is true with the word "row" replaced with the word "column".

Proof: By Proposition 8.2 .3 when two rows are switched, the determinant of the resulting matrix is $(-1)$ times the determinant of the original matrix. By Corollary 8.3.1 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if $A_{1}$ is the matrix obtained from $A$ by switching two columns,

$$
\operatorname{det}(A)=\operatorname{det}\left(A^{T}\right)=-\operatorname{det}\left(A_{1}^{T}\right)=-\operatorname{det}\left(A_{1}\right)
$$

If $A$ has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, $\operatorname{det}(A)=-\operatorname{det}(A)$ and so $\operatorname{det}(A)=0$.

It remains to verify the last assertion.

$$
\begin{aligned}
& \begin{aligned}
\operatorname{det}(A) \equiv & \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots\left(x a_{r k_{i}}+y b_{r k_{i}}\right) \cdots a_{n k_{n}} \\
& =x \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots a_{r k_{i}} \cdots a_{n k_{n}}
\end{aligned} \\
& +y \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) a_{1 k_{1}} \cdots b_{r k_{i}} \cdots a_{n k_{n}} \equiv x \operatorname{det}\left(A_{1}\right)+y \operatorname{det}\left(A_{2}\right) .
\end{aligned}
$$

The same is true of columns because $\operatorname{det}\left(A^{T}\right)=\operatorname{det}(A)$ and the rows of $A^{T}$ are the columns of $A$.

### 8.4 Basic Properties of the Determinant

Definition 8.4.1 A vector, $\boldsymbol{w}$, is a linear combination of the vectors $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}\right\}$ if there exist scalars $c_{1}, \cdots c_{r}$ such that $\boldsymbol{w}=\sum_{k=1}^{r} c_{k} \boldsymbol{v}_{k}$. That is, $\boldsymbol{w} \in \operatorname{span}\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}\right)$.

The following corollary is also of great use.
Corollary 8.4.2 Suppose $A$ is an $n \times n$ matrix and some column (row) is a linear combination of $r$ other columns (rows). Then $\operatorname{det}(A)=0$.

Proof: Let $A=\left(\begin{array}{lll}a_{1} & \cdots & a_{n}\end{array}\right)$ be the columns of $A$ and suppose the condition that one column is a linear combination of $r$ of the others is satisfied. Say $\boldsymbol{a}_{i}=\sum_{j \neq i} c_{j} \boldsymbol{a}_{j}$. Then by Corollary 8.3.2, $\operatorname{det}(A)=$

$$
\operatorname{det}\left(\begin{array}{lllll}
\boldsymbol{a}_{1} & \cdots & \sum_{j \neq i} c_{j} \boldsymbol{a}_{j} & \cdots & \boldsymbol{a}_{n}
\end{array}\right)=\sum_{j \neq i} c_{j} \operatorname{det}\left(\begin{array}{lllll}
\boldsymbol{a}_{1} & \cdots & \boldsymbol{a}_{j} & \cdots & \boldsymbol{a}_{n}
\end{array}\right)=0
$$

because each of these determinants in the sum has two equal rows.
Recall the following definition of matrix multiplication.
Definition 8.4.3 If $A$ and $B$ are $n \times n$ matrices, $A=\left(a_{i j}\right)$ and $B=\left(b_{i j}\right), A B=\left(c_{i j}\right)$ where $c_{i j} \equiv \sum_{k=1}^{n} a_{i k} b_{k j}$.

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

Theorem 8.4.4 Let $A$ and $B$ be $n \times n$ matrices. Then

$$
\operatorname{det}(A B)=\operatorname{det}(A) \operatorname{det}(B)
$$

Proof: Let $c_{i j}$ be the $i j^{\text {th }}$ entry of $A B$. Then by Proposition 8.2.3,

$$
\begin{aligned}
\operatorname{det}(A B) & =\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) c_{1 k_{1}} \cdots c_{n k_{n}} \\
& =\sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right)\left(\sum_{r_{1}} a_{1 r_{1}} b_{r_{1} k_{1}}\right) \cdots\left(\sum_{r_{n}} a_{n r_{n}} b_{r_{n} k_{n}}\right)
\end{aligned}
$$

$$
\begin{aligned}
& =\sum_{\left(r_{1} \cdots, r_{n}\right)} \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}\left(k_{1}, \cdots, k_{n}\right) b_{r_{1} k_{1}} \cdots b_{r_{n} k_{n}}\left(a_{1 r_{1}} \cdots a_{n r_{n}}\right) \\
& =\sum_{\left(r_{1} \cdots, r_{n}\right)} \operatorname{sgn}\left(r_{1} \cdots r_{n}\right) a_{1 r_{1}} \cdots a_{n r_{n}} \operatorname{det}(B)=\operatorname{det}(A) \operatorname{det}(B) .
\end{aligned}
$$

Note that this shows that if two matrices are similar, then they have the same determinant and also the same characteristic polynomial, $\operatorname{det}(\lambda I-A)$.

### 8.4.1 Binet Cauchy Formula

The Binet Cauchy formula is a generalization of the theorem which says the determinant of a product is the product of the determinants. The situation is illustrated in the following picture where $A, B$ are matrices.


Theorem 8.4.5 Let $A$ be an $n \times m$ matrix with $n \geq m$ and let $B$ be a $m \times n$ matrix. Also let $A_{i}, i=1, \cdots, C(n, m)$ be the $m \times m$ submatrices of $A$ which are obtained by deleting $n-m$ rows and let $B_{i}$ be the $m \times m$ submatrices of $B$ which are obtained by deleting corresponding $n-m$ columns. Then

$$
\operatorname{det}(B A)=\sum_{k=1}^{C(n, m)} \operatorname{det}\left(B_{k}\right) \operatorname{det}\left(A_{k}\right)
$$

Proof: This follows from a computation. By Corollary 8.3.1 on Page 175, $\operatorname{det}(B A)=$

$$
\begin{gathered}
\frac{1}{m!} \sum_{\left(i_{1} \cdots i_{m}\right)} \sum_{\left(j_{1} \cdots j_{m}\right)} \operatorname{sgn}\left(i_{1} \cdots i_{m}\right) \operatorname{sgn}\left(j_{1} \cdots j_{m}\right)(B A)_{i_{1} j_{1}}(B A)_{i_{2} j_{2}} \cdots(B A)_{i_{m} j_{m}} \\
\frac{1}{m!} \sum_{\left(i_{1} \cdots i_{m}\right)} \sum_{\left(j_{1} \cdots j_{m}\right)} \operatorname{sgn}\left(i_{1} \cdots i_{m}\right) \operatorname{sgn}\left(j_{1} \cdots j_{m}\right) \\
\sum_{r_{1}=1}^{n} B_{i_{1} r_{1}} A_{r_{1} j_{1}} \sum_{r_{2}=1}^{n} B_{i_{2} r_{2}} A_{r_{2} j_{2}} \cdots \sum_{r_{m}=1}^{n} B_{i_{m} r_{m}} A_{r_{m} j_{m}}
\end{gathered}
$$

Now denote by $I_{k}$ one of the subsets of $\{1, \cdots, n\}$ which has $m$ elements. Thus there are $C(n, m)$ of these.

$$
\begin{aligned}
= & \sum_{k=1}^{C(n, m)} \sum_{\left\{r_{1}, \cdots, r_{m}\right\}=I_{k}} \frac{1}{m!} \sum_{\left(i_{1} \cdots i_{m}\right)} \sum_{\left(j_{1} \cdots j_{m}\right)} \operatorname{sgn}\left(i_{1} \cdots i_{m}\right) \operatorname{sgn}\left(j_{1} \cdots j_{m}\right) . \\
& B_{i_{1} r_{1}} A_{r_{1} j_{1}} B_{i_{2} r_{2}} A_{r_{2} j_{2}} \cdots B_{i_{m} r_{m}} A_{r_{m} j_{m}} \\
& \sum_{k=1}^{C(n, m)} \sum_{\left\{r_{1}, \cdots, r_{m}\right\}=I_{k}} \frac{1}{m!} \sum_{\left(i_{1} \cdots i_{m}\right)} \operatorname{sgn}\left(i_{1} \cdots i_{m}\right) B_{i_{1} r_{1}} B_{i_{2} r_{2}} \cdots B_{i_{m} r_{m}} . \\
& \sum_{\left(j_{1} \cdots j_{m}\right)} \operatorname{sgn}\left(j_{1} \cdots j_{m}\right) A_{r_{1} j_{1}} A_{r_{2} j_{2}} \cdots A_{r_{m} j_{m}}
\end{aligned}
$$

$$
=\sum_{k=1}^{C(n, m)} \sum_{\left\{r_{1}, \cdots, r_{m}\right\}=I_{k}} \frac{1}{m!} \operatorname{sgn}\left(r_{1} \cdots r_{m}\right)^{2} \operatorname{det}\left(B_{k}\right) \operatorname{det}\left(A_{k}\right)=\sum_{k=1}^{C(n, m)} \operatorname{det}\left(B_{k}\right) \operatorname{det}\left(A_{k}\right)
$$

since there are $m$ ! ways of arranging the indices $\left\{r_{1}, \cdots, r_{m}\right\}$.

### 8.5 Expansion Using Cofactors

Lemma 8.5.1 Suppose a matrix is of the form

$$
M=\left(\begin{array}{cc}
A & *  \tag{8.10}\\
\mathbf{0} & a
\end{array}\right) \text { or }\left(\begin{array}{ll}
A & \mathbf{0} \\
* & a
\end{array}\right)
$$

where $a$ is a number and $A$ is an $(n-1) \times(n-1)$ matrix and $*$ denotes either a column or a row having length $n-1$ and the 0 denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then $\operatorname{det}(M)=a \operatorname{det}(A)$.

Proof: Denote $M$ by $\left(m_{i j}\right)$. Thus in the first case, $m_{n n}=a$ and $m_{n i}=0$ if $i \neq n$ while in the second case, $m_{n n}=a$ and $m_{i n}=0$ if $i \neq n$. From the definition of the determinant,

$$
\operatorname{det}(M) \equiv \sum_{\left(k_{1}, \cdots, k_{n}\right)} \operatorname{sgn}_{n}\left(k_{1}, \cdots, k_{n}\right) m_{1 k_{1}} \cdots m_{n k_{n}}
$$

Letting $\theta$ denote the position of $n$ in the ordered list, $\left(k_{1}, \cdots, k_{n}\right)$ then using the earlier conventions used to prove Lemma 8.1.1, $\operatorname{det}(M)$ equals

$$
\sum_{\left(k_{1}, \cdots, k_{n}\right)}(-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_{1}, \cdots, k_{\theta-1}, k_{\theta+1}^{\theta}, \cdots, \stackrel{n-1}{k_{n}}\right) m_{1 k_{1}} \cdots m_{n k_{n}}
$$

Now suppose the second case. Then if $k_{n} \neq n$, the term involving $m_{n k_{n}}$ in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta=n$ or in other words, those for which $k_{n}=n$. Therefore, the above expression reduces to

$$
a \sum_{\left(k_{1}, \cdots, k_{n-1}\right)} \operatorname{sgn}_{n-1}\left(k_{1}, \cdots k_{n-1}\right) m_{1 k_{1}} \cdots m_{(n-1) k_{n-1}}=a \operatorname{det}(A) .
$$

To get the assertion in the first case, use Corollary 8.3.1 to write

$$
\operatorname{det}(M)=\operatorname{det}\left(M^{T}\right)=\operatorname{det}\left(\left(\begin{array}{cc}
A^{T} & \mathbf{0} \\
* & a
\end{array}\right)\right)=a \operatorname{det}\left(A^{T}\right)=a \operatorname{det}(A)
$$

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

Definition 8.5.2 Let $A=\left(a_{i j}\right)$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix $\operatorname{cof}(A)$ is defined by $\operatorname{cof}(A)=\left(c_{i j}\right)$ where to obtain $c_{i j}$ delete the $i^{\text {th }}$ row and the $j^{\text {th }}$ column of $A$, take the determinant of the $(n-1) \times(n-1)$ matrix which results, (This is called the $i j^{\text {th }}$ minor of $A$. ) and then multiply this number by $(-1)^{i+j}$. To make the formulas easier to remember, $\operatorname{cof}(A)_{i j}$ will denote the $i j^{\text {th }}$ entry of the cofactor matrix.

The following is the main result.

Theorem 8.5.3 Let $A$ be an $n \times n$ matrix where $n \geq 2$. Then

$$
\begin{equation*}
\operatorname{det}(A)=\sum_{j=1}^{n} a_{i j} \operatorname{cof}(A)_{i j}=\sum_{i=1}^{n} a_{i j} \operatorname{cof}(A)_{i j} \tag{8.11}
\end{equation*}
$$

The first formula consists of expanding the determinant along the $i^{\text {th }}$ row and the second expands the determinant along the $j^{\text {th }}$ column.

Proof: Let $\left(a_{i 1}, \cdots, a_{i n}\right)$ be the $i^{t h}$ row of $A$. Let $B_{j}$ be the matrix obtained from $A$ by leaving every row the same except the $i^{\text {th }}$ row which in $B_{j}$ equals $\left(0, \cdots, 0, a_{i j}, 0, \cdots, 0\right)$. Then by Corollary 8.3.2, $\operatorname{det}(A)=\sum_{j=1}^{n} \operatorname{det}\left(B_{j}\right)$. For example if

$$
A=\left(\begin{array}{lll}
a & b & c \\
d & e & f \\
h & i & j
\end{array}\right)
$$

and $i=2$, then

$$
B_{1}=\left(\begin{array}{lll}
a & b & c \\
d & 0 & 0 \\
h & i & j
\end{array}\right), B_{2}=\left(\begin{array}{lll}
a & b & c \\
0 & e & 0 \\
h & i & j
\end{array}\right), B_{3}=\left(\begin{array}{lll}
a & b & c \\
0 & 0 & f \\
h & i & j
\end{array}\right)
$$

Denote by $A^{i j}$ the $(n-1) \times(n-1)$ matrix obtained by deleting the $i^{t h}$ row and the $j^{\text {th }}$ column of $A$. Thus $\operatorname{cof}(A)_{i j} \equiv(-1)^{i+j} \operatorname{det}\left(A^{i j}\right)$. At this point, recall that from Proposition 8.2.3, when two rows or two columns in a matrix $M$, are switched, this results in multiplying the determinant of the old matrix by -1 to get the determinant of the new matrix. Therefore, by Lemma 8.5.1,

$$
\begin{aligned}
\operatorname{det}\left(B_{j}\right) & =(-1)^{n-j}(-1)^{n-i} \operatorname{det}\left(\left(\begin{array}{cc}
A^{i j} & * \\
0 & a_{i j}
\end{array}\right)\right) \\
& =(-1)^{i+j} \operatorname{det}\left(\left(\begin{array}{cc}
A^{i j} & * \\
0 & a_{i j}
\end{array}\right)\right)=a_{i j} \operatorname{cof}(A)_{i j}
\end{aligned}
$$

Therefore, $\operatorname{det}(A)=\sum_{j=1}^{n} a_{i j} \operatorname{cof}(A)_{i j}$ which is the formula for expanding $\operatorname{det}(A)$ along the $i^{\text {th }}$ row. Also,

$$
\operatorname{det}(A)=\operatorname{det}\left(A^{T}\right)=\sum_{j=1}^{n} a_{i j}^{T} \operatorname{cof}\left(A^{T}\right)_{i j}=\sum_{j=1}^{n} a_{j i} \operatorname{cof}(A)_{j i}
$$

which is the formula for expanding $\operatorname{det}(A)$ along the $i^{\text {th }}$ column.

### 8.6 A Formula for the Inverse

Note that this gives an easy way to write a formula for the inverse of an $n \times n$ matrix.
Theorem 8.6.1 $A^{-1}$ exists if and only if $\operatorname{det}(A) \neq 0$. If $\operatorname{det}(A) \neq 0$, then the $i j$ th entry of $A^{-1}$ is given by $a_{i j}^{-1}$ where $a_{i j}^{-1}=\operatorname{det}(A)^{-1} \operatorname{cof}(A)_{j i}$ for $\operatorname{cof}(A)_{i j}$ the $i j^{\text {th }}$ cofactor of $A$.

Proof: By Theorem 8.5.3 and letting $\left(a_{i r}\right)=A$, if $\operatorname{det}(A) \neq 0$,

$$
\sum_{i=1}^{n} a_{i r} \operatorname{cof}(A)_{i r} \operatorname{det}(A)^{-1}=\operatorname{det}(A) \operatorname{det}(A)^{-1}=1
$$

Now in the matrix $A$, replace the $k^{t h}$ column with the $r^{\text {th }}$ column and then expand along the $k^{t h}$ column. This yields for $k \neq r, \sum_{i=1}^{n} a_{i r} \operatorname{cof}(A)_{i k} \operatorname{det}(A)^{-1}=0$ by Corollary 8.3.2 because there are two equal columns. Summarizing, $\sum_{i=1}^{n} a_{i r} \operatorname{cof}(A)_{i k} \operatorname{det}(A)^{-1}=\delta_{r k}$. Using the other formula in Theorem 8.5.3, and similar reasoning, $\sum_{j=1}^{n} a_{r j} \operatorname{cof}(A)_{k j} \operatorname{det}(A)^{-1}=$ $\delta_{r k}$. This proves that if $\operatorname{det}(A) \neq 0$, then $A^{-1}$ exists with $A^{-1}=\left(a_{i j}^{-1}\right)$, where $a_{i j}^{-1}=$ $\operatorname{cof}(A)_{j i} \operatorname{det}(A)^{-1}$.

Now suppose $A^{-1}$ exists. Then by Theorem 8.4.4,

$$
1=\operatorname{det}(I)=\operatorname{det}\left(A A^{-1}\right)=\operatorname{det}(A) \operatorname{det}\left(A^{-1}\right)
$$

so $\operatorname{det}(A) \neq 0$.
The next corollary points out that if an $n \times n$ matrix $A$ has a right or a left inverse, then it has an inverse.

Corollary 8.6.2 Let $A$ be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix $B$ such that $B A=I$. Then $A^{-1}$ exists and $A^{-1}=B$. Also, if there exists $C$ an $n \times n$ matrix such that $A C=I$, then $A^{-1}$ exists and $A^{-1}=C$.

Proof: Since $B A=I$, Theorem 8.4.4 implies $\operatorname{det} B \operatorname{det} A=1$ and so $\operatorname{det} A \neq 0$. Therefore from Theorem 8.6.1, $A^{-1}$ exists. Therefore,

$$
A^{-1}=(B A) A^{-1}=B\left(A A^{-1}\right)=B I=B
$$

The case where $C A=I$ is handled similarly.
The conclusion of this corollary is that left inverses, right inverses and inverses are all the same in the context of $n \times n$ matrices.

Theorem 8.6.1 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix $A$. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, $A^{-1}$ is equal to one over the determinant of $A$ times the adjugate matrix of $A$.

### 8.6.1 Cramer's Rule

In case you are solving a system of equations, $\boldsymbol{A x}=\boldsymbol{y}$ for $\boldsymbol{x}$, it follows that if $A^{-1}$ exists,

$$
\boldsymbol{x}=\left(A^{-1} A\right) \boldsymbol{x}=A^{-1}(A \boldsymbol{x})=A^{-1} \boldsymbol{y}
$$

thus solving the system. Now in the case that $A^{-1}$ exists, there is a formula for $A^{-1}$ given above. Using this formula,

$$
x_{i}=\sum_{j=1}^{n} a_{i j}^{-1} y_{j}=\sum_{j=1}^{n} \frac{1}{\operatorname{det}(A)} \operatorname{cof}(A)_{j i} y_{j} .
$$

By the formula for the expansion of a determinant along a column,

$$
x_{i}=\frac{1}{\operatorname{det}(A)} \operatorname{det}\left(\begin{array}{ccccc}
* & \cdots & y_{1} & \cdots & * \\
\vdots & & \vdots & & \vdots \\
* & \cdots & y_{n} & \cdots & *
\end{array}\right)
$$

where here the $i^{\text {th }}$ column of $A$ is replaced with the column vector, $\left(y_{1} \cdots, y_{n}\right)^{T}$, and the determinant of this modified matrix is taken and divided by $\operatorname{det}(A)$. This formula is known as Cramer's rule.

Definition 8.6.3 A matrix $M$, is upper triangular if $M_{i j}=0$ whenever $i>j$. Thus such a matrix equals zero below the main diagonal, the entries of the form $M_{i i}$ as shown.

$$
\left(\begin{array}{cccc}
* & * & \cdots & * \\
0 & * & \ddots & \vdots \\
\vdots & \ddots & \ddots & * \\
0 & \cdots & 0 & *
\end{array}\right)
$$

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

With this definition, here is a simple corollary of Theorem 8.5.3.
Corollary 8.6.4 Let $M$ be an upper (lower) triangular matrix. Then $\operatorname{det}(M)$ is obtained by taking the product of the entries on the main diagonal.

### 8.6.2 An Identity of Cauchy

Theorem 8.6.5 Both the left and the right sides in the following yield the same polynomial in the variables $a_{i}, b_{i}$ for $i \leq n$.

$$
\prod_{i, j}\left(a_{i}+b_{j}\right)\left|\begin{array}{ccc}
\frac{1}{a_{1}+b_{1}} & \cdots & \frac{1}{a_{1}+b_{n}}  \tag{8.12}\\
\vdots & & \vdots \\
\frac{1}{a_{n}+b_{1}} & \cdots & \frac{1}{a_{n}+b_{n}}
\end{array}\right|=\prod_{j<i}\left(a_{i}-a_{j}\right)\left(b_{i}-b_{j}\right)
$$

Proof: The theorem is true if $n=2$. This follows from some computations. Suppose it is true for $n-1, n \geq 3$.

$$
\begin{aligned}
& \left|\begin{array}{cccc}
\frac{1}{a_{1}+b_{1}} & \frac{1}{a_{1}+b_{2}} & \cdots & \frac{1}{a_{1}+b_{n}} \\
\vdots & \vdots & \cdots & \vdots \\
\frac{1}{a_{n-1}+b_{1}} & \frac{1}{a_{n-1}+b_{2}} & & \frac{1}{a_{n-1}+b_{n}} \\
\frac{1}{a_{n}+b_{1}} & \frac{1}{a_{n}+b_{2}} & \cdots & \frac{1}{a_{n}+b_{n}}
\end{array}\right| \\
& =\left|\begin{array}{cccc}
\frac{a_{n}-a_{1}}{\left(a_{1}+b_{1}\right)\left(b_{1}+a_{n}\right)} & \frac{a_{n}-a_{1}}{\left(a_{1}+b_{2}\right)\left(b_{2}+a_{n}\right)} & \cdots & \frac{a_{n}-a_{1}}{\left(a_{1}+b_{n}\right)\left(a_{n}+b_{n}\right)} \\
\vdots & \vdots & \ldots & \vdots \\
\frac{a_{n}-a_{n-1}}{\left(a_{n-1}+b_{1}\right)\left(a_{n}+b_{1}\right)} & \frac{a_{n}-a_{n-1}}{\left(b_{2}+a_{n}\right)\left(b_{2}+a_{n-1}\right)} & & \frac{a_{n}-a_{n-1}}{\left(a_{n}+b_{n}\right)\left(b_{n}+a_{n-1}\right)} \\
\frac{1}{a_{n}+b_{1}} & \frac{1}{a_{n}+b_{2}} & \cdots & \frac{1}{a_{n}+b_{n}}
\end{array}\right|
\end{aligned}
$$

Continuing to use the multilinear properties of determinants, this equals

$$
\left.\left|\begin{array}{cccc}
\frac{1}{\left(a_{1}+b_{1}\right)\left(b_{1}+a_{n}\right)} & \frac{1}{\left(a_{1}+b_{2}\right)\left(b_{2}+a_{n}\right)} & \cdots & \frac{1}{\left(a_{1}+b_{n}\right)\left(a_{n}+b_{n}\right)} \\
\vdots & \vdots & \cdots & \vdots \\
\frac{1}{\left(a_{n-1}+b_{1}\right)\left(a_{n}+b_{1}\right)} & \frac{1}{\left(b_{2}+a_{n}\right)\left(b_{2}+a_{n-1}\right)} & & \frac{1}{\left(a_{n}+b_{n}\right)\left(b_{n}+a_{n-1}\right)} \\
\frac{1}{a_{n}+b_{1}} & \frac{1}{a_{n}+b_{2}} & \cdots & \frac{1}{a_{n}+b_{n}}
\end{array}\right| \begin{aligned}
& n=1
\end{aligned} \right\rvert\,
$$

and this equals

$$
\left|\begin{array}{cccc}
\frac{1}{\left(a_{1}+b_{1}\right)} & \frac{1}{\left(a_{1}+b_{2}\right)} & \cdots & \frac{1}{\left(a_{1}+b_{n}\right)} \\
\vdots & \vdots & \cdots & \vdots \\
\frac{1}{\left(a_{n-1}+b_{1}\right)} & \frac{1}{\left(b_{2}+a_{n-1}\right)} & & \frac{1}{\left(b_{n}+a_{n-1}\right)} \\
1 & 1 & \cdots & 1
\end{array}\right| \begin{aligned}
& \prod_{k=1}^{n-1}\left(a_{n}-a_{k}\right) \\
& \prod_{k=1}^{n}\left(a_{n}+b_{k}\right) \\
&
\end{aligned}
$$

Now take -1 times the last column and add to each previous column. Thus it equals

$$
\left|\begin{array}{cccc|}
\frac{b_{n}-b_{1}}{\left(a_{1}+b_{1}\right)\left(a_{1}+b_{n}\right)} & \frac{b_{n}-b_{2}}{\left(a_{1}+b_{2}\right)\left(a_{1}+b_{n}\right)} & \cdots & \frac{1}{\left(a_{1}+b_{n}\right)} \\
\vdots & \vdots & \cdots & \vdots \\
\frac{b_{n}-b_{1}}{\left(b_{1}+a_{n-1}\right)\left(b_{n}+a_{n-1}\right)} & \frac{b_{n}-b_{2}}{\left(b_{2}+a_{n-1}\right)\left(b_{n}+a_{n-1}\right)} & & \frac{1}{\left(a_{n-1}+b_{n}\right)} \\
0 & 0 & \cdots & 1
\end{array}\right| \begin{aligned}
& \prod_{k=1}^{n-1}\left(a_{n}-a_{k}\right) \\
& \prod_{k=1}^{n}\left(a_{n}+b_{k}\right) \\
&
\end{aligned}
$$

Now continue simplifying using the multilinear property of the determinant.

$$
\left|\begin{array}{cccc|}
\frac{1}{\left(a_{1}+b_{1}\right)} & \frac{1}{\left(a_{1}+b_{2}\right)} & \cdots & 1 \\
\vdots & \vdots & \cdots & \vdots \\
\frac{1}{\left(b_{1}+a_{n-1}\right)} & \frac{1}{\left(b_{2}+a_{n-1}\right)} & & 1 \\
0 & 0 & \cdots & 1
\end{array}\right| \frac{\prod_{k=1}^{n-1}\left(a_{n}-a_{k}\right)}{\prod_{k=1}^{n}\left(a_{n}+b_{k}\right)} \frac{\prod_{k=1}^{n-1}\left(b_{n}-b_{k}\right)}{\prod_{k=1}^{n-1}\left(a_{k}+b_{n}\right)}
$$

Now, expanding along the bottom row, what has just resulted is

$$
\left|\begin{array}{ccc}
\frac{1}{a_{1}+b_{1}} & \cdots & \frac{1}{a_{1}+b_{n-1}} \\
\vdots & \cdots & \vdots \\
\frac{1}{a_{n-1}+b_{1}} & \cdots & \frac{1}{a_{n-1}+b_{n-1}}
\end{array}\right| \frac{\prod_{k=1}^{n-1}\left(a_{n}-a_{k}\right)}{\prod_{k=1}^{n}\left(a_{n}+b_{k}\right)} \frac{\prod_{k=1}^{n-1}\left(b_{n}-b_{k}\right)}{\prod_{k=1}^{n-1}\left(a_{k}+b_{n}\right)}
$$

By induction this equals

$$
\begin{gathered}
\frac{\prod_{k=1}^{n-1}\left(a_{n}-a_{k}\right)}{\prod_{k=1}^{n}\left(a_{n}+b_{k}\right)} \frac{\prod_{k=1}^{n-1}\left(b_{n}-b_{k}\right)}{\prod_{k=1}^{n-1}\left(a_{k}+b_{n}\right)} \frac{\prod_{j<i \leq n-1}\left(a_{i}-a_{j}\right)\left(b_{i}-b_{j}\right)}{\prod_{i, j \leq n-1}\left(a_{i}+b_{j}\right)} \\
=\frac{\prod_{j<i \leq n}\left(a_{i}-a_{j}\right)\left(b_{i}-b_{j}\right)}{\prod_{i, j \leq n}\left(a_{i}+b_{j}\right)}
\end{gathered}
$$

### 8.7 Rank of a Matrix

Definition 8.7.1 A submatrix of a matrix $A$ is the rectangular array of numbers obtained by deleting some rows and columns of $A$. Let $A$ be an $m \times n$ matrix. The determinant rank of the matrix equals $r$ where $r$ is the largest number such that some $r \times r$ submatrix of $A$ has a non zero determinant. The row rank is defined to be the dimension of the span of the rows. The column rank is defined to be the dimension of the span of the columns.

Theorem 8.7.2 If $A$, an $m \times n$ matrix has determinant rank $r$, then there exist $r$ rows of the matrix such that every other row is a linear combination of these $r$ rows.

Proof: Suppose the determinant rank of $A=\left(a_{i j}\right)$ equals $r$. Thus some $r \times r$ submatrix has non zero determinant and there is no larger square submatrix which has non zero determinant. Suppose such a submatrix is determined by the $r$ columns whose indices are $j_{1}<\cdots<j_{r}$ and the $r$ rows whose indices are $i_{1}<\cdots<i_{r}$. I want to show that every row is a linear combination of these rows. Consider the $l^{t h}$ row and let $p$ be an index between 1 and $n$. Form the following $(r+1) \times(r+1)$ matrix

$$
\left(\begin{array}{llll}
a_{i_{1} j_{1}} & \cdots & a_{i_{1} j_{r}} & a_{i_{1} p} \\
\vdots & & \vdots & \vdots \\
a_{i_{r} j_{1}} & \cdots & a_{i_{r} j_{r}} & a_{i_{r} p} \\
a_{l j_{1}} & \cdots & a_{l j_{r}} & a_{l p}
\end{array}\right)
$$

Of course you can assume $l \notin\left\{i_{1}, \cdots, i_{r}\right\}$ because there is nothing to prove if the $l^{\text {th }}$ row is one of the chosen ones. The above matrix has determinant 0 . This is because if $p \notin$ $\left\{j_{1}, \cdots, j_{r}\right\}$ then the above would be a submatrix of $A$ which is too large to have non zero determinant. On the other hand, if $p \in\left\{j_{1}, \cdots, j_{r}\right\}$ then the above matrix has two columns which are equal so its determinant is still 0 .

Expand the determinant of the above matrix along the last column. Let $C_{k}$ denote the cofactor associated with the entry $a_{i_{k}} p$. This is not dependent on the choice of $p$. Remember, you delete the column and the row the entry is in and take the determinant of what is left and multiply by -1 raised to an appropriate power. Let $C$ denote the cofactor associated with $a_{l p}$. This is given to be nonzero, it being the determinant of the matrix $r \times r$ matrix in the upper left corner. Thus

$$
0=a_{l p} C+\sum_{k=1}^{r} C_{k} a_{i_{k} p}
$$

which implies

$$
a_{l p}=\sum_{k=1}^{r} \frac{-C_{k}}{C} a_{i_{k} p} \equiv \sum_{k=1}^{r} m_{k} a_{i_{k} p}
$$

Since this is true for every $p$ and since $m_{k}$ does not depend on $p$, this has shown the $l^{\text {th }}$ row is a linear combination of the $i_{1}, i_{2}, \cdots, i_{r}$ rows.

Corollary 8.7.3 The determinant rank equals the row rank.
Proof: From Theorem 8.7.2, every row is in the span of $r$ rows where $r$ is the determinant rank. Therefore, the row rank (dimension of the span of the rows) is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, it
follows from Theorem 8.7.2 that there exist $p$ rows for $p<r \equiv$ determinant rank, such that the span of these $p$ rows equals the row space. But then you could consider the $r \times r$ sub matrix which determines the determinant rank and it would follow that each of these rows would be in the span of the restrictions of the $p$ rows just mentioned. By Theorem 3.1.5, the exchange theorem, the rows of this sub matrix would not be linearly independent and so some row is a linear combination of the others. By Corollary 8.4.2 the determinant would be 0 , a contradiction.

Corollary 8.7.4 If A has determinant rank $r$, then there exist $r$ columns of the matrix such that every other column is a linear combination of these $r$ columns. Also the column rank equals the determinant rank.

Proof: This follows from the above by considering $A^{T}$. The rows of $A^{T}$ are the columns of $A$ and the determinant rank of $A^{T}$ and $A$ are the same. Therefore, from Corollary 8.7.3, column rank of $A=$ row rank of $A^{T}=$ determinant rank of $A^{T}=$ determinant rank of $A$.

The following theorem is of fundamental importance and ties together many of the ideas presented above.

Theorem 8.7.5 Let A be an $n \times n$ matrix. Then the following are equivalent.

1. $\operatorname{det}(A)=0$.
2. $A, A^{T}$ are not one to one.
3. A is not onto.

Proof: Suppose $\operatorname{det}(A)=0$. Then the determinant rank of $A=r<n$. Therefore, there exist $r$ columns such that every other column is a linear combination of these columns by Theorem 8.7.2. In particular, it follows that for some $m$, the $m^{\text {th }}$ column is a linear combination of all the others. Thus letting $A=\left(\begin{array}{lllll}a_{1} & \cdots & a_{m} & \cdots & a_{n}\end{array}\right)$ where the columns are denoted by $\boldsymbol{a}_{i}$, there exists scalars $\alpha_{i}$ such that

$$
\boldsymbol{a}_{m}=\sum_{k \neq m} \alpha_{k} \boldsymbol{a}_{k}
$$

Now consider the column vector, $\boldsymbol{x} \equiv\left(\begin{array}{lllll}\alpha_{1} & \cdots & -1 & \cdots & \alpha_{n}\end{array}\right)^{T}$. Then

$$
A \boldsymbol{x}=-\boldsymbol{a}_{m}+\sum_{k \neq m} \alpha_{k} \boldsymbol{a}_{k}=\mathbf{0}
$$

Since also $A \mathbf{0}=\mathbf{0}$, it follows $A$ is not one to one. Similarly, $A^{T}$ is not one to one by the same argument applied to $A^{T}$. This verifies that 1.) implies 2.).

Now suppose 2.). Then since $A^{T}$ is not one to one, it follows there exists $\boldsymbol{x} \neq \mathbf{0}$ such that $A^{T} \boldsymbol{x}=\mathbf{0}$. Taking the transpose of both sides yields $\boldsymbol{x}^{T} A=\mathbf{0}^{T}$ where the $\mathbf{0}^{T}$ is a $1 \times n$ matrix or row vector. Now if $A \boldsymbol{y}=\boldsymbol{x}$, then

$$
|\boldsymbol{x}|^{2}=\boldsymbol{x}^{T}(A \boldsymbol{y})=\left(\boldsymbol{x}^{T} A\right) \boldsymbol{y}=\mathbf{0} \boldsymbol{y}=0
$$

contrary to $\boldsymbol{x} \neq \mathbf{0}$. Consequently there can be no $\boldsymbol{y}$ such that $A \boldsymbol{y}=\boldsymbol{x}$ and so $A$ is not onto. This shows that 2.) implies 3.).

Finally, suppose 3.). If 1.) does not hold, then $\operatorname{det}(A) \neq 0$ but then from Theorem 8.6.1 $A^{-1}$ exists and so for every $\boldsymbol{y} \in \mathbb{F}^{n}$ there exists a unique $\boldsymbol{x} \in \mathbb{F}^{n}$ such that $A \boldsymbol{x}=\boldsymbol{y}$. In fact $\boldsymbol{x}=A^{-1} \boldsymbol{y}$. Thus $A$ would be onto contrary to 3 .). This shows 3 .) implies 1.).

Corollary 8.7.6 Let A be an $n \times n$ matrix. Then the following are equivalent.

1. $\operatorname{det}(A) \neq 0$.
2. A and $A^{T}$ are one to one.
3. A is onto.

Proof: This follows immediately from the above theorem.

### 8.8 Summary of Determinants

In all the following $A, B$ are $n \times n$ matrices

1. $\operatorname{det}(A)$ is a number.
2. $\operatorname{det}(A)$ is linear in each row and in each column.
3. If you switch two rows or two columns, the determinant of the resulting matrix is -1 times the determinant of the unswitched matrix. (This and the previous one say

$$
\left(\boldsymbol{a}_{1} \cdots \boldsymbol{a}_{n}\right) \rightarrow \operatorname{det}\left(\boldsymbol{a}_{1} \cdots \boldsymbol{a}_{n}\right)
$$

is an alternating multilinear function or alternating tensor.
4. $\operatorname{det}\left(e_{1}, \cdots, e_{n}\right)=1$.
5. $\operatorname{det}(A B)=\operatorname{det}(A) \operatorname{det}(B)$
6. det $(A)$ can be expanded along any row or any column and the same result is obtained.
7. $\operatorname{det}(A)=\operatorname{det}\left(A^{T}\right)$
8. $A^{-1}$ exists if and only if $\operatorname{det}(A) \neq 0$ and in this case

$$
\begin{equation*}
\left(A^{-1}\right)_{i j}=\frac{1}{\operatorname{det}(A)} \operatorname{cof}(A)_{j i} \tag{8.13}
\end{equation*}
$$

9. Determinant rank, row rank and column rank are all the same number for any $m \times n$ matrix.

### 8.9 The Cayley Hamilton Theorem

Here is a simple proof of the Cayley Hamilton theorem in the special case that the field of scalars is $\mathbb{R}, \mathbb{Q}$, or $\mathbb{C}$. This proof does not work for arbitrary fields. A proof of this theorem valid for every field will be outlined in exercises. See Problem 21 on Page 191. The cases considered here comprise most major applications of the Cayley Hamilton theorem.

Definition 8.9.1 Let $A$ be an $n \times n$ matrix. The characteristic polynomial is defined as

$$
q_{A}(t) \equiv \operatorname{det}(t I-A)
$$

and the solutions to $q_{A}(t)=0$ are called eigenvalues. For $A$ a matrix and $p(t)=t^{n}+$ $a_{n-1} t^{n-1}+\cdots+a_{1} t+a_{0}$, denote by $p(A)$ the matrix defined by

$$
p(A) \equiv A^{n}+a_{n-1} A^{n-1}+\cdots+a_{1} A+a_{0} I .
$$

The explanation for the last term is that $A^{0}$ is interpreted as $I$, the identity matrix. This is always the characteristic polynomial, but in this section, the field will be one of those mentioned above.

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by $q_{A}(t)=0$. It is one of the most important theorems in linear algebra ${ }^{1}$. The proof in this section is not the most general proof, but works well when the field of scalars is $\mathbb{R}$ or $\mathbb{C}$. The following lemma will help with its proof.

Lemma 8.9.2 Suppose for all $|\lambda|$ large enough,

$$
A_{0}+A_{1} \lambda+\cdots+A_{m} \lambda^{m}=0
$$

where the $A_{i}$ are $n \times n$ matrices. Then each $A_{i}=0$.
Proof: Suppose some $A_{i} \neq 0$. Let $p$ be the largest index of those which are non zero. Then multiply by $\lambda^{-p}$.

$$
A_{0} \lambda^{-p}+A_{1} \lambda^{-p+1}+\cdots+A_{p-1} \lambda^{-1}+A_{p}=0
$$

Now let $\lambda \rightarrow \infty$. Thus $A_{p}=0$ after all. Hence each $A_{i}=0$.
With the lemma, here is a simple corollary.
Corollary 8.9.3 Let $A_{i}$ and $B_{i}$ be $n \times n$ matrices and suppose

$$
A_{0}+A_{1} \lambda+\cdots+A_{m} \lambda^{m}=B_{0}+B_{1} \lambda+\cdots+B_{m} \lambda^{m}
$$

for all $|\lambda|$ large enough. Then $A_{i}=B_{i}$ for all $i$. If $A_{i}=B_{i}$ for each $A_{i}, B_{i}$ then one can substitute an $n \times n$ matrix $M$ for $\lambda$ and the identity will continue to hold.

Proof: Subtract and use the result of the lemma. The last claim is obvious by matching terms.

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.
Theorem 8.9.4 Let $A$ be an $n \times n$ matrix and let $q(\lambda) \equiv \operatorname{det}(\lambda I-A)$ be the characteristic polynomial. Then $q(A)=0$.

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I-A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then $\lambda$ cannot be in the finite list of eigenvalues of $A$ and so for such $\lambda,(\lambda I-A)^{-1}$ exists.) Therefore, by Theorem 8.6.1, $C(\lambda)=q(\lambda)(\lambda I-A)^{-1}$. Say $q(\lambda)=a_{0}+a_{1} \lambda+\cdots+\lambda^{n}$. Note that each entry in $C(\lambda)$ is a polynomial in $\lambda$ having degree no more than $n-1$. For example, you might have something like

$$
C(\lambda)=\left(\begin{array}{ccc}
\lambda^{2}-6 \lambda+9 & 3-\lambda & 0 \\
2 \lambda-6 & \lambda^{2}-3 \lambda & 0 \\
\lambda-1 & \lambda-1 & \lambda^{2}-3 \lambda+2
\end{array}\right)
$$

[^6]\[

=\left($$
\begin{array}{ccc}
9 & 3 & 0 \\
-6 & 0 & 0 \\
-1 & -1 & 2
\end{array}
$$\right)+\lambda\left($$
\begin{array}{ccc}
-6 & -1 & 0 \\
2 & -3 & 0 \\
1 & 1 & -3
\end{array}
$$\right)+\lambda^{2}\left($$
\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}
$$\right)
\]

Therefore, collecting the terms in the general case, $C(\lambda)=C_{0}+C_{1} \lambda+\cdots+C_{n-1} \lambda^{n-1}$ for $C_{j}$ some $n \times n$ matrix. Then $C(\lambda)(\lambda I-A)=\left(C_{0}+C_{1} \lambda+\cdots+C_{n-1} \lambda^{n-1}\right)(\lambda I-A)=$ $q(\lambda) I$. Then multiplying out the middle term, it follows that for all $|\lambda|$ sufficiently large,

$$
\begin{gathered}
a_{0} I+a_{1} I \lambda+\cdots+I \lambda^{n}=C_{0} \lambda+C_{1} \lambda^{2}+\cdots+C_{n-1} \lambda^{n} \\
-\left[C_{0} A+C_{1} A \lambda+\cdots+C_{n-1} A \lambda^{n-1}\right] \\
=-C_{0} A+\left(C_{0}-C_{1} A\right) \lambda+\left(C_{1}-C_{2} A\right) \lambda^{2}+\cdots+\left(C_{n-2}-C_{n-1} A\right) \lambda^{n-1}+C_{n-1} \lambda^{n}
\end{gathered}
$$

Then, using Corollary 8.9.3, one can replace $\lambda$ on both sides with $A$. Then the right side is seen to equal 0 . Hence the left side, $q(A) I$ is also equal to 0 .

Here is an interesting and significant application of block multiplication. In this theorem, $q_{M}(t)$ denotes the characteristic polynomial, $\operatorname{det}(t I-M)$. The zeros of this polynomial will be shown later to be eigenvalues of the matrix $M$. First note that from block multiplication, for the following block matrices consisting of square blocks of an appropriate size,

$$
\begin{gathered}
\left(\begin{array}{cc}
A & 0 \\
B & C
\end{array}\right)=\left(\begin{array}{cc}
A & 0 \\
B & I
\end{array}\right)\left(\begin{array}{cc}
I & 0 \\
0 & C
\end{array}\right) \text { so } \\
\operatorname{det}\left(\begin{array}{cc}
A & 0 \\
B & C
\end{array}\right)=\operatorname{det}\left(\begin{array}{cc}
A & 0 \\
B & I
\end{array}\right) \operatorname{det}\left(\begin{array}{cc}
I & 0 \\
0 & C
\end{array}\right)=\operatorname{det}(A) \operatorname{det}(C)
\end{gathered}
$$

Theorem 8.9.5 Let $A$ be an $m \times n$ matrix and let $B$ be an $n \times m$ matrix for $m \leq n$. Then $q_{B A}(t)=t^{n-m} q_{A B}(t)$, so the eigenvalues of $B A$ and $A B$ are the same including multiplicities except that BA has $n-m$ extra zero eigenvalues. Here $q_{A}(t)$ denotes the characteristic polynomial of the matrix $A$.

Proof: Use block multiplication to write

$$
\begin{gathered}
\left(\begin{array}{cc}
A B & 0 \\
B & 0
\end{array}\right)\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)=\left(\begin{array}{cc}
A B & A B A \\
B & B A
\end{array}\right) \\
\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)\left(\begin{array}{cc}
0 & 0 \\
B & B A
\end{array}\right)=\left(\begin{array}{cc}
A B & A B A \\
B & B A
\end{array}\right) \\
\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)\left(\begin{array}{cc}
0 & 0 \\
B & B A
\end{array}\right)=\left(\begin{array}{cc}
A B & 0 \\
B & 0
\end{array}\right)\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)
\end{gathered}
$$

Therefore,

$$
\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)^{-1}\left(\begin{array}{cc}
A B & 0 \\
B & 0
\end{array}\right)\left(\begin{array}{cc}
I & A \\
0 & I
\end{array}\right)=\left(\begin{array}{cc}
0 & 0 \\
B & B A
\end{array}\right)
$$

Since the two matrices above are similar, it follows that

$$
\left(\begin{array}{cc}
0_{m \times m} & 0 \\
B & B A
\end{array}\right),\left(\begin{array}{cc}
A B & 0 \\
B & 0_{n \times n}
\end{array}\right)
$$

have the same characteristic polynomials.Thus

$$
\operatorname{det}\left(\begin{array}{cc}
t I_{m \times m} & 0  \tag{8.14}\\
-B & t I-B A
\end{array}\right)=\operatorname{det}\left(\begin{array}{cc}
t I-A B & 0 \\
-B & t I_{n \times n}
\end{array}\right)
$$

Therefore, $t^{m} \operatorname{det}(t I-B A)=t^{n} \operatorname{det}(t I-A B)$ and so

$$
\operatorname{det}(t I-B A)=q_{B A}(t)=t^{n-m} \operatorname{det}(t I-A B)=t^{n-m} q_{A B}(t)
$$

### 8.10 Exercises

1. Let $m<n$ and let $A$ be an $m \times n$ matrix. Show that $A$ is not one to one. Hint: Consider the $n \times n$ matrix $A_{1}$ which is of the form $A_{1} \equiv\binom{A}{0}$ where the 0 denotes an $(n-m) \times n$ matrix of zeros. Thus $\operatorname{det} A_{1}=0$ and so $A_{1}$ is not one to one. Now observe that $A_{1} \boldsymbol{x}$ is the vector, $A_{1} \boldsymbol{x}=\binom{A \boldsymbol{x}}{0}$ which equals zero if and only if $A x=0$.
2. Let $\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}$ be vectors in $\mathbb{F}^{n}$ and let $M\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right)$ denote the matrix whose $i^{\text {th }}$ column equals $\boldsymbol{v}_{i}$. Define $d\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right) \equiv \operatorname{det}\left(M\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right)\right)$. Prove that $d$ is linear in each variable, (multilinear), that

$$
\begin{equation*}
d\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{i}, \cdots, \boldsymbol{v}_{j}, \cdots, \boldsymbol{v}_{n}\right)=-d\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{j}, \cdots, \boldsymbol{v}_{i}, \cdots, \boldsymbol{v}_{n}\right) \tag{8.15}
\end{equation*}
$$

and

$$
\begin{equation*}
d\left(e_{1}, \cdots, e_{n}\right)=1 \tag{8.16}
\end{equation*}
$$

where here $\boldsymbol{e}_{j}$ is the vector in $\mathbb{F}^{n}$ which has a zero in every position except the $j^{\text {th }}$ position in which it has a one.
3. If $A, B$ are similar matrices, show that they have the same determinant. Also show that they have the same characteristic polynomial.
4. Suppose $f: \mathbb{F}^{n} \times \cdots \times \mathbb{F}^{n} \rightarrow \mathbb{F}$ satisfies 8.15 and 8.16 and is linear in each variable. Show that $f=d$.
5. Use row operations to evaluate by hand the determinant

$$
\operatorname{det}\left(\begin{array}{cccc}
1 & 2 & 3 & 2 \\
-6 & 3 & 2 & 3 \\
5 & 2 & 2 & 3 \\
3 & 4 & 6 & 4
\end{array}\right)
$$

6. Find the inverse if it exists of the matrix

$$
\left(\begin{array}{ccc}
e^{t} & \cos t & \sin t \\
e^{t} & -\sin t & \cos t \\
e^{t} & -\cos t & -\sin t
\end{array}\right)
$$

7. Let $L y=y^{(n)}+a_{n-1}(x) y^{(n-1)}+\cdots+a_{1}(x) y^{\prime}+a_{0}(x) y$ where the $a_{i}$ are given continuous functions defined on an interval, $(a, b)$ and $y$ is some function which has $n$ derivatives so it makes sense to write $L y$. Suppose $L y_{k}=0$ for $k=1,2, \cdots, n$. The Wronskian of these functions, $y_{i}$ is defined as

$$
W\left(y_{1}, \cdots, y_{n}\right)(x) \equiv \operatorname{det}\left(\begin{array}{ccc}
y_{1}(x) & \cdots & y_{n}(x) \\
y_{1}^{\prime}(x) & \cdots & y_{n}^{\prime}(x) \\
\vdots & & \vdots \\
y_{1}^{(n-1)}(x) & \cdots & y_{n}^{(n-1)}(x)
\end{array}\right)
$$

Show that for $W(x)=W\left(y_{1}, \cdots, y_{n}\right)(x)$ to save space,

$$
W^{\prime}(x)=\operatorname{det}\left(\begin{array}{ccc}
y_{1}(x) & \cdots & y_{n}(x) \\
\vdots & \cdots & \vdots \\
y_{1}^{(n-2)}(x) & & y_{n}^{(n-2)}(x) \\
y_{1}^{(n)}(x) & \cdots & y_{n}^{(n)}(x)
\end{array}\right)
$$

Now use the differential equation, $L y=0$ which is satisfied by each of these functions, $y_{i}$ to verify that $W^{\prime}+a_{n-1}(x) W=0$. Give an explicit solution of this linear differential equation, Abel's formula, and use your answer to verify that the Wronskian of these solutions to the equation, $L y=0$ either vanishes identically on $(a, b)$ or never.
8. Show that the identity matrix is not similar to any other matrix.
9. Two $n \times n$ matrices, $A$ and $B$, are similar if $B=S^{-1} A S$ for some invertible $n \times n$ matrix $S$. Prove a theorem which is illustrated by the following picture.

| same trace,characteristic polynomial, determinant |
| :---: |
| similar |

Give an example of two matrices which are not similar but they have the same trace, characteristic polynomial and determinant.
10. Suppose the characteristic polynomial of an $n \times n$ matrix $A$ is of the form

$$
t^{n}+a_{n-1} t^{n-1}+\cdots+a_{1} t+a_{0}
$$

and that $a_{0} \neq 0$. Find a formula $A^{-1}$ in terms of powers of the matrix $A$. Show that $A^{-1}$ exists if and only if $a_{0} \neq 0$. In fact, show that $a_{0}=(-1)^{n} \operatorname{det}(A)$. Note how similar this is to what we did with algebraic numbers earlier on.
11. $\uparrow$ Letting $p(t)$ denote the characteristic polynomial of $A$, show that $p_{\varepsilon}(t) \equiv p(t-\varepsilon)$ is the characteristic polynomial of $A+\varepsilon I$. Then show that if $\operatorname{det}(A)=0$, it follows that $\operatorname{det}(A+\varepsilon I) \neq 0$ whenever $|\varepsilon|$ is sufficiently small but nonzero.
12. In constitutive modeling of the stress and strain tensors, one sometimes considers sums of the form $\sum_{k=0}^{\infty} a_{k} A^{k}$ where $A$ is a $3 \times 3$ matrix. Show using the Cayley Hamilton theorem that if such a thing makes any sense, you can always obtain it as a finite sum having no more than 3 terms.
13. Recall you can find the determinant from expanding along the $j^{\text {th }} \operatorname{column} . \operatorname{det}(A)=$ $\sum_{i} A_{i j}(\operatorname{cof}(A))_{i j}$ Think of $\operatorname{det}(A)$ as a function of the entries, $A_{i j}$. Explain why the $i j^{\text {th }}$ cofactor is really just $\frac{\partial \operatorname{det}(A)}{\partial A_{i j}}$.
14. Let $U$ be an open set in $\mathbb{R}^{n}$ and let $\boldsymbol{g}: U \rightarrow \mathbb{R}^{n}$ be such that all the first partial derivatives of all components of $\boldsymbol{g}$ exist and are continuous. Under these conditions form the matrix $D \boldsymbol{g}(\boldsymbol{x})$ given by $D \boldsymbol{g}(\boldsymbol{x})_{i j} \equiv \frac{\partial g_{i}(\boldsymbol{x})}{\partial x_{j}} \equiv g_{i, j}(\boldsymbol{x})$ The best kept secret in calculus courses is that the linear transformation determined by this matrix $D \boldsymbol{g}(\boldsymbol{x})$ is called the derivative of $\boldsymbol{g}$ and is the correct generalization of the concept of derivative of a function of one variable. Suppose the second partial derivatives also exist and are continuous. Then show that $\sum_{j}(\operatorname{cof}(D \boldsymbol{g}))_{i j, j}=0$. Hint: First explain why $\sum_{i} g_{i, k} \operatorname{cof}(D \boldsymbol{g})_{i j}=\delta_{j k} \operatorname{det}(D \boldsymbol{g})$. Next differentiate with respect to $x_{j}$ and sum on $j$ using the equality of mixed partial derivatives. Assume $\operatorname{det}(D \boldsymbol{g}) \neq 0$ to prove the identity in this special case. Then explain using Problem 11 why there exists a sequence $\varepsilon_{k} \rightarrow 0$ such that for $\boldsymbol{g}_{\varepsilon_{k}}(\boldsymbol{x}) \equiv \boldsymbol{g}(\boldsymbol{x})+\boldsymbol{\varepsilon}_{k} \boldsymbol{x}$, $\operatorname{det}\left(D \boldsymbol{g}_{\varepsilon_{k}}\right) \neq 0$ and so the identity holds for $\boldsymbol{g}_{\varepsilon_{k}}$. Then take a limit to get the desired result in general. This is an extremely important identity which has surprising implications. One can build degree theory on it for example. It also leads to simple proofs of the Brouwer fixed point theorem from topology.
15. A determinant of the form

$$
\begin{array}{|cccc}
1 & 1 & \cdots & 1 \\
a_{0} & a_{1} & \cdots & a_{n} \\
a_{0}^{2} & a_{1}^{2} & \cdots & a_{n}^{2} \\
\vdots & \vdots & & \vdots \\
a_{0}^{n-1} & a_{1}^{n-1} & \cdots & a_{n}^{n-1} \\
a_{0}^{n} & a_{1}^{n} & \cdots & a_{n}^{n}
\end{array}
$$

is called a Vandermonde determinant. Show it equals $\prod_{0 \leq i<j \leq n}\left(a_{j}-a_{i}\right)$. By this is meant to take the product of all terms of the form $\left(a_{j}-a_{i}\right)$ such that $j>i$. Hint: Show it works if $n=1$ so you are looking at $\left|\begin{array}{cc}1 & 1 \\ a_{0} & a_{1}\end{array}\right|$. Then suppose it holds for $n-1$ and consider the case $n$. Consider the polynomial in $t, p(t)$ which is obtained from the above by replacing the last column with the column $\left(\begin{array}{llll}1 & t & \cdots & t^{n}\end{array}\right)^{T}$. Explain why $p\left(a_{j}\right)=0$ for $i=0, \cdots, n-1$. Explain why $p(t)=c \prod_{i=0}^{n-1}\left(t-a_{i}\right)$. Of course $c$ is the coefficient of $t^{n}$. Find this coefficient from the above description of
$p(t)$ and the induction hypothesis. Then plug in $t=a_{n}$ and observe the formula is valid for $n$.
16. The example in this exercise was shown to me by Marc van Leeuwen and it helped to correct a misleading proof of the Cayley Hamilton theorem presented in this chapter. If $p(\lambda)=q(\lambda)$ for all $\lambda$ or for all $\lambda$ large enough where $p(\lambda), q(\lambda)$ are polynomials having matrix coefficients, then it is not necessarily the case that $p(A)=q(A)$ for $A$ a matrix of an appropriate size. The proof in question read as though it was using this incorrect argument. Let

$$
E_{1}=\left(\begin{array}{ll}
1 & 0 \\
0 & 0
\end{array}\right), E_{2}=\left(\begin{array}{ll}
0 & 0 \\
0 & 1
\end{array}\right), N=\left(\begin{array}{ll}
0 & 1 \\
0 & 0
\end{array}\right)
$$

Show that for all $\lambda,\left(\lambda I+E_{1}\right)\left(\lambda I+E_{2}\right)=\left(\lambda^{2}+\lambda\right) I=\left(\lambda I+E_{2}\right)\left(\lambda I+E_{1}\right)$. However,

$$
\left(N I+E_{1}\right)\left(N I+E_{2}\right) \neq\left(N I+E_{2}\right)\left(N I+E_{1}\right) .
$$

Explain why this can happen. In the proof of the Cayley-Hamilton theorem given in the chapter, show that the matrix $A$ does commute with the matrices $C_{i}$ in that argument. Hint: Multiply both sides out with $N$ in place of $\lambda$. Does $N$ commute with $E_{i}$ ?
17. Explain why the proof of the Cayley-Hamilton theorem given in this chapter cannot possibly hold for arbitrary fields of scalars.
18. Suppose $A$ is $m \times n$ and $B$ is $n \times m$. Letting $I$ be the identity of the appropriate size, is it the case that $\operatorname{det}(I+A B)=\operatorname{det}(I+B A)$ ? Explain why or why not.
19. Suppose $A$ is a linear transformation and let the characteristic polynomial be

$$
\operatorname{det}(\lambda I-A)=\prod_{j=1}^{q} \phi_{j}(\lambda)^{n_{j}}
$$

where the $\phi_{j}(\lambda)$ are irreducible. Explain using Corollary 1.13 .10 why the irreducible factors of the minimum polynomial are $\phi_{j}(\lambda)$ and why the minimum polynomial is of the form $\prod_{j=1}^{q} \phi_{j}(\lambda)^{r_{j}}$ where $r_{j} \leq n_{j}$. You can use the Cayley Hamilton theorem if you like.
20. $M=\left(\begin{array}{ccc}B_{1} & & \\ & \ddots & \\ & & B_{r}\end{array}\right)$ is a block diagonal matrix. Show $\operatorname{det}(M)=\prod_{k=1}^{r} \operatorname{det}\left(B_{k}\right)$.
21. Use the existence of the Jordan canonical form for a linear transformation whose minimum polynomial factors completely to give a proof of the Cayley Hamilton theorem which is valid for any field of scalars. Hint: First assume the minimum polynomial factors completely into linear factors. In this case, note that the characteristic polynomial is of degree $n$ and is the product of $(\lambda-\mu)$ where $\mu$ is an eigenvalue and listed according to algebraic multiplicity. However, if there are multiple blocks corresponding to some $\mu$, then the minimum polynomial will have such terms but
fewer of them. If the minimum polynomial does not split, consider a splitting field of the minimum polynomial. Then consider the minimum polynomial with respect to this larger field. How will the two minimum polynomials be related? The two characteristic polynomials will be exactly the same, being defined in terms of the determinant of $\lambda I-A$. Show the minimum polynomial always divides the characteristic polynomial for any field $\mathbb{F}$.
22. Let $q(\lambda)=\lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}$ be a polynomial and $C$ its companion matrix.

$$
C=\left(\begin{array}{cccc}
0 & & & -a_{0} \\
1 & & & -a_{1} \\
& \ddots & & \vdots \\
0 & & 1 & -a_{n-1}
\end{array}\right)
$$

Show that $\operatorname{det}(\lambda I-C)=q(\lambda)$ and that the minimum polynomial has degree $n$. It fill follow from the next problem or the preceding one that characteristic polynomial will coincide with the minimal polynomial for this.
23. $\uparrow$ Use the existence of the rational canonical form $M$ to give a proof of the Cayley Hamilton theorem valid for any field, even fields like the integers mod $p$ for $p$ a prime. The proof in this chapter on determinants was fine for fields like $\mathbb{Q}$ or $\mathbb{R}$ where you could let $\lambda \rightarrow \infty$ but it is not clear the same result holds in general. Show the minimum polynomial $\prod_{k=1}^{p} \phi_{k}(\lambda)^{m_{k}} \operatorname{divides} \operatorname{det}(\lambda I-M)$. Hint: Recall that for a linear transformation, it has a rational canonical form $M$ which is block diagonal $M=\left(\begin{array}{ccc}C_{1} & & 0 \\ & \ddots & \\ 0 & & C_{r}\end{array}\right)$ and one of the blocks has det $\left(\lambda I-C_{s_{k}}\right)=\phi_{k}(\lambda)^{m_{k}}$ since one of the companion matrices comes from $\phi_{k}(\lambda)^{m_{k}}$, this for each $k$. There may be other blocks for which $\operatorname{det}\left(\lambda I-\hat{C}_{j}\right)=\phi_{k}(\lambda)^{l}, l<m_{k}$. However, $\operatorname{det}(\lambda I-M)=$ $\prod_{k=1}^{r} \operatorname{det}\left(\lambda I-C_{k}\right)$ which is a polynomial divisible by the minimum polynomial.
24. Show that to find the eigenvalues of a matrix, it suffices to consider the roots of the characteristic polynomial. Hint: Use Cayley Hamilton theorem. This gives another way to find eigenvalues.
25. Recall that a matrix was diagonalizable if it was similar to a diagonal matrix. Suppose you have a matrix $A$ whose entries are in $\mathbb{F}$ and the characteristic polynomial is the same as the minimum polynomial but the characteristic polynomial of the matrix has a repeated root. Can you show that the matrix cannot be diagonalizable in any field containing $\mathbb{F}$ ?
26. For $W$ a subspace of $V, W$ is said to have a complementary subspace [24] $W^{\prime}$ if $W \oplus W^{\prime}=V$. Suppose that both $W, W^{\prime}$ are invariant with respect to $A \in \mathscr{L}(V, V)$. Show that for any polynomial $f(\lambda)$, if $f(A) x \in W$, then there exists $w \in W$ such that $f(A) x=f(A) w$. A subspace $W$ is called $A$ admissible if it is $A$ invariant and the condition of this problem holds.
27. When you have an Abelian group $V$ and a commutative ring with unity $K$ such that the usual vector space operations hold

$$
\begin{gathered}
k\left(v_{1}+v_{2}\right)=k v_{1}+k v_{2},\left(k_{1}+k_{2}\right) v=k_{1} v+k_{2} v \\
k_{1}\left(k_{2} v\right)=\left(k_{1} k_{2}\right) v, 1 v=v
\end{gathered}
$$

then you call this $V$ a $K$ module. Thus, it is just a vector space except you have a ring of scalars rather than a field of scalars. Now suppose $K=\mathbb{Z}$ the integers and $V=\mathbb{Z}_{m}$ where $m$ is some positive integer. Then if $k \in K$ and $\bar{a} \in \mathbb{Z}_{m}$, you define $k \bar{a}$ in the usual way. Just add $\bar{a}$ to itself $k$ times or if $k$ is negative, you just add $(-\bar{a})=\overline{m-a}$ to itself $|k|$ times. Explain why this is a $\mathbb{Z}$ module. More generally, explain why an arbitrary Abelian group is a $\mathbb{Z}$ module. However, show that in general, there is no linearly independent set of elements of $\mathbb{Z}_{m}$ which spans $\mathbb{Z}_{m}$, although it is certainly true that $\overline{1}$ spans $\mathbb{Z}_{m}$. Thus, when you replace a field with a ring, you loose the theorem that gives you a linearly independent subset of a spanning set. Hint: If $\overline{1}$ is in the span of your supposed basis, you have problems. If not in the span of your supposed basis, then you don't have a spanning set.
28. Now suppose you have $K$ a commutative ring with unity and consider $K^{n}$. Show $\left\{\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{n}\right\}$ spans $K^{n}$ and if you have $\left\{\boldsymbol{a}_{1}, \cdots, \boldsymbol{a}_{m}\right\}$ for $m<n$, then $\left\{\boldsymbol{a}_{1}, \cdots, \boldsymbol{a}_{m}\right\}$ does not span $K^{n}$. Hint: If it does span, then explain why you could get the following

$$
A_{n \times m} P_{m \times n}=\left(\begin{array}{lll}
a_{1} & \cdots & a_{m}
\end{array}\right) P=I_{n \times n} .
$$

Then consider this: $\left(\begin{array}{ll}A_{n \times m} & 0\end{array}\right)\binom{P_{m \times n}}{0}=I_{n \times n}$ Consider Theorem 8.4.4 which still works if the entries of the matrix are from a commutative ring with unity. Is $\left\{e_{1}, \cdots, e_{n}\right\}$ also linearly independent? By this is meant one of the definitions given earlier that if you have a linear combination of these vectors equal to 0 , then all of the scalars are zero. Since the scalars only come from a ring, you can't conclude that this is the same thing as saying that no vector is a linear combination of the others.
29. If $A(t)$ is an $n \times n$ matrix and $A(t)=\left(\begin{array}{lll}\boldsymbol{a}_{1}(t) & \cdots & \boldsymbol{a}_{n}(t)\end{array}\right)$ show that

$$
\frac{d}{d t} \operatorname{det}(A(t))=\sum_{i=1}^{n} \operatorname{det}\left(A_{i}(t)\right)
$$

where $A_{i}(t)$ has the same columns except for the $i^{t h}$ column which is $\boldsymbol{a}_{i}^{\prime}(t)$.
30. You have vectors $\boldsymbol{x}_{i}^{\prime}(t)=J \boldsymbol{x}_{i}(t)$ where $J$ is a Jordan canonical form and is $n \times n$. Form the matrix $\Phi(t) \equiv\left(\begin{array}{lll}x_{1}(t) & \cdots & x_{n}(t)\end{array}\right)$. Explain why $\Phi^{\prime}(t)=J \Phi(t)$. Now consider $\boldsymbol{y}_{i}(t)^{T}$ to be the $i^{\text {th }}$ row of this matrix $\Phi(t)$. Explain why

$$
\begin{equation*}
\boldsymbol{y}_{i}^{\prime}(t)^{T}=\lambda_{i} \boldsymbol{y}_{i}(t)^{T}+a_{i} \boldsymbol{y}_{i+1}(t)^{T} \tag{*}
\end{equation*}
$$

Now, using the result of Problem 29 explain why

$$
\operatorname{det}(\Phi(t))^{\prime}=\sum_{i=1}^{n} \operatorname{det}\left(\begin{array}{c}
\boldsymbol{y}_{1}(t)^{T}  \tag{**}\\
\vdots \\
\lambda_{i} \boldsymbol{y}_{i}(t)^{T}+a_{i} \boldsymbol{y}_{i+1}(t)^{T} \\
\vdots \\
\boldsymbol{y}_{n}(t)^{T}
\end{array}\right)
$$

where $a_{i}=0$ or 1 and $a_{n}=0$. Next explain, using elementary ODE why $\operatorname{det}(\Phi(t))=$ $C e^{\operatorname{trace}(J) t}$ for some constant $C$, showing that $\operatorname{det}(\Phi(t))$ either vanishes for all $t$ or for no $t$.
31. Obtain exactly the same result as the above Problem 30 for an arbitrary $A$ an $n \times n$ matrix. Use the result on the existence of Jordan canonical form along with properties of determinants to make this easy. Recall also the earlier problem that the trace is an invariant meaning that it does not change under similarity transformations. The formula you get is Abel's formula for first order systems. A few other simple problems using Jordan form will wipe out almost the entire typical undergraduate differential equations course. There is actually an easier, but trickier way to get this result of this problem.

## Chapter 9

## Some Items Which Resemble Linear AIgebra

This chapter is on some topics which don't usually appear in linear algebra texts but which seem to be related to linear algebra in the sense that the ideas are similar.

### 9.1 The Symmetric Polynomial Theorem

First here is a definition of polynomials in many variables which have coefficients in a commutative ring. A commutative ring would be a field except you don't know that every nonzero element has a multiplicative inverse. A good example of a commutative ring is the integers. In particular, every field is a commutative ring. Thus, a commutative ring satisfies the following axioms. They are just the field axioms with one omission mentioned above. You don't have $x^{-1}$ if $x \neq 0$. We will assume that the ring has 1 , the multiplicative identity.

Axiom 9.1.1 Here are the axioms for a commutative ring.

1. $x+y=y+x$, (commutative law for addition)
2. There exists 0 such that $x+0=x$ for all $x$, (additive identity).
3. For each $x \in \mathbb{F}$, there exists $-x \in \mathbb{F}$ such that $x+(-x)=0$, (existence of additive inverse).
4. $(x+y)+z=x+(y+z),($ associative law for addition $)$.
5. $x y=y x$, (commutative law for multiplication). You could write this as $x \times y=y \times x$.
6. $(x y) z=x(y z),($ associative law for multiplication).
7. There exists 1 such that $1 x=x$ for all $x$,(multiplicative identity).
8. $x(y+z)=x y+x z .($ distributive law $)$.

Definition 9.1.2 Let $\boldsymbol{k} \equiv\left(k_{1}, k_{2}, \cdots, k_{n}\right)$ where each $k_{i}$ is a nonnegative integer. Let

$$
|\boldsymbol{k}| \equiv \sum_{i} k_{i}
$$

Polynomials of degree $p$ in the variables $x_{1}, x_{2}, \cdots, x_{n}$ are expressions of the form

$$
g\left(x_{1}, x_{2}, \cdots, x_{n}\right)=\sum_{|\boldsymbol{k}| \leq p} a_{\boldsymbol{k}} x_{1}^{k_{1}} \cdots x_{n}^{k_{n}}
$$

where each $a_{\boldsymbol{k}}$ is in a commutative ring. If all $a_{\boldsymbol{k}}=0$, the polynomial has no degree. Such a polynomial is said to be symmetric if whenever $\sigma$ is a permutation of $\{1,2, \cdots, n\}$,

$$
g\left(x_{\sigma(1)}, x_{\sigma(2)}, \cdots, x_{\sigma(n)}\right)=g\left(x_{1}, x_{2}, \cdots, x_{n}\right)
$$

An example of a symmetric polynomial is

$$
s_{1}\left(x_{1}, x_{2}, \cdots, x_{n}\right) \equiv \sum_{i=1}^{n} x_{i}
$$

Another one is

$$
s_{n}\left(x_{1}, x_{2}, \cdots, x_{n}\right) \equiv x_{1} x_{2} \cdots x_{n}
$$

Definition 9.1.3 The elementary symmetric polynomial

$$
s_{k}\left(x_{1}, x_{2}, \cdots, x_{n}\right), k=1, \cdots, n
$$

is the coefficient of $(-1)^{k} x^{n-k}$ in the following polynomial.

$$
\begin{aligned}
& \left(x-x_{1}\right)\left(x-x_{2}\right) \cdots\left(x-x_{n}\right) \\
= & x^{n}-s_{1} x^{n-1}+s_{2} x^{n-2}-\cdots \pm s_{n}
\end{aligned}
$$

Thus

$$
\begin{aligned}
s_{1} & =x_{1}+x_{2}+\cdots+x_{n} \\
s_{2}=\sum_{i<j} x_{i} x_{j}, s_{3} & =\sum_{i<j<k} x_{i} x_{j} x_{k}, \ldots, s_{n}=x_{1} x_{2} \cdots x_{n}
\end{aligned}
$$

Note that it follows from the above definition that

$$
\alpha^{k} s_{k}\left(x_{1}, x_{2}, \cdots, x_{n}\right)=s_{k}\left(\alpha x_{1}, \cdots, \alpha x_{n}\right)
$$

Then the following result is the fundamental theorem in the subject. It is the symmetric polynomial theorem. This is a very remarkable theorem. It says that if you know a polynomial in some variables is symmetric, then it is the sum of polynomials in the elementary symmetric polynomials which are the coefficients of $p(x)=\prod_{k}\left(x-x_{k}\right)$.

What is an example of a polynomial which is NOT symmetric? These are not hard to find. Consider $g(x, y)=x+2 y$ for example. So when do we encounter them? It is often in coefficients which result from expanding something like $\prod_{k=1}^{n}\left(x-x_{k}\right)$.

Theorem 9.1.4 Every symmetric polynomial $g\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ equals a polynomial in the elementary symmetric polynomials.

$$
g\left(x_{1}, x_{2}, \cdots, x_{n}\right)=\sum_{\boldsymbol{k}} a_{\boldsymbol{k}} s_{1}^{k_{1}} \cdots s_{n}^{k_{n}}
$$

and the $a_{k}$ in the commutative ring are unique with all but finitely many of the coefficients $a_{k}$ being 0 .

Proof: The proof is by induction on the number of variables. If $n=1$, it is obviously true because $s_{1}=x_{1}$ and $g\left(x_{1}\right)$ can only be a polynomial in $x_{1}$. Suppose the theorem is true for $n-1$ variables and $g\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ has degree $d$. Thus in the sum for the polynomial, $|\boldsymbol{k}| \leq d$. By induction, there is a polynomial

$$
\begin{equation*}
Q\left(\tilde{s}_{1}, \cdots, \tilde{s}_{n-1}\right)=\sum_{|\boldsymbol{k}| \leq p} a_{\boldsymbol{k}} \tilde{s}_{1}^{k_{1}} \cdots \tilde{s}_{n-1}^{k_{n-1}}=g\left(x_{1}, x_{2}, \cdots, x_{n-1}, 0\right) \tag{9.1}
\end{equation*}
$$

where $\tilde{s}_{k}$ is a symmetric polynomial for the variables $\left\{x_{1}, x_{2}, \cdots, x_{n-1}\right\}$. Now let

$$
\begin{equation*}
p\left(x_{1}, x_{2}, \cdots, x_{n}\right) \equiv g\left(x_{1}, x_{2}, \cdots, x_{n}\right)-Q\left(s_{1}, \cdots, s_{n-1}\right) \tag{9.2}
\end{equation*}
$$

Thus $p\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ is a symmetric polynomial because each $s_{j}$ is symmetric and $g$ is given to be symmetric. Notice how $\tilde{s}_{k}$ was replaced with $s_{k}$.

If $x_{n}$ is set equal to 0 , the right side of 9.2 reduces to 0 because for $k \leq n-1$,

$$
s_{k}\left(x_{1}, x_{2}, \cdots, x_{n-1}, 0\right)=\tilde{s}_{k}\left(x_{1}, x_{2}, \cdots, x_{n-1}\right)
$$

This follows from the definition of these symmetric polynomials or their description in Definition 9.1.3. Indeed, the coefficient of $x^{n-k}$ in

$$
\left(x-x_{1}\right)\left(x-x_{2}\right) \cdots\left(x-x_{n-1}\right)(x-0)
$$

is the same as the coefficient of $x^{(n-1)-k}$ in $\left(x-x_{1}\right)\left(x-x_{2}\right) \cdots\left(x-x_{n-1}\right)$. Thus, the right side of 9.2 reduces to $g\left(x_{1}, x_{2}, \cdots, x_{n-1}, 0\right)-Q\left(\tilde{s}_{1}, \cdots, \tilde{s}_{n-1}\right)=0$ from 9.1 when $x_{n}=0$.

Thus $x_{n}$ divides $p\left(x_{1}, x_{2}, \cdots, x_{n}\right)$. In other words, each term in $p\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ has a factor of $x_{n}$. The same must be true with $x_{j}$ since otherwise, the symmetric polynomial $p\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ would change if you switched $x_{j}$ and $x_{n}$. Hence there exists a symmetric polynomial $g_{1}\left(x_{1}, x_{2}, \cdots, x_{n}\right)$ such that

$$
s_{n} g_{1}\left(x_{1}, x_{2}, \cdots, x_{n}\right)=g\left(x_{1}, x_{2}, \cdots, x_{n}\right)-Q\left(s_{1}, \cdots, s_{n-1}\right)
$$

Recall $s_{n}=x_{1} x_{2} \cdots x_{n}$. Thus

$$
g\left(x_{1}, x_{2}, \cdots, x_{n}\right)=s_{n} g_{1}\left(x_{1}, x_{2}, \cdots, x_{n}\right)+Q\left(s_{1}, \cdots, s_{n-1}\right) .
$$

Now if $g_{1}$ is not constant, do for $g_{1}$ what was just done for $g$. Obtain

$$
\begin{aligned}
g\left(x_{1}, x_{2}, \cdots, x_{n}\right)= & s_{n}\left(s_{n} g_{2}\left(x_{1}, x_{2}, \cdots, x_{n}\right)+Q_{2}\left(s_{1}, \cdots, s_{n-1}\right)\right) \\
& +Q\left(s_{1}, \cdots, s_{n-1}\right)
\end{aligned}
$$

Continue this way, obtaining a sequence of $g_{k}$ till the process stops with some $g_{m}$ being a constant. This must happen because the degree of $g_{k}$ becomes strictly smaller with each iteration. This yields a polynomial in the elementary symmetric polynomials for $\left\{x_{1}, x_{2}, \cdots, x_{n}\right\}$.

Example 9.1.5 Let $g(x, y)=x^{3}+y^{3}$. It is clear that $g(x, y)=g(y, x)$ so $g$ is a symmetric polynomial. Write as a polynomial in the elementary functions.

The above proof tells how to do this. First note that $x^{3}=\tilde{s}_{1}^{3}$ where $s_{1}$ is the symmetric polynomial associated with the single variable $x$. Thus $p(x, y)=x^{3}+y^{3}-s_{1}^{3}$ where this $s_{1}$ is $x+y$. Then $p(x, y)=x^{3}+y^{3}-(x+y)^{3}=-3 x^{2} y-3 x y^{2}$ and this equals $(-x y)(3 x+3 y)=$ $-3 s_{2} s_{1}$. Thus $-3 s_{1} s_{2}=x^{3}+y^{3}-s_{1}^{3}$ and so $g(x, y)=s_{1}^{3}-3 s_{1} s_{2}$.

You can see that if you have a symmetric polynomial in more variables, you could use a process of reducing one variable at a time in $g\left(x_{1}, \ldots, x_{n-1}, 0\right)$ to eventually obtain this function as a polynomial in the symmetric polynomials in variables $\left\{x_{1}, \ldots, x_{n-1}\right\}$.

Note that if you have $\prod_{i=1}^{m}\left(x-x_{i}\right)$ then by definition, it is the sum of terms like

$$
g\left(x_{1}, \cdots, x_{m}\right) x^{m-k}
$$

If you replace $x$ with $x_{i}$ and sum over all $i$, you would get $\sum_{i=1}^{m} g\left(x_{1}, \cdots, x_{m}\right) x_{i}^{m-k}$ which would also be a symmetric polynomial. It is of the form

$$
g\left(x_{1}, \cdots, x_{m}\right) x_{1}^{m-k}+g\left(x_{1}, \cdots, x_{m}\right) x_{2}^{m-k}+\cdots+g\left(x_{1}, \cdots, x_{m}\right) x_{m}^{m-k}
$$

so when you switch some variable in this, you get the same thing.
Here is a very interesting result which I saw claimed in a paper by Steinberg and Redheffer on Lindermannn's theorem which follows from the above theorem. It is a very surprising property of symmetric polynomials (surprising for me anyway) and is the main tool for proving the Lindermann Weierstrass theorem.

Theorem 9.1.6 Let $\alpha_{1}, \cdots, \alpha_{n}$ be roots of the polynomial equation

$$
\begin{equation*}
p(x) \equiv a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}=0 \tag{*}
\end{equation*}
$$

where each $a_{i}$ is an integer. Then any symmetric polynomial in the quantities

$$
a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}
$$

having integer coefficients is also an integer. Also any symmetric polynomial with rational coefficients in the quantities $\alpha_{1}, \cdots, \alpha_{n}$ is a rational number.

Proof: Let $f\left(x_{1}, \cdots, x_{n}\right)$ be the symmetric polynomial having integer coefficients. From Theorem 9.1.4 it follows there are integers $a_{k_{1} \cdots k_{n}}$ such that

$$
\begin{equation*}
f\left(x_{1}, \cdots, x_{n}\right)=\sum_{k_{1}+\cdots+k_{n} \leq m} a_{k_{1} \cdots k_{n}} p_{1}^{k_{1}} \cdots p_{n}^{k_{n}} \tag{9.3}
\end{equation*}
$$

where the $p_{i}$ are elementary symmetric polynomials defined as the coefficients of $\hat{p}(x)=$ $\prod_{j=1}^{n}\left(x-x_{j}\right)$ with $p_{k}\left(x_{1}, \ldots, x_{n}\right)$ of degree $k$ since it is the coefficient of $x^{n-k}$. Earlier we had them $\pm$ these coefficients. Thus

$$
\begin{aligned}
& f\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \\
= & \sum_{k_{1}+\cdots+k_{n}=d} a_{k_{1} \cdots k_{n}} p_{1}^{k_{1}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \cdots p_{n}^{k_{n}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right)
\end{aligned}
$$

Now the given polynomial in $*, p(x)$ is of the form

$$
\begin{gathered}
a_{n} \prod_{j=1}^{n}\left(x-\alpha_{j}\right) \equiv a_{n}\left(\sum_{k=0}^{n} p_{k}\left(\alpha_{1}, \cdots, \alpha_{n}\right) x^{n-k}\right) \\
=a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}
\end{gathered}
$$

Thus, equating coefficients, $a_{n} p_{k}\left(\alpha_{1}, \cdots, \alpha_{n}\right)=a_{n-k}$. Multiply both sides by $a_{n}^{k-1}$. Thus $p_{k}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right)=a_{n}^{k-1} a_{n-k}$ an integer. Therefore,

$$
\begin{aligned}
& f\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \\
= & \sum_{k_{1}+\cdots+k_{n}=d} a_{k_{1} \cdots k_{n}} p_{1}^{k_{1}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \cdots p_{n}^{k_{n}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right)
\end{aligned}
$$

and each $p_{k}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right)$ is an integer. Thus $f\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right)$ is an integer. From this, it is obvious that $f\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ is rational. Indeed, from 9.3,

$$
f\left(\alpha_{1}, \cdots, \alpha_{n}\right)=\sum_{k_{1}+\cdots+k_{n}=d} a_{k_{1} \cdots k_{n}} p_{1}^{k_{1}}\left(\alpha_{1}, \cdots, \alpha_{n}\right) \cdots p_{n}^{k_{n}}\left(\alpha_{1}, \cdots, \alpha_{n}\right)
$$

Now multiply both sides by $a_{n}^{M}$, an integer where $M$ is chosen large enough that

$$
\begin{gathered}
a_{n}^{M} f\left(\alpha_{1}, \cdots, \alpha_{n}\right)= \\
\sum_{k_{1}+\cdots+k_{n}=d} a_{n}^{h\left(k_{1}, \ldots, k_{n}\right)} a_{k_{1} \cdots k_{n}} p_{1}^{k_{1}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \cdots p_{n}^{k_{n}}\left(a_{n} \alpha_{1}, \cdots, a_{n} \alpha_{n}\right) \text { an integer. }
\end{gathered}
$$

where $h\left(k_{1}, \ldots, k_{n}\right)$ is some nonnegative integer. Thus $f\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ is rational. If the $f$ had rational coefficients, then $m f$ would have integer coefficients for a suitable $m$ and so $m f\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ would be rational which yields $f\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ is rational.

Nothing would change in the last claim of this theorem if $\mathbb{Q}$ were a general field. You would get $f\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ is in the general field.

Corollary 9.1.7 Let $\alpha_{1}, \cdots, \alpha_{n}$ be roots of the polynomial equation

$$
p(x) \equiv x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}=0
$$

where each $a_{i}$ is in a field $\mathbb{F}$. Then any symmetric polynomial in $\alpha_{1}, \cdots, \alpha_{n}$ which has coefficients in $\mathbb{F}$ is in $\mathbb{F}$.

Proof: Let $f\left(x_{1}, \ldots, x_{n}\right)$ be a symmetric polynomial. Then by the symmetric polynomial theorem,

$$
f\left(\alpha_{1}, \cdots, \alpha_{n}\right)=\sum_{k} b_{\boldsymbol{k}} s_{1}^{k_{1}} s_{2}^{k_{2}} \cdots s_{n}^{k_{n}}
$$

where the $s_{k}\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ is $\pm$ the coefficient of $x^{k}$ in $\prod_{i=1}^{n}\left(x-\alpha_{i}\right)$. Thus $a_{k}= \pm s_{k}$ and so the above sum is in $\mathbb{F}$.

### 9.2 Transcendental Numbers

Most numbers are like this, transcendental. Here the algebraic numbers are those which are roots of a polynomial equation having rational numbers as coefficients, equivalently integer coefficients. By the fundamental theorem of algebra, all these numbers are in $\mathbb{C}$ and they constitute a countable collection of numbers in $\mathbb{C}$. Therefore, most numbers in $\mathbb{C}$ are transcendental. Nevertheless, it is very hard to prove that a particular number is transcendental. Probably the most famous theorem about this is the Lindermannn Weierstrass theorem, 1884.

Theorem 9.2.1 Let the $\alpha_{i}$ be distinct nonzero algebraic numbers and let the $a_{i}$ be nonzero algebraic numbers. Then $\sum_{i=1}^{n} a_{i} e^{\alpha_{i}} \neq 0$.

I am following the interesting Wikepedia article on this subject. You can also look at the book by Baker [5], Transcendental Number Theory, Cambridge University Press. There are also many other treatments which you can find on the web including an interesting article by Steinberg and Redheffer, already mentioned, which appeared in about 1950.

The proof makes use of the following identity. For $f(x)$ a polynomial,

$$
\begin{equation*}
I(s) \equiv \int_{0}^{s} e^{s-x} f(x) d x=e^{s} \sum_{j=0}^{\operatorname{deg}(f)} f^{(j)}(0)-\sum_{j=0}^{\operatorname{deg}(f)} f^{(j)}(s) \tag{9.4}
\end{equation*}
$$

where $f^{(j)}$ denotes the $j^{t h}$ derivative. It is like the convolution integral discussed earlier with Laplace transforms. In this formula, $s \in \mathbb{C}$ and the integral is defined in the natural way as

$$
\begin{equation*}
\int_{0}^{1} s f(t s) e^{s-t s} d t \tag{9.5}
\end{equation*}
$$

The identity follows from integration by parts.

$$
\begin{aligned}
\int_{0}^{1} s f(t s) e^{s-t s} d t & =s e^{s} \int_{0}^{1} f(t s) e^{-t s} d t \\
& =s e^{s}\left[-\left.\frac{e^{-t s}}{s} f(t s)\right|_{0} ^{1}+\int_{0}^{1} \frac{e^{-t s}}{s} s f^{\prime}(s t) d t\right] \\
& =s e^{s}\left[-\frac{e^{-s}}{s} f(s)+\frac{1}{s} f(0)+\int_{0}^{1} e^{-t s} f^{\prime}(s t) d t\right] \\
& =e^{s} f(0)-f(s)+\int_{0}^{1} s e^{s-t s} f^{\prime}(s t) d t \\
& \equiv e^{s} f(0)-f(s)+\int_{0}^{s} e^{s-x} f^{\prime}(x) d x
\end{aligned}
$$

Continuing this way establishes the identity since the right end looks just like what we started with except with a derivative on the $f$.

Lemma 9.2.2 Let $\left(x_{1}, \ldots, x_{n}\right) \rightarrow g\left(x, x_{1}, \ldots, x_{n}\right)$ be symmetric and let

$$
x \rightarrow g\left(x, x_{1}, \ldots, x_{n}\right)
$$

be a polynomial. Then

$$
\frac{d^{m}}{d x^{m}} g\left(x, x_{1}, \ldots, x_{n}\right)
$$

is symmetric in the variables $\left\{x_{1}, \ldots, x_{n}\right\}$. If $\left(x_{1}, \ldots, x_{n}\right) \rightarrow h\left(x, x_{1}, \ldots, x_{n}\right)$ is symmetric, then for $r$ some nonnegative integer,

$$
\sum_{k=1}^{n} h\left(x_{k}, x_{1}, \ldots, x_{n}\right) x_{k}^{r}
$$

is symmetric. In particular,

$$
\sum_{k=1}^{n} \frac{d^{l}}{d x^{l}} g\left(\cdot, x_{1}, \ldots, x_{n}\right)\left(x_{k}\right) x_{k}^{r}
$$

is symmetric in $\left\{x_{1}, \ldots, x_{n}\right\}$.
Proof: The coefficients of the polynomial $x \rightarrow g\left(x, x_{1}, \ldots, x_{n}\right)$ are symmetric functions of $\left\{x_{1}, \ldots, x_{n}\right\}$. Differentiating with respect to $x$ multiple times just gives another polynomial in $x$ having coefficients which are symmetric functions. Thus the first part is proved. For the second part, the sum is of the form

$$
h\left(x_{1}, x_{1}, \ldots, x_{n}\right) x_{1}^{r}+h\left(x_{2}, x_{1}, \ldots, x_{n}\right) x_{2}^{r}+\cdots+h\left(x_{n}, x_{1}, \ldots, x_{n}\right) x_{n}^{r}
$$

You see that this is unchanged from switching two variables. For example, switch $x_{1}$ and $x_{2}$. By assumption, nothing changes in the terms after the first two. The first term then becomes

$$
h\left(x_{2}, x_{2}, x_{1} \ldots, x_{n}\right) x_{2}^{r}=h\left(x_{2}, x_{1}, x_{2}, \ldots, x_{n}\right) x_{2}^{r}
$$

and the second term becomes

$$
h\left(x_{1}, x_{2}, x_{1}, \ldots, x_{n}\right) x_{1}^{r}=h\left(x_{1}, x_{1}, x_{2}, \ldots, x_{n}\right) x_{1}^{r}
$$

which are the same two terms, just added in a different order. The situation works the same way with any other pair of variables.

Recall that every algebraic number is a root of a polynomial having integer coefficients.
Lemma 9.2.3 Let $Q(x)=v x^{m}+\cdots+u$ have roots $\beta_{1}, \ldots, \beta_{m}$ listed according to multiplicity and let the coefficients be integers. Let

$$
\begin{equation*}
f(x) \equiv \frac{v^{(m+1) p} Q^{p}(x) x^{p-1}}{(p-1)!} \tag{9.6}
\end{equation*}
$$

a polynomial of degree $n=p m+p-1$. Then

$$
\begin{gather*}
\sum_{j=0}^{n} f^{(j)}(0)=v^{p(m+1)} u^{p}+m_{1}(p) p  \tag{9.7}\\
\sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)=m_{2}(p) p \tag{9.8}
\end{gather*}
$$

where $m_{1}(p), m_{2}(p)$ are integers and $p$ will be a large prime.
Proof: First consider 9.7. $f(x)=\frac{v^{(m+1) p}\left(v x^{m}+\cdots+u\right)^{p} x^{p-1}}{(p-1)!}$. Then $f^{j}(0)=0$ unless $j \geq$ $p-1$ because otherwise, that $x^{p-1}$ term will result in some $x^{r}, r>0$ and everything is zero when you plug in $x=0$. Now say $j=p-1$. Then it is clear that you get a $(p-1)$ ! which cancels the denominator and letting $x=0$, you get the integer $f^{(p-1)}(0)=u^{p} v^{(m+1) p}$. So what if $j>p-1$ ?

$$
\begin{aligned}
& \frac{d^{j}}{d x^{j}}\left(\left(v x^{m}+\cdots+u\right)^{p} x^{p-1}\right) \\
= & \sum_{r=0}^{j}\binom{j}{i} \frac{d^{i}}{d x^{i}}\left(\left(v x^{m}+\cdots+u\right)^{p}\right) \frac{d^{j-i}}{d x^{j-i}} x^{p-1}
\end{aligned}
$$

and, since eventually $x=0$, only $j-i=p-1$ is of interest, so $i=j-p+1$ where $j \geq p$ as just mentioned. Since $i \geq 1$, there will be a factor of $p$ and a factor of $(p-1)$ ! from $\frac{d^{j-i}}{d x^{j-i}} x^{p-1}$. Thus when $x=0$, this reduces to $m_{1}(p) p(p-1)!$ and so this yields 9.7.

Next consider 9.8 which says that $\sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)=m_{2}(p) p$. The factorization of $Q(x)$ is $v\left(x-\beta_{1}\right) \cdots\left(x-\beta_{m}\right)$. Replace $Q(x)$ with its factorization in 9.6 to get

$$
\begin{equation*}
f(x)(p-1)!=v^{p} v^{(m+1) p}\left(\left(x-\beta_{1}\right)\left(x-\beta_{2}\right) \cdots\left(x-\beta_{m}\right)\right)^{p} x^{p-1} \tag{9.9}
\end{equation*}
$$

First notice that $(p-1)!f^{(j)}\left(\beta_{i}\right)=0$ unless $j \geq p$. Thus all terms in computing

$$
f^{(j)}\left(\beta_{i}\right)(p-1)!
$$

for $j \geq p$ have a factor of $p!$. If you have

$$
g\left(x, \beta_{1}, \cdots, \beta_{m}\right) \equiv v^{p} v^{(m+1) p}\left(\left(x-\beta_{1}\right)\left(x-\beta_{2}\right) \cdots\left(x-\beta_{m}\right)\right)^{p} x^{p-1}
$$

it is symmetric in the $\beta_{i}$ so all derivatives with respect to $x$ are also symmetric in these $\beta_{i}$ by Lemma 9.2.2. By the same lemma, for $j \geq p$

$$
\sum_{i=1}^{m} \frac{d^{j}}{d x^{j}}\left(g\left(\cdot, \beta_{1}, \cdots, \beta_{m}\right)\left(\beta_{i}\right) \frac{1}{(p-1)!}\right)=\sum_{i=1}^{m} f^{(j)}\left(\beta_{i}\right)
$$

is symmetric in the $\beta_{1}, \cdots, \beta_{m}$. Thanks to the factor $v^{p} v^{(m+1) p}$ and the factor $p$ ! coming from $j \geq p$, it is a symmetric polynomial in the $v \beta_{i}$ with integer coefficients, each multiplied by $p$ with the $\beta_{i}$ roots of $Q(x)=v x^{m}+\cdots+u$. By Theorem 9.1.6 this is an integer. As noted earlier, it equals 0 unless $j \geq p$ when it contains a factor of $p$. Thus the sum of these integers is also an integer times $p$. It follows that

$$
\sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)=m_{2}(p) p, m_{2}(p) \text { an integer. }
$$

Note that no use was made of $p$ being a large prime number. This will come next.
Lemma 9.2.4 If $K$ and $c$ are nonzero integers, and

$$
\beta_{1}, \cdots, \beta_{m}
$$

are the roots of a single polynomial with integer coefficients,

$$
Q(x)=v x^{m}+\cdots+u
$$

where $v, u \neq 0$, then,

$$
K+c\left(e^{\beta_{1}}+\cdots+e^{\beta_{m}}\right) \neq 0
$$

Letting

$$
f(x) \equiv \frac{v^{(m+1) p} Q^{p}(x) x^{p-1}}{(p-1)!}
$$

and $I(s)$ be defined in terms of $f(x)$ as above,

$$
I(s) \equiv \int_{0}^{s} e^{s-x} f(x) d x=e^{s} \sum_{j=0}^{\operatorname{deg}(f)} f^{(j)}(0)-\sum_{j=0}^{\operatorname{deg}(f)} f^{(j)}(s)
$$

it follows,

$$
\begin{equation*}
\lim _{p \rightarrow \infty} \sum_{i=1}^{m} I\left(\beta_{i}\right)=0 \tag{9.10}
\end{equation*}
$$

and for $n$ the degree of $f(x), n=p m+p-1$, where $m_{i}(p)$ is some integer for $p$ a large prime number.

Proof: The first step is to verify 9.10 for $f(x)$ as given in 9.6 for $p$ large prime numbers. Let $p$ be a large prime number. Then 9.10 follows right away from the definition of $I\left(\beta_{j}\right)$ and the definition of $f(x)$.

$$
\left|I\left(\beta_{j}\right)\right| \leq \int_{0}^{1}\left|\beta_{j} f\left(t \beta_{j}\right) e^{\beta_{j}-t \beta_{j}}\right| d t \leq \int_{0}^{1}\left|\frac{|v|^{(m-1) p}\left|Q\left(t \beta_{j}\right)\right|^{p} t^{p-1}\left|\beta_{j}\right|^{p-1}}{(p-1)!}\right| d t
$$

which clearly converges to 0 using considerations involving convergent series which show the integrand converges uniformly to 0 . The degree of $f(x)$ is $n \equiv p m+p-1$ where $p$ will be a sufficiently large prime number from now on.

From 9.4,

$$
\begin{gather*}
c \sum_{i=1}^{m} I\left(\beta_{i}\right)=c \sum_{i=1}^{m}\left(e^{\beta_{i}} \sum_{j=0}^{n} f^{(j)}(0)-\sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)\right) \\
=\left(K+c \sum_{i=1}^{m} e^{\beta_{i}}\right) \sum_{j=0}^{n} f^{(j)}(0)-\left(K \sum_{j=0}^{n} f^{(j)}(0)+c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)\right) \tag{9.11}
\end{gather*}
$$

Here $K \sum_{j=0}^{n} f^{(j)}(0)$ is added and subtracted. From Lemma 9.2.3,

$$
v^{p(m+1)} u^{p}+m_{1}(p) p+m_{2}(p) p=K \sum_{j=0}^{n} f^{(j)}(0)+c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}\left(\beta_{i}\right)
$$

Thus, if $p$ is very large,

$$
c \sum_{i=1}^{m} I\left(\beta_{i}\right)=\text { small }=K v^{p(m+1)} u^{p}+M(p) p+\left(K+c \sum_{i=1}^{m} e^{\beta_{i}}\right) \sum_{j=0}^{n} f^{(j)}(0)
$$

Let $p$ be prime and larger than $\max (K, v, u)$. If $K+c \sum_{i=1}^{m} e^{\beta_{i}}=0$, the above is impossible because it would require

$$
\text { small }=K v^{p(m+1)} u^{p}+M(p) p
$$

Now the right side is a nonzero integer because $p$ cannot divide $K v^{p(m+1)} u^{p}$ so the right side cannot equal something small.

Note that this shows $\pi$ is irrational. If $\pi=k / m$ where $k, m$ are integers, then both $i \pi$ and $-i \pi$ are roots of the polynomial with integer coefficients, $m^{2} x^{2}+k^{2}$ which would require, from what was just shown that

$$
0 \neq 2+e^{i \pi}+e^{-i \pi}
$$

which is not the case since the sum on the right equals 0 .
The following corollary follows from this. It is like the above lemma except it involves several polynomials. First is a lemma.

Lemma 9.2.5 Let $v_{k}, u_{k}, m_{k}$ be integers for $k=1,2 \ldots, m, u_{k}, v_{k}$ nonzero. Then for each $k$ there exists $\alpha_{k}$ an integer such that $\alpha_{k}^{m_{k}+2} v_{k}^{m_{k}+1} u_{k}$ is $U$ for some non zero integer.

Proof: Let $U \equiv\left(\prod_{j=1}^{m} v_{j}^{m_{j}+1} u_{j}\right)^{\prod_{j=1}^{m}\left(m_{j}+2\right)^{2}} \equiv \alpha_{k}^{m_{k}+2} v_{k}^{m_{k}+1} u_{k}$ where $\alpha_{k}$ is an integer chosen to make this so.

Corollary 9.2.6 Let $K$ and $c_{i}$ for $i=1, \cdots, n$ be nonzero integers. For each $k$ between 1 and $n$ let $\left\{\beta(k)_{i}\right\}_{i=1}^{m_{k}}$ be the roots of a polynomial with integer coefficients,

$$
Q_{k}(x) \equiv v_{k} x^{m_{k}}+\cdots+u_{k}
$$

where $v_{k}, u_{k} \neq 0$. Then

$$
\begin{equation*}
K+c_{1}\left(\sum_{j=1}^{m_{1}} e^{\beta(1)_{j}}\right)+c_{2}\left(\sum_{j=1}^{m_{2}} e^{\beta(2)_{j}}\right)+\cdots+c_{n}\left(\sum_{j=1}^{m_{n}} e^{\beta(n)_{j}}\right) \neq 0 \tag{*}
\end{equation*}
$$

Proof: Let $K_{k}$ be nonzero integers which add to $K$. It is certainly possible to obtain this since the $K_{k}$ are allowed to change sign. They only need to be nonzero. Also let $\alpha_{k}$ be as in the above lemma such that $\alpha_{k}^{m_{k}+2} v_{k}^{m_{k}+1} u_{k}=U$ some integer. Thus, replacing each $Q_{k}(x)$ with $\alpha_{k} v_{k} x^{m_{k}}+\cdots+\alpha_{k} u_{k}$, it follows that for each large prime $p,\left(\alpha_{k} v\right)^{p\left(m_{k}+1\right)}\left(\alpha_{k} u\right)^{p}=$ $\left(\alpha_{k}^{m_{k}+2} \nu^{m_{k}+1}\right)^{p}=U^{p}$. From now on, use the new $Q_{k}(x)$.

Defining $f_{k}(x)$ and $I_{k}(s)$ as in Lemma 9.2.4,

$$
f_{k}(x) \equiv \frac{v^{(m+1) p} Q_{k}^{p}(x) x^{p-1}}{(p-1)!}
$$

and as before, let $p$ be a very large prime number. It follows from Lemma 9.2.4 that for each $k=1, \cdots, n$,

$$
\begin{aligned}
c_{k} \sum_{i=1}^{m_{k}} I_{k}\left(\beta(k)_{i}\right)= & \left(K_{k}+c_{k} \sum_{i=1}^{m_{k}} e^{\beta(k)_{i}}\right) \sum_{j=0}^{\operatorname{deg}\left(f_{k}\right)} f_{k}^{(j)}(0) \\
& -\left(K_{k} \sum_{j=0}^{\operatorname{deg}\left(f_{k}\right)} f_{k}^{(j)}(0)+c_{k} \sum_{i=1}^{m_{k}} \sum_{j=0}^{\operatorname{deg}\left(f_{k}\right)} f_{k}^{(j)}\left(\beta(k)_{i}\right)\right)
\end{aligned}
$$

This is exactly the same computation as in the beginning of that lemma except one adds and subtracts $K_{k} \sum_{j=0}^{\operatorname{deg}\left(f_{k}\right)} f_{k}^{(j)}(0)$ rather than $K \sum_{j=0}^{\operatorname{deg}\left(f_{k}\right)} f_{k}^{(j)}(0)$ where the $K_{k}$ are chosen such that their sum equals $K$ and the term on the left converges to 0 as $p \rightarrow \infty$. By Lemma 9.2.4,

$$
\begin{gathered}
c_{k} \sum_{i=1}^{m_{k}} I_{k}\left(\beta(k)_{i}\right)=\left(K_{k}+c_{k} \sum_{i=1}^{m_{k}} e^{\beta(k)_{i}}\right)\left(U^{p}+N_{k} p\right) \\
-K_{k}\left(U^{p}+N_{k} p\right)-c_{k} N_{k}^{\prime} p \\
=\left(K_{k}+c_{k} \sum_{i=1}^{m_{k}} e^{\beta(k)_{i}}\right) U^{p}-K U^{p}+M_{k} p
\end{gathered}
$$

where $M_{k}$ is some integer. Now add.

$$
\sum_{k=1}^{m} c_{k} \sum_{i=1}^{m_{k}} I_{k}\left(\beta(k)_{i}\right)=U^{p}\left(K+\sum_{k=1}^{m} c_{k} \sum_{i=1}^{m_{k}} e^{\beta(k)_{i}}\right)-K m U^{p}+M p
$$

If $K+\sum_{k=1}^{m} c_{k} \sum_{i=1}^{m_{k}} e^{\beta(k)_{i}}=0$, then if $p>\max (K, m, U)$ you would have $-K m U^{p}+M p$ an integer so it cannot equal the left side which will be small if $p$ is large. Therefore, * follows.

Next is an even more interesting Lemma which follows from the above corollary.

Lemma 9.2.7 If $b_{0}, b_{1}, \cdots, b_{n}$ are non zero integers, and $\gamma_{1}, \cdots, \gamma_{n}$ are distinct algebraic numbers, then

$$
b_{0} e^{\gamma_{0}}+b_{1} e^{\gamma_{1}}+\cdots+b_{n} e^{\gamma_{n}} \neq 0
$$

Proof: Assume

$$
\begin{equation*}
b_{0} e^{\gamma_{0}}+b_{1} e^{\gamma_{1}}+\cdots+b_{n} e^{\gamma_{n}}=0 \tag{9.12}
\end{equation*}
$$

Divide by $e^{\gamma_{0}}$ and letting $K=b_{0}$,

$$
\begin{equation*}
K+b_{1} e^{\alpha(1)}+\cdots+b_{n} e^{\alpha(n)}=0 \tag{9.13}
\end{equation*}
$$

where $\alpha(k)=\gamma_{k}-\gamma_{0}$. These are still distinct algebraic numbers. Therefore, $\alpha(k)$ is a root of a polynomial

$$
\begin{equation*}
Q_{k}(x)=v_{k} x^{m_{k}}+\cdots+u_{k} \tag{9.14}
\end{equation*}
$$

having integer coefficients, $v_{k}, u_{k} \neq 0$. Recall algebraic numbers were defined as roots of polynomial equations having rational coefficients. Just multiply by the denominators to get one with integer coefficients. Let the roots of this polynomial equation be

$$
\left\{\alpha(k)_{1}, \cdots, \alpha(k)_{m_{k}}\right\}
$$

and suppose they are listed in such a way that $\alpha(k)_{1}=\alpha(k)$. Thus, by Theorem 9.1.6 every symmetric polynomial in these roots is rational.

Letting $i_{k}$ be an integer in $\left\{1, \cdots, m_{k}\right\}$ it follows from the assumption 9.12 that

$$
\begin{equation*}
\prod_{\substack{\left(i_{1}, \cdots, i_{n}\right) \\ i_{k} \in\left\{1, \cdots, m_{k}\right\}}}\left(K+b_{1} e^{\alpha(1)_{i_{1}}}+b_{2} e^{\alpha(2)_{i_{2}}}+\cdots+b_{n} e^{\alpha(n)_{i_{n}}}\right)=0 \tag{9.15}
\end{equation*}
$$

This is because one of the factors is the one occurring in 9.13 when $i_{k}=1$ for every $k$. The product is taken over all distinct ordered lists $\left(i_{1}, \cdots, i_{n}\right)$ where $i_{k}$ is as indicated. Expand this possibly huge product. This will yield something like the following.

$$
\begin{gather*}
K^{\prime}+c_{1}\left(e^{\beta(1)_{1}}+\cdots+e^{\beta(1)_{\mu(1)}}\right)+c_{2}\left(e^{\beta(2)_{1}}+\cdots+e^{\beta(2)_{\mu(2)}}\right)+\cdots+ \\
c_{N}\left(e^{\beta(N)_{1}}+\cdots+e^{\beta(N)_{\mu(N)}}\right)=0 \tag{9.16}
\end{gather*}
$$

These integers $c_{j}$ come from products of the $b_{i}$ and $K$. You group these exponentials according to which $c_{i}$ they multiply. The $\beta(i)_{j}$ are the distinct exponents which result, each being a sum of some of the $\alpha(r)_{i_{r}}$. Since the product included all roots for each $Q_{k}(x)$, interchanging their order does not change the distinct exponents $\beta(i)_{j}$ which result. They might occur in a different order however, but you would still have the same distinct exponents associated with each $c_{s}$ as shown in the sum. Thus any symmetric polynomial in the $\beta(s)_{1}, \beta(s)_{2}, \cdots, \beta(s)_{\mu(s)}$ is also a symmetric polynomial in the roots of $Q_{k}(x)$, $\alpha(k)_{1}, \alpha(k)_{2}, \cdots, \alpha(k)_{m_{k}}$ for each $k$.

Doesn't this contradict Corollary 9.2.6? This is not yet clear because we don't know that the $\beta(i)_{1}, \ldots, \beta(i)_{\mu(i)}$ are roots of a polynomial having rational coefficients. For a given $r, \beta(r)_{1}, \cdots, \beta(r)_{\mu(r)}$ are roots of the polynomial

$$
\begin{equation*}
\left(x-\beta(r)_{1}\right)\left(x-\beta(r)_{2}\right) \cdots\left(x-\beta(r)_{\mu(r)}\right) \tag{9.17}
\end{equation*}
$$

the coefficients of which are elementary symmetric polynomials in the

$$
\beta(r)_{i}, i \leq \mu(r)
$$

Thus the coefficients are symmetric polynomials in the $\alpha(k)_{1}, \alpha(k)_{2}, \cdots, \alpha(k)_{m_{k}}$ for each $k$. Say the polynomial is of the form

$$
\sum_{l=0}^{\mu(r)} x^{n-l} B_{l}(A(1), \cdots, A(n))
$$

where $A(k)$ signifies the roots of $Q_{k}(x),\left\{\alpha(k)_{1}, \cdots, \alpha(k)_{m_{k}}\right\}$. Thus, by the symmetric polynomial theorem applied to the commutative ring $\mathbb{Q}[A(1), \cdots, A(n-1)]$, the above polynomial is of the form

$$
\sum_{l=0}^{\mu(r)} x^{\mu(r)-l} \sum_{\boldsymbol{k}_{l}} B_{\boldsymbol{k}_{l}}(A(1), \cdots, A(n-1)) s_{1}^{k_{1}^{l}} \cdots s_{\mu(r)}^{k_{n}^{l}}
$$

where the $s_{k}$ is one of the elementary symmetric polynomials in $\left\{\alpha(n)_{1}, \cdots, \alpha(n)_{m_{n}}\right\}$ and $B_{k_{l}}$ is symmetric in $\alpha(k)_{1}, \alpha(k)_{2}, \cdots, \alpha(k)_{m_{k}}$ for each $k \leq n-1$ and

$$
B_{\boldsymbol{k}_{l}} \in \mathbb{Q}[A(1), \cdots, A(n-1)] .
$$

Now do to $B_{\boldsymbol{k}_{l}}$ what was just done to $B_{l}$ featuring $A(n-1)$ this time, and continue till eventually you obtain for the coefficient of $x^{\mu(r)-l}$ a large sum of rational numbers times a product of symmetric polynomials in $A(1), A(2)$, etc. By Theorem 9.1.6 applied repeatedly, beginning with $A(1)$ and then to $A(2)$ and so forth, one finds that the coefficient of $x^{\mu(r)-l}$ is a rational number and so the $\beta(r)_{j}$ for $j \leq \mu(r)$ are algebraic numbers and roots of a polynomial which has rational coefficients, namely the one in 9.17 , hence roots of a polynomial with integer coefficients. Now 9.16 contradicts Corollary 9.2.6.

Note this lemma is sufficient to prove Lindermann's theorem that $\pi$ is transcendental. Here is why. If $\pi$ is algebraic, then so is $i \pi$ and so from this lemma, $e^{0}+e^{i \pi} \neq 0$ but this is not the case because $e^{i \pi}=-1$.

The next theorem is the main result, the Lindermann Weierstrass theorem. It replaces the integers $b_{i}$ in the above lemma with algebraic numbers.

Theorem 9.2.8 Suppose $a(1), \cdots, a(n)$ are nonzero algebraic numbers and suppose

$$
\alpha(1), \cdots, \alpha(n)
$$

are distinct algebraic numbers. Then

$$
a(1) e^{\alpha(1)}+a(2) e^{\alpha(2)}+\cdots+a(n) e^{\alpha(n)} \neq 0
$$

Proof: Suppose $a(j) \equiv a(j)_{1}$ is a root of the polynomial

$$
v_{j} x^{m_{j}}+\cdots+u_{j}
$$

where $v_{j}, u_{j} \neq 0$. Let the roots of this polynomial be $a(j)_{1}, \cdots, a(j)_{m_{j}}$. Suppose to the contrary that

$$
a(1)_{1} e^{\alpha(1)}+a(2)_{1} e^{\alpha(2)}+\cdots+a(n)_{1} e^{\alpha(n)}=0
$$

Then consider the big product

$$
\begin{equation*}
\prod_{\substack{\left(i_{1}, \cdots, i_{n}\right) \\ i_{k} \in\left\{1, \cdots, m_{k}\right\}}}\left(a(1)_{i_{1}} e^{\alpha(1)}+a(2)_{i_{2}} e^{\alpha(2)}+\cdots+a(n)_{i_{n}} e^{\alpha(n)}\right) \tag{9.18}
\end{equation*}
$$

the product taken over all ordered lists $\left(i_{1}, \cdots, i_{n}\right)$. Since one of the factors in this product equals 0 , this product equals

$$
\begin{equation*}
0=b_{1} e^{\beta(1)}+b_{2} e^{\beta(2)}+\cdots+b_{N} e^{\beta(N)} \tag{9.19}
\end{equation*}
$$

where the $\beta(j)$ are the distinct exponents which result and the $b_{k}$ result from combining terms corresponding to a single $\beta(k)$. The $\beta(i)$ are clearly algebraic because they are the sum of the $\alpha(i)$. I want to show that the $b_{k}$ are actually rational numbers. Since the product in 9.18 is taken for all ordered lists as described above, it follows that for a given $k$, if $a(k)_{i}$ is switched with $a(k)_{j}$, that is, two of the roots of $v_{k} x^{m_{k}}+\cdots+u_{k}$ are switched, then the product is unchanged and so 9.19 is also unchanged. Thus each $b_{l}$ is a symmetric polynomial in the $a(k)_{j}, j=1, \cdots, m_{k}$ for each $k$. Consider then a particular $b_{k}$.It follows

$$
b_{k}=\sum_{\left(j_{1}, \cdots, j_{m_{n}}\right)} A_{j_{1}, \cdots, j_{m_{n}}} a(n)_{1}^{j_{1}} \cdots a(n)_{m_{n}}^{j_{m_{n}}}
$$

and this is symmetric in the $\left\{a(n)_{1}, \cdots, a(n)_{m_{n}}\right\}$ (note $n$ is distinguished) the coefficients $A_{j_{1}, \cdots, j_{m_{n}}}$ being in the commutative ring $\mathbb{Q}[A(1), \cdots, A(n-1)]$ where $A(p)$ denotes

$$
a(k)_{1}, \cdots, a(k)_{m_{p}}
$$

and so from Theorem 9.1.4,

$$
b_{k}=\sum_{\left(j_{1}, \cdots, j_{m_{n}}\right)} B_{j_{1}, \cdots, j_{m_{n}}} p_{1}^{j_{1}}\left(a(n)_{1} \cdots a(n)_{m_{n}}\right) \cdots p_{m_{n}}^{j_{m_{n}}}\left(a(n)_{1} \cdots a(n)_{m_{n}}\right)
$$

where the $B_{j_{1}, \cdots, j_{m_{n}}}$ are symmetric in $\left\{a(k)_{j}\right\}_{j=1}^{m_{k}}$ for each $k \leq n-1$. and the $p_{k}^{l}$ are elementary symmetric polynomials. Now doing to $B_{j_{1}, \cdots, j_{m_{n}}}$ what was just done to $b_{k}$ and continuing this way, it follows $b_{k}$ is a finite sum of rational numbers times powers of elementary polynomials in the various $\left\{a(k)_{j}\right\}_{j=1}^{m_{k}}$ for $k \leq n$. By Theorem 9.1.6 this is a rational number. Thus $b_{k}$ is a rational number as desired. Multiplying by the product of all the denominators, it follows there exist integers $c_{i}$ such that

$$
0=c_{1} e^{\beta(1)}+c_{2} e^{\beta(2)}+\cdots+c_{N} e^{\beta(N)}
$$

which contradicts Lemma 9.2.7.
This theorem is sufficient to show $e$ is transcendental. If it were algebraic, then

$$
e e^{-1}+(-1) e^{0} \neq 0
$$

but this is not the case. If $a \neq 1$ is algebraic, then $\ln (a)$ is transcendental. To see this, note that

$$
1 e^{\ln (a)}+(-1) a e^{0}=0
$$

which cannot happen if $\ln (a)$ is algebraic according to the above theorem. If $a$ is algebraic and $\sin (a) \neq 0$, then $\sin (a)$ is transcendental because

$$
\frac{1}{2 i} e^{i a}-\frac{1}{2 i} e^{-i a}+(-1) \sin (a) e^{0}=0
$$

which cannot occur if $\sin (a)$ is algebraic. There are doubtless other examples of numbers which are transcendental by this amazing theorem. For example, $\pi$ is also transcendental. This is because $1+e^{i \pi}=0$. This couldn't happen if $\pi$ were algebraic because then so would be $i \pi$.

### 9.3 The Fundamental Theorem of Algebra

This is devoted to a mostly algebraic proof of the fundamental theorem of algebra. It depends on the interesting results about symmetric polynomials which are presented above. I found it on the Wikipedia article about the fundamental theorem of algebra. You google "fundamental theorem of algebra" and go to the Wikipedia article. It gives several other proofs in addition to this one. According to this article, the first completely correct proof of this major theorem is due to Argand in 1806. Gauss and others did it earlier but their arguments had gaps in them.

You can't completely escape analysis when you prove this theorem. The necessary analysis due to Bolzano in about 1817 is in the following lemma.

Lemma 9.3.1 Suppose $p(x)=x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ where $n$ is odd and the coefficients are real. Then $p(x)$ has a real root.

Proof: This follows from the intermediate value theorem from calculus.
Next is an algebraic consideration. First recall some notation.

$$
\prod_{i=1}^{m} a_{i} \equiv a_{1} a_{2} \cdots a_{m}
$$

Recall a polynomial in $\left\{z_{1}, \cdots, z_{n}\right\}$ is symmetric only if it can be written as a sum of elementary symmetric polynomials raised to various powers multiplied by constants.

The following is the main part of the theorem. In fact this is one version of the fundamental theorem of algebra which people studied earlier in the 1700's.

Lemma 9.3.2 Let $p(x)=x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ be a polynomial with real coefficients. Then it has a complex root.

Proof: It is possible to write $n=2^{k} m$. where $m$ is odd. If $n$ is odd, $k=0$. If $n$ is even, keep dividing by 2 until you are left with an odd number. If $k=0$ so that $n$ is odd, it follows from Lemma 9.3.1 that $p(x)$ has a real, hence complex root. The proof will be by induction on $k$, the case $k=0$ being done. Suppose then that it works for $n=2^{l} m$ where $m$ is odd and $l \leq k-1$ and let $n=2^{k} m$ where $m$ is odd. Let $\left\{z_{1}, \cdots, z_{n}\right\}$ be the roots of the polynomial $p(x)$ in a splitting field, the existence of this field being given by the above proposition. I need to show that at least one of these $z_{j}$ is complex. Then

$$
\begin{equation*}
p(x)=\prod_{j=1}^{n}\left(x-z_{j}\right)=\sum_{k=0}^{n}(-1)^{k} p_{k}\left(z_{1}, \cdots, z_{n}\right) x^{k} \tag{9.20}
\end{equation*}
$$

where $p_{k}\left(z_{1}, \cdots, z_{n}\right)$ is the $k^{\text {th }}$ elementary symmetric polynomial. Note this shows

$$
\begin{equation*}
a_{n-k}=p_{k}\left(z_{1}, \cdots, z_{n}\right)(-1)^{k}, \text { a real number. } \tag{9.21}
\end{equation*}
$$

There is another polynomial which has coefficients which are sums of real numbers times the $p_{k}$ raised to various powers and it is

$$
q_{t}(x) \equiv \prod_{1 \leq i<j \leq n}\left(x-\left(z_{i}+z_{j}+t z_{i} z_{j}\right)\right), t \in \mathbb{R}
$$

I need to verify this is really the case for $q_{t}(x)$. When you switch any two of the $z_{i}$ in $q_{t}(x)$ the polynomial does not change. Thus the coefficients of $q_{t}(x)$ must be symmetric polynomials in the $z_{i}$ with real coefficients. Hence by Proposition 9.1.4 these coefficients are real polynomials in terms of the elementary symmetric polynomials $p_{k}$. Thus by 9.21 the coefficients of $q_{t}(x)$ are real polynomials in terms of the $a_{k}$ of the original polynomial. Recall these were all real. It follows, and this is what was wanted, that $q_{t}(x)$ has all real coefficients.

Note that the degree of $q_{t}(x)$ is $\binom{n}{2}$ because there are this number of ways to pick $i<j$ out of $\{1, \cdots, n\}$. Now for some $m$,

$$
\binom{n}{2}=\frac{n(n-1)}{2}=2^{k-1} m\left(2^{k} m-1\right)=2^{k-1}(\text { odd })
$$

and so by induction, for each $t \in \mathbb{R}, q_{t}(x)$ has a complex root.
There must exist $s \neq t$ such that for a single pair of indices $i, j$, with $i<j$,

$$
\left(z_{i}+z_{j}+t z_{i} z_{j}\right),\left(z_{i}+z_{j}+s z_{i} z_{j}\right)
$$

are both complex. Here is an explanation why. Let $A(i, j)$ denote those $t \in \mathbb{R}$ such that $\left(z_{i}+z_{j}+t z_{i} z_{j}\right)$ is complex. It was just shown that every $t \in \mathbb{R}$ must be in some $A(i, j)$. There are infinitely many $t \in \mathbb{R}$ and so some $A(i, j)$ contains two of them.

Now for that $t, s$,

$$
\begin{aligned}
z_{i}+z_{j}+t z_{i} z_{j} & =a \in \mathbb{C} \\
z_{i}+z_{j}+s z_{i} z_{j} & =b \in \mathbb{C}
\end{aligned}
$$

where $t \neq s$ and so by Cramer's rule,

$$
z_{i}+z_{j}=\frac{\left|\begin{array}{cc}
a & t \\
b & s
\end{array}\right|}{\left|\begin{array}{cc}
1 & t \\
1 & s
\end{array}\right|} \in \mathbb{C}
$$

and also

$$
z_{i} z_{j}=\frac{\left|\begin{array}{ll}
1 & a \\
1 & b
\end{array}\right|}{\left|\begin{array}{ll}
1 & t \\
1 & s
\end{array}\right|} \in \mathbb{C}
$$

At this point, note that the roots of $p(x)$ in the splitting field, $z_{i}, z_{j}$ are both solutions to the equation

$$
x^{2}-\left(z_{1}+z_{2}\right) x+z_{1} z_{2}=0
$$

which from the above has complex coefficients. By the quadratic formula the $z_{i}, z_{j}$ are both complex. Thus the original polynomial has a complex root.

With this lemma, it is easy to prove the fundamental theorem of algebra. The difference between the lemma and this theorem is that in the theorem, the coefficients are only assumed to be complex. What this means is that if you have any polynomial with complex coefficients it has a complex root and so it is not irreducible. Hence the field extension is the same field. Another way to say this is that for every complex polynomial there exists a factorization into linear factors or in other words a splitting field for a complex polynomial is the field of complex numbers.

Theorem 9.3.3 Let $p(x) \equiv a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ be any complex polynomial, $n \geq 1, a_{n} \neq 0$. Then it has a complex root. Furthermore, there exist complex numbers $z_{1}, \cdots, z_{n}$ such that

$$
p(x)=a_{n} \prod_{k=1}^{n}\left(x-z_{k}\right)
$$

Proof: First suppose $a_{n}=1$. Consider the polynomial $q(x) \equiv p(x) \overline{p(\bar{x})}$

$$
\begin{aligned}
& \left(x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}\right) . \\
& \left(x^{n}+\overline{a_{n-1}} x^{n-1}+\cdots+\overline{a_{1}} x+\overline{a_{0}}\right)
\end{aligned}
$$

This polynomial has real coefficients because the coefficient of $x^{m}$ is of the form

$$
\sum_{k=0}^{m} a^{m-k} \overline{a_{k}}
$$

and the sum involves adding terms of the form

$$
a_{k} \overline{a_{j}}+\overline{a_{k}} a_{j}=a_{k} \overline{a_{j}}+\overline{a_{k}} a_{j}=a_{k} \overline{a_{j}}+\overline{a_{k} \overline{a_{j}}}
$$

so it is of the form of a complex number added to its conjugate. Hence $q(x)$ has real coefficients as claimed. Therefore, by by Lemma 9.3.2 it has a complex root $z$. Hence either $p(z)=0$ or $p(\bar{z})=0$. Thus $p(x)$ has a complex root.

Next suppose $a_{n} \neq 1$. Then simply divide by it and get a polynomial in which $a_{n}=1$. Denote this modified polynomial as $q(x)$. Then by what was just shown and the Euclidean algorithm, there exists $z_{1} \in \mathbb{C}$ such that

$$
q(x)=\left(x-z_{1}\right) q_{1}(x)
$$

where $q_{1}(x)$ has complex coefficients. Now do the same thing for $q_{1}(x)$ to obtain

$$
q(x)=\left(x-z_{1}\right)\left(x-z_{2}\right) q_{2}(x)
$$

and continue this way. Thus

$$
\frac{p(x)}{a_{n}}=\prod_{j=1}^{n}\left(x-z_{j}\right)
$$

Why use this more elaborate proof? I think it is because you can give other examples of algebraically complete fields. For example, begin with $\mathbb{Q}$ and let the algebraic numbers be denoted by $\mathbb{A}$. These are those numbers which are roots of a polynomial having rational coefficients. Then consider $\mathbb{A}_{2}$ to be those complex numbers which are roots of a polynomial having coefficients in $\mathbb{A}$. In general, let $\mathbb{A}_{n}$ be roots of polynomials with coefficients in $\mathbb{A}_{n-1}$. In general, if $\mathbb{A}_{n-1}$ is countable, then so is $\mathbb{A}_{n}$. This is routine to show using the fact that there are countably many polynomials of degree $m$ for each $m \in \mathbb{N}$. Each has at most $m$ roots. Thus $\mathbb{A}_{\infty} \equiv \cup_{n=1}^{\infty} \mathbb{A}_{n}$ is countable because $\mathbb{Q}$ is. Now recall also that it was shown that the algebraic numbers over a field are a field. Therefore, $\mathbb{A}_{\infty}$ is also a field because any finite number of elements of $\mathbb{A}_{\infty}$ must be in a single one of the fields $\mathbb{A}_{n}$ for large enough $n$. Now consider Lemma 9.3.1 applied to a polynomial having real coefficients in $\mathbb{A}_{\infty}$. These coefficients are in some $\mathbb{A}_{n}$ and so the root from $\mathbb{C}$ having these coefficients is in $\mathbb{A}_{n+1} \subseteq \mathbb{A}_{\infty}$. Now the rest of the argument goes similarly. You show using the same considerations that every polynomial having real coefficients in $\mathbb{A}_{\infty}$ has a root in $\mathbb{A}_{\infty}$. Then you do the easy extension to the case where the coefficients in $\mathbb{A}_{\infty}$ are complex. This field is clearly much smaller than $\mathbb{C}$ because it is countable, and yet it is algebraically complete. The standard analysis proof given earlier will obviously not work because it is based on compactness considerations.

### 9.4 More on Algebraic Field Extensions

This is on field extensions. There are many linear algebra techniques which are used in this discussion and it seems to me to be very interesting. I am following various algebra books in assembling this material. I hope it is useful and that I have not diminished it too much by my attempts to write it down, because it is clear to me that, even though it has nothing to do with my own interests, it is some of the most wonderful mathematics I have ever seen.

Consider the notion of splitting fields. It is desired to show that any two are isomorphic, meaning that there exists a one to one and onto mapping from one to the other which preserves all the algebraic structure. To begin with, is a theorem about extending homomorphisms. [26]

Definition 9.4.1 Suppose $\mathbb{F}, \overline{\mathbb{F}}$ are two fields and that $f: \mathbb{F} \rightarrow \overline{\mathbb{F}}$ is a homomorphism. This means that

$$
f(x y)=f(x) f(y), f(x+y)=f(x)+f(y)
$$

An isomorphism is a homomorphism which is one to one and onto. A monomorphism is a homomorphism which is one to one. An automorphism is an isomorphism of a single field. Sometimes people use the symbol $\simeq$ to indicate something is an isomorphism. Then if $p(x) \in \mathbb{F}[x]$, say

$$
p(x)=\sum_{k=0}^{n} a_{k} x^{k}
$$

$\bar{p}(x)$ will be the polynomial in $\overline{\mathbb{F}}[x]$ defined as

$$
\bar{p}(x) \equiv \sum_{k=0}^{n} f\left(a_{k}\right) x^{k}
$$

Also consider $f$ as a homomorphism of $\mathbb{F}[x]$ and $\overline{\mathbb{F}}[x]$ in the obvious way.

$$
f(p(x))=\bar{p}(x)
$$

If $f$ defined on $\mathbb{F}[x]$ is as just described, then is indeed a homomorphism of $\mathbb{F}[x]$ and $\overline{\mathbb{F}}[x]$ as claimed. This follows from an elementary computation.

The following is a nice theorem which will be useful.
Theorem 9.4.2 Let $\mathbb{F}$ be a field and let $r$ be algebraic over $\mathbb{F}$. Let $p(x)$ be the minimum polynomial of $r$. Thus $p(r)=0$ and $p(x)$ is monic and no nonzero polynomial having coefficients in $\mathbb{F}$ of smaller degree has $r$ as a root. In particular, $p(x)$ is irreducible over $\mathbb{F}$. Then define $f: \mathbb{F}[x] \rightarrow \mathbb{F}(r)$, the polynomials in $r$ by

$$
f\left(\sum_{i=0}^{m} a_{i} x^{i}\right) \equiv \sum_{i=0}^{m} a_{i} r^{i}
$$

Then $f$ is a homomorphism. Also, defining $g: \mathbb{F}[x] /(p(x))$ by

$$
g([q(x)]) \equiv f(q(x)) \equiv q(r)
$$

it follows that $g$ is an isomorphism from the field $\mathbb{F}[x] /(p(x))$ to $\mathbb{F}(r)$.
Proof: First of all, consider why $f$ is a homomorphism. The preservation of sums is obvious. Consider products.

$$
\begin{aligned}
f\left(\sum_{i} a_{i} x^{i} \sum_{j} b_{j} x^{j}\right) & =f\left(\sum_{i, j} a_{i} b_{j} x^{i+j}\right)=\sum_{i j} a_{i} b_{j} r^{i+j} \\
& =\sum_{i} a_{i} r^{i} \sum_{j} b_{j} r^{j}=f\left(\sum_{i} a_{i} x^{i}\right) f\left(\sum_{j} b_{j} x^{j}\right)
\end{aligned}
$$

Thus it is clear that $f$ is a homomorphism.
First consider why $g$ is even well defined. If $[q(x)]=\left[q_{1}(x)\right]$, this means that

$$
q_{1}(x)-q(x)=p(x) l(x)
$$

for some $l(x) \in \mathbb{F}[x]$. Therefore,

$$
\begin{aligned}
f\left(q_{1}(x)\right) & =f(q(x))+f(p(x) l(x)) \\
& =f(q(x))+f(p(x)) f(l(x)) \\
& \equiv q(r)+p(r) l(r)=q(r)=f(q(x))
\end{aligned}
$$

Now from this, it is obvious that $g$ is a homomorphism.

$$
\begin{aligned}
g\left([q(x)]\left[q_{1}(x)\right]\right) & =g\left(\left[q(x) q_{1}(x)\right]\right)=f\left(q(x) q_{1}(x)\right)=q(r) q_{1}(r) \\
g([q(x)]) g\left(\left[q_{1}(x)\right]\right) & \equiv q(r) q_{1}(r)
\end{aligned}
$$

Similarly, $g$ preserves sums. Now why is $g$ one to one? It suffices to show that if $g([q(x)])=$ 0 , then $[q(x)]=0$. Suppose then that $g([q(x)]) \equiv q(r)=0$ Then $q(x)=p(x) l(x)+\rho(x)$ where the degree of $\rho(x)$ is less than the degree of $p(x)$ or else $\rho(x)=0$. If $\rho(x) \neq 0$, then it follows that $\rho(r)=0$ and $\rho(x)$ has smaller degree than that of $p(x)$ which contradicts the definition of $p(x)$ as the minimum polynomial of $r$. Hence, $[q(x)]=0$ and $g$ is one to one. Since $p(x)$ is irreducible, $\mathbb{F}[x] /(p(x))$ is a field. It is clear that $g$ is onto. Therefore, $\mathbb{F}(r)$ is a field also. (This was shown earlier by different reasoning.)

Here is a diagram of what the following theorem says.

## Extending $f$ to $g$

| $\mathbb{F}$ | $\stackrel{f}{\rightarrow}$ | $\overline{\mathbb{F}}$ |
| :---: | :---: | :---: |
| $p(x) \in \mathbb{F}[x]$ | $\xrightarrow{\sim}$ | $\bar{p}(x) \in \overline{\mathbb{F}}[x]$ |
| $p(x)=\sum_{k=0}^{n} a_{k} x^{k}$ | $\rightarrow$ | $\sum_{k=0}^{n} f\left(a_{k}\right) x^{k}=\bar{p}(x)$ |
| $p(r)=0$ |  | $\bar{p}(\bar{r})=0$ |
| $\mathbb{F}(r)$ | $\xrightarrow{g}$ | $\overline{\mathbb{F}}(\bar{r})$ |
| $r$ | $\xrightarrow{\sim}$ | $\bar{r}$ |

## One such $g$ for each $\bar{r}$

Definition 9.4.3 Let $f: \mathbb{F} \rightarrow \overline{\mathbb{F}}$ be an isomorphism. For the sake of convenience, if $q(x)$ is a polynomial $b_{m} x^{m}+b_{n-1} x^{m-1}+\cdots+b_{1} x+b_{0}$ in $\mathbb{F}[x]$, then $\bar{q}(x)$ will denote the polynomial

$$
f\left(b_{m}\right) x^{m}+f\left(b_{n-1}\right) x^{m-1}+\cdots+f\left(b_{1}\right) x+f\left(b_{0}\right) \equiv f(q(x))
$$

Then $f$ defined in this way on $\mathbb{F}[x]$ is a homomorphism.
Recall that if $p(x)$ is a monic polynomial of degree at least 1 , then $\mathbb{F}[x] /(p(x))$ is a commutative ring. This is from Lemma 3.4.8. I will assume $p(x)$ has degree at least 1 because otherwise there isn't anything new being shown in what follows. Here is a simple lemma.

Lemma 9.4.4 Let $F$ be a field and let $G$ be a commutative ring. Then the multiplicative identity and additive identities are unique in both and if there is an isomorphism $h: F \rightarrow G$, then $G$ is also a field.

Proof: If $1, \overline{1}$ are multiplicative identities in a commutative ring, then $1=1 \overline{1}=\overline{1}$ so it is unique. Also the additive identity is unique. This is because if you have $0,0^{\prime}$ both additive identities, then $0=0+0^{\prime}=0^{\prime} . h(0)+h(x)=h(0+x)=h(x)$ and so $h(0)$ is the additive identity in $G$ since a typical thing in $G$ is $h(x)$. Also $h(1) h(x)=h(1 x)=h(x)$ so $h(1)=1$ in $G$ by what was just shown. Now suppose $h(x) \neq 0$. Then $x \neq 0$ because $h(0)=0$ in $G$. Hence $h\left(x^{-1}\right) h(x)=h\left(x^{-1} x\right)=h(1)$, the multiplicative identity in $G$. Hence if something in $G$ is not 0 , then it has a multiplicative inverse and so $G$ is a field.

If $\mathbb{K}$ is a finite field extension of $\mathbb{F}$, this means that $[\mathbb{K}: \mathbb{F}]<\infty$. Recall $[\mathbb{K}: \mathbb{F}]$ is the dimension of the vector space $\mathbb{K}$ having field of scalars $\mathbb{F}$. Then if a basis for $\mathbb{K}$ is $\left\{k_{1}, \ldots, k_{n}\right\}$, each of these $k_{i}$ is algebraic and so $\mathbb{K}=\mathbb{F}\left(k_{1}, \cdots, k_{n}\right)$. The next theorem considers the case where you have a simple extension so the basis is of the form $\left\{1, r, \ldots, r^{n-1}\right\}$.

Theorem 9.4.5 Let $f: \mathbb{F} \rightarrow \overline{\mathbb{F}}$ be an isomorphism of the two fields. Let $r$ be algebraic over $\mathbb{F}$ with minimum polynomial $p(x) \equiv x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}$ and suppose there exists $\bar{r}$ a root of $\bar{p}(x) \equiv f(p(x)), \bar{p}(\bar{r})=0$. Then

1. If $h: \mathbb{F}[x] /(p(x)) \rightarrow \overline{\mathbb{F}}[x] /(\bar{p}(x))$ is defined by

$$
h([q(x)]) \equiv[\bar{q}(x)] \equiv[f(q(x))]
$$

then $h$ is an isomorphism.
2. $\bar{p}(x)$ is the minimum polynomial for $\bar{r}$
3. There exists a unique isomorphism $g: \mathbb{F}(r) \rightarrow \mathbb{F}(\bar{r})$ which agrees with $f$ on $\mathbb{F}$ and $g(r)=g(\bar{r})$.

Proof: Define $h: \mathbb{F}[x] /(p(x)) \rightarrow \overline{\mathbb{F}}[x] /(\bar{p}(x))$ by $h([q(x)]) \equiv[\bar{q}(x)] \equiv[f(q(x))]$. I claim this is an isomorphism.

First, why is $h$ well defined? If $[q(x)]=\left[q_{1}(x)\right]$ is $[\bar{q}(x)]=\left[\bar{q}_{1}(x)\right]$ ? This is equivalent to verifying that if $[q(x)]=0$, then $[\bar{q}(x)]=0$. Does this happen? If $q(x)=p(x) l(x)$, is $\bar{q}(x)=\bar{p}(x) \bar{l}(x)$ ? This is true since $f$ is a homomorphism. Thus this $h$ is well defined.

Also, $h$ is a homomorphism because

$$
h\left([q(x)]\left[q_{1}(x)\right]\right)=h\left(\left[q(x) q_{1}(x)\right]\right) \equiv\left[f\left(q(x) q_{1}(x)\right)\right]=[f(q(x))]\left[f\left(q_{1}(x)\right)\right]
$$

$h$ is clearly onto because $f$ is onto $\overline{\mathbb{F}}$.
Is $h$ one to one? If $h([q(x)]) \equiv[\bar{q}(x)] \equiv[f(q(x))]=0$, does it follow that $[q(x)]=0$ ? To say that $[\bar{q}(x)]=0$ is to say that $f(g(x))=\bar{q}(x)=\bar{l}(x) \bar{p}(x)=f(l(x) p(x))$. Thus $f(l(x) p(x))=f(q(x))$. However, since $f$ is one to one on $\mathbb{F}$, this requires $l(x) p(x)=$ $q(x)$ and so $[q(x)]=0$ showing that $h$ is one to one. Hence $h$ is an isomorphism.

Thus $\overline{\mathbb{F}}[x] /(\bar{p}(x))$ is a field by Lemma 9.4.4 because $\mathbb{F}[x] /(p(x))$ is a field due to $p(x)$ being the minimum polynomial for $r$ which forces $p(x)$ to be irreducible. Indeed, if $p(x)=$ $k(x) l(x)$ where each of these two have smaller degree than $p(x)$, then $0=p(r)=k(r) l(r)$ and since $k(r), l(r)$ are in the field $\mathbb{F}(r)$, it follows that one of these is 0 which contradicts $p(x)$ being the minimum polynomial for $r$. It follows from Lemma 3.4.8 that $\bar{p}(x)$ must also be irreducible. Hence $\bar{p}(x)$ is the minimum polynomial for $\bar{r}$.

Then from Theorem 9.4.2, the following diagram holds

$$
\mathbb{F}(r) \xrightarrow{\alpha^{-1}} \mathbb{F}[x] /(p(x)) \xrightarrow{h} \overline{\mathbb{F}}[x] /(\bar{p}(x)) \xrightarrow{\bar{\alpha}} \overline{\mathbb{F}}(\bar{r})
$$

where $\bar{\alpha}$ is the isomorphism described by $\bar{\alpha}([\bar{q}(x)]) \equiv \bar{q}(\bar{r})$. Thus all mappings are isomorphisms and so you can let $g=\bar{\alpha} \circ h \circ \alpha^{-1}$ and this shows the existence of an extension of $f$ as an isomorphism from $\mathbb{F}(r)$ to $\overline{\mathbb{F}}(\bar{r})$ which satisfies

$$
g(q(r)) \equiv \bar{\alpha} \circ h \circ \alpha^{-1}(q(r)) \equiv \bar{\alpha} \circ h([q(x)]) \equiv \bar{\alpha}([\bar{q}(x)]) \equiv \bar{q}(\bar{r})
$$

In particular, if $q(r)=r$, then since $f(1)$ is the multiplicative identity in $\overline{\mathbb{F}}$,

$$
g(r)=g(q(r)) \equiv \bar{\alpha} \circ h([q(x)]) \equiv \bar{\alpha} \circ h([x]) \equiv \bar{\alpha}([f(1) x])=\bar{\alpha}([x])=\bar{r}
$$

If $q(r) \in \mathbb{F}$, then $q(r)=q_{0} \in \mathbb{F}$ and so $g(q(r))=g\left(q_{0}\right)=$

$$
\bar{\alpha} \circ h \circ \alpha^{-1}\left(q_{0}\right) \equiv \bar{\alpha} \circ h\left(\left[q_{0}\right]\right) \equiv \bar{\alpha}\left(\left[\bar{q}_{0}\right]\right) \equiv \bar{q}_{0} \equiv f\left(q_{0}\right) .
$$

so $g$ agrees with $f$ on $\mathbb{F}$.
Now suppose $g: \mathbb{F}(r) \rightarrow \overline{\mathbb{F}}(\bar{r})$ is an isomorphism which agrees with $f$ on $\mathbb{F}$ where $\bar{p}(\bar{r})=0$ and $g(r)=\bar{r}$. How many can there be? There is one by the above. I need to show it must have the above form. To do this, note that

$$
g \circ \alpha=\bar{\alpha} \circ h \text { on } \mathbb{F}[x] /(p(x))
$$

This follows from the definition and the fact that $g$ agrees with $f$ on $\mathbb{F}$. Indeed, if $q(x)=$ $a_{m} x^{m}+\cdots+a_{1} x^{1}+a_{0}$, then, since $g$ is an isomorphism which agrees with $f$ on $\mathbb{F}$, and $g(r)=\bar{r}$,

$$
g(q(r))=f\left(a_{m}\right) \bar{r}^{m}+\cdots+f\left(a_{1}\right) \bar{r}^{1}+f\left(a_{0}\right) \equiv \bar{q}(\bar{r})
$$

Therefore,

$$
g \circ \alpha([q(x)]) \equiv g(q(r))=\bar{q}(\bar{r}), \bar{\alpha} \circ h([q(x)])=\bar{\alpha}([f(q(x))]) \equiv \bar{\alpha}([\bar{q}])=\bar{q}(\bar{r})
$$

It follows that $g=\bar{\alpha} \circ h \circ \alpha^{-1}$. Thus, there is only one such homomorphism and it is what was just obtained.

The following corollary emphasizes the main content of the above theorem.
Corollary 9.4.6 Let $f: \mathbb{F} \rightarrow \overline{\mathbb{F}}$ be an isomorphism and let $\bar{p}(x) \equiv f(p(x))$ where $p(x)$ is the minimum polynomial for algebraic $r$. Then for each root $\bar{r}$ of $\bar{p}(x)$ there is an isomorphism from $\mathbb{F}(r)$ to $\overline{\mathbb{F}}(\bar{r})$ which extends $f$ and satisfies $g(r)=\bar{r}$. Also, if $g$ is a monomophism extending from $\mathbb{F}(r)$ to $\overline{\mathbb{K}}$ where $\overline{\mathbb{K}}$ is a field which contains all roots of $\bar{p}(x)$, then there is a root $\bar{r}$ of $\bar{p}(x)$ such that $g(r)=\bar{r}$.

Proof: It only remains to verify the last claim. Let $g$ be a monomorphism. I need to find a root $\bar{r}$ of $\bar{p}(x)$. Let $\bar{r} \equiv g(r)$. Then if

$$
p(x)=x^{m}+a_{m-1} x^{m-1}+\cdots+a_{1} x^{1}+a_{0},
$$

it follows that

$$
\begin{aligned}
\bar{p}(x) & =x^{m}+f\left(a_{m-1}\right) x^{m-1}+\cdots+f\left(a_{1}\right) x^{1}+f\left(a_{0}\right) \\
& =x^{m}+g\left(a_{m-1}\right) x^{m-1}+\cdots+g\left(a_{1}\right) x^{1}+g\left(a_{0}\right)
\end{aligned}
$$

and so,

$$
\begin{aligned}
\bar{p}(g(r)) & =g(r)^{m}+g\left(a_{m-1}\right) g(r)^{m-1}+\cdots+g\left(a_{1}\right) g(r)^{1}+g\left(a_{0}\right) \\
& =g(p(r))=g(0)=0
\end{aligned}
$$

Lemma 9.4.7 If $p(x)$ is a monic irreducible polynomial, then it is the minimum polynomial for each of its roots.

Proof: If $r$ is a root of $p(x)$, then let $q(x)$ be the minimum polynomial for $r$. Then

$$
p(x)=q(x) k(x)+R(x)
$$

where $R(x)$ is 0 or else has smaller degree than $q(x)$. However, $R(r)=0$ and this contradicts $q(x)$ being the minimum polynomial of $r$. Hence $q(x)$ divides $p(x)$ or else $k(x)=1$. The latter possibility must be the case because $p(x)$ is irreducible.

This lemma is about to be used in the proof of the following theorem. It involves the splitting fields $\left.\mathbb{K}=\mathbb{F}\left[r_{1}, \ldots, r_{m}\right], \overline{\mathbb{K}}=\overline{\mathbb{F}}^{[ } \bar{r}_{1}, \ldots, \bar{r}_{m}\right]$ of $p(x), \bar{p}(x)$ where $\eta$ is an isomorphism of $\mathbb{F}$ and $\overline{\mathbb{F}}$ as described above. See [26]. Here is a little diagram which describes what this theorem says. It is about isomorphisms of $\mathbb{K}$ and $\overline{\mathbb{K}}$ which extend a given isomorphism $\eta$ : $\mathbb{F} \rightarrow \overline{\mathbb{F}}$.

Definition 9.4.8 Recall that the symbol $[\mathbb{K}: \mathbb{F}]$ where $\mathbb{K}$ is a field extension of $\mathbb{F}$ means the dimension of the vector space $\mathbb{K}$ with field of scalars $\mathbb{F}$.

| F | $\xrightarrow{\square}$ | $\overline{\mathbb{F}}$ |
| :---: | :---: | :---: |
| $p(x)$ | $\eta p(x)=\bar{p}(x)$ | $\bar{p}(x)$ |
| $\mathbb{F}\left(r_{1}, \cdots, r_{n}\right)$ | $\xrightarrow{\text { 洨 }}$ | $\overline{\mathbb{F}}\left(\bar{r}_{1}, \cdots, \bar{r}_{n}\right)$ |
|  | $i=1, \cdots, m,\left\{\begin{array}{l}m \leq[\mathbb{K}: \mathbb{F}] \\ m=[\mathbb{K}: \mathbb{F}], \bar{r}_{i} \neq \bar{r}_{j}\end{array}\right.$ |  |

In the next theorem, the polynomials $p(x), \bar{p}(x)$ are not necessarily irreducible.
Theorem 9.4.9 Let $\eta$ be an isomorphism from $\mathbb{F}$ to $\overline{\mathbb{F}}$ and let $\mathbb{K}, \overline{\mathbb{K}}$ be splitting fields of $p(x)$ and $\bar{p}(x)$. If the roots of $p(x)$ are $\left\{r_{1}, \cdots, r_{m}\right\}$, recall that $\mathbb{F}\left(r_{1}, \cdots, r_{m}\right) \equiv \mathbb{K}$ is a field since each root of the polynomials is algebraic. Thus, this must be the splitting field of $p(x)$, the smallest field which contains each of the roots of $p(x)$. The case is similar for $\overline{\mathbb{K}}$. Then

1. There exist at most $[\mathbb{K}: \mathbb{F}]$ isomorphisms $\zeta_{i}: \mathbb{K} \rightarrow \mathbb{K}$ which extend $\eta$.
2. If $\left\{\bar{r}_{1}, \cdots, \bar{r}_{n}\right\}$ are distinct, then there exist exactly $[\mathbb{K}: \mathbb{F}]$ isomorphisms of the above sort.
3. In either case, the two splitting fields $\mathbb{K}, \overline{\mathbb{K}}$ are isomorphic with any of these $\zeta_{i}$ serving as an isomorphism.

Proof: Suppose $[\mathbb{K}: \mathbb{F}]=1$. Say a basis for $\mathbb{K}$ is $\{r\}$. Then $\{1, r\}$ is dependent and so there exist $a, b \in \mathbb{F}$, not both zero such that $a+b r=0$. Then it follows that $r \in \mathbb{F}$ and so in this case $\mathbb{F}=\mathbb{K}$. Then the isomorphism which extends $\eta$ is just $\eta$ itself and there is exactly 1 isomorphism.

Next suppose $[\mathbb{K}: \mathbb{F}]>1$. Then $p(x)$ has a factor $q(x)$ irreducible over $\mathbb{F}$ which has degree larger than 1. If not, you could factor $p(x)$ as linear factors and so all the roots would be in $\mathbb{F}$ so the dimension $[\mathbb{K}: \mathbb{F}]$ would equal 1 . Without loss of generality, let the roots of $q(x)$ in $\mathbb{K}$ be $\left\{r_{1}, \cdots, r_{m}\right\}$. Thus

$$
q(x)=\prod_{i=1}^{m}\left(x-r_{i}\right), \quad p(x)=\prod_{i=1}^{n}\left(x-r_{i}\right)
$$

Now $\bar{q}(x) \equiv \eta(q(x))$ defined analogously to $\bar{p}(x)$, also has degree at least 2. Furthermore, it divides $\bar{p}(x)$ all of whose roots are in $\overline{\mathbb{K}}$. This is obvious because $\eta$ is an isomorphism. You have

$$
l(x) q(x)=p(x) \text { so } \bar{l}(x) \bar{q}(x)=\bar{p}(x)
$$

Denote the roots of $\bar{q}(x)$ in $\overline{\mathbb{K}}$ as $\left\{\bar{r}_{1}, \cdots, \bar{r}_{m}\right\}$ where they are counted according to multiplicity.

Recall why $\left[\mathbb{F}\left(r_{1}\right): \mathbb{F}\right]=m$. It is because $q(x)$ is irreducible and monic so by Lemma 9.4.7, it is the minimum polynomial for each of the $r_{i}$. Since $q(x)$ is irreducible, it follows that $1, r_{1}, r_{1}^{2}, \ldots, r_{1}^{m-1}$ must be independent so the dimension is at least $m$. However, it is not more than $m$ because $q(x)$ is of degree $m$. Thus, using the division algorithm, everything in $\mathbb{F}\left(r_{1}\right)$ is expressible as a polynomial in $r_{1}$ of degree less than $m$.

Then from Corollary 9.4.6, using $q(x)$ and $\bar{q}(x)$ in place of the $p(x)$ and $\bar{p}(x)$ in this corollary, there exist $k \leq m$ one to one homomorphisms (monomorphisms) $\zeta_{i}$ mapping $\mathbb{F}\left(r_{1}\right)$ to $\overline{\mathbb{K}} \equiv \overline{\mathbb{F}}\left(\bar{r}_{1}, \cdots, \bar{r}_{n}\right)$, one for each distinct root of $\bar{q}(x)$ in $\overline{\mathbb{K}}$. These are $\left\{\xi_{1}, \ldots, \xi_{k}\right\}$ where $k \leq m$. If the roots of $\bar{p}(x)$ are distinct, then this is sufficient to imply that the roots of $\bar{q}(x)$ are also distinct, and $k=m=\left[\mathbb{F}\left(r_{1}\right): \mathbb{F}\right]$. Otherwise, maybe $k<m$. (It is conceivable that $\bar{q}(x)$ might have repeated roots in $\overline{\mathbb{K}}$.) Then by Proposition 3.4.13,

$$
[\mathbb{K}: \mathbb{F}]=\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right] \overbrace{\left.\mathbb{F}\left(r_{1}\right): \mathbb{F}\right]}^{>1}
$$

and so $\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right]<[\mathbb{K}: \mathbb{F}]$.
Therefore, by induction, two things happen:
1.) Each of these one to one homomorphisms mapping $\mathbb{F}\left(r_{1}\right)$ to $\overline{\mathbb{K}}$ called $\xi_{i}$ for $i \leq$ $k \leq m=\left[\mathbb{F}\left(r_{1}\right): \mathbb{F}\right]$ extends to an isomorphism from $\mathbb{K}$ to $\overline{\mathbb{K}}$.
2.) For each of these $\zeta_{i}$, there are no more than $\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right]$ extensions of these isomorphisms, exactly $\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right]$ in case the roots of $\bar{p}(x)$ are distinct.

Therefore, if the roots of $\bar{p}(x)$ are distinct, this has shown that there are

$$
\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right] m=\left[\mathbb{K}: \mathbb{F}\left(r_{1}\right)\right]\left[\mathbb{F}\left(r_{1}\right): \mathbb{F}\right]=[\mathbb{K}: \mathbb{F}]
$$

isomorphisms of $\mathbb{K}$ to $\overline{\mathbb{K}}$ which agree with $\eta$ on $\mathbb{F}$. If the roots of $\bar{p}(x)$ are not distinct, then maybe there are fewer than $[\mathbb{K}: \mathbb{F}]$ extensions of $\eta$.

Is this all of the isomorphisms? Suppose $\zeta$ is such an isomorphism of $\mathbb{K}$ and $\overline{\mathbb{K}}$. Then consider its restriction to $\mathbb{F}\left(r_{1}\right)$. By Corollary 9.4.6, this restriction must coincide with one of the $\zeta_{i}$ chosen earlier. Then by induction, $\zeta$ is one of the extensions of the $\zeta_{i}$ just mentioned. Thus, in particular, $\mathbb{K}$ and $\overline{\mathbb{K}}$ are isomorphic.

### 9.4.1 The Galois Group

First, here is the definition of a Group.
Definition 9.4.10 A group $G$ is a nonempty set with an operation, denoted here as $\cdot$ such that the following axioms hold. (Often the operation is composition.)

1. For $\alpha, \beta, \gamma \in G,(\alpha \cdot \beta) \cdot \gamma=\alpha \cdot(\beta \cdot \gamma)$. We usually don't bother to write the $\cdot$.
2. There exists $\imath \in G$ such that $\alpha \imath=\imath \alpha=\alpha$
3. For every $\alpha \in G$, there exists $\alpha^{-1} \in G$ such that $\alpha \alpha^{-1}=\alpha^{-1} \alpha=1$.

In Theorem 9.4.9, consider the case where $\mathbb{F}=\overline{\mathbb{F}}$ and the isomorphism of $\mathbb{F}$ with itself is just the identity.

Definition 9.4.11 When $\mathbb{K}$ is a finite extension of $\mathbb{L}$, denote by $G(\mathbb{K}, \mathbb{L})$ the automorphisms of $\mathbb{K}$ which leave $\mathbb{L}$ fixed. For a finite set $S$, denote by $|S|$ as the number of elements of $S$.

Most of the following theorem was shown earlier in Theorem 9.4.9.
Theorem 9.4.12 Let $\mathbb{K}$ be the splitting field of $p(x)$ over the field $\mathbb{F}$. Thus $\mathbb{K}$ consists of $\mathbb{F}\left[a_{1}, \ldots, a_{n}\right]$ where $\left\{a_{1}, \ldots, a_{n}\right\}$ are the roots of $p(x)$. Then

$$
\begin{equation*}
|G(\mathbb{K}, \mathbb{F})| \leq[\mathbb{K}: \mathbb{F}] \tag{9.22}
\end{equation*}
$$

When the roots of $\bar{p}(x)=p(x)$ are distinct, equality holds in the above. If the roots are listed according to multiplicity, the automorphisms are determined by the permutations of the roots. When the roots are distinct, $|G(\mathbb{K}, \mathbb{F})|=n!$. Also, $G(\mathbb{K}, \mathbb{F})$ is a group for the operation being composition.

Proof: So how large is $|G(\mathbb{K}, \mathbb{F})|$ in case $p(x)$ is a polynomial of degree $n$ which has $n$ distinct roots? Let $p(x)$ be a monic polynomial with roots in $\mathbb{K},\left\{r_{1}, \cdots, r_{n}\right\}$ and suppose that none of the $r_{i}$ is in $\mathbb{F}$. Thus

$$
p(x)=x^{n}+a_{1} x^{n-1}+a_{2} x^{n-2}+\cdots+a_{n}=\prod_{k=1}^{n}\left(x-r_{k}\right), a_{i} \in \mathbb{F}
$$

Thus $\mathbb{K}=\mathbb{F}\left[r_{1}, \cdots, r_{n}\right]$. Let $\sigma$ be a mapping from $\left\{r_{1}, \cdots, r_{n}\right\}$ to $\left\{r_{1}, \cdots, r_{n}\right\}$, say $r_{j} \rightarrow r_{i_{j}}$. In other words $\sigma$ produces a permutation of these roots. Consider the following way of obtaining something in $G(\mathbb{K}, \mathbb{F})$ from $\sigma$. If you have a typical thing in $\mathbb{K}$, you can obtain another thing in $\mathbb{K}$ by replacing each $r_{j}$ with $r_{i_{j}}$ in an element of $\mathbb{F}\left[r_{1}, \cdots, r_{n}\right]$, a polynomial which has coefficients in $\mathbb{F}$. Furthermore, if you do this, then the resulting map from $\mathbb{K}$ to $\mathbb{K}$ is an automorphism, preserving the operations of multiplication and addition. Does it keep $\mathbb{F}$ fixed? Of course it does because you don't change the coefficients of the polynomials which are always in $\mathbb{F}$. Thus every permutation of the roots determines an automorphism of $\mathbb{K}$.

Now suppose $\sigma$ is an automorphism of $\mathbb{K}$ and the roots of $p(x)$ are distinct. Does $\sigma$ determine a permutation of the roots? If $r_{i}$ is a root, what of $\sigma\left(r_{i}\right)$ ? Is it also a root simply due to $\sigma$ being an automorphism? Note that $\sigma(0)=0$ and so $\sigma(0)=0=\sigma\left(p\left(r_{i}\right)\right)=$ $p\left(\sigma\left(r_{i}\right)\right)$, the last from the assumption that $\sigma$ is an automorphism. Thus $\sigma$ maps roots to roots. Since it is one to one and the roots are distinct, it must be a permutation. It follows that $|G(\mathbb{K}, \mathbb{F})|$ equals the number of permutations of $\left\{r_{1}, \cdots, r_{n}\right\}$ which is $n$ ! and that there is a one to one correspondence between the permutations of the roots and $G(\mathbb{K}, \mathbb{F})$. It is always the case that an automorphism takes roots to roots, but if the roots are repeated, then there may be fewer than $n$ ! of these automorphisms.

Now consider the claim about $G(\mathbb{K}, \mathbb{F})$ being a group.
The associative law $(\alpha \cdot \beta) \cdot \gamma=\alpha \cdot(\beta \cdot \gamma)$ is obvious. This is just the way composition acts.

The identity $\imath$ is just the identity map, clearly an automorphism which fixes $\mathbb{F}$.
Each automorphism is, by definition one to one and onto. Therefore, the inverse must also be an automorphism. Indeed, if $\sigma(x), \sigma(y)$ are two generic things in $\mathbb{K}$, then

$$
\sigma^{-1}(\sigma(x) \sigma(y))=\sigma^{-1}(\sigma(x y))=x y=\sigma^{-1}(\sigma(x)) \sigma^{-1}(\sigma(y))
$$

That $\sigma^{-1}$ is an automorphism with respect to addition goes the same way. The estimate on the size is from 9.22 . Does the inverse fix $\mathbb{F}$ ? Consider $\alpha, \alpha^{2}, \cdots$. Because of the estimate on the size of $G(\mathbb{K}, \mathbb{F})$, you must have $\alpha^{m}=\alpha^{n}$ for some $m<n$. Hence multiply on the left by $\left(\alpha^{-1}\right)^{m}$ to get $l=\alpha^{n-m}$. Thus $\alpha^{-1}=\alpha^{(n-1)-m}$ which is $\alpha$ raised to a nonnegative power. The right leaves $\mathbb{F}$ fixed and so the left does also.

In the above, there is a field which is a finite extension of a smaller field and the group of automorphisms which leave the given smaller field fixed was discussed. Next is a more general notion in which there is given a group of automorphisms. This group will determine a smaller field called a fixed field.

Definition 9.4.13 Let $G$ be a group of automorphisms of a field $\mathbb{K}$. Then denote by $\mathbb{K}_{G}$ the fixed field of $G$. Thus

$$
\mathbb{K}_{G} \equiv\{x \in \mathbb{K}: \sigma(x)=x \text { for all } \sigma \in G\}
$$

Lemma 9.4.14 Let $G$ be a group of automorphisms of a field $\mathbb{K}$. Then $\mathbb{K}_{G}$ is a field.
Proof: It suffices to show that $\mathbb{K}_{G}$ is closed with respect to the operations of the field $\mathbb{K}$. Suppose then $x, y \in \mathbb{K}_{G}$. Is $x+y \in \mathbb{K}_{G}$ ? Is $x y \in \mathbb{K}_{G}$ ? This is obviously so because the things in $G$ are automorphisms. Thus if $\theta \in G, \theta(x+y)=\theta x+\theta y=x+y$. It is similar with multiplication.

There is another fundamental estimate due to Artin and is certainly not obvious. I also found this in [26]. There is more there about some of these things than what I am including. Above it was shown that $|G(\mathbb{K}, \mathbb{F})| \leq[\mathbb{K}: \mathbb{F}]$. This fundamental estimate goes the other direction when $\mathbb{F}$ is a fixed field.

Theorem 9.4.15 Let $\mathbb{K}$ be a field and let $G$ be a finite group of automorphisms of $\mathbb{K}$. Then

$$
\begin{equation*}
\left[\mathbb{K}: \mathbb{K}_{G}\right] \leq|G| \tag{9.23}
\end{equation*}
$$

Proof: Let $G=\left\{\sigma_{1}, \cdots, \sigma_{n}\right\}, \sigma_{1}=\imath$ the identity map and suppose $\left\{u_{1}, \cdots, u_{m}\right\}$ is a linearly independent set in $\mathbb{K}$ with respect to the field $\mathbb{K}_{G}$. These $\sigma_{i}$ are the automorphisms of $\mathbb{K}$. Suppose $m>n$. Then consider the system of equations

$$
\begin{gather*}
\sigma_{1}\left(u_{1}\right) x_{1}+\sigma_{1}\left(u_{2}\right) x_{2}+\cdots+\sigma_{1}\left(u_{m}\right) x_{m}=0 \\
\sigma_{2}\left(u_{1}\right) x_{1}+\sigma_{2}\left(u_{2}\right) x_{2}+\cdots+\sigma_{2}\left(u_{m}\right) x_{m}=0 \\
\vdots  \tag{9.24}\\
\sigma_{n}\left(u_{1}\right) x_{1}+\sigma_{n}\left(u_{2}\right) x_{2}+\cdots+\sigma_{n}\left(u_{m}\right) x_{m}=0
\end{gather*}
$$

which is of the form $M \boldsymbol{x}=\mathbf{0}$ for $\boldsymbol{x} \in \mathbb{K}^{m}$. Since $M$ has more columns than rows, there exists a nonzero solution $\boldsymbol{x} \in \mathbb{K}^{m}$ to the above system. Let the solution $\boldsymbol{x}$ be one which has the least possible number of nonzero entries. Without loss of generality, some $x_{k}=1$ for some $k$.

If $\sigma_{r}\left(x_{k}\right)=x_{k}$ for all $x_{k}$ and for each $r$, then the $x_{k}$ are each in $\mathbb{K}_{G}$ and so the first equation would say

$$
u_{1} x_{1}+u_{2} x_{2}+\cdots+u_{m} x_{m}=0
$$

with not all $x_{i}=0$ and this contradicts the linear independence of the $u_{i}$. Therefore, there exists $l \neq k$ and $\sigma_{r}$ such that $\sigma_{r}\left(x_{l}\right) \neq x_{l}$. For purposes of illustration, say $l>k$. Now do $\sigma_{r}$ to both sides of all the above equations. This yields, after re ordering the resulting equations a list of equations of the form

$$
\begin{gathered}
\sigma_{1}\left(u_{1}\right) \sigma_{r}\left(x_{1}\right)+\cdots+\sigma_{1}\left(u_{k}\right) 1+\cdots+\sigma_{1}\left(u_{l}\right) \sigma_{r}\left(x_{l}\right)+\cdots+\sigma_{1}\left(u_{m}\right) \sigma_{r}\left(x_{m}\right)=0 \\
\sigma_{2}\left(u_{1}\right) \sigma_{r}\left(x_{1}\right)+\cdots+\sigma_{2}\left(u_{k}\right) 1+\cdots+\sigma_{2}\left(u_{l}\right) \sigma_{r}\left(x_{l}\right)+\cdots+\sigma_{2}\left(u_{m}\right) \sigma_{r}\left(x_{m}\right)=0 \\
\vdots \\
\sigma_{n}\left(u_{1}\right) \sigma_{r}\left(x_{1}\right)+\cdots+\sigma_{n}\left(u_{k}\right) 1+\cdots+\sigma_{n}\left(u_{l}\right) \sigma_{r}\left(x_{l}\right)+\cdots+\sigma_{n}\left(u_{m}\right) \sigma_{r}\left(x_{m}\right)=0
\end{gathered}
$$

This is because $\sigma(1)=1$ if $\sigma$ is an automorphism. It is of the form $M \sigma_{r}(\boldsymbol{x})=\mathbf{0}$. The original system in 9.24 is of the form

$$
\begin{gather*}
\sigma_{1}\left(u_{1}\right) x_{1}+\cdots+\sigma_{1}\left(u_{k}\right) 1+\cdots+\sigma_{1}\left(u_{l}\right) x_{l}+\cdots+\sigma_{1}\left(u_{m}\right) x_{m}=0 \\
\sigma_{2}\left(u_{1}\right) x_{1}+\cdots+\sigma_{2}\left(u_{k}\right) 1+\cdots+\sigma_{2}\left(u_{l}\right) x_{l}+\cdots+\sigma_{2}\left(u_{m}\right) x_{m}=0 \\
\vdots \\
\sigma_{n}\left(u_{1}\right) x_{1}+\cdots+\sigma_{n}\left(u_{k}\right) 1+\cdots+\sigma_{n}\left(u_{l}\right) x_{l}+\cdots+\sigma_{n}\left(u_{m}\right) x_{m}=0
\end{gather*}
$$

which will be denoted as $M \boldsymbol{x}=\mathbf{0}$. Thus $M\left(\sigma_{r}(\boldsymbol{x})-\boldsymbol{x}\right)=\mathbf{0}$ where $\boldsymbol{y} \equiv \sigma_{r}(\boldsymbol{x})-\boldsymbol{x} \neq \mathbf{0}$. If any $x_{k}$ is 0 , then $\sigma_{r}\left(x_{k}\right)=0$. Thus all zero entries of $\boldsymbol{x}$ remain 0 in $\boldsymbol{y}$ and $y_{k}=0$ whereas $x_{k} \neq 0$ so $\boldsymbol{y}$ has fewer nonzero entries than $\boldsymbol{x}$ contradicting the choice of $\boldsymbol{x}$ as the one with fewest nonzero entries such that $M \boldsymbol{x}=\mathbf{0}$.

With the above estimate, here is another relation between the fixed fields and subgroups of automorphisms.

Proposition 9.4.16 Let $H$ be a finite group of automorphisms defined on a field $\mathbb{K}$. Then for $\mathbb{K}_{H}$ the fixed field,

$$
G\left(\mathbb{K}, \mathbb{K}_{H}\right)=H
$$

Proof: If $\sigma \in H$, then by definition of $\mathbb{K}_{H}, \sigma \in G\left(\mathbb{K}, \mathbb{K}_{H}\right)$ so $H \subseteq G\left(\mathbb{K}, \mathbb{K}_{H}\right)$. Then by Theorem 9.4.15 and Theorem 9.4.12,

$$
|H| \geq\left[\mathbb{K}: \mathbb{K}_{H}\right] \geq\left|G\left(\mathbb{K}, \mathbb{K}_{H}\right)\right| \geq|H|
$$

and so $H=G\left(\mathbb{K}, \mathbb{K}_{H}\right)$.
For $H$ a group of automorphisms of $G(\mathbb{K}, \mathbb{F})$, let $H x$ be all $h x$ for $h \in H$. Thus $H x=x$ means $h x=x$ for all $h \in H . \mathbb{K}_{H}=\{x \in \mathbb{K}: H x=x\}$.

Note how this proposition shows $G(\mathbb{K}, \mathbb{F})=G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}\right)$. Thus

$$
|G(\mathbb{K}, \mathbb{F})|=\left|G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}\right)\right|=\left[\mathbb{K}: \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}\right]
$$

Is $\mathbb{K}_{G(\mathbb{K}, \mathbb{F})}=\mathbb{F}$ ? If $x \in \mathbb{F}, G(\mathbb{K}, \mathbb{F}) x=x$ so by definition, $x \in \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}$ and so $\mathbb{F} \subseteq \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}$. However, if $x \in \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}$ so $G(\mathbb{K}, \mathbb{F})$ fixes $x$, it is not at all clear that $x \in \mathbb{F}$. Maybe $G(\mathbb{K}, \mathbb{F})$ fixes more things than $\mathbb{F}$. Later a situation is given in which $\mathbb{K}_{G(\mathbb{K}, \mathbb{F})}=\mathbb{F}$.

Summary 9.4.17 The following are now available.

1. Let $\mathbb{K}$ be the splitting field of $p(x)$. Then $|G(\mathbb{K}, \mathbb{F})| \leq[\mathbb{K}: \mathbb{F}]$. If the roots of $p(x)$ are unique, then these are equal.
2. $|G(\mathbb{K}, \mathbb{F})| \leq n!$ and when the roots of $p(x)$ are distinct, $|G(\mathbb{K}, \mathbb{F})|=n$ !.
3. If $H$ is a finite group of automorphisms on an arbitrary field $\mathbb{K}$, then it follows that $G\left(\mathbb{K}, \mathbb{K}_{H}\right)=H$ where $\mathbb{K}_{H}$ is the fixed field of $H$.
4. $\mathbb{F} \subseteq \mathbb{K}_{G(\mathbb{K}, \mathbb{F})}$ but it is not clear that these are equal.

### 9.4.2 Normal Field Extensions

The following is the definition of a normal field extension.
Definition 9.4.18 Let $\mathbb{K}$ be a finite dimensional extension of a field $\mathbb{F}$ such that every element of $\mathbb{K}$ is algebraic over $\mathbb{F}$, that is, each element of $\mathbb{K}$ is a root of some polynomial in $\mathbb{F}[x]$. Then $\mathbb{K}$ is called a normal extension if for every $k \in \mathbb{K}$ all roots of the minimum polynomial of $k$ are contained in $\mathbb{K}$.

So what are some ways to tell that a field is a normal extension? It turns out that if $\mathbb{K}$ is a splitting field of $f(x) \in \mathbb{F}[x]$, then $\mathbb{K}$ is a normal extension. I found this in [26]. This is an amazing result.

Proposition 9.4.19 The following are valid

1. Let $\mathbb{K}$ be a splitting field of $f(x) \in \mathbb{F}[x]$. Then $\mathbb{K}$ is a normal extension.
2. If $\mathbb{L}$ is an intermediate field between $\mathbb{F}$ and $\mathbb{K}$ where $\mathbb{K}$ is a normal field extension of $\mathbb{F}$, then $\mathbb{L}$ is also a normal extension of $\mathbb{F}$.

Proof: 1.) Let $r \in \mathbb{K} \equiv \mathbb{F}\left(a_{1}, \ldots, a_{q}\right)$ where $\left\{a_{1}, \ldots, a_{q}\right\}$ are the roots of $f(x)$ and let $g(x)$ be the minimum polynomial of $r$ with coefficients in $\mathbb{F}$. Thus, $g(x)$ is an irreducible monic polynomial in $\mathbb{F}[x]$ having $r$ as a root. It is required to show that every other root of $g(x)$ is in $\mathbb{K}$. Let the roots of $g(x)$ in a splitting field be $\left\{r_{1}=r, r_{2}, \cdots, r_{m}\right\}$. Now $g(x)$ is the minimum polynomial of $r_{j}$ over $\mathbb{F}$ because $g(x)$ is irreducible by Lemma 9.4.7.

By Theorem 9.4.5, there exists an isomorphism $\eta$ of $\mathbb{F}\left(r_{1}\right)$ and $\mathbb{F}\left(r_{j}\right)$ which fixes $\mathbb{F}$ and maps $r_{1}$ to $r_{j}$. Thus $\eta$ is an extension of the identity on $\mathbb{F}$. Now $\mathbb{K}\left(r_{1}\right)$ and $\mathbb{K}\left(r_{j}\right)$ are splitting fields of $f(x)$ over $\mathbb{F}\left(r_{1}\right)$ and $\mathbb{F}\left(r_{j}\right)$ respectively. By Theorem 9.4.9, the two fields $\mathbb{K}\left(r_{1}\right)$ and $\mathbb{K}\left(r_{j}\right)$ are isomorphic, the isomorphism, $\zeta$ extending $\eta$. Hence

$$
\left[\mathbb{K}\left(r_{1}\right): \mathbb{K}\right]=\left[\mathbb{K}\left(r_{j}\right): \mathbb{K}\right]
$$

But $r_{1} \in \mathbb{K}$ and so $\mathbb{K}\left(r_{1}\right)=\mathbb{K}$. Therefore, $\left[\mathbb{K}\left(r_{j}\right): \mathbb{K}\right]=1$ and so $\mathbb{K}=\mathbb{K}\left(r_{j}\right)$ and so $r_{j}$ is also in $\mathbb{K}$. Thus all the roots of $g(x)$ are in $\mathbb{K}$.
2.) Consider the last assertion. Suppose $r=r_{1} \in \mathbb{L}$ where the minimum polynomial for $r$ is denoted by $q(x)$. Then since $\mathbb{K}$ is a normal extension, all the roots of $q(x)$ are in $\mathbb{K}$. Let them be $\left\{r_{1}, \cdots, r_{m}\right\}$. By Theorem 9.4.5 applied to the identity map on $\mathbb{L}$, there exists an isomorphism $\theta: \mathbb{L}\left(r_{1}\right) \rightarrow \mathbb{L}\left(r_{j}\right)$ which fixes $\mathbb{L}$ and takes $r_{1}$ to $r_{j}$. But this implies that

$$
1=\left[\mathbb{L}\left(r_{1}\right): \mathbb{L}\right]=\left[\mathbb{L}\left(r_{j}\right): \mathbb{L}\right]
$$

Hence $r_{j} \in \mathbb{L}$ also. If $r_{j} \notin \mathbb{L}$, then $\left\{1, r_{j}\right\}$ is independent and so the dimension would be at least 2 . Since $r$ was an arbitrary element of $\mathbb{L}$, this shows that $\mathbb{L}$ is normal.

### 9.4.3 Normal Subgroups and Quotient Groups

When you look at groups, one of the first things to consider is the notion of a normal subgroup. The word "normal" is greatly over used in math. Its meaning in this context is given next.

Definition 9.4.20 Let $G$ be a group. A subset $N$ of a group $G$ is called a subgroup if it contains 1 the identity and is closed with respect to the operation on $G$. That is, if $\alpha, \beta \in N$, then $\alpha \beta \in N$. Then a subgroup $N$ is said to be a normal subgroup if whenever $\alpha \in G$,

$$
\alpha^{-1} N \alpha \subseteq N
$$

The important thing about normal subgroups is that you can define the quotient group $G / N$.

Definition 9.4.21 Let $N$ be a subgroup of $G$. Define an equivalence relation $\sim$ as follows.

$$
\alpha \sim \beta \text { means } \alpha^{-1} \beta \in N
$$

Why is this an equivalence relation? It is clear that $\alpha \sim \alpha$ because $\alpha^{-1} \alpha=\imath \in N$ since $N$ is a subgroup. If $\alpha \sim \beta$, then $\alpha^{-1} \beta \in N$ and so, since $N$ is a subgroup,

$$
\left(\alpha^{-1} \beta\right)^{-1}=\beta^{-1} \alpha \in N
$$

which shows that $\beta \sim \alpha$. Now suppose $\alpha \sim \beta$ and $\beta \sim \gamma$. Then $\alpha^{-1} \beta \in N$ and $\beta^{-1} \gamma \in N$. Then since $N$ is a subgroup

$$
\alpha^{-1} \beta \beta^{-1} \gamma=\alpha^{-1} \gamma \in N
$$

and so $\alpha \sim \gamma$ which shows that it is an equivalence relation as claimed. Denote by $[\alpha]$ the equivalence class determined by $\alpha$.

Now in the case of $N$ a normal subgroup, you can consider the quotient group.
Definition 9.4.22 Let $N$ be a normal subgroup of a group $G$ and define $G / N$ as the set of all equivalence classes with respect to the above equivalence relation. Also define

$$
[\alpha][\beta] \equiv[\alpha \beta]
$$

Proposition 9.4.23 The above definition is well defined and it also makes $G / N$ into a group.

Proof: First consider the claim that the definition is well defined. Suppose then that $\alpha \sim \bar{\alpha}$ and $\beta \sim \bar{\beta}$. It is required to show that

$$
[\alpha \beta]=[\bar{\alpha} \bar{\beta}]
$$

Is $(\alpha \beta)^{-1} \bar{\alpha} \bar{\beta} \in N$ ? Is $\beta^{-1} \alpha^{-1} \bar{\alpha} \bar{\beta} \in N$ ?

$$
\begin{aligned}
(\alpha \beta)^{-1} \bar{\alpha} \bar{\beta} & =\beta^{-1} \alpha^{-1} \bar{\alpha} \bar{\alpha}=\beta^{-1} \overbrace{\alpha^{-1} \bar{\alpha} \bar{\beta}}^{\in N} \\
& =\overbrace{\beta^{-1}\left(\alpha^{-1} \bar{\alpha}\right) \beta \beta^{-1} \bar{\beta}}^{\in N}=n_{1} n_{2} \in N
\end{aligned}
$$

Thus the operation is well defined. Clearly the identity is $[\imath]$ where $l$ is the identity in $G$ and the inverse is $\left[\alpha^{-1}\right]$ where $\alpha^{-1}$ is the inverse for $\alpha$ in $G$. The associative law is also obvious.

Note that it was important to have the subgroup be normal in order to have the operation defined on the quotient group consisting of the set of equivalence classes.

### 9.4.4 Separable Polynomials

This is a good time to make a very important observation about irreducible polynomials.
Lemma 9.4.24 Suppose $q(x) \neq p(x)$ are both irreducible polynomials over a field $\mathbb{F}$. Then there is no root common to both $p(x)$ and $q(x)$.

Proof: If $l(x)$ is a monic polynomial which divides them both, then $l(x)$ must equal 1. Otherwise, it would equal $p(x)$ and $q(x)$ which would require these two to be equal. Thus $p(x)$ and $q(x)$ are relatively prime and there exist polynomials $a(x), b(x)$ having coefficients in $\mathbb{F}$ such that

$$
a(x) p(x)+b(x) q(x)=1
$$

Now if $p(x)$ and $q(x)$ share a root $r$, then $(x-r)$ divides both sides of the above in $\mathbb{K}[x]$ where $\mathbb{K}$ is a field which contains all roots of both polynomials. But this is impossible.

Now here is an important definition of a class of polynomials which yield equality in the inequality of Theorem 9.4.12. We know that if $p(x)$ of this theorem has distinct roots, then equality holds. However, there is a more general kind of polynomial which also gives equality.

Definition 9.4.25 Let $p(x)$ be a polynomial having coefficients in a field $\mathbb{F}$. Also let $\mathbb{K}$ be a splitting field. Then $p(x)$ is separable if it is of the form

$$
p(x)=\prod_{i=1}^{m} q_{i}(x)^{k_{i}}
$$

where each $q_{i}(x)$ is irreducible over $\mathbb{F}$ and each $q_{i}(x)$ has distinct roots in $\mathbb{K}$. From the above lemma, no two $q_{i}(x)$ share a root. Thus

$$
p_{1}(x) \equiv \prod_{i=1}^{m} q_{i}(x)
$$

has distinct roots in $\mathbb{K}$.
Example 9.4.26 For example, consider the case where $\mathbb{F}=\mathbb{Q}$ and the polynomial is of the form

$$
\left(x^{2}+1\right)^{2}\left(x^{2}-2\right)^{2}=x^{8}-2 x^{6}-3 x^{4}+4 x^{2}+4
$$

Then let $\mathbb{K}$ be the splitting field over $\mathbb{Q}, \mathbb{Q}[i, \sqrt{2}]$.The polynomials $x^{2}+1$ and $x^{2}-2$ are irreducible over $\mathbb{Q}$ and each has distinct roots in $\mathbb{K}$.

Then the following corollary is the reason why separable polynomials are so important. Also, one can show that if $\mathbb{F}$ contains a field which is isomorphic to $\mathbb{Q}$ then every polynomial with coefficients in $\mathbb{F}$ is separable. This will be done later after presenting the big results. This is equivalent to saying that the field has characteristic zero. In addition, the property of being separable holds in other situations.

Corollary 9.4.27 Let $\mathbb{K}$ be a splitting field of $p(x)$ over the field $\mathbb{F}$. Assume $p(x)$ is separable. Then

$$
|G(\mathbb{K}, \mathbb{F})|=[\mathbb{K}: \mathbb{F}]
$$

Proof: Just note that $\mathbb{K}$ is also the splitting field of $p_{1}(x)$, the product of the distinct irreducible factors and that from Lemma 9.4.24, $p_{1}(x)$ has distinct roots. Thus the conclusion follows from Theorem 9.4.9 or 9.4.12.

What if $\mathbb{L}$ is an intermediate field between $\mathbb{F}$ and $\mathbb{K}$ ? Then $p_{1}(x)$ still has coefficients in $\mathbb{L}$ and distinct roots in $\mathbb{K}$ and so it also follows that

$$
|G(\mathbb{K}, \mathbb{L})|=[\mathbb{K}: \mathbb{L}]
$$

Now the following says that you can start with $\mathbb{L}$, go to the group $G(\mathbb{K}, \mathbb{L})$ and then to the fixed field of this group and end up back where you started. More precisely,

Proposition 9.4.28 If $\mathbb{K}$ is a splitting field of $p(x)$ over the field $\mathbb{F}$ for separable $p(x)$, and if $\mathbb{L}$ is a field between $\mathbb{K}$ and $\mathbb{F}$, then $\mathbb{K}$ is also a splitting field of $p(x)$ over $\mathbb{L}$ and also

$$
\mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}
$$

In every case, even if $p(x)$ is not separable, $\mathbb{L} \subseteq \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$.
Proof: First of all, I claim that $\mathbb{L} \subseteq \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$ in any case. This is because of the definition. If $l \in \mathbb{L}$, then it is in the fixed field of $G(\mathbb{K}, \mathbb{L})$ since by definition, $G(\mathbb{K}, \mathbb{L})$ fixes everything in $\mathbb{L}$.

Now suppose $p(x)$ is separable. By the above Lemma 9.4.14 and Corollary 9.4.27,

$$
\begin{aligned}
|G(\mathbb{K}, \mathbb{L})| & =[\mathbb{K}: \mathbb{L}]=\left[\mathbb{K}: \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}\right]\left[\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}: \mathbb{L}\right] \\
& =\left|G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}\right)\right|\left[\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}: \mathbb{L}\right]=|G(\mathbb{K}, \mathbb{L})|\left[\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}: \mathbb{L}\right]
\end{aligned}
$$

which shows that $\left[\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}: \mathbb{L}\right]=1$ and so, it follows that $\mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$.
It is obvious that $\mathbb{K}$ is a splitting field of $p(x)$ over $\mathbb{L}$ because $\mathbb{L} \supseteq \mathbb{F}$ so the coefficients of $p(x)$ are in $\mathbb{L}$.

This has shown that in the context of $\mathbb{K}$ being a splitting field of a separable polynomial over $\mathbb{F}$ and $\mathbb{L}$ being an intermediate field, $\mathbb{L}$ is a fixed field of a subgroup of $G(\mathbb{K}, \mathbb{F})$, namely $G(\mathbb{K}, \mathbb{L})$.

$$
\mathbb{F} \subseteq \mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})} \subseteq \mathbb{K}
$$

In the above context, it is clear that $G(\mathbb{K}, \mathbb{L}) \subseteq G(\mathbb{K}, \mathbb{F})$ because if it fixes everything in $\mathbb{L}$ then it fixes everything in the smaller field $\mathbb{F}$. Then an obvious question is whether every subgroup of $G(\mathbb{K}, \mathbb{F})$ is obtained in the form $G(\mathbb{K}, \mathbb{L})$ for some intermediate field $\mathbb{L}$ ?

This leads to the following interesting correspondence in the case where $\mathbb{K}$ is a splitting field of a separable polynomial over a field $\mathbb{F}$.

$$
\quad \text { Subgroups of } G(\mathbb{K}, \mathbb{F})
$$

Then $\alpha \beta \mathbb{L}=\mathbb{L}$ and $\beta \alpha H=H$. Thus there exists a one to one correspondence between the fixed fields and the subgroups of $G(\mathbb{K}, \mathbb{F})$. The following theorem summarizes the above result.

Theorem 9.4.29 Let $\mathbb{K}$ be a splitting field of a separable polynomial $p(x)$ over a field $\mathbb{F}$. Then there exists a one to one correspondence between the fixed fields $\mathbb{K}_{H}$ for $H$ a subgroup of $G(\mathbb{K}, \mathbb{F})$ and the intermediate fields as described in the above. $H_{1} \subseteq H_{2}$ if and only if $\mathbb{K}_{H_{1}} \supseteq \mathbb{K}_{H_{2}}$. Also $|H|=\left[\mathbb{K}: \mathbb{K}_{H}\right]$.

Proof: The one to one correspondence is established above in Proposition 9.4.16 because $G\left(\mathbb{K}, \mathbb{K}_{H}\right)=H$ whenever $H$ is a subgroup of $G(\mathbb{K}, \mathbb{F})$. Thus each subgroup $H$ determines an intermediate field $\mathbb{K}_{H}$. Going the other direction, if $\mathbb{L}$ is an intermediate field, it comes from a sub-group because $G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}\right)=G(\mathbb{K}, \mathbb{L})$ so $\mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$ as mentioned earlier. The claim about the fixed fields is obvious because if the group is larger, then the fixed field must get harder because it is more difficult to fix everything using more automorphisms than with fewer automorphisms. Consider the estimate. From Theorem 9.4.15, $|H| \geq\left[\mathbb{K}: \mathbb{K}_{H}\right]$. But also, $H=G\left(\mathbb{K}, \mathbb{K}_{H}\right)$ from Proposition 9.4.16 $G\left(\mathbb{K}, \mathbb{K}_{H}\right)=H$ and from Theorem 9.4.12, and what was just shown, $|H|=\left|G\left(\mathbb{K}, \mathbb{K}_{H}\right)\right| \leq\left[\mathbb{K}: \mathbb{K}_{H}\right] \leq|H|$

Note that from the above discussion, when $\mathbb{K}$ is a splitting field of $p(x) \in \mathbb{F}[x]$, this implies that if $\mathbb{L}$ is an intermediate field, then it is also a fixed field of a subgroup of $G(\mathbb{K}, \mathbb{F})$. In fact, from the above, $\mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$. If $H$ is a subgroup, then it is also the Galois group $H=G\left(\mathbb{K}, \mathbb{K}_{H}\right)$. By Proposition 9.4.19, each of these intermediate fields $\mathbb{L}$ is also a normal extension of $\mathbb{F}$. Here is a summary of the principal items obtained up till now.

Summary 9.4.30 When $\mathbb{K}$ is the splitting field of a separable polynomial with coefficients in $\mathbb{F}$, the following are obtained.

1. There is a one to one correspondence between the fixed fields $\mathbb{K}_{H}$ and the subgroups $H$ of $G(\mathbb{K}, \mathbb{F})$. This is given by $\theta(H) \equiv \mathbb{K}_{H} . \quad \theta^{-1}(\mathbb{L})=G(\mathbb{K}, \mathbb{L})$. that is $H=$ $G\left(\mathbb{K}, \mathbb{K}_{H}\right)$ whenever $H$ is a subgroup of $G(\mathbb{K}, \mathbb{F})$.
2. All the intermediate fields are normal field extensions of $\mathbb{F}$ and are fixed fields

$$
\mathbb{L}=\mathbb{K}_{G(\mathbb{K}, \mathbb{L})}
$$

3. For $H$ a subgroup of $G(\mathbb{K}, \mathbb{F}),|H|=\left[\mathbb{K}: \mathbb{K}_{H}\right], H=G\left(\mathbb{K}, \mathbb{K}_{H}\right)$.

Are the Galois groups $G(\mathbb{L}, \mathbb{F})$ for $\mathbb{L}$ an intermediate field between $\mathbb{F}$ and $\mathbb{K}$ for $\mathbb{K}$ the splitting field of a separable polynomial normal subgroups of $G(\mathbb{K}, \mathbb{F})$ ? It might seem like a normal expectation to have. One would hope this is the case.

### 9.4.5 Intermediate Fields and Normal Subgroups

When $\mathbb{K}$ is a splitting field of a separable polynomial having coefficients in $\mathbb{F}$, the intermediate fields are each normal extensions from the above Proposition 9.4.19 which says that splitting fields are normal extensions. If $\mathbb{L}$ is one of these intermediate fields, what about $G(\mathbb{L}, \mathbb{F})$ ? is this a normal subgroup of $G(\mathbb{K}, \mathbb{F})$ ? More generally, consider the following diagram which has now been established in the case that $\mathbb{K}$ is a splitting field of a separable polynomial in $\mathbb{F}[x]$.

$$
\begin{array}{llllll}
\mathbb{F} \equiv \mathbb{L}_{0} & \subseteq \mathbb{L}_{1} & \subseteq \mathbb{L}_{2} & \cdots & \subseteq \mathbb{L}_{k-1} & \subseteq \mathbb{L}_{k} \equiv \mathbb{K}  \tag{9.25}\\
G(\mathbb{F}, \mathbb{F})=\{\imath\} & \subseteq G\left(\mathbb{L}_{1}, \mathbb{F}\right) & \subseteq G\left(\mathbb{L}_{2}, \mathbb{F}\right) & \cdots & \subseteq G\left(\mathbb{L}_{k-1}, \mathbb{F}\right) & \subseteq G(\mathbb{K}, \mathbb{F})
\end{array}
$$

The intermediate fields $\mathbb{L}_{i}$ are each normal extensions of $\mathbb{F}$ each element of $\mathbb{L}_{i}$ being algebraic. As implied in the diagram, there is a one to one correspondence between the intermediate fields and the Galois groups displayed. Is $G\left(\mathbb{L}_{j-1}, \mathbb{F}\right)$ a normal subgroup of $G\left(\mathbb{L}_{j}, \mathbb{F}\right)$ ?

Lemma 9.4.31 $G\left(\mathbb{K}, \mathbb{L}_{j}\right)$ is a normal subgroup of $G(\mathbb{K}, \mathbb{F})$. Here $\mathbb{K}$ is a splitting field for some polynomial having coefficients in $\mathbb{F}$ or more generally a normal extension of $\mathbb{F}$.

Proof: Let $\eta \in G(\mathbb{K}, \mathbb{F})$ and let $\sigma \in G\left(\mathbb{K}, \mathbb{L}_{j}\right)$. Is $\eta^{-1} \sigma \eta \in G\left(\mathbb{K}, \mathbb{L}_{j}\right)$ ? First I need to verify it is a automorphism on $\mathbb{K}$. After this, I need to show that it fixes $\mathbb{L}_{j} . \eta^{-1} \sigma \eta$ is obviously an automorphism on $\mathbb{K}$ because each in the product is. Does $\eta^{-1} \sigma \eta$ fix $\mathbb{L}_{j}$ ? Let $r \in \mathbb{L}_{j}$ with minimum polynomial $f(x)$ having roots $r_{i}$ and coefficients in $\mathbb{F}$. Then $0=$ $\eta f(r)=f(\eta(r))$ and so $\eta(r)$ is one of the roots of $f(x)$. It follows that $\eta(r) \in \mathbb{L}_{j}$ because $\mathbb{K}$ is a normal extension and $\mathbb{L}_{j}$ is an intermediate field so is also a normal extension. See Proposition 9.4.19. Therefore, $\sigma$ fixes $\eta(r)$ and so $\eta^{-1} \sigma \eta(r)=\eta^{-1} \eta(r)=r$.

Because of this lemma, it makes sense to consider the quotient group

$$
G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{j}\right)
$$

This leads to the following fundamental theorem of Galois theory.
Theorem 9.4.32 Let $\mathbb{K}$ be a splitting field of a separable polynomial $p(x)$ having coefficients in a field $\mathbb{F}$. Let $\left\{\mathbb{L}_{i}\right\}_{i=0}^{k}$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$ as shown above in 9.25. Then each of these is a normal extension of $\mathbb{F}$ (Proposition 9.4.19) and the Galois group $G\left(\mathbb{K}, \mathbb{L}_{j}\right)$ is a normal subgroup of $G(\mathbb{K}, \mathbb{F})$ and

$$
G\left(\mathbb{L}_{j}, \mathbb{F}\right) \simeq G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{j}\right)
$$

where the symbol $\simeq$ indicates the two groups are isomorphic.
Proof: All that remains is to check that the above isomorphism is valid. Let

$$
\theta: G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{j}\right) \rightarrow G\left(\mathbb{L}_{j}, \mathbb{F}\right),\left.\theta[\sigma] \equiv \sigma\right|_{\mathbb{L}_{j}}
$$

In other words, this is just the restriction of $\sigma$ to $\mathbb{L}_{j}$. Thus the quotient group is well defined by Proposition 9.4.23. Is $\theta$ well defined? First of all, does it have values in $G\left(\mathbb{L}_{j}, \mathbb{F}\right)$ ? In other words, if $\sigma \in G(\mathbb{K}, \mathbb{F})$, does its restriction to $\mathbb{L}_{j}$ send $\mathbb{L}_{j}$ to $\mathbb{L}_{j}$ ? If $r \in \mathbb{L}_{j}$ it has a minimum polynomial $q(x)$ with coefficients in $\mathbb{F}$. $\sigma(r)$ is one of the other roots of $q(x)$ (Theorem 9.4.12) so, since $\mathbb{K}$ is a normal extension, being a splitting field of a separable polynomial, $\sigma(r) \in \mathbb{K}$. But these subfields are all normal extensions so $\sigma(r) \in \mathbb{L}_{j}$.

Thus $\theta$ has values in $G\left(\mathbb{L}_{j}, \mathbb{F}\right)$. Is $\theta$ well defined? If $\left[\sigma_{1}\right]=\left[\sigma_{2}\right]$, then by definition, $\sigma_{1}^{-1} \sigma_{2} \in G\left(\mathbb{K}, \mathbb{L}_{j}\right)$ so $\sigma_{1}^{-1} \sigma_{2}$ fixes everything in $\mathbb{L}_{1}$. Thus if $r \in \mathbb{L}_{1}, \sigma_{1}^{-1} \sigma_{2} r=r$ and so $\sigma_{2} r=\sigma_{1} r$. It follows that the restrictions of $\sigma_{1}$ and $\sigma_{2}$ to $\mathbb{L}_{j}$ are equal. Therefore, $\theta$ is well defined. It is obvious that $\theta$ is a homomorphism. Why is $\theta$ onto? This follows right away from Theorem 9.4.9. Note that $\mathbb{K}$ is the splitting field of $p(x)$ over $\mathbb{L}_{j}$ since $L_{j} \supseteq \mathbb{F}$. Also if $\sigma \in G\left(\mathbb{L}_{j}, \mathbb{F}\right)$ so it is an automorphism of $\mathbb{L}_{j}$, then, since it fixes $\mathbb{F}, p(x)=\bar{p}(x)$ in that theorem. Thus $\sigma$ extends to $\zeta$, an automorphism of $\mathbb{K}$. Thus $\theta \zeta=\sigma$. Why is $\theta$ one to one? If $\theta[\sigma]=\theta[\alpha]$, this means $\sigma=\alpha$ on $\mathbb{L}_{j}$. Thus $\sigma \alpha^{-1}$ is the identity on $\mathbb{L}_{j}$. Hence $\sigma \alpha^{-1} \in G\left(\mathbb{K}, \mathbb{L}_{j}\right)$ which is what it means for $[\sigma]=[\alpha]$.

The following picture is a summary of what has just been shown.

$$
\begin{array}{ccc}
\mathbb{L}_{k} \equiv \mathbb{K}=\mathbb{K}_{G(\mathbb{K}, \mathbb{F})} & G(\mathbb{K}, \mathbb{F}) & \simeq G(\mathbb{K}, \mathbb{F}) / G(\mathbb{K}, \mathbb{K}) \\
\vdots & \vdots & \vdots \\
\mathbb{L}_{j}=\mathbb{K}_{G\left(\mathbb{L}_{j}, \mathbb{F}\right)} & G\left(\mathbb{L}_{j}, \mathbb{F}\right) & \simeq G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{j}\right) \\
\vdots & \vdots & \vdots \\
\mathbb{L}_{1}=\mathbb{K}_{G\left(\mathbb{L}_{1}, \mathbb{F}\right)} & G\left(\mathbb{L}_{1}, \mathbb{F}\right) & \simeq G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{1}\right) \\
\mathbb{F} \equiv \mathbb{L}_{0} & G\left(\mathbb{L}_{0}, \mathbb{F}\right)=\{\imath\} & \simeq G(\mathbb{K}, \mathbb{F}) / G(\mathbb{K}, \mathbb{F})
\end{array}
$$

### 9.4.6 Permutations

As explained above, the automorphisms of a splitting field $\mathbb{K}$ of $p(x) \in \mathbb{F}[x]$ are determined by the permutations of the roots of $p(x)$. Thus it makes sense to consider permutations.

Let $\left\{a_{1}, \cdots, a_{n}\right\}$ be a set of distinct elements. Then a permutation of these elements is usually thought of as a list in a particular order. Thus there are exactly $n$ ! permutations of a set having $n$ distinct elements. With this definition, here is a simple lemma.

Lemma 9.4.33 Every permutation can be obtained from every other permutation by a finite number of switches.

Proof: This is obvious if $n=1$ or 2 . Suppose then that it is true for sets of $n-1$ elements. Take two permutations of $\left\{a_{1}, \cdots, a_{n}\right\}, P_{1}, P_{2}$. To get from $P_{1}$ to $P_{2}$ using switches, first make a switch to obtain the last element in the list coinciding with the last element of $P_{2}$. By induction, there are switches which will arrange the first $n-1$ to the right order.

It is customary to consider permutations in terms of the set $I_{n} \equiv\{1, \cdots, n\}$ to be more specific. Then one can think of a given permutation as a mapping $\sigma$ from this set $I_{n}$ to itself which is one to one and onto. In fact, $\sigma(i) \equiv j$ where $j$ is in the $i^{t h}$ position. Often people write such a $\sigma$ in the following form

$$
\left(\begin{array}{cccc}
1 & 2 & \cdots & n  \tag{9.26}\\
i_{1} & i_{2} & \cdots & i_{n}
\end{array}\right)
$$

meaning $1 \rightarrow i_{1}, 2 \rightarrow i_{2}, \ldots$ where $\left\{i_{1}, i_{2}, \ldots, i_{n}\right\}=\{1,2, \ldots, n\}$. An easy way to understand the above permutation is through the use of matrix multiplication by permutation matrices. The above vector $\left(i_{1}, \cdots, i_{n}\right)^{T}$ is obtained by

$$
\left(\begin{array}{cccc}
e_{i_{1}} & e_{i_{2}} & \cdots & e_{i_{n}}
\end{array}\right)\left(\begin{array}{c}
1  \tag{9.27}\\
2 \\
\vdots \\
n
\end{array}\right)
$$

This can be seen right away from looking at a simple example or by using the definition of matrix multiplication directly.

Definition 9.4.34 The sign of the permutation 9.26 is defined as the determinant of the above matrix in 9.27.

In other words, the sign of the permutation

$$
\left(\begin{array}{cccc}
1 & 2 & \cdots & n \\
i_{1} & i_{2} & \cdots & i_{n}
\end{array}\right)
$$

equals $\operatorname{sgn}\left(i_{1}, \cdots, i_{n}\right)$ defined earlier in Lemma 8.1.1.
Note that from the fact that the determinant is well defined and its properties, the sign of a permutation is 1 if and only if the permutation is produced by an even number of switches and that the number of switches used to produce a given permutation must be either even or odd. Of course a switch is a permutation itself and this is called a transposition. Note also that all these matrices are orthogonal matrices so to take the inverse, it suffices to take a transpose, the inverse also being a permutation matrix.

The resulting group consisting of the permutations of $I_{n}$ is called $S_{n}$. An important idea is the notion of a cycle. Let $\sigma$ be a permutation, a one to one and onto function defined on $I_{n}$. A cycle is of the form

$$
\left(k, \sigma(k), \sigma^{2}(k), \sigma^{3}(k), \cdots, \sigma^{m-1}(k)\right), \sigma^{m}(k)=k
$$

The last condition must hold for some $m$ because $I_{n}$ is finite. Then a cycle can be considered as a permutation as follows. Let $\left(i_{1}, i_{2}, \cdots, i_{m}\right)$ be a cycle. Then define $\sigma$ by $\sigma\left(i_{1}\right)=$ $i_{2}, \sigma\left(i_{2}\right)=i_{3}, \cdots, \sigma\left(i_{m}\right)=i_{1}$, and if $k \notin\left\{i_{1}, i_{2}, \cdots, i_{m}\right\}$, then $\sigma(k)=k$.

Note that if you have two cycles, $\left(i_{1}, i_{2}, \cdots, i_{m}\right),\left(j_{1}, j_{2}, \cdots, j_{m}\right)$ which are disjoint in the sense that

$$
\left\{i_{1}, i_{2}, \cdots, i_{m}\right\} \cap\left\{j_{1}, j_{2}, \cdots, j_{m}\right\}=\emptyset
$$

then they commute. It is then clear that every permutation can be represented in a unique way by disjoint cycles. Start with 1 and form the cycle determined by 1 . Then start with the smallest $k \in I_{n}$ which was not included and begin a cycle starting with this. Continue this way. Use the convention that $(k)$ is just the identity sending $k$ to $k$ and all other indices to themselves. This representation is unique up to order of the cycles which does not matter because they commute. Note that a transposition can be written as $(a, b), a \rightarrow b$ and $b \rightarrow a$.

A cycle can be written as a product of non disjoint transpositions.

$$
\left(i_{1}, i_{2}, \cdots, i_{m}\right)=\left(i_{m-1}, i_{m}\right) \cdots\left(i_{3}, i_{m}\right)\left(i_{2}, i_{m}\right)\left(i_{1}, i_{m}\right)
$$

Thus if $m$ is odd, the permutation has sign 1 and if $m$ is even, the permutation has sign -1 . Also, it is clear the inverse of the above permutation is $\left(i_{1}, i_{2}, \cdots, i_{m}\right)^{-1}=\left(i_{m}, \cdots, i_{2}, i_{1}\right)$. For example, $(1,2,3)=(2,3)(1,3)$.

Definition 9.4.35 $A_{n}$ is the subgroup of $S_{n}$ such that for $\sigma \in A_{n}, \sigma$ is the product of an even number of transpositions. It is called the alternating group.

Since each transposition switches a pair of columns in the above permutation matrix, the sign of the determinant which is the sign of the permutation is always 1 for permutations in $A_{n}$. This is another way to describe $A_{n}$, those permutations with sign 1 . If $n=1$, there is only one permutation and it is the identity so $A_{1}=$ identity. If $n=2$, you would have two permutations, the identity and the transposition (1,2). Thus $A_{2}=$ identity. It might be useful to think of the identity map as having zero transpositions.

The following important result is useful in describing $A_{n}$.

Proposition 9.4.36 Let $n \geq 3$. Then every permutation in $A_{n}$ is the product of 3 cycles and the identity.

Proof: In case $n=3$, you can list all of the permutations in $A_{n}$.

$$
\left(\begin{array}{lll}
1 & 2 & 3 \\
1 & 2 & 3
\end{array}\right),\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 3 & 1
\end{array}\right),\left(\begin{array}{lll}
1 & 2 & 3 \\
3 & 1 & 2
\end{array}\right)
$$

In terms of cycles, these are

$$
\text { identity, }(1,2,3),(1,3,2)
$$

You can easily check that the the last two are inverses of each other.
Now suppose $n \geq 4$. The permutations in $A_{n}$ are defined as the product of an even number of transpositions. There are two cases. The first case is where you have two transpositions which share a number,

$$
(a, c)(c, b)=(a, c, b)
$$

Thus when they share a number, the product is just a 3 cycle. Next suppose you have the product of two transpositions which are disjoint. This can happen because $n \geq 4$. First note that

$$
(a, b)=(c, b)(b, a, c)=(c, b, a)(c, a)
$$

Therefore,

$$
\begin{aligned}
(a, b)(c, d) & =(c, b, a)(c, a)(a, d)(d, c, a) \\
& =(c, b, a)(c, a, d)(d, c, a)
\end{aligned}
$$

and so every product of disjoint transpositions is the product of 3 cycles.
Lemma 9.4.37 If $n \geq 5$, then if $B$ is a normal subgroup of $A_{n}$, and $B$ is not the identity, then $B$ must contain a 3 cycle.

Proof: Let $\alpha$ be the permutation in $B$ which is "closest" to the identity without being the identity. That is, out of all permutations which are not the identity, this is one which has the most fixed points or equivalently moves the fewest numbers. Then $\alpha$ is the product of disjoint cycles. Suppose that the longest cycle is the first one and it has at least four numbers. Thus

$$
\alpha=\left(i_{1}, i_{2}, i_{3}, i_{4}, \cdots, m\right) \gamma_{1} \cdots \gamma_{p}
$$

Since $B$ is normal,

$$
\alpha_{1} \equiv\left(i_{3}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}, i_{4}, \cdots, m\right)\left(i_{1}, i_{2}, i_{3}\right) \gamma_{1} \cdots \gamma_{p} \in A_{m}
$$

Then since the various cycles are disjoint, $\alpha_{1} \alpha^{-1}=$

$$
\begin{aligned}
& \left(i_{3}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}, i_{4}, \cdots, m\right)\left(i_{1}, i_{2}, i_{3}\right) \gamma_{1} \\
& \cdots \gamma_{p}\left(m, \cdots, i_{4}, i_{3}, i_{2}, i_{1}\right) \gamma_{p}^{-1} \cdots \gamma_{1}^{-1} \\
= & \left(i_{3}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}, i_{4}, \cdots, m\right)\left(i_{1}, i_{2}, i_{3}\right)\left(m, \cdots, i_{4}, i_{3}, i_{2}, i_{1}\right) \gamma_{1} \\
& \cdots \gamma_{p} \gamma_{p}^{-1} \cdots \gamma_{1}^{-1} \\
= & \left(i_{3}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}, i_{4}, \cdots, m\right)\left(i_{1}, i_{2}, i_{3}\right)\left(m, \cdots, i_{4}, i_{3}, i_{2}, i_{1}\right)
\end{aligned}
$$

Then for this permutation, $i_{1} \rightarrow i_{3}, i_{2} \rightarrow i_{2}, i_{3} \rightarrow i_{4}, i_{4} \rightarrow i_{1}$. The other numbers not in $\left\{i_{1}, i_{2}, i_{3}, i_{4}\right\}$ are fixed, and in addition $i_{2}$ is fixed which did not happen with $\alpha$. Therefore, this new permutation moves only 3 numbers. Since it is assumed that $m \geq 4$, this is a contradiction to $\alpha$ fixing the most points. It follows that

$$
\begin{equation*}
\alpha=\left(i_{1}, i_{2}, i_{3}\right) \gamma_{1} \cdots \gamma_{p} \tag{9.28}
\end{equation*}
$$

or else

$$
\begin{equation*}
\alpha=\left(i_{1}, i_{2}\right) \gamma_{1} \cdots \gamma_{p} \tag{9.29}
\end{equation*}
$$

In the first case 9.28 , say $\gamma_{1}=\left(i_{4}, i_{5}, \cdots\right)$. Multiply as follows $\alpha_{1}=$

$$
\left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}\right)\left(i_{4}, i_{5}, \cdots\right) \gamma_{2} \cdots \gamma_{p}\left(i_{1}, i_{2}, i_{4}\right) \in B
$$

Then form $\alpha_{1} \alpha^{-1} \in B$ given by

$$
\begin{gathered}
\left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}\right)\left(i_{4}, i_{5}, \cdots\right) \gamma_{2} \cdots \gamma_{p}\left(i_{1}, i_{2}, i_{4}\right) \gamma_{p}^{-1} \cdots \gamma_{1}^{-1}\left(i_{3}, i_{2}, i_{1}\right) \\
=\left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}, i_{3}\right)\left(i_{4}, i_{5}, \cdots\right)\left(i_{1}, i_{2}, i_{4}\right)\left(\cdots, i_{5}, i_{4}\right)\left(i_{3}, i_{2}, i_{1}\right)
\end{gathered}
$$

Then $i_{1} \rightarrow i_{4}, i_{2} \rightarrow i_{3}, i_{3} \rightarrow i_{5}, i_{4} \rightarrow i_{2}, i_{5} \rightarrow i_{1}$ and other numbers are fixed. Thus $\alpha_{1} \alpha^{-1}$ moves 5 points. However, $\alpha$ moves more than 5 if $\gamma_{i}$ is not the identity for any $i \geq 2$. It follows that

$$
\alpha=\left(i_{1}, i_{2}, i_{3}\right) \gamma_{1}
$$

and $\gamma_{1}$ can only be a transposition. However, this cannot happen because then the above $\alpha$ would not even be in $A_{n}$. Therefore, $\gamma_{1}=\imath$ and so

$$
\alpha=\left(i_{1}, i_{2}, i_{3}\right)
$$

Thus in this case, $B$ contains a 3 cycle.
Now consider case 9.29 . None of the $\gamma_{i}$ can be a cycle of length more than 4 since the above argument would eliminate this possibility. If any has length 3 then the above argument implies that $\alpha$ equals this 3 cycle. It follows that each $\gamma_{i}$ must be a 2 cycle. Say

$$
\alpha=\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right) \gamma_{2} \cdots \gamma_{p}
$$

Thus it moves at least four numbers, greater than four if any of $\gamma_{i}$ for $i \geq 2$ is not the identity. As before, $\alpha_{1} \equiv$

$$
\begin{aligned}
& \left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right) \gamma_{2} \cdots \gamma_{p}\left(i_{1}, i_{2}, i_{4}\right) \\
=\quad & \left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}, i_{4}\right) \gamma_{2} \cdots \gamma_{p} \in B
\end{aligned}
$$

Then $\alpha_{1} \alpha^{-1}=$

$$
\begin{aligned}
& \left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}, i_{4}\right) \gamma_{2} \cdots \gamma_{p} \gamma_{p}^{-1} \cdots \gamma_{2}^{-1} \gamma_{1}^{-1}\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}\right) \\
= & \left(i_{4}, i_{2}, i_{1}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}, i_{4}\right)\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}\right) \in B
\end{aligned}
$$

Then $i_{1} \rightarrow i_{3}, i_{2} \rightarrow i_{4}, i_{3} \rightarrow i_{1}, i_{4} \rightarrow i_{3}$ so this moves exactly four numbers. Therefore, none of the $\gamma_{i}$ is different than the identity for $i \geq 2$. It follows that

$$
\begin{equation*}
\alpha=\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right) \tag{9.30}
\end{equation*}
$$

and $\alpha$ moves exactly four numbers. Then since $B$ is normal, $\alpha_{1} \equiv$

$$
\left(i_{5}, i_{4}, i_{3}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right)\left(i_{3}, i_{4}, i_{5}\right) \in B
$$

Then $\alpha_{1} \alpha^{-1}=$

$$
\left(i_{5}, i_{4}, i_{3}\right)\left(i_{1}, i_{2}\right)\left(i_{3}, i_{4}\right)\left(i_{3}, i_{4}, i_{5}\right)\left(i_{3}, i_{4}\right)\left(i_{1}, i_{2}\right) \in B
$$

Then $i_{1} \rightarrow i_{1}, i_{2} \rightarrow i_{2}, i_{3} \rightarrow i_{4}, i_{4} \rightarrow i_{5}, i_{5} \rightarrow i_{3}$. Thus this permutation moves only three numbers and so $\alpha$ cannot be of the form given in 9.30. It follows that case 9.29 does not occur.

Definition 9.4.38 A group $G$ is said to be simple if its only normal subgroups are itself and the identity.

The following major result is due to Galois [26].
Proposition 9.4.39 Let $n \geq 5$. Then $A_{n}$ is simple.
Proof: From Lemma 9.4.37, if $B$ is a normal subgroup of $A_{n}, B \neq\{l\}$, then it contains a 3 cycle $\alpha=\left(i_{1}, i_{2}, i_{3}\right)$,

$$
\left(\begin{array}{ccc}
i_{1} & i_{2} & i_{3} \\
i_{2} & i_{3} & i_{1}
\end{array}\right)
$$

Now let $\left(j_{1}, j_{2}, j_{3}\right)$ be another 3 cycle.

$$
\left(\begin{array}{lll}
j_{1} & j_{2} & j_{3} \\
j_{2} & j_{3} & j_{1}
\end{array}\right)
$$

Let $\sigma$ be a permutation which satisfies

$$
\sigma\left(i_{k}\right)=j_{k}
$$

Then

$$
\begin{aligned}
& \sigma \alpha \sigma^{-1}\left(j_{1}\right)=\sigma \alpha\left(i_{1}\right)=\sigma\left(i_{2}\right)=j_{2} \\
& \sigma \alpha \sigma^{-1}\left(j_{2}\right)=\sigma \alpha\left(i_{2}\right)=\sigma\left(i_{3}\right)=j_{3} \\
& \sigma \alpha \sigma^{-1}\left(j_{3}\right)=\sigma \alpha\left(i_{3}\right)=\sigma\left(i_{1}\right)=j_{1}
\end{aligned}
$$

while $\sigma \alpha \sigma^{-1}$ leaves all other numbers fixed. Thus $\sigma \alpha \sigma^{-1}$ is the given 3 cycle. It follows that $B$ contains every 3 cycle not just a particular one. By Proposition 9.4.36, this implies $B=A_{n}$. The only problem is that it is not known whether $\sigma$ is in $A_{n}$ a product of an even number of transpositions. This is where $n \geq 5$ is used. If necessary, you can modify $\sigma$ on two numbers not equal to any of the $\left\{i_{1}, i_{2}, i_{3}\right\}$ by multiplying by a transposition so that the possibly modified $\sigma$ is expressed as an even number of transpositions.

### 9.4.7 Solvable Groups

Recall the fundamental theorem of Galois theory which established a correspondence between the normal subgroups of $G(\mathbb{K}, \mathbb{F})$ and normal field extensions whenever $\mathbb{K}$ is the splitting field of a separable polynomial $p(x)$. Also recall that if $H$ is one of these normal subgroups, then there was an isomorphism between $G\left(\mathbb{K}_{H}, \mathbb{F}\right)$ and the quotient group $G(\mathbb{K}, \mathbb{F}) / H$. The general idea of a solvable group is given next.

Definition 9.4.40 A group $G$ is solvable if there exists a decreasing sequence of subgroups $\left\{H_{i}\right\}_{i=0}^{m}$ such that $H_{i}$ is a normal subgroup of $H_{(i-1)}$,

$$
G=H_{0} \supseteq H_{1} \supseteq \cdots \supseteq H_{m}=\{\imath\},
$$

and each quotient group $H_{i-1} / H_{i}$ is Abelian. That is, for $[a],[b] \in H_{i-1} / H_{i}$,

$$
[a b]=[a][b]=[b][a]=[b a]
$$

Note that if $G$ is an Abelian group, then it is automatically solvable. In fact you can just consider $H_{0}=G, H_{1}=\{\imath\}$. In this case $H_{0} / H_{1}$ is just the group $G$ which is Abelian. Also, the definition requires $H_{m-1}$ to be Abelian.

There is another idea which helps in understanding whether a group is solvable. It involves the commutator subgroup. This is a very good idea.

Definition 9.4.41 Let $a, b \in G$ a group. Then the commutator is

$$
a b a^{-1} b^{-1}
$$

The commutator subgroup, denoted by $G^{\prime}$, is the smallest subgroup which contains all the commutators.

The nice thing about the commutator subgroup is that it is a normal subgroup. There are also many other amazing properties.

Theorem 9.4.42 Let $G$ be a group and let $G^{\prime}$ be the commutator subgroup. Then $G^{\prime}$ is a normal subgroup. Also the quotient group $G / G^{\prime}$ is Abelian. If $H$ is any normal subgroup of $G$ such that $G / H$ is Abelian, then $H \supseteq G^{\prime}$. If $G^{\prime}=\{\imath\}$, then $G$ must be Abelian.

Proof: The elements of $G^{\prime}$ are just finite products of things like $a b a^{-1} b^{-1}$. Note that the inverse of something like this is also one of these.

$$
\left(a b a^{-1} b^{-1}\right)^{-1}=b a b^{-1} a^{-1}
$$

Thus the collection of finite products is indeed a subgroup. Now consider $h \in G$. Then

$$
\begin{gathered}
h a b a^{-1} b^{-1} h^{-1}=h a h^{-1} h b h^{-1} h a^{-1} h^{-1} h b^{-1} h^{-1} \\
=h a h^{-1} h b h^{-1}\left(h a h^{-1}\right)^{-1}\left(h b h^{-1}\right)^{-1}
\end{gathered}
$$

which is another one of those commutators. Thus for $c$ a commutator and $h \in G$,

$$
h c h^{-1}=c_{1}
$$

another commutator. If you have a product of commutators $c_{1} c_{2} \cdots c_{m}$, then

$$
h c_{1} c_{2} \cdots c_{m} h^{-1}=\prod_{i=1}^{m} h c_{i} h^{-1}=\prod_{i=1}^{m} d_{i} \in G^{\prime}
$$

where the $d_{i}$ are each commutators. Hence $G^{\prime}$ is a normal subgroup.
Consider now the quotient group. Is $[g][h]=[h][g]$ ? In other words, is $[g h]=[h g]$ ? In other words, is $g h(h g)^{-1}=g h g^{-1} h^{-1} \in G^{\prime}$ ? Of course. This is a commutator and $\dot{G}^{\prime}$ consists of products of these things. Thus the quotient group is Abelian.

Now let $H$ be a normal subgroup of $G$ such that $G / H$ is Abelian. Then if $g, h \in G$,

$$
[g h]=[h g], g h(h g)^{-1}=g h g^{-1} h^{-1} \in H
$$

Thus every commutator is in $H$ and so $H \supseteq G$.
The last assertion is obvious because $G /\{\imath\}$ is isomorphic to $G$. Also, to say that $G^{\prime}=\{\imath\}$ is to say that

$$
a b a^{-1} b^{-1}=\imath
$$

which implies that $a b=b a$.
Let $G$ be a group and let $G^{\prime}$ be its commutator subgroup. Then the commutator subgroup of $G^{\prime}$ is $G^{\prime \prime}$ and so forth. To save on notation, denote by $G^{(k)}$ the $k^{\text {th }}$ commutator subgroup. Thus you have the sequence

$$
G \equiv G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq G^{(3)} \ldots
$$

each $G^{(i)}$ being a normal subgroup of $G^{(i-1)}$ although this does not say that $G^{(i)}$ is a normal subgroup of $G$. Then there is a useful criterion for a group to be solvable.

Theorem 9.4.43 If $G$ is a solvable group and $\hat{G}$ is a subgroup of $G$ then $\hat{G}$ is also solvable.
Proof: Suppose $G=H_{0} \supseteq H_{1} \supseteq \cdots \supseteq H_{m}=\{\imath\}$ where the quotient groups are Abelian and the $H_{i}$ are normal. Consider $H_{k} \cap \hat{G}$. Would this be a normal subgroup of $\hat{G}$ ? Let $a \in \hat{G}$ and $x \in H_{k} \cap \hat{G}$. Is $a x a^{-1} \in H_{k} \cap \hat{G}$ ? We know this product is in $\hat{G}$ because $\hat{G}$ is a subgroup. We know it is in $H_{k}$ because $H_{k}$ is normal. Thus the $H_{k} \cap \hat{G}$ are normal. What of the quotient groups $\left(H_{k} \cap \hat{G}\right) /\left(H_{k+1} \cap \hat{G}\right)$ ? Are these Abelian? If $[x],[y]$ are in $\left(H_{k} \cap \hat{G}\right) /\left(H_{k+1} \cap \hat{G}\right)$ is $[x y]=[y x]$ ? Is $(x y)^{-1} y x \in H_{k+1} \cap \hat{G}$ ? This equals $y^{-1} x^{-1} y x$. However, $x y, y x$ are both in $H_{k}$ and the quotient groups $H_{k} / H_{k+1}$ are Abelian, so $(x y)^{-1} y x \in H_{k+1}$. But also $(x y)^{-1} y x \in \hat{G}$ because $\hat{G}$ is a subgroup. Hence $[x][y]=[x y]=[y x]=[y][x]$ and so the quotient groups are Abelian. Hence $\hat{G}=H_{0} \cap \hat{G} \supseteq \hat{G} \cap H_{1} \supseteq \cdots \supseteq \hat{G} \cap H_{m}=\{\imath\}$ and so $\hat{G}$ is solvable.
Theorem 9.4.44 Let $G$ be a group. It is solvable if and only if $G^{\prime}$ is solvable so $G^{(k)}=\{1\}$ for some $k$.

Proof: If $G^{(k)}=\{\imath\}$ then $G$ is clearly solvable because $G^{(k-1)} / G^{(k)}$ is Abelian by Theorem 9.4.42. The sequence of commutator subgroups provides the necessary sequence of subgroups.

Next suppose that you have

$$
G=H_{0} \supseteq H_{1} \supseteq \cdots \supseteq H_{m}=\{\imath\}
$$

where each is normal in the preceding and the quotient groups are Abelian. Then from Theorem 9.4.42, $G^{(1)} \subseteq H_{1}$. Thus $H_{1}^{\prime} \supseteq G^{(2)}$. But also, from Theorem 9.4.42, since $H_{1} / H_{2}$ is Abelian,

$$
H_{2} \supseteq H_{1}^{\prime} \supseteq G^{(2)}
$$

Continuing this way $G^{(k)}=\{\imath\}$ for some $k \leq m$.
Alternatively, you could let $\hat{G}=G^{\prime}$. This is solvable by Theorem 9.4.43 since it is a subgroup of $G$.

Theorem 9.4.45 If $G$ is a solvable group and if $H$ is a homomorphic image of $G$, then $H$ is also solvable.

Proof: By the above theorem, it suffices to show that $H^{(k)}=\{\imath\}$ for some $k$. Let $f$ be the homomorphism. Then $H^{\prime}=f\left(G^{\prime}\right)$. To see this, consider a commutator of $H, f(a) f(b) f(a)^{-1} f(b)^{-1}=f\left(a b a^{-1} b^{-1}\right)$. It follows that $H^{(1)}=f\left(G^{(1)}\right)$. Now continue this way, letting $G^{(1)}$ play the role of $G$ and $H^{(1)}$ the role of $H$. Thus, since $G$ is solvable, some $G^{(k)}=\{\imath\}$ and so $H^{(k)}=\{\imath\}$ also.

Now as an important example, of a group which is not solvable, here is a theorem.
Theorem 9.4.46 For $n \geq 5, S_{n}$ is not solvable.
Proof: It is clear that $A_{n}$ is a normal subgroup of $S_{n}$ because if $\sigma$ is a permutation, then it has the same sign as $\sigma^{-1}$. Thus $\sigma \alpha \sigma^{-1} \in A_{n}$ if $\alpha \in A_{n}$ because both $\alpha$ and $\sigma \alpha \sigma^{-1}$ are a product of an even number of transpositions. If $H$ is a normal subgroup of $S_{n}$, for which $S_{n} / H$ is Abelian, then $H$ contains the commutator $S_{n}^{\prime}$. However, $\alpha \sigma \alpha^{-1} \sigma^{-1} \in A_{n}$ obviously so $A_{n} \supseteq S_{n}^{\prime}$. By Proposition 9.4.39 ( $A_{n}$ is simple), this forces $S_{n}^{\prime}=A_{n}$. So what is $S_{n}^{\prime \prime}$ ? If it is $S_{n}$, then $S_{n}^{(k)} \neq\{\imath\}$ for any $k$ and it follows that $S_{n}$ is not solvable. If $S_{n}^{\prime \prime}=\{\imath\}$, the only other possibility, then $A_{n} /\{\imath\}$ is Abelian and so $A_{n}$ is Abelian, but this is obviously false because the cycles $(1,2,3),(2,1,4)$ are both in $A_{n}$. However, $(1,2,3)(2,1,4)$ is

$$
\left(\begin{array}{llll}
1 & 2 & 3 & 4 \\
4 & 2 & 1 & 3
\end{array}\right)
$$

while $(2,1,4)(1,2,3)$ is

$$
\left(\begin{array}{llll}
1 & 2 & 3 & 4 \\
1 & 3 & 4 & 2
\end{array}\right)
$$

Alternatively, by Theorem 9.4.43, if $S_{n}$ is solvable, then so is $A_{n}$. However, $A_{n}$ is simple so there is no normal subgroup other than $A_{n}$ and $\imath$. Now $A_{n} /\{\imath\}=A_{n}$ is not commutative for $n \geq 4$.

Note that the above shows that $A_{n}$ is not Abelian for $n=4$ also.

### 9.4.8 Solvability by Radicals

The idea here is to begin with a big field $\mathbb{F}$ and show that there is no way to solve the polynomial in terms of radicals of things in the big field. It will then follow that there is no way to get a solution in terms of radicals of things in a smaller field like the rational numbers. The most interesting conclusion is what will be presented here, that you can't do it. This amazing conclusion is due to Abel and Galois and dates from the 1820's. It put a stop to the search for formulas which would solve polynomial equations.

First of all, in the case where all fields are contained in $\mathbb{C}$, there exists a field which has all the $n^{\text {th }}$ roots of 1 . You could simply define it to be the smallest sub field of $\mathbb{C}$ such that it contains these roots. You could also enlarge it by including some other numbers. For example, you could include $\mathbb{Q}$. Observe that if $\xi \equiv e^{i 2 \pi / n}$, then $\xi^{n}=1$ but $\xi^{k} \neq 1$ if $k<n$ and that if $k<l<n, \xi^{k} \neq \xi^{l}$. The following is from Herstein [20]. This is the kind of field considered here.

Lemma 9.4.47 Suppose a field $\mathbb{F}$ has all the $n^{\text {th }}$ roots of 1 for a particular $n$ and suppose there exists $\xi$ such that the $n^{\text {th }}$ roots of 1 are of the form $\xi^{k}$ for $k=1, \cdots, n$, the $\xi^{k}$ being distinct, as is the case when all fields are in $\mathbb{C}$. Let $a \in \mathbb{F}$ be nonzero. Let $\mathbb{K}$ denote the
splitting field of $x^{n}-$ a over $\mathbb{F}$, thus $\mathbb{K}$ is a normal extension of $\mathbb{F}$. Then $\mathbb{K}=\mathbb{F}(u)$ where $u$ is any root of $x^{n}-a$. The Galois group $G(\mathbb{K}, \mathbb{F})$ is Abelian.

Proof: Let $u$ be a root of $x^{n}-a$ and let $\mathbb{K}$ equal $\mathbb{F}(u)$. Then let $\xi$ be the $n^{\text {th }}$ root of unity mentioned. Then

$$
\left(\xi^{k} u\right)^{n}=\left(\xi^{n}\right)^{k} u^{n}=a
$$

and so each $\xi^{k} u$ is a root of $x^{n}-a$ and these are distinct. It follows that

$$
\left\{u, \xi u, \cdots, \xi^{n-1} u\right\}
$$

are the roots of $x^{n}-a$ and all are in $\mathbb{F}(u)$. Thus $\mathbb{F}(u)=\mathbb{K}$. Let $\sigma \in G(\mathbb{K}, \mathbb{F})$ and observe that since $\sigma$ fixes $\mathbb{F}$,

$$
0=\sigma\left(\left(\xi^{k} u\right)^{n}-a\right)=\left(\sigma\left(\xi^{k} u\right)\right)^{n}-a
$$

It follows that $\sigma$ maps roots of $x^{n}-a$ to roots of $x^{n}-a$. Therefore, if $\sigma, \alpha$ are two elements of $G(\mathbb{K}, \mathbb{F})$, there exist $i, j$ each no larger than $n-1$ such that

$$
\sigma(u)=\xi^{i} u, \alpha(u)=\xi^{j} u
$$

A typical thing in $\mathbb{F}(u)$ is $p(u)$ where $p(x) \in \mathbb{F}[x]$. Then

$$
\begin{aligned}
& \sigma \alpha(p(u))=p\left(\xi^{j} \xi^{i} u\right)=p\left(\xi^{i+j} u\right) \\
& \alpha \sigma(p(u))=p\left(\xi^{i} \xi^{j} u\right)=p\left(\xi^{i+j} u\right)
\end{aligned}
$$

Therefore, $G(\mathbb{K}, \mathbb{F})$ is Abelian.
Thus this one is clearly solvable as noted above. To say a polynomial is solvable by radicals is expressed precisely in the following definition.

Definition 9.4.48 For $\mathbb{F}$ a field, a polynomial $p(x) \in \mathbb{F}[x]$ is solvable by radicals over $\mathbb{F} \equiv \mathbb{F}_{0}$ if there are algebraic numbers $a_{i}, i=1,2, \ldots, k$, and a sequence of fields $\mathbb{F}_{1}=$ $\mathbb{F}\left(a_{1}\right), \mathbb{F}_{2}=\mathbb{F}_{1}\left(a_{2}\right), \cdots, \mathbb{F}_{k}=\mathbb{F}_{k-1}\left(a_{k}\right)$ such that for each $i \geq 1, a_{i}^{k_{i}} \in \mathbb{F}_{i-1}$ and $\mathbb{F}_{k}$ contains a splitting field $\mathbb{K}$ for $p(x)$ over $\mathbb{F}$.

Actually, the only case of interest here is included in the following lemma.
Lemma 9.4.49 In Definition 9.4.48 when the roots of unity are of the form $\xi^{k}$ as described in Lemma 9.4.47, $\mathbb{F}_{k}$ is a splitting field provided you assume $\mathbb{F}$ contains all the $n^{\text {th }}$ roots of 1 for all $n \leq \max \left\{k_{i}\right\}_{i=1}^{k}$.

Proof: by Lemma 9.4.47,

$$
\mathbb{F}_{k}=\mathbb{F}\left(a_{1}, a_{2}, \cdots, a_{k}\right)=\mathbb{F}\left(\left\{a_{1}^{j}\right\}_{j=1}^{k_{1}-1}, \ldots,\left\{a_{1}^{j}\right\}_{j=1}^{k_{k}-1}\right)
$$

and so $\mathbb{F}_{k}$ is the splitting field of $\prod_{i=1}^{k}\left(x^{k_{i}}-a_{i}^{k_{i}}\right)$. Each $a_{i}$ is a single root of $x^{k_{i}}-a_{i}^{k_{i}}$ where $a_{i}^{k_{i}} \in \mathbb{F}$.

At this point, it is a good idea to recall the big fundamental theorem mentioned above which gives the correspondence between normal subgroups and normal field extensions since it is about to be used again.

$$
\begin{array}{llllll}
\mathbb{F} \equiv \mathbb{F}_{0} & \subseteq \mathbb{F}_{1} & \subseteq \mathbb{F}_{2} & \cdots & \subseteq \mathbb{F}_{k-1} & \subseteq \mathbb{F}_{k} \equiv \mathbb{K} \\
G(\mathbb{F}, \mathbb{F})=\{\imath\} & \subseteq G\left(\mathbb{F}_{1}, \mathbb{F}\right) & \subseteq G\left(\mathbb{F}_{2}, \mathbb{F}\right) & \cdots & \subseteq G\left(\mathbb{F}_{k-1}, \mathbb{F}\right) & \subseteq G\left(\mathbb{F}_{k}, \mathbb{F}\right)
\end{array}
$$

Theorem 9.4.50 Let $\mathbb{K}$ be a splitting field for a separable polynomial $p(x) \in \mathbb{F}[x]$. Let $\left\{\mathbb{F}_{i}\right\}_{i=0}^{k}$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$. Then each of these is a normal extension of $\mathbb{F}$ and the Galois group $G\left(\mathbb{F}_{j-1}, \mathbb{F}\right)$ is a normal subgroup of $G\left(\mathbb{F}_{j}, \mathbb{F}\right)$. In addition to this,

$$
G\left(\mathbb{F}_{j}, \mathbb{F}\right) \simeq G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{F}_{j}\right)
$$

where the symbol $\simeq$ indicates the two spaces are isomorphic.
Theorem 9.4.51 Let $f(x)$ be a separable polynomial in $\mathbb{F}[x]$ where $\mathbb{F}$ contains all $n^{\text {th }}$ roots of unity for each $n \in \mathbb{N}$ or for all $n \leq m$ and the roots of unity are of the form $\xi^{k}$ as described in Lemma 9.4.47. Let $\mathbb{K}$ be a splitting field of $f(x)$. If $f(x)$ is solvable by radicals over $\mathbb{F}$, or solvable by radicals over $\mathbb{F}$ with the $k_{i} \leq m$ in Definition 9.4.48, then the Galois group $G(\mathbb{K}, \mathbb{F})$ is a solvable group.

Proof: Using the definition given above for $f(x)$ to be solvable by radicals, there is a sequence of fields

$$
\mathbb{F}_{0}=\mathbb{F} \subseteq \mathbb{F}_{1} \subseteq \cdots \subseteq \mathbb{F}_{k}, \mathbb{K} \subseteq \mathbb{F}_{k}
$$

where $\mathbb{F}_{i}=\mathbb{F}_{i-1}\left(a_{i}\right), a_{i}^{k_{i}} \in \mathbb{F}_{i-1}$, and each field extension is a normal extension of the preceding one. By Lemma 9.4.49, $\mathbb{F}_{k}$ is the splitting field of a polynomial having coefficients in $\mathbb{F}_{j-1}$. This follows from the Lemma 9.4.49 above. Then it follows from Theorem 9.4.50, letting $\mathbb{F}_{j-1}$ play the role of $\mathbb{F}$, that

$$
G\left(\mathbb{F}_{j}, \mathbb{F}_{j-1}\right) \simeq G\left(\mathbb{F}_{k}, \mathbb{F}_{j-1}\right) / G\left(\mathbb{F}_{k}, \mathbb{F}_{j}\right)
$$

By Lemma 9.4.47, the Galois group $G\left(\mathbb{F}_{j}, \mathbb{F}_{j-1}\right)$ is Abelian and so this and the above isomorphism requires that $G\left(\mathbb{F}_{k}, \mathbb{F}\right)$ is a solvable group since the quotient groups are Abelian.

By Theorem 9.4.43, it follows that, since $G(\mathbb{K}, \mathbb{F})$ is a subgroup of $G\left(\mathbb{F}_{k}, \mathbb{F}\right)$, it must also be solvable.

Now consider the equation

$$
p(x)=x^{n}-a_{1} x^{n-1}+a_{2} x^{n-2}+\cdots \pm a_{n}, p(x) \in \mathbb{F}[x], n \geq 5
$$

and suppose that $p(x)$ has distinct roots, none of them in $\mathbb{F}$. Let $\mathbb{K}$ be a splitting field for $p(x)$ over $\mathbb{F}$ so that $p(x)=\prod_{k=1}^{n}\left(x-r_{i}\right)$. Then it follows that $a_{i}=s_{i}\left(r_{1}, \cdots, r_{n}\right)$ where the $s_{i}$ are the elementary symmetric functions defined in Definition 9.1.3. For $\sigma \in G(\mathbb{K}, \mathbb{F})$ you can define $\bar{\sigma} \in S_{n}$ by the rule $\bar{\sigma}(k) \equiv j$ where $\sigma\left(r_{k}\right)=r_{j}$. Recall that the automorphisms of $G(\mathbb{K}, \mathbb{F})$ take roots of $p(x)$ to roots of $p(x)$. This mapping $\sigma \rightarrow \bar{\sigma}$ is onto, a homomorphism, and one to one and onto because the symmetric functions $s_{i}$ are unchanged when the roots are permuted. Thus a rational function in $s_{1}, s_{2}, \cdots, s_{n}$ is unaffected when the roots $r_{k}$ are permuted. It follows that $G(\mathbb{K}, \mathbb{F})$ cannot be solvable if $n \geq 5$ because $S_{n}$ is not solvable.

For example, consider $3 x^{5}-25 x^{3}+45 x+1$ or equivalently $x^{5}-\frac{25}{3} x^{3}+15 x+\frac{1}{3} \in \mathbb{Q}(x)$. It clearly has no rational roots and a graph will show it has 5 real roots. Let $\mathbb{F}=\mathbb{Q}(\boldsymbol{\omega})$ where $\boldsymbol{\omega}$ denotes all $k^{\text {th }}$ roots of unity for $k \leq 5$. Then some computations show that none of these roots of the polynomial are in $\mathbb{F}$ and they are all distinct. Thus the polynomial cannot be solved by radicals involving $k^{t h}$ roots for $k \leq 5$ of numbers in $\mathbb{Q}$. In fact, it can't be solved by radicals involving $k^{t h}$ roots for $k \leq 5$ of numbers in $\mathbb{Q}(\boldsymbol{\omega})$.

Recall that $\mathbb{Q}(\sqrt{2})$ can be written as $a+b \sqrt{2}$ where $a, b$ are rational. However, algebraic numbers are roots of polynomials having rational coefficients. Can each of these be written in this way in terms of radicals. It was just shown that, surprisingly, this is not the case. It is a little like the fact from real analysis that it is extremely difficult to give an explicit description of a generic Borel set, except that the present situation seems even worse because in the case of Borel sets, you can sort of do it provided you use enough hard set theory. Thus you must use the definition of algebraic numbers described above. It is also pointless to search for the equivalent of the quadratic formula for polynomials of degree 5 or more.

### 9.5 A Few Generalizations

Sometimes people consider things which are more general. Also, it is worthwhile identifying situations when all polynomials are separable to generalize Theorem 9.4.51.

### 9.5.1 The Normal Closure of a Field Extension

An algebraic extension $\mathbb{F}\left(a_{1}, a_{2}, \cdots, a_{m}\right)$ is contained in a field which is a normal extension of $\mathbb{F}$. To begin with, recall the following definition.

Definition 9.5.1 When you have $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ with each $a_{i}$ algebraic so $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is a field, you could consider $f(x) \equiv \prod_{i=1}^{m} f_{i}(x)$. where $f_{i}(x)$ is the minimum polynomial of $a_{i}$. Then if $\mathbb{K}$ is a splitting field for $f(x)$, this $\mathbb{K}$ is called the normal closure. It is at least as large as $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ and it has the advantage of being a normal extension.

Let $G(\mathbb{K}, \mathbb{F})=\left\{\eta_{1}, \eta_{2}, \cdots, \eta_{q}\right\}$. The conjugate fields are defined as the fields

$$
\eta_{j}\left(\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)\right)
$$

Thus each of these fields is isomorphic to any other and they are all contained in $\mathbb{K}$. Let $\mathbb{K}^{\prime}$ denote the smallest field contained in $\mathbb{K}$ which contains all of these conjugate fields. Note that if $k \in \mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ so that $\eta_{i}(k)$ is in one of these conjugate fields, then $\eta_{j} \eta_{i}(k)$ is also in a conjugate field because $\eta_{j} \eta_{i}$ is one of the automorphisms of $G(\mathbb{K}, \mathbb{F})$. Let

$$
S=\left\{k \in \mathbb{K}^{\prime}: \eta_{j}(k) \in \mathbb{K}^{\prime} \text { each } j\right\}
$$

Then from what was just shown, each conjugate field is in $S$. Suppose $k \in S$. What about $k^{-1}$ ?

$$
\eta_{j}(k) \eta_{j}\left(k^{-1}\right)=\eta_{j}\left(k k^{-1}\right)=\eta_{j}(1)=1
$$

and so $\left(\eta_{j}(k)\right)^{-1}=\eta_{j}\left(k^{-1}\right)$. Now $\left(\eta_{j}(k)\right)^{-1} \in \mathbb{K}^{\prime}$ because $\mathbb{K}^{\prime}$ is a field. Therefore, $\eta_{j}\left(k^{-1}\right) \in \mathbb{K}^{\prime}$. Thus $S$ is closed with respect to taking inverses. It is also closed with respect to products. Thus it is clear that $S$ is a field which contains each conjugate field. However,
$\mathbb{K}^{\prime}$ was defined as the smallest field which contains the conjugate fields. Therefore, $S=\mathbb{K}^{\prime}$ and so this shows that each $\eta_{j}$ maps $\mathbb{K}^{\prime}$ to itself while fixing $\mathbb{F}$. Thus $G(\mathbb{K}, \mathbb{F}) \subseteq G\left(\mathbb{K}^{\prime}, \mathbb{F}\right)$ because each of the $\eta_{i}$ is in $G\left(\mathbb{K}^{\prime}, \mathbb{F}\right)$. This is what was just shown. However, since $\mathbb{K}^{\prime} \subseteq \mathbb{K}$, it follows that also $G\left(\mathbb{K}^{\prime}, \mathbb{F}\right) \subseteq G(\mathbb{K}, \mathbb{F})$. Therefore, $G\left(\mathbb{K}^{\prime}, \mathbb{F}\right)=G(\mathbb{K}, \mathbb{F})$ and by the one to one correspondence between the intermediate fields and the Galois groups, it follows that $\mathbb{K}^{\prime}=\mathbb{K}$. If $\mathbb{K}^{\prime}$ is a proper subset of $\mathbb{K}$ then you would need to have $G\left(\mathbb{K}^{\prime}, \mathbb{F}\right)$ a proper subgroup of $G(\mathbb{K}, \mathbb{F})$ but these are equal. This proves the following lemma.

Lemma 9.5.2 Let $\mathbb{K}$ denote the normal extension of $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ with each $a_{i}$ algebraic so that $\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)$ is a field. Thus $\mathbb{K}$ is the splitting field of the product of the minimum polynomials of the $a_{i}$. Then $\mathbb{K}$ is also the smallest field containing the conjugate fields $\eta_{j}\left(\mathbb{F}\left(a_{1}, \cdots, a_{m}\right)\right)$ for $\left\{\eta_{1}, \eta_{2}, \cdots, \eta_{q}\right\}=G(\mathbb{K}, \mathbb{F})$.

Lemma 9.5.3 In Definition 9.4.48, you can assume that $\mathbb{F}_{k}$ is a normal extension of $\mathbb{F}$.
Proof: First note that $\mathbb{F}_{k}=\mathbb{F}\left[a_{1}, a_{2}, \cdots, a_{k}\right]$. Let $\mathbb{G}$ be the normal extension of $\mathbb{F}_{k}$. By Lemma 9.5.2, $\mathbb{G}$ is the smallest field which contains the conjugate fields

$$
\begin{gathered}
\eta_{j}\left(\mathbb{F}\left(a_{1}, a_{2}, \cdots, a_{k}\right)\right)=\mathbb{F}\left(\eta_{j} a_{1}, \eta_{j} a_{2}, \cdots, \eta_{j} a_{k}\right) \\
\text { for }\left\{\eta_{1}, \eta_{2}, \cdots, \eta_{m}\right\}=G\left(\mathbb{F}_{k}, \mathbb{F}\right) . \text { Also, }\left(\eta_{j} a_{i}\right)^{k_{i}}=\eta_{j}\left(a_{i}^{k_{i}}\right) \in \eta_{j} \mathbb{F}_{i-1}, \eta_{j} \mathbb{F}=\mathbb{F} \text {. Then } \\
\mathbb{G}=\mathbb{F}\left(\eta_{1}\left(a_{1}\right), \eta_{1}\left(a_{2}\right), \cdots, \eta_{1}\left(a_{k}\right), \eta_{2}\left(a_{1}\right), \eta_{2}\left(a_{2}\right), \cdots, \eta_{2}\left(a_{k}\right) \cdots\right)
\end{gathered}
$$

and this is a splitting field so is a normal extension. Thus $\mathbb{G}$ could be the new $\mathbb{F}_{k}$ with respect to a longer sequence of $a_{i}$ but would now be a splitting field.

### 9.5.2 Conditions for Separability

So when is it that a polynomial having coefficients in a field $\mathbb{F}$ is separable? It turns out that this is always the case for fields which are enough like the rational numbers. It involves considering the derivative of a polynomial. In doing this, there will be no analysis used, just the rule for differentiation which we all learned in calculus. Thus the derivative is defined as follows.

$$
\begin{aligned}
& \left(a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}\right)^{\prime} \\
\equiv & n a_{n} x^{n-1}+a_{n-1}(n-1) x^{n-2}+\cdots+a_{1}
\end{aligned}
$$

This kind of formal manipulation is what most students do anyway, never thinking about where it comes from. Here $n a_{n}$ means to add $a_{n}$ to itself $n$ times. With this definition, it is clear that the usual rules such as the product rule hold. This discussion follows [26].

Definition 9.5.4 A field has characteristic 0 if $n a \neq 0$ for all $n \in \mathbb{N}$ and $a \neq 0$. Otherwise a field $\mathbb{F}$ has characteristic $p$ if $p \cdot 1=0$ for $p \cdot 1$ defined as 1 added to itself $p$ times and $p$ is the smallest positive integer for which this takes place.

Note that with this definition, some of the terms of the derivative of a polynomial could vanish in the case that the field has characteristic $p$. I will go ahead and write them anyway. For example, if the field has characteristic $p$, then $\left(x^{p}-a\right)^{\prime}=0$. because formally it equals $p \cdot 1 x^{p-1}=0 x^{p-1}$, the 1 being the 1 in the field.

Note that the field $\mathbb{Z}_{p}$ does not have characteristic 0 because $p \cdot 1=0$. Thus not all fields have characteristic 0 .

How can you tell if a polynomial has no repeated roots? This is the content of the next theorem.

Theorem 9.5.5 Let $p(x)$ be a monic polynomial having coefficients in a field $\mathbb{F}$, and let $\mathbb{K}$ be a field in which $p(x)$ factors

$$
p(x)=\prod_{i=1}^{n}\left(x-r_{i}\right), \quad r_{i} \in \mathbb{K} .
$$

Then the $r_{i}$ are distinct if and only if $p(x)$ and $p^{\prime}(x)$ are relatively prime over $\mathbb{F}$.
Proof: Suppose first that $p^{\prime}(x)$ and $p(x)$ are relatively prime over $\mathbb{F}$. Since they are not both zero, there exists polynomials $a(x), b(x)$ having coefficients in $\mathbb{F}$ such that

$$
a(x) p(x)+b(x) p^{\prime}(x)=1
$$

Now suppose $p(x)$ has a repeated root $r$. Then in $\mathbb{K}[x], p(x)=(x-r)^{2} g(x)$ and so $p^{\prime}(x)=$ $2(x-r) g(x)+(x-r)^{2} g^{\prime}(x)$. Then in $\mathbb{K}[x]$,

$$
a(x)(x-r)^{2} g(x)+b(x)\left(2(x-r) g(x)+(x-r)^{2} g^{\prime}(x)\right)=1
$$

Then letting $x=r$, it follows that $0=1$. Hence $p(x)$ has no repeated roots.
Next suppose there are no repeated roots of $p(x)$. Then $p^{\prime}(x)=\sum_{i=1}^{n} \prod_{j \neq i}\left(x-r_{j}\right)$. $p^{\prime}(x)$ cannot be zero in this case because $p^{\prime}\left(r_{n}\right)=\prod_{j=1}^{n-1}\left(r_{n}-r_{j}\right) \neq 0$ because it is the product of nonzero elements of $\mathbb{K}$. Similarly no term in the sum for $p^{\prime}(x)$ can equal zero because $\prod_{j \neq i}\left(r_{i}-r_{j}\right) \neq 0$. Then if $q(x)$ is a monic polynomial of degree larger than 1 which divides $p(x)$, then the roots of $q(x)$ in $\mathbb{K}$ are a subset of $\left\{r_{1}, \cdots, r_{n}\right\}$. Without loss of generality, suppose these roots of $q(x)$ are $\left\{r_{1}, \cdots, r_{k}\right\}, k \leq n-1$, since $q(x)$ divides $p^{\prime}(x)$ which has degree at most $n-1$. Then $q(x)=\prod_{i=1}^{k}\left(x-r_{i}\right)$ but this fails to divide $p^{\prime}(x)$ as polynomials in $\mathbb{K}[x]$ and so $q(x)$ fails to divide $p^{\prime}(x)$ as polynomials in $\mathbb{F}[x]$ either. Therefore, $q(x)=1$ and so the two are relatively prime.

The following lemma says that the usual calculus result holds in case you are looking at polynomials with coefficients in a field of characteristic 0 .

Lemma 9.5.6 Suppose that $\mathbb{F}$ has characteristic 0 . Then if $f^{\prime}(x)=0$, it follows that $f(x)$ is a constant.

Proof: Suppose

$$
f(x)=a_{n} x^{n}+a_{n-1} x^{n-1}+\cdots+a_{1} x+a_{0}
$$

Then

$$
0 x^{n}+0 x^{n-1}+\cdots+0 x+0=n a_{n} x^{n-1}+a_{n-1}(n-1) x^{n-2}+\cdots+a_{1}
$$

Therefore, each coefficient on the right is 0 . Since the field has characteristic 0 it follows that each $a_{k}=0$ for $k \geq 1$. Thus $f(x)=a_{0} \in \mathbb{F}$.

If $\mathbb{F}$ has characteristic $p$ as in $\mathbb{Z}_{p}$ for $p$ prime, this is not true. Indeed, $x^{p}-1$ is not constant but has derivative equal to 0 .

Now here is a major result which applies to fields of characteristic 0 .

Theorem 9.5.7 If $\mathbb{F}$ is a field of characteristic 0, then every polynomial $p(x)$, having coefficients in $\mathbb{F}$ is separable.

Proof: It is required to show that the irreducible factors of $p(x)$ have distinct roots in $\mathbb{K}$ a splitting field for $p(x)$. So let $q(x)$ be an irreducible, non constant, monic polynomial. Thus $q^{\prime}(x) \neq 0$ because the field has characteristic 0 . If $l(x)$ is a monic polynomial of positive degree which divides both $q(x)$ and $q^{\prime}(x)$, then since $q(x)$ is irreducible, it must be the case that $l(x)=q(x)$ or $l(x)=1$. If $l(x)=q(x)$, then this forces $q(x)$ to divide $q^{\prime}(x)$, a nonzero polynomial having smaller degree than $q(x)$. This is impossible. Hence $l(x)=1$ and so $q^{\prime}(x)$ and $q(x)$ are relatively prime which implies that $q(x)$ has distinct roots.

It follows that the above theory all holds for any field of characteristic 0. For example, if the field is $\mathbb{Q}$ then everything holds.

Proposition 9.5.8 If a field $\mathbb{F}$ has characteristic $p$, then $p$ is a prime.
Proof: First note that if $n \cdot 1=0$, if and only if for all $a \neq 0, n \cdot a=0$ also. This just follows from the distributive law and the definition of what is meant by $n \cdot 1$, meaning that you add 1 to itself $n$ times. Suppose then that there are positive integers, each larger than $1 n, m$ such that $n m \cdot 1=0$. Then grouping the terms in the sum associated with $n m \cdot 1$, it follows that $n(m \cdot 1)=0$. If the characteristic of the field is $n m$, this is a contradiction because then $m \cdot 1 \neq 0$ but $n$ times it is, implying that $n<n m$ but $n \cdot a=0$ for a nonzero $a$. Hence $n \cdot 1=0$ showing that $m n$ is not the characteristic of the field after all.

Definition 9.5.9 A field $\mathbb{F}$ is called perfect if every polynomial $p(x)$ having coefficients in $\mathbb{F}$ is separable.

The above shows that fields of characteristic 0 are perfect. The above theory about Galois groups and fixed fields all works for perfect fields. What about fields of characteristic $p$ where $p$ is a prime? The following interesting lemma has to do with a nonzero $a \in \mathbb{F}$ having a $p^{\text {th }}$ root in $\mathbb{F}$.

Lemma 9.5.10 Let $\mathbb{F}$ be a field of characteristic $p$. Let $a \neq 0$ where $a \in \mathbb{F}$. Then either $x^{p}-a$ is irreducible or there exists $b \in \mathbb{F}$ such that $x^{p}-a=(x-b)^{p}$.

Proof: Suppose that $x^{p}-a$ is not irreducible. Then $x^{p}-a=g(x) f(x)$ where the degree of $g(x), k$ is less than $p$ and at least as large as 1 . Then let $b$ be a root of $g(x)$. Then $b^{p}-a=0$. Therefore,

$$
x^{p}-a=x^{p}-b^{p}=(x-b)^{p} .
$$

That is right. $x^{p}-b^{p}=(x-b)^{p}$ just like many beginning calculus students believe. It happens because of the binomial theorem and the fact that the other terms have a factor of p. Hence

$$
x^{p}-a=(x-b)^{p}=g(x) f(x)
$$

and so $g(x)$ divides $(x-b)^{p}$ which requires that $g(x)=(x-b)^{k}$ since $g(x)$ has degree $k$. It follows, since $g(x)$ is given to have coefficients in $\mathbb{F}$, that $b^{k} \in \mathbb{F}$. Also $b^{p} \in \mathbb{F}$. Since $k, p$ are relatively prime, due to the fact that $k<p$ with $p$ prime, there are integers $m, n$ such that $1=m k+n p$. Then from what you mean by raising $b$ to an integer power and the usual rules of exponents for integer powers, $b=\left(b^{k}\right)^{m}\left(b^{p}\right)^{n} \in \mathbb{F}$.

So when is a field of characteristic $p$ perfect? As observed above, for a field of characteristic $p,(a+b)^{p}=a^{p}+b^{p}$. Also, $(a b)^{p}=a^{p} b^{p}$. It follows that $a \rightarrow a^{p}$ is a homomorphism. This is also one to one because, as mentioned above $(a-b)^{p}=a^{p}-b^{p}$. Therefore, if $a^{p}=b^{p}$, it follows that $a=b$. Therefore, this homomorphism is also one to one.

Let $\mathbb{F}^{p}$ be the collection of $a^{p}$ where $a \in \mathbb{F}$. Then clearly $\mathbb{F}^{p}$ is a subfield of $\mathbb{F}$ because it is the image of a one to one homomorphism. What follows is the condition for a field of characteristic $p$ to be perfect.

Theorem 9.5.11 Let $\mathbb{F}$ be a field of characteristic $p$. Then $\mathbb{F}$ is perfect if and only if $\mathbb{F}=\mathbb{F}^{p}$.
Proof: Suppose $\mathbb{F}=\mathbb{F}^{p}$ first. Let $f(x)$ be an irreducible polynomial over $\mathbb{F}$. By Theorem 9.5.5, if $f^{\prime}(x)$ and $f(x)$ are relatively prime over $\mathbb{F}$ then $f(x)$ has no repeated roots. Suppose then that the two polynomials are not relatively prime. If $d(x)$ divides both $f(x)$ and $f^{\prime}(x)$ with degree of $d(x) \geq 1$. Then, since $f(x)$ is irreducible, it follows that $d(x)$ is a multiple of $f(x)$ and so $f(x)$ divides $f^{\prime}(x)$ which is impossible unless $f^{\prime}(x)=0$. But if $f^{\prime}(x)=0$, then $f(x)$ must be of the form

$$
a_{0}+a_{1} x^{p}+a_{2} x^{2 p}+\cdots+a_{n} x^{n p}
$$

since if it had some other nonzero term with exponent not a multiple of $p$ then $f^{\prime}(x)$ could not equal zero since you would have something surviving in the expression for the derivative after taking out multiples of $p$ which is like $k a x^{k-1}$ where $a \neq 0$ and $k<p$. Thus $k a \neq 0$. Hence the form of $f(x)$ is as indicated above.

If $a_{k}=b_{k}^{p}$ for some $b_{k} \in \mathbb{F}$, then the expression for $f(x)$ is

$$
b_{0}^{p}+b_{1}^{p} x^{p}+b_{2}^{p} x^{2 p}+\cdots+b_{n}^{p} x^{n p}=\left(b_{0}+b_{1} x+b_{x} x^{2}+\cdots+b_{n} x^{n}\right)^{p}
$$

because of the fact noted earlier that $a \rightarrow a^{p}$ is a homomorphism. However, this says that $f(x)$ is not irreducible after all. It follows that there exists $a_{k}$ such that $a_{k} \notin \mathbb{F}^{p}$ contrary to the assumption that $\mathbb{F}=\mathbb{F}^{p}$. Hence the greatest common divisor of $f^{\prime}(x)$ and $f(x)$ must be 1.

Next consider the other direction. Suppose $\mathbb{F} \neq \mathbb{F}^{p}$. Then there exists $a \in \mathbb{F} \backslash \mathbb{F}^{p}$. Consider the polynomial $x^{p}-a$. As noted, its derivative equals 0 . Therefore, $x^{p}-a$ and its derivative cannot be relatively prime. In fact, $x^{p}-a$ would divide both.

Now suppose $\mathbb{F}$ is a finite field. If $n \cdot 1$ is never equal to 0 then, since the field is finite, $k \cdot 1=m \cdot 1$, for some $k<m . m>k$, and $(m-k) \cdot 1=0$ which is a contradiction. Hence $\mathbb{F}$ is a field of characteristic $p$ for some prime $p$, by Proposition 9.5.8. The mapping $a \rightarrow a^{p}$ was shown to be a homomorphism which is also one to one. Therefore, $\mathbb{F}^{p}$ is a subfield of $\mathbb{F}$. It follows that it has characteristic $q$ for some $q$ a prime. However, this requires $q=p$ and so $\mathbb{F}^{p}=\mathbb{F}$. Then the following corollary is obtained from the above theorem.

With this information, here is a convenient version of the fundamental theorem of Galois theory.

Theorem 9.5.12 Let $\mathbb{K}$ be a splitting field of any polynomial $p(x) \in \mathbb{F}[x]$ where $\mathbb{F}$ is either of characteristic 0 or of characteristic $p$ with $\mathbb{F}^{p}=\mathbb{F}$. Let $\left\{\mathbb{L}_{i}\right\}_{i=0}^{k}$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$. Then each of these is a normal extension of $\mathbb{F}$ and the Galois group $G\left(\mathbb{L}_{j-1}, \mathbb{F}\right)$ is a normal subgroup of $G\left(\mathbb{L}_{j}, \mathbb{F}\right)$. In addition to this,

$$
G\left(\mathbb{L}_{j}, \mathbb{F}\right) \simeq G(\mathbb{K}, \mathbb{F}) / G\left(\mathbb{K}, \mathbb{L}_{j}\right)
$$

where the symbol $\simeq$ indicates the two spaces are isomorphic.

## Part II

## Linear Algebra as Baby Functional Analysis

## Chapter 10

## Normed Linear Spaces

In addition to the algebraic aspects of linear algebra presented earlier, there are many analytical and geometrical concepts which are usually included. This material involves the special fields $\mathbb{R}$ and $\mathbb{C}$ instead of general fields. It is these things which are typically generalized in functional analysis. The main new idea is that the notion of distance is included. This allows one to consider continuity, compactness, and many other topics from calculus. First is a general treatment of the notion of distance which has nothing to do with linear algebra but is a useful part of the vocabulary leading most efficiently to the inclusion of analytical topics.

### 10.1 Metric Spaces

This section is here to provide definitions and main theorems about fundamental analytical ideas and terminology. The first part is on metric spaces which really have absolutely nothing to do with linear algebra but they provide a convenient framework for discussion of the analytical aspects of linear algebra.

### 10.1.1 Limits

It is most efficient to discus things in terms of abstract metric spaces to begin with.
Definition 10.1.1 A non empty set $X$ is called a metric space if there is a function $d$ : $X \times X \rightarrow[0, \infty)$ which satisfies the following axioms.

1. $d(x, y)=d(y, x)$
2. $d(x, y) \geq 0$ and equals 0 if and only if $x=y$
3. $d(x, y)+d(y, z) \geq d(x, z)$

This function $d$ is called the metric. We often refer to it as the distance.
Definition 10.1.2 An open ball, denoted as $B(x, r)$ is defined as follows.

$$
B(x, r) \equiv\{y: d(x, y)<r\}
$$

A set $U$ is said to be open if whenever $x \in U$, it follows that there is $r>0$ such that $B(x, r) \subseteq U$. More generally, a point $x$ is said to be an interior point of $U$ if there exists such a ball. In words, an open set is one for which every point is an interior point.

For example, you could have $X$ be a subset of $\mathbb{R}$ and $d(x, y)=|x-y|$.
Then the first thing to show is the following.
Proposition 10.1.3 An open ball is an open set.
Proof: Suppose $y \in B(x, r)$. We need to verify that $y$ is an interior point of $B(x, r)$. Let $\delta=r-d(x, y)$. Then if $z \in B(y, \delta)$, it follows that

$$
d(z, x) \leq d(z, y)+d(y, x)<\delta+d(y, x)=r-d(x, y)+d(y, x)=r
$$

Thus $y \in B(y, \delta) \subseteq B(x, r)$.

Definition 10.1.4 Let $S$ be a nonempty subset of a metric space. Then $p$ is a limit point (accumulation point) of S iffor every $r>0$ there exists a point different than $p$ in $B(p, r) \cap S$. Sometimes people denote the set of limit points as $S^{\prime}$.

A related idea is the notion of the limit of a sequence. Recall that a sequence is really just a mapping from $\mathbb{N}$ to $X$. We write them as $\left\{x_{n}\right\}$ or $\left\{x_{n}\right\}_{n=1}^{\infty}$ if we want to emphasize the values of $n$. Then the following definition is what it means for a sequence to converge.

Definition 10.1.5 We say that $x=\lim _{n \rightarrow \infty} x_{n}$ when for every $\varepsilon>0$ there exists $N$ such that if $n \geq N$, then

$$
d\left(x, x_{n}\right)<\varepsilon
$$

Often we write $x_{n} \rightarrow x$ for short. This is equivalent to saying

$$
\lim _{n \rightarrow \infty} d\left(x, x_{n}\right)=0
$$

Proposition 10.1.6 The limit is well defined. That is, if $x, x^{\prime}$ are both limits of a sequence, then $x=x^{\prime}$.

Proof: From the definition, there exist $N, N^{\prime}$ such that if $n \geq N$, then $d\left(x, x_{n}\right)<\varepsilon / 2$ and if $n \geq N^{\prime}$, then $d\left(x, x_{n}\right)<\varepsilon / 2$. Then let $M \geq \max \left(N, N^{\prime}\right)$. Let $n>M$. Then

$$
d\left(x, x^{\prime}\right) \leq d\left(x, x_{n}\right)+d\left(x_{n}, x^{\prime}\right)<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon
$$

Since $\varepsilon$ is arbitrary, this shows that $x=x^{\prime}$ because $d\left(x, x^{\prime}\right)=0$.
Next there is an important theorem about limit points and convergent sequences.
Theorem 10.1.7 Let $S \neq \emptyset$. Then $p$ is a limit point of $S$ if and only if there exists a sequence of distinct points of $S,\left\{x_{n}\right\}$ none of which equal $p$ such that $\lim _{n \rightarrow \infty} x_{n}=p$.

Proof: $\Longrightarrow$ Suppose $p$ is a limit point. Why does there exist the promised convergent sequence? Let $x_{1} \in B(p, 1) \cap S$ such that $x_{1} \neq p$. If $x_{1}, \cdots, x_{n}$ have been chosen, let $x_{n+1} \neq p$ be in

$$
B\left(p, \delta_{n+1}\right) \cap S
$$

where $\delta_{n+1}=\min \left\{\frac{1}{n+1}, d\left(x_{i}, p\right), i=1,2, \cdots, n\right\}$. Then this constructs the necessary convergent sequence.
$\Longleftarrow$ Conversely, if such a sequence $\left\{x_{n}\right\}$ exists, then for every $r>0, B(p, r)$ contains $x_{n} \in S$ for all $n$ large enough. Hence, $p$ is a limit point because none of these $x_{n}$ are equal to $p$.

Definition 10.1.8 $A$ set $H$ is closed means $H^{C}$ is open.
Note that this says that the complement of an open set is closed. If $V$ is open, then the complement of its complement is itself. Thus $\left(V^{C}\right)^{C}=V$ an open set. Hence $V^{C}$ is closed. Then the following theorem gives the relation between closed sets and limit points.

Theorem 10.1.9 A set $H$ is closed if and only if it contains all of its limit points.

Proof: $\Longrightarrow$ Let $H$ be closed and let $p$ be a limit point. We need to verify that $p \in H$. If it is not, then since $H$ is closed, its complement is open and so there exists $\delta>0$ such that $B(p, \delta) \cap H=\emptyset$. However, this prevents $p$ from being a limit point.
$\Longleftarrow$ Next suppose $H$ has all of its limit points. Why is $H^{C}$ open? If $p \in H^{C}$ then it is not a limit point and so there exists $\delta>0$ such that $B(p, \delta)$ has no points of $H$. In other words, $H^{C}$ is open. Hence $H$ is closed.

Corollary 10.1.10 $A$ set $H$ is closed if and only if whenever $\left\{h_{n}\right\}$ is a sequence of points of $H$ which converges to a point $x$, it follows that $x \in H$.

Proof: $\Longrightarrow$ Suppose $H$ is closed and $h_{n} \rightarrow x$. If $x \in H$ there is nothing left to show. If $x \notin H$, then from the definition of limit, it is a limit point of $H$. Hence $x \in H$ after all.
$\Longleftarrow$ Suppose the limit condition holds, why is $H$ closed? Let $x \in H^{\prime}$ the set of limit points of $H$. By Theorem 10.1.7 there exists a sequence of points of $H,\left\{h_{n}\right\}$ such that $h_{n} \rightarrow x$. Then by assumption, $x \in H$. Thus $H$ contains all of its limit points and so it is closed by Theorem 10.1.9.

Next is the important concept of a subsequence.
Definition 10.1.11 Let $\left\{x_{n}\right\}_{n=1}^{\infty}$ be a sequence. Then if $n_{1}<n_{2}<\cdots$ is a strictly increasing sequence of indices, we say $\left\{x_{n_{k}}\right\}_{k=1}^{\infty}$ is a subsequence of $\left\{x_{n}\right\}_{n=1}^{\infty}$.

The really important thing about subsequences is that they preserve convergence.
Theorem 10.1.12 Let $\left\{x_{n_{k}}\right\}$ be a subsequence of a convergent sequence $\left\{x_{n}\right\}$ where $x_{n} \rightarrow$ $x$. Then $\lim _{k \rightarrow \infty} x_{n_{k}}=x$ also.

Proof: Let $\varepsilon>0$ be given. Then there exists $N$ such that $d\left(x_{n}, x\right)<\varepsilon$ if $n \geq N$. It follows that if $k \geq N$, then $n_{k} \geq N$ and so $d\left(x_{n_{k}}, x\right)<\varepsilon$ if $k \geq N$. This is what it means to say $\lim _{k \rightarrow \infty} x_{n_{k}}=x$.

Another useful idea is the distance to a set.
Definition 10.1.13 Let $(X, d)$ be a metric space and let $S$ be a nonempty set in $X$. Then

$$
\operatorname{dist}(x, S) \equiv \inf \{d(x, y): y \in S\}
$$

The following lemma is the fundamental result.
Lemma 10.1.14 The function, $x \rightarrow \operatorname{dist}(x, S)$ is continuous and in fact satisfies

$$
|\operatorname{dist}(x, S)-\operatorname{dist}(y, S)| \leq d(x, y)
$$

Proof: Suppose dist $(x, S)$ is as least as large as $\operatorname{dist}(y, S)$. Then pick $z \in S$ such that $d(y, z) \leq \operatorname{dist}(y, S)+\varepsilon$. Then

$$
\begin{aligned}
& |\operatorname{dist}(x, S)-\operatorname{dist}(y, S)|=\operatorname{dist}(x, S)-\operatorname{dist}(y, S) \leq d(x, z)-(d(y, z)-\varepsilon) \\
& \quad=d(x, z)-d(y, z)+\varepsilon \leq d(x, y)+d(y, z)-d(y, z)+\varepsilon=d(x, y)+\varepsilon
\end{aligned}
$$

Since $\varepsilon>0$ is arbitrary, this proves the lemma. It is similar if $\operatorname{dist}(x, S) \leq \operatorname{dist}(y, S)$. Just switch the roles of $x$ and $y$.

### 10.1.2 Cauchy Sequences, Completeness

Of course it does not go the other way. For example, you could let $x_{n}=(-1)^{n}$ and it has a convergent subsequence but fails to converge. Here $d(x, y)=|x-y|$ and the metric space is just $\mathbb{R}$.

However, there is a kind of sequence for which it does go the other way. This is called a Cauchy sequence.

Definition 10.1.15 $\left\{x_{n}\right\}$ is called a Cauchy sequence iffor every $\varepsilon>0$ there exists $N$ such that if $m, n \geq N$, then $d\left(x_{n}, x_{m}\right)<\varepsilon$.

Now the major theorem about this is the following.
Theorem 10.1.16 Let $\left\{x_{n}\right\}$ be a Cauchy sequence. Then it converges if and only if any subsequence converges.

Proof: $\Longrightarrow$ This was just done above.
$\Longleftarrow$ Suppose now that $\left\{x_{n}\right\}$ is a Cauchy sequence and $\lim _{k \rightarrow \infty} x_{n_{k}}=x$. Then there exists $N_{1}$ such that if $k>N_{1}$, then $d\left(x_{n_{k}}, x\right)<\varepsilon / 2$. From the definition of what it means to be Cauchy, there exists $N_{2}$ such that if $m, n \geq N_{2}$, then $d\left(x_{m}, x_{n}\right)<\varepsilon / 2$. Let $N \geq \max \left(N_{1}, N_{2}\right)$. Then if $k \geq N$, then $n_{k} \geq N$ and so

$$
\begin{equation*}
d\left(x, x_{k}\right) \leq d\left(x, x_{n_{k}}\right)+d\left(x_{n_{k}}, x_{k}\right)<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon \tag{10.1}
\end{equation*}
$$

It follows from the definition that $\lim _{k \rightarrow \infty} x_{k}=x$.
Definition 10.1.17 A metric space is said to be complete if every Cauchy sequence converges.

Another nice thing to note is this.
Proposition 10.1.18 If $\left\{x_{n}\right\}$ is a sequence and if $p$ is a limit point of the set $S=\cup_{n=1}^{\infty}\left\{x_{n}\right\}$ then there is a subsequence $\left\{x_{n_{k}}\right\}$ such that $\lim _{k \rightarrow \infty} x_{n_{k}}=x$.

Proof: By Theorem 10.1.7, there exists a sequence of distinct points of $S$ denoted as $\left\{y_{k}\right\}$ such that none of them equal $p$ and $\lim _{k \rightarrow \infty} y_{k}=p$. Thus $B(p, r)$ contains infinitely many different points of the set $D$, this for every $r$. Let $x_{n_{1}} \in B(p, 1)$ where $n_{1}$ is the first index such that $x_{n_{1}} \in B(p, 1)$. Suppose $x_{n_{1}}, \cdots, x_{n_{k}}$ have been chosen, the $n_{i}$ increasing and let $1>\delta_{1}>\delta_{2}>\cdots>\delta_{k}$ where $x_{n_{i}} \in B\left(p, \delta_{i}\right)$. Then let

$$
\delta_{k+1} \leq \min \left\{\frac{1}{2^{k+1}}, d\left(p, x_{n_{j}}\right), \delta_{j}, j=1,2 \cdots, k\right\}
$$

Let $x_{n_{k+1}} \in B\left(p, \delta_{k+1}\right)$ where $n_{k+1}$ is the first index such that $x_{n_{k+1}}$ is contained $B\left(p, \delta_{k+1}\right)$. Then $\lim _{k \rightarrow \infty} x_{n_{k}}=p$.

Another useful result is the following.
Lemma 10.1.19 Suppose $x_{n} \rightarrow x$ and $y_{n} \rightarrow y$. Then $d\left(x_{n}, y_{n}\right) \rightarrow d(x, y)$.

Proof: Consider the following.

$$
d(x, y) \leq d\left(x, x_{n}\right)+d\left(x_{n}, y\right) \leq d\left(x, x_{n}\right)+d\left(x_{n}, y_{n}\right)+d\left(y_{n}, y\right)
$$

so $d(x, y)-d\left(x_{n}, y_{n}\right) \leq d\left(x, x_{n}\right)+d\left(y_{n}, y\right)$. Similarly

$$
d\left(x_{n}, y_{n}\right)-d(x, y) \leq d\left(x, x_{n}\right)+d\left(y_{n}, y\right)
$$

and so

$$
\left|d\left(x_{n}, y_{n}\right)-d(x, y)\right| \leq d\left(x, x_{n}\right)+d\left(y_{n}, y\right)
$$

and the right side converges to 0 as $n \rightarrow \infty$.
First are some simple lemmas featuring one dimensional considerations. In these, the metric space is $\mathbb{R}$ and the distance is given by $d(x, y) \equiv|x-y|$.First recall the nested interval lemma. You should have seen something like it in calculus, but this is often not the case because there is much more interest in trivialities like integration techniques.

Lemma 10.1.20 Let $\left[a_{k}, b_{k}\right] \supseteq\left[a_{k+1}, b_{k+1}\right]$ for all $k=1,2,3, \cdots$. Then there exists a point $p$ in $\cap_{k=1}^{\infty}\left[a_{k}, b_{k}\right]$.

Proof: We note that for any $k, l, a_{k} \leq b_{l}$. Here is why. If $k \leq l$, then $a_{k} \leq a_{l} \leq b_{l}$. If $k>l$, then $b_{l} \geq b_{k} \geq a_{k}$. It follows that for each $l, \sup _{k} a_{k} \leq b_{l}$. Hence $\sup _{k} a_{k}$ is a lower bound to the set of all $b_{l}$ and so it is no larger than the greatest lower bound. It follows that $\sup _{k} a_{k} \leq \inf _{l} b_{l}$. Pick $x \in\left[\sup _{k} a_{k}, \inf _{l} b_{l}\right]$. Then for every $k, a_{k} \leq x \leq b_{k}$. Hence $x \in \cap_{k=1}^{\infty}\left[a_{k}, b_{k}\right]$.
Lemma 10.1.21 The closed interval $[a, b]$ is compact. This means that if there is a collection of open intervals of the form $(a, b)$ whose union includes all of $[a, b]$, then in fact $[a, b]$ is contained in the union of finitely many of these open intervals.

Proof: Let $\mathscr{C}$ be a set of open intervals the union of which includes all of $[a, b]$ and suppose $[a, b]$ fails to admit a finite subcover. That is, no finite subset of $\mathscr{C}$ has union which contains $[a, b]$. Then this must be the case for one of the two intervals $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$. Let $I_{1}$ be the one for which this is so. Then split it into two equal pieces like what was just done and let $I_{2}$ be a half for which there is no finite subcover of sets of $\mathscr{C}$. Continue this way. This yields a nested sequence of closed intervals $I_{1} \supseteq I_{2} \supseteq \cdots$ and by the above lemma, there exists a point $x$ in all of these intervals. There exists $U \in \mathscr{C}$ such that $x \in U$. Thus $x \in(a, b) \in \mathscr{C}$. However, for all $n$ large enough, the length of $I_{n}$ is less than $\min (|x-a|,|x-b|)$. Hence $I_{n}$ is actually contained in $(a, b) \in \mathscr{C}$ contrary to the construction. Hence $[a, b]$ is compact after all.

As a useful corollary, this shows that $\mathbb{R}$ is complete.
Corollary 10.1.22 The real line $\mathbb{R}$ is complete.
Proof: Suppose $\left\{x_{k}\right\}$ is a Cauchy sequence in $\mathbb{R}$. Then there exists $M$ such that

$$
\left\{x_{k}\right\}_{k=1}^{\infty} \subseteq[-M, M]
$$

Why? If there is no convergent subsequence, then for each $x \in[-M, M]$, there is an open set $\left(x-\delta_{x}, x+\delta_{x}\right)$ which contains $x_{k}$ for only finitely many values of $k$. Since $[-M, M]$ is compact, there are finitely many of these open sets whose union includes $[-M, M]$. This is a contradiction because $[-M, M]$ contains $x_{k}$ for all $k \in \mathbb{N}$ so at least one of the open sets must contain $x_{k}$ for infinitely many $k$. Thus there is a convergent subsequence. Therefore, using Theorem 10.1.16, the original Cauchy sequence converges to some $x \in[-M, M]$.

Example 10.1.23 Let $n \in \mathbb{N}$. $\mathbb{C}^{n}$ with distance given by $d(\boldsymbol{x}, \boldsymbol{y}) \equiv \max _{j \in\{1, \cdots, n\}}\left\{\left|x_{j}-y_{j}\right|\right\}$ is a complete space. Recall that $|a+j b| \equiv \sqrt{a^{2}+b^{2}}$. Then $\mathbb{C}^{n}$ is complete. Similarly $\mathbb{R}^{n}$ is complete.

To see that this is complete, let $\left\{\boldsymbol{x}^{k}\right\}_{k=1}^{\infty}$ be a Cauchy sequence. Observe that for each $j,\left\{x_{j}^{k}\right\}_{k=1}^{\infty}$. That is, each component is a Cauchy sequence in $\mathbb{C}$. Next,

$$
\left|\operatorname{Re} x_{j}^{k}-\operatorname{Re} x_{j}^{k+p}\right| \leq\left|x_{j}^{k}-x_{j}^{k+p}\right|
$$

Therefore, $\left\{\operatorname{Re} x_{j}^{k}\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Similarly $\left\{\operatorname{Im} x_{j}^{k}\right\}_{k=1}^{\infty}$ is a Cauchy sequence. It follows from completeness of $\mathbb{R}$ shown above, that these converge. Thus there exists $a_{j}, b_{j}$ such that

$$
\lim _{k \rightarrow \infty} \operatorname{Re} x_{j}^{k}+i \operatorname{Im} x_{j}^{k}=a_{j}+i b_{j} \equiv \boldsymbol{x}
$$

and so $\boldsymbol{x}^{k} \rightarrow \boldsymbol{x}$ showing that $\mathbb{C}^{n}$ is complete. The same argument shows that $\mathbb{R}^{n}$ is complete. It is easier because you don't need to fuss with real and imaginary parts.

### 10.1.3 Closure of a Set

Next is the topic of the closure of a set.
Definition 10.1.24 Let A be a nonempty subset of $(X, d)$ a metric space. Then $\bar{A}$ is defined to be the intersection of all closed sets which contain A. Note the whole space, $X$ is one such closed set which contains $A$. The whole space $X$ is closed because its complement is open, its complement being $\emptyset$. It is certainly true that every point of the empty set is an interior point because there are no points of $\emptyset$.

Lemma 10.1.25 Let $A$ be a nonempty set in $(X, d)$. Then $\bar{A}$ is a closed set and $\bar{A}=A \cup A^{\prime}$ where $A^{\prime}$ denotes the set of limit points of $A$.

Proof: First of all, denote by $\mathscr{C}$ the set of closed sets which contain $A$. Then $\bar{A}=\cap \mathscr{C}$ and this will be closed if its complement is open. However, $(\bar{A})^{C}=\cup\left\{H^{C}: H \in \mathscr{C}\right\}$.Each $H^{C}$ is open and so the union of all these open sets must also be open. This is because if $x$ is in this union, then it is in at least one of them. Hence it is an interior point of that one. But this implies it is an interior point of the union of them all which is an even larger set. Thus $\bar{A}$ is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \bar{A}$ so if $x \in A$, then $x \in \bar{A}$. Now consider $y \in A^{\prime}$ but $y \notin A$. If $y \notin \bar{A}$, a closed set, then there exists $B(y, r) \subseteq \bar{A}^{C}$. Thus $y$ cannot be a limit point of $A$, a contradiction. Therefore, $A \cup A^{\prime} \subseteq \bar{A}$

Next suppose $x \in \bar{A}$ and suppose $x \notin A$. Then if $B(x, r)$ contains no points of $A$ different than $x$, since $x$ itself is not in $A$, it would follow that $B(x, r) \cap A=\emptyset$ and so recalling that open balls are open, $B(x, r)^{C}$ is a closed set containing $A$ so from the definition, it also contains $\bar{A}$ which is contrary to the assertion that $x \in \bar{A}$. Hence if $x \notin A$, then $x \in A^{\prime}$ and so $A \cup A^{\prime} \supseteq \bar{A}$

### 10.1.4 Continuous Functions

The following is a fairly general definition of what it means for a function to be continuous. It includes everything seen in typical calculus classes as a special case.

Definition 10.1.26 Let $f: X \rightarrow Y$ be a function where $(X, d)$ and $(Y, \rho)$ are metric spaces. Then $f$ is continuous at $x \in X$ if and only if the following condition holds. For every $\varepsilon>0$, there exists $\delta>0$ such that if $d(\hat{x}, x)<\boldsymbol{\delta}$, then $\rho(f(\hat{x}), f(x))<\varepsilon$. If $f$ is continuous at every $x \in X$ we say that $f$ is continuous on $X$.

For example, you could have a real valued function $f(x)$ defined on an interval $[0,1]$. In this case you would have $X=[0,1]$ and $Y=\mathbb{R}$ with the distance given by $d(x, y)=|x-y|$. Then the following theorem is the main result.

Theorem 10.1.27 Let $f: X \rightarrow Y$ where $(X, d)$ and $(Y, \rho)$ are metric spaces. Then the following are equivalent.
a $f$ is continuous at $x$.
$b$ Whenever $x_{n} \rightarrow x$, it follows that $f\left(x_{n}\right) \rightarrow f(x)$.
Also, the following are equivalent.
c $f$ is continuous on $X$.
$d$ Whenever $V$ is open in $Y$, it follows that $f^{-1}(V) \equiv\{x: f(x) \in V\}$ is open in $X$.
$e$ Whenever $H$ is closed in $Y$, it follows that $f^{-1}(H)$ is closed in $X$.
Proof: $\mathrm{a} \Longrightarrow \mathrm{b}$ : Let $f$ be continuous at $x$ and suppose $x_{n} \rightarrow x$. Then let $\varepsilon>0$ be given. By continuity, there exists $\delta>0$ such that if $d(\hat{x}, x)<\delta$, then $\rho(f(\hat{x}), f(x))<\varepsilon$. Since $x_{n} \rightarrow x$, it follows that there exists $N$ such that if $n \geq N$, then $d\left(x_{n}, x\right)<\delta$ and so, if $n \geq N$, it follows that $\rho\left(f\left(x_{n}\right), f(x)\right)<\varepsilon$. Since $\varepsilon>0$ is arbitrary, it follows that $f\left(x_{n}\right) \rightarrow f(x)$.
$\mathrm{b} \Longrightarrow \mathrm{a}$ : Suppose b holds but $f$ fails to be continuous at $x$. Then there exists $\varepsilon>0$ such that for all $\delta>0$, there exists $\hat{x}$ such that $d(\hat{x}, x)<\delta$ but $\rho(f(\hat{x}), f(x)) \geq \varepsilon$. Letting $\delta=1 / n$, there exists $x_{n}$ such that $d\left(x_{n}, x\right)<1 / n$ but $\rho\left(f\left(x_{n}\right), f(x)\right) \geq \varepsilon$. Now this is a contradiction because by assumption, the fact that $x_{n} \rightarrow x$ implies that $f\left(x_{n}\right) \rightarrow f(x)$. In particular, for large enough $n, \rho\left(f\left(x_{n}\right), f(x)\right)<\varepsilon$ contrary to the construction.
$\mathrm{c} \Longrightarrow \mathrm{d}$ : Let $V$ be open in $Y$. Let $x \in f^{-1}(V)$ so that $f(x) \in V$. Since $V$ is open, there exists $\varepsilon>0$ such that $B(f(x), \varepsilon) \subseteq V$. Since $f$ is continuous at $x$, it follows that there exists $\delta>0$ such that if $\hat{x} \in B(x, \delta)$, then $f(\hat{x}) \in B(f(x), \varepsilon) \subseteq V .(f(B(x, \delta)) \subseteq B(f(x), \varepsilon))$ In other words, $B(x, \delta) \subseteq f^{-1}(B(f(x), \varepsilon)) \subseteq f^{-1}(V)$ which shows that, since $x$ was an arbitrary point of $f^{-1}(V)$, every point of $f^{-1} \overline{(V)}$ is an interior point which implies $f^{-1}(V)$ is open.
$\mathrm{d} \Longrightarrow$ e: Let $H$ be closed in $Y$. Then $f^{-1}(H)^{C}=f^{-1}\left(H^{C}\right)$ which is open by assumption. Hence $f^{-1}(H)$ is closed because its complement is open.
$\mathrm{e} \Longrightarrow \mathrm{d}$ : Let $V$ be open in $Y$. Then $f^{-1}(V)^{C}=f^{-1}\left(V^{C}\right)$ which is assumed to be closed. This is because the complement of an open set is a closed set.
$\mathrm{d} \Longrightarrow \mathrm{c}$ : Let $x \in X$ be arbitrary. Is it the case that $f$ is continuous at $x$ ? Let $\varepsilon>0$ be given. Then $B(f(x), \varepsilon)$ is an open set in $V$ and so $x \in f^{-1}(B(f(x), \varepsilon))$ which is given to be open. Hence there exists $\delta>0$ such that $x \in B(x, \delta) \subseteq f^{-1}(B(f(x), \varepsilon))$. Thus, $f(B(x, \boldsymbol{\delta})) \subseteq B(f(x), \varepsilon)$ so $\rho(f(\hat{x}), f(x))<\varepsilon$. Thus $f$ is continuous at $x$ for every $x$.

### 10.1.5 Separable Metric Spaces

Definition 10.1.28 A metric space is called separable if there exists a countable dense subset $D$. This means two things. First, $D$ is countable, and second that if $x$ is any point and $r>0$, then $B(x, r) \cap D \neq \emptyset$. A metric space is called completely separable if there exists a countable collection of nonempty open sets $\mathscr{B}$ such that every open set is the union of some subset of $\mathscr{B}$. This collection of open sets is called a countable basis.

For those who like to fuss about empty sets, the empty set is open and it is indeed the union of a subset of $\mathscr{B}$ namely the empty subset.

Theorem 10.1.29 A metric space is separable if and only if it is completely separable.
Proof: $\Longleftarrow$ Let $\mathscr{B}$ be the special countable collection of open sets and for each $B \in \mathscr{B}$, let $p_{B}$ be a point of $B$. Then let $\mathscr{P} \equiv\left\{p_{B}: B \in \mathscr{B}\right\}$. If $B(x, r)$ is any ball, then it is the union of sets of $\mathscr{B}$ and so there is a point of $\mathscr{P}$ in it. Since $\mathscr{B}$ is countable, so is $\mathscr{P}$.
$\Longrightarrow$ Let $D$ be the countable dense set and let

$$
\mathscr{B} \equiv\{B(d, r): d \in D, r \in \mathbb{Q} \cap[0, \infty)\}
$$

Then $\mathscr{B}$ is countable because the Cartesian product of countable sets is countable. It suffices to show that every ball is the union of these sets. Let $B(x, R)$ be a ball. Let $y \in B(y, \delta) \subseteq B(x, R)$. Then there exists $d \in B\left(y, \frac{\delta}{10}\right)$. Let $\varepsilon \in \mathbb{Q}$ and $\frac{\delta}{10}<\varepsilon<\frac{\delta}{5}$. Then $y \in B(d, \varepsilon) \in \mathscr{B}$. Is $B(d, \varepsilon) \subseteq B(x, R)$ ? If so, then the desired result follows because this would show that every $y \in B(x, R)$ is contained in one of these sets of $\mathscr{B}$ which is contained in $B(x, R)$ showing that $B(x, R)$ is the union of sets of $\mathscr{B}$. Let $z \in B(d, \varepsilon) \subseteq B\left(d, \frac{\delta}{5}\right)$. Then

$$
d(y, z) \leq d(y, d)+d(d, z)<\frac{\delta}{10}+\varepsilon<\frac{\delta}{10}+\frac{\delta}{5}<\delta
$$

Hence $B(d, \varepsilon) \subseteq B(y, \delta) \subseteq B(x, r)$. Therefore, every ball is the union of sets of $\mathscr{B}$ and, since every open set is the union of balls, it follows that every open set is the union of sets of $\mathscr{B}$.

Definition 10.1.30 Let $S$ be a nonempty set. Then a set of open sets $\mathscr{C}$ is called an open cover of $S$ if $\cup \mathscr{C} \supseteq \mathscr{S}$. (It covers up the set $S$. Think lilly pads covering the surface of $a$ pond.)

One of the important properties possessed by separable metric spaces is the Lindeloff property.

Definition 10.1.31 A metric space has the Lindeloff property if whenever $\mathscr{C}$ is an open cover of a set $S$, there exists a countable subset of $\mathscr{C}$ denoted here by $\mathscr{B}$ such that $\mathscr{B}$ is also an open cover of $S$.

Theorem 10.1.32 Every separable metric space has the Lindeloff property.
Proof: Let $\mathscr{C}$ be an open cover of a set $S$. Let $\mathscr{B}$ be a countable basis. Such exists by Theorem 10.1.29. Let $\hat{\mathscr{B}}$ denote those sets of $\mathscr{B}$ which are contained in some set of $\mathscr{C}$. Thus $\hat{\mathscr{B}}$ is a countable open cover of $S$. Now for $B \in \mathscr{B}$, let $U_{B}$ be a set of $\mathscr{C}$ which contains $B$. Letting $\widehat{\mathscr{C}}$ denote these sets $U_{B}$ it follows that $\widehat{\mathscr{C}}$ is countable and is an open cover of $S$.

Definition 10.1.33 A Polish space is a complete separable metric space. These things turn out to be very useful in probability theory and in other areas.

### 10.1.6 Compact Sets in Metric Space

As usual, we are not worrying about empty sets.
Definition 10.1.34 A metric space $K$ is compact if whenever $\mathscr{C}$ is an open cover of

$$
K,(\cup \mathscr{C} \supseteq K, \text { each set of } \mathscr{C} \text { is open })
$$

there exists a finite subset of $\mathscr{C}\left\{U_{1}, \cdots, U_{n}\right\}$ such that $K \subseteq \cup_{k=1}^{n} U_{k}$. In words, every open cover admits a finite sub-cover.

The above definition is equivalent to the same statement with the provision that each open set in $\mathscr{C}$ is an open ball. See Problem 15 on Page 283.

This is the real definition given above. However, in metric spaces, it is equivalent to another definition called sequentially compact.

Definition 10.1.35 A metric space $K$ is sequentially compact means that whenever $\left\{x_{n}\right\} \subseteq$ $K$, there exists a subsequence $\left\{x_{n_{k}}\right\}$ such that $\lim _{k \rightarrow \infty} x_{n_{k}}=x \in K$ for some point $x$. In words, every sequence has a subsequence which converges to a point in the set.

Definition 10.1.36 Let $X$ be a metric space. Then a finite set of points $\left\{x_{1}, \cdots, x_{n}\right\}$ is called an $\varepsilon$ net if

$$
X \subseteq \cup_{k=1}^{n} B\left(x_{k}, \varepsilon\right)
$$

If, for every $\varepsilon>0$ a metric space has an $\varepsilon$ net, then we say that the metric space is totally bounded.

Lemma 10.1.37 If a metric space $(K, d)$ is sequentially compact, then it is separable and totally bounded.

Proof: Pick $x_{1} \in K$. If $B\left(x_{1}, \varepsilon\right) \supseteq K$, then stop. Otherwise, pick $x_{2} \notin B\left(x_{1}, \varepsilon\right)$. Continue this way. If $\left\{x_{1}, \cdots, x_{n}\right\}$ have been chosen, either $K \subseteq \cup_{k=1}^{n} B\left(x_{k}, \varepsilon\right)$ in which case, you have found an $\varepsilon$ net or this does not happen in which case, you can pick $x_{n+1} \notin \cup_{k=1}^{n} B\left(x_{k}, \varepsilon\right)$. The process must terminate since otherwise, the sequence would need to have a convergent subsequence which is not possible because every pair of terms is farther apart than $\varepsilon$. Thus for every $\varepsilon>0$, there is an $\varepsilon$ net. Thus the metric space is totally bounded. Let $N_{\varepsilon}$ denote an $\varepsilon$ net. Let $D=\cup_{k=1}^{\infty} N_{1 / 2^{k}}$. Then this is a countable dense set. It is countable because it is the countable union of finite sets and it is dense because given a point, there is a point of $D$ within $1 / 2^{k}$ of it.

Also recall that a complete metric space is one for which every Cauchy sequence converges to a point in the metric space.

The following is the main theorem which relates these concepts.
Theorem 10.1.38 For $(X, d)$ a metric space, the following are equivalent.

1. $(X, d)$ is compact.
2. $(X, d)$ is sequentially compact.
3. $(X, d)$ is complete and totally bounded.

Proof: $1 . \Longrightarrow 2$. Let $\left\{x_{n}\right\}$ be a sequence. Suppose it fails to have a convergent subsequence. Then it follows right away that no value of the sequence is repeated infinitely often. If $\cup_{n=1}^{\infty}\left\{x_{n}\right\}$ has a limit point in $X$, then it follows from Proposition 10.1.18 there would be a convergent subsequence converging to this limit point. Therefore, assume $\cup_{k=1}^{\infty}\left\{x_{n}\right\}$ has no limit point. This is equivalent to saying that $\cup_{k=m}^{\infty}\left\{x_{k}\right\}$ has no limit point for each $m$. Thus these are closed sets by Theorem 10.1.9 because they contain all of their limit points due to the fact that they have none. Hence the open sets

$$
\left(\cup_{k=m}^{\infty}\left\{x_{n}\right\}\right)^{C}
$$

yield an open cover. This is an increasing sequence of open sets and none of them contain all the values of the sequence because no value is repeated for infinitely many indices. Thus this is an open cover which has no finite subcover contrary to 1 .
$2 . \Longrightarrow 3$. If $(X, d)$ is sequentially compact, then by Lemma 10.1.37, it is totally bounded. If $\left\{x_{n}\right\}$ is a Cauchy sequence, then there is a subsequence which converges to $x \in X$ by assumption. However, from Theorem 10.1.16 this requires the original Cauchy sequence to converge.
$3 . \Longrightarrow 1$. Since $(X, d)$ is totally bounded, there must be a countable dense subset of $X$. Just take the union of $1 / 2^{k}$ nets for each $k \in \mathbb{N}$. Thus $(X, d)$ is completely separable by Theorem 10.1.32 has the Lindeloff property. Hence, if $X$ is not compact, there is a countable set of open sets $\left\{U_{i}\right\}_{i=1}^{\infty}$ which covers $X$ but no finite subset does. Consider the nonempty closed sets $F_{n}$ and pick $x_{n} \in F_{n}$ where

$$
X \backslash \cup_{i=1}^{n} U_{i} \equiv X \cap\left(\cup_{i=1}^{n} U_{i}\right)^{C} \equiv F_{n}
$$

Let $\left\{x_{m}^{k}\right\}_{m=1}^{M_{k}}$ be a $1 / 2^{k}$ net for $X$. We have for some $m, B\left(x_{m_{k}}^{k}, 1 / 2^{k}\right)$ contains $x_{n}$ for infinitely many values of $n$ because there are only finitely many balls and infinitely many indices. Then of the finitely many $\left\{x_{m}^{k+1}\right\}$ for which $B\left(x_{m}^{k+1}, 1 / 2^{k+1}\right)$ has nonempty intersection with $B\left(x_{m_{k}}^{k}, 1 / 2^{k}\right)$, pick one $x_{m_{k+1}}^{k+1}$ such that $B\left(x_{m_{k+1}}^{k+1}, 1 / 2^{k+1}\right)$ contains $x_{n}$ for infinitely many $n$. Then obviously $\left\{x_{m_{k}}^{k}\right\}_{k=1}^{\infty}$ is a Cauchy sequence because

$$
d\left(x_{m_{k}}^{k}, x_{m_{k+1}}^{k+1}\right) \leq \frac{1}{2^{k}}+\frac{1}{2^{k+1}} \leq \frac{1}{2^{k-1}}
$$

Hence for $p<q$,

$$
d\left(x_{m_{p}}^{p}, x_{m_{q}}^{q}\right) \leq \sum_{k=p}^{q-1} d\left(x_{m_{k}}^{k}, x_{m_{k+1}}^{k+1}\right)<\sum_{k=p}^{\infty} \frac{1}{2^{k-1}}=\frac{1}{2^{p-2}}
$$

Now take a subsequence $x_{n_{k}} \in B\left(x_{m_{k}}^{k}, 2^{-k}\right)$ and it follows that $\lim _{k \rightarrow \infty} x_{n_{k}}=\lim _{k \rightarrow \infty} x_{m_{k}}^{k}=$ $x \in X$. However, $x \in F_{n}$ for each $n$ since each $F_{n}$ is closed and these sets are nested. Thus it follows that $x \in \cap_{n} F_{n}$ contrary to the claim that $\left\{U_{i}\right\}_{i=1}^{\infty}$ covers $X$.

One of the important theorems about compactness is the extreme value theorem.
Theorem 10.1.39 Let $(K, d)$ be a compact metric space. Let $f: K \rightarrow \mathbb{R}$ be continuous. Then $f$ achieves a maximum value and a minimum value on $K$.

Proof: Let $\lambda \equiv \sup \{f(x): x \in X\}$. Let $x_{n} \in K$ such that $\lim _{n \rightarrow \infty} f\left(x_{n}\right)=\lambda$. This is called a maximalizing sequence. By compactness, there is a subsequence $\left\{x_{n_{k}}\right\}$ such that $\lim _{k \rightarrow \infty} x_{n_{k}}=x \in K$. Then by continuity, $f(x)=\lim _{k \rightarrow \infty} f\left(x_{n_{k}}\right)=\lambda$. Similarly $f$ achieves its minimum on $K$. Do the same argument with a minimizing sequence.

### 10.1.7 Lipschitz Continuity and Contraction Maps

The following is of more interest in the case of normed vector spaces, but there is no harm in stating it in this more general setting. You should verify that the functions described in the following definition are all continuous.

Definition 10.1.40 Let $f: X \rightarrow Y$ where $(X, d)$ and $(Y, \rho)$ are metric spaces. Then $f$ is said to be Lipschitz continuous if for every $x, \hat{x} \in X, \rho(f(x), f(\hat{x})) \leq r d(x, \hat{x})$. The function is called a contraction map if $r<1$.

The big theorem about contraction maps is the following.
Theorem 10.1.41 Let $f:(X, d) \rightarrow(X, d)$ be a contraction map and let $(X, d)$ be a complete metric space. Thus Cauchy sequences converge and also $d(f(x), f(\hat{x})) \leq r d(x, \hat{x})$ where $r<1$. Then $f$ has a unique fixed point. This is a point $x \in X$ such that $f(x)=x$. Also, if $x_{0}$ is any point of $X$, then

$$
d\left(x, x_{0}\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

Also, for each n,

$$
d\left(f^{n}\left(x_{0}\right), x_{0}\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

and $x=\lim _{n \rightarrow \infty} f^{n}\left(x_{0}\right)$.
Proof: Pick $x_{0} \in X$ and consider the sequence of iterates of the map,

$$
x_{0}, f\left(x_{0}\right), f^{2}\left(x_{0}\right), \cdots
$$

We argue that this is a Cauchy sequence. For $m<n$, it follows from the triangle inequality,

$$
d\left(f^{m}\left(x_{0}\right), f^{n}\left(x_{0}\right)\right) \leq \sum_{k=m}^{n-1} d\left(f^{k+1}\left(x_{0}\right), f^{k}\left(x_{0}\right)\right) \leq \sum_{k=m}^{\infty} r^{k} d\left(f\left(x_{0}\right), x_{0}\right)
$$

The reason for this last is as follows.

$$
\begin{gathered}
d\left(f^{2}\left(x_{0}\right), f\left(x_{0}\right)\right) \leq r d\left(f\left(x_{0}\right), x_{0}\right) \\
d\left(f^{3}\left(x_{0}\right), f^{2}\left(x_{0}\right)\right) \leq r d\left(f^{2}\left(x_{0}\right), f\left(x_{0}\right)\right) \leq r^{2} d\left(f\left(x_{0}\right), x_{0}\right)
\end{gathered}
$$

and so forth. Therefore,

$$
d\left(f^{m}\left(x_{0}\right), f^{n}\left(x_{0}\right)\right) \leq d\left(f\left(x_{0}\right), x_{0}\right) \frac{r^{m}}{1-r}
$$

which shows that this is indeed a Cauchy sequence. Therefore, there exists $x$ such that

$$
\lim _{n \rightarrow \infty} f^{n}\left(x_{0}\right)=x
$$

By continuity,

$$
f(x)=f\left(\lim _{n \rightarrow \infty} f^{n}\left(x_{0}\right)\right)=\lim _{n \rightarrow \infty} f^{n+1}\left(x_{0}\right)=x
$$

Also note that this estimate yields

$$
d\left(x_{0}, f^{n}\left(x_{0}\right)\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

Now $d\left(x_{0}, x\right) \leq d\left(x_{0}, f^{n}\left(x_{0}\right)\right)+d\left(f^{n}\left(x_{0}\right), x\right)$ and so

$$
d\left(x_{0}, x\right)-d\left(f^{n}\left(x_{0}\right), x\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

Letting $n \rightarrow \infty$, it follows that

$$
d\left(x_{0}, x\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

It only remains to verify that there is only one fixed point. Suppose then that $x, x^{\prime}$ are two. Then

$$
d\left(x, x^{\prime}\right)=d\left(f(x), f\left(x^{\prime}\right)\right) \leq r d\left(x^{\prime}, x\right)
$$

and so $d\left(x, x^{\prime}\right)=0$ because $r<1$.
The above is the usual formulation of this important theorem, but we actually proved a better result.

Corollary 10.1.42 Let $B$ be a closed subset of the complete metric space $(X, d)$ and let $f: B \rightarrow X$ be a contraction map

$$
d(f(x), f(\hat{x})) \leq r d(x, \hat{x}), r<1
$$

Also suppose there exists $x_{0} \in B$ such that the sequence of iterates $\left\{f^{n}\left(x_{0}\right)\right\}_{n=1}^{\infty}$ remains in $B$. Then $f$ has a unique fixed point in $B$ which is the limit of the sequence of iterates. This is a point $x \in B$ such that $f(x)=x$. In the case that $B=\overline{B\left(x_{0}, \delta\right)}$, the sequence of iterates satisfies the inequality

$$
d\left(f^{n}\left(x_{0}\right), x_{0}\right) \leq \frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}
$$

and so it will remain in $B$ if

$$
\frac{d\left(x_{0}, f\left(x_{0}\right)\right)}{1-r}<\delta
$$

Proof: By assumption, the sequence of iterates stays in $B$. Then, as in the proof of the preceding theorem, for $m<n$, it follows from the triangle inequality,

$$
\begin{aligned}
d\left(f^{m}\left(x_{0}\right), f^{n}\left(x_{0}\right)\right) & \leq \sum_{k=m}^{n-1} d\left(f^{k+1}\left(x_{0}\right), f^{k}\left(x_{0}\right)\right) \\
& \leq \sum_{k=m}^{\infty} r^{k} d\left(f\left(x_{0}\right), x_{0}\right)=\frac{r^{m}}{1-r} d\left(f\left(x_{0}\right), x_{0}\right)
\end{aligned}
$$

Hence the sequence of iterates is Cauchy and must converge to a point $x$ in $X$. However, $B$ is closed and so it must be the case that $x \in B$. Then as before,

$$
x=\lim _{n \rightarrow \infty} f^{n}\left(x_{0}\right)=\lim _{n \rightarrow \infty} f^{n+1}\left(x_{0}\right)=f\left(\lim _{n \rightarrow \infty} f^{n}\left(x_{0}\right)\right)=f(x)
$$

As to the sequence of iterates remaining in $B$ where $B$ is a ball as described, the inequality above in the case where $m=0$ yields

$$
d\left(x_{0}, f^{n}\left(x_{0}\right)\right) \leq \frac{1}{1-r} d\left(f\left(x_{0}\right), x_{0}\right)
$$

and so, if the right side is less than $\delta$, then the iterates remain in $B$. As to the fixed point being unique, it is as before. If $x, x^{\prime}$ are both fixed points in $B$, then $d\left(x, x^{\prime}\right)=d\left(f(x), f\left(x^{\prime}\right)\right) \leq$ $r d\left(x, x^{\prime}\right)$ and so $x=x^{\prime}$.

The contraction mapping theorem has an extremely useful generalization. In order to get a unique fixed point, it suffices to have some power of $f$ a contraction map.

Theorem 10.1.43 Let $f:(X, d) \rightarrow(X, d)$ have the property that for some $n \in \mathbb{N}, f^{n}$ is a contraction map and let $(X, d)$ be a complete metric space. Then there is a unique fixed point for $f$. As in the earlier theorem the sequence of iterates $\left\{f^{n}\left(x_{0}\right)\right\}_{n=1}^{\infty}$ also converges to the fixed point.

Proof: From Theorem 10.1.41 there is a unique fixed point for $f^{n}$. Thus

$$
f^{n}(x)=x
$$

Then

$$
f^{n}(f(x))=f^{n+1}(x)=f(x)
$$

By uniqueness, $f(x)=x$.
Now consider the sequence of iterates. Suppose it fails to converge to $x$. Then there is $\varepsilon>0$ and a subsequence $n_{k}$ such that

$$
d\left(f^{n_{k}}\left(x_{0}\right), x\right) \geq \varepsilon
$$

Now $n_{k}=p_{k} n+r_{k}$ where $r_{k}$ is one of the numbers $\{0,1,2, \cdots, n-1\}$. It follows that there exists one of these numbers which is repeated infinitely often. Call it $r$ and let the further subsequence continue to be denoted as $n_{k}$. Thus

$$
d\left(f^{p_{k} n+r}\left(x_{0}\right), x\right) \geq \varepsilon
$$

In other words,

$$
d\left(f^{p_{k} n}\left(f^{r}\left(x_{0}\right)\right), x\right) \geq \varepsilon
$$

However, from Theorem 10.1.41, as $k \rightarrow \infty, f^{p_{k} n}\left(f^{r}\left(x_{0}\right)\right) \rightarrow x$ which contradicts the above inequality. Hence the sequence of iterates converges to $x$, as it did for $f$ a contraction map.

Now with the above material on analysis, it is time to begin using the ideas from linear algebra in this special case where the field of scalars is $\mathbb{R}$ or $\mathbb{C}$.

### 10.1.8 Convergence Of Functions

Next is to consider the meaning of convergence of sequences of functions. There are two main ways of convergence of interest here, pointwise and uniform convergence.

Definition 10.1.44 Let $f_{n}: X \rightarrow Y$ where $(X, d),(Y, \rho)$ are two metric spaces. Then $\left\{f_{n}\right\}$ is said to converge pointwise to a function $f: X \rightarrow Y$ if for every $x \in X$,

$$
\lim _{n \rightarrow \infty} f_{n}(x)=f(x)
$$

$\left\{f_{n}\right\}$ is said to converge uniformly iffor all $\varepsilon>0$, there exists $N$ such that if $n \geq N$, then

$$
\sup _{x \in X} \rho\left(f_{n}(x), f(x)\right)<\varepsilon
$$

Here is a well known example illustrating the difference between pointwise and uniform convergence.
Example 10.1.45 Let $f_{n}(x)=x^{n}$ on the metric space $[0,1]$. Then this function converges pointwise to

$$
f(x)=\left\{\begin{array}{l}
0 \text { on }[0,1) \\
1 \text { at } 1
\end{array}\right.
$$

but it does not converge uniformly on this interval to $f$.
Note how the target function $f$ in the above example is not continuous even though each function in the sequence is. The nice thing about uniform convergence is that it takes continuity of the functions in the sequence and imparts it to the target function. It does this for both continuity at a single point and uniform continuity. Thus uniform convergence is a very superior thing.

Theorem 10.1.46 Let $f_{n}: X \rightarrow Y$ where $(X, d),(Y, \rho)$ are two metric spaces and suppose each $f_{n}$ is continuous at $x \in X$ and also that $f_{n}$ converges uniformly to $f$ on $X$. Then $f$ is also continuous at $x$. In addition to this, if each $f_{n}$ is uniformly continuous on $X$, then the same is true for $f$.

Proof: Let $\varepsilon>0$ be given. Then

$$
\rho(f(x), f(\hat{x})) \leq \rho\left(f(x), f_{n}(x)\right)+\rho\left(f_{n}(x), f_{n}(\hat{x})\right)+\rho\left(f_{n}(\hat{x}), f(\hat{x})\right)
$$

By uniform convergence, there exists $N$ such that $\rho\left(f(x), f_{n}(x)\right), \rho\left(f_{n}(\hat{x}), f(\hat{x})\right)$ are each less than $\varepsilon / 3$ provided $n \geq N$. Thus picking such an $n$,

$$
\rho(f(x), f(\hat{x})) \leq \frac{2 \varepsilon}{3}+\rho\left(f_{n}(x), f_{n}(\hat{x})\right)
$$

Now from the continuity of $f_{n}$, there exists $\delta>0$ such that if $d(x, \hat{x})<\delta$, then

$$
\rho\left(f_{n}(x), f_{n}(\hat{x})\right)<\varepsilon / 3 .
$$

Hence, if $d(x, \hat{x})<\delta$, then

$$
\rho(f(x), f(\hat{x})) \leq \frac{2 \varepsilon}{3}+\rho\left(f_{n}(x), f_{n}(\hat{x})\right)<\frac{2 \varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon
$$

Hence, $f$ is continuous at $x$.
Next consider uniform continuity. It follows from the uniform convergence that if $x, \hat{x}$ are any two points of $X$, then if $n \geq N$, then, picking such an $n$,

$$
\rho(f(x), f(\hat{x})) \leq \frac{2 \varepsilon}{3}+\rho\left(f_{n}(x), f_{n}(\hat{x})\right)
$$

By uniform continuity of $f_{n}$ there exists $\delta$ such that if $d(x, \hat{x})<\delta$, then the term on the right in the above is less than $\varepsilon / 3$. Hence if $d(x, \hat{x})<\delta$, then $\rho(f(x), f(\hat{x}))<\varepsilon$ and so $f$ is uniformly continuous as claimed.

### 10.2 Connected Sets

This has absolutely nothing to do with linear algebra but is here to provide convenient results to be used later when linear algebra will occur as part of some topics in analysis.

Stated informally, connected sets are those which are in one piece. In order to define what is meant by this, I will first consider what it means for a set to not be in one piece. This is called separated. Connected sets are defined in terms of not being separated. This is why theorems about connected sets sometimes seem a little tricky.

Definition 10.2.1 A set, $S$ in a metric space, is separated if there exist sets $A, B$ such that

$$
S=A \cup B, A, B \neq \emptyset, \text { and } \bar{A} \cap B=\bar{B} \cap A=\emptyset
$$

In this case, the sets $A$ and $B$ are said to separate $S$. A set is connected if it is not separated. Remember $\bar{A}$ denotes the closure of the set $A$.

Note that the concept of connected sets is defined in terms of what it is not. This makes it somewhat difficult to understand. One of the most important theorems about connected sets is the following.

Theorem 10.2.2 Suppose $\mathscr{U}$ is a set of connected sets and that there exists a point $p$ which is in all of these connected sets. Then $K \equiv \cup \mathscr{U}$ is connected.

Proof: Suppose $K=A \cup B$ where $\bar{A} \cap B=\bar{B} \cap A=\emptyset, A \neq \emptyset, B \neq \emptyset$. Let $U \in \mathscr{U}$. Then $U=(U \cap A) \cup(U \cap B)$ and this would separate $U$ if both sets in the union are nonempty since the limit points of $U \cap B$ are contained in the limit points of $B$. It follows that every set of $\mathscr{U}$ is contained in one of $A$ or $B$. Suppose then that some $U \subseteq A$. Then all $U \in \mathscr{U}$ must be contained in $A$ because if one is contained in $B$, this would violate the assumption that they all have a point $p$ in common. Thus $K$ is connected after all because this requires $B=\emptyset$. Alternatively, $p$ is in one of these sets. Say $p \in A$. Then by the above argument every $U$ must be in $A$ because if not, the above would be a separation of $U$. Thus $B=\emptyset$.

The intersection of connected sets is not necessarily connected as is shown by the following picture.


Theorem 10.2.3 Let $f: X \rightarrow Y$ be continuous where $Y$ is a metric space and $X$ is connected. Then $\boldsymbol{f}(X)$ is also connected.

Proof: To do this you show $\boldsymbol{f}(X)$ is not separated. Suppose to the contrary that $\boldsymbol{f}(X)=$ $A \cup B$ where $A$ and $B$ separate $f(X)$. Then consider the sets $f^{-1}(A)$ and $f^{-1}(B)$. If $\boldsymbol{z}$ $\in \boldsymbol{f}^{-1}(B)$, then $\boldsymbol{f}(\boldsymbol{z}) \in B$ and so $\boldsymbol{f}(\boldsymbol{z})$ is not a limit point of $A$. Therefore, there exists an open set, $U$ containing $f(\boldsymbol{z})$ such that $U \cap A=\emptyset$. But then, the continuity of $\boldsymbol{f}$ and Theorem
10.1.27 implies that $\boldsymbol{f}^{-1}(U)$ is an open set containing $\boldsymbol{z}$ such that $\boldsymbol{f}^{-1}(U) \cap \boldsymbol{f}^{-1}(A)=\emptyset$. Therefore, $\boldsymbol{f}^{-1}(B)$ contains no limit points of $\boldsymbol{f}^{-1}(A)$. Similar reasoning implies $\boldsymbol{f}^{-1}(A)$ contains no limit points of $\boldsymbol{f}^{-1}(B)$. It follows that $X$ is separated by $\boldsymbol{f}^{-1}(A)$ and $\boldsymbol{f}^{-1}(B)$, contradicting the assumption that $X$ was connected.

An arbitrary set can be written as a union of maximal connected sets called connected components. This is the concept of the next definition.

Definition 10.2.4 Let $S$ be a set and let $\boldsymbol{p} \in S$. Denote by $C_{\boldsymbol{p}}$ the union of all connected subsets of $S$ which contain $\boldsymbol{p}$. This is called the connected component determined by $\boldsymbol{p}$.

Theorem 10.2.5 Let $C_{p}$ be a connected component of a set $S$ in a metric space. Then $C_{p}$ is a connected set and if $C_{p} \cap C_{q} \neq \emptyset$, then $C_{p}=C_{q}$.

Proof: Let $\mathscr{C}$ denote the connected subsets of $S$ which contain $\boldsymbol{p}$. By Theorem 10.2.2, $\cup \mathscr{C}=C_{\boldsymbol{p}}$ is connected. If $\boldsymbol{x} \in C_{\boldsymbol{p}} \cap C_{\boldsymbol{q}}$, then from Theorem 10.2.2, $C_{\boldsymbol{p}} \supseteq C_{\boldsymbol{p}} \cup C_{\boldsymbol{q}}$ and so $C_{\boldsymbol{p}} \supseteq C_{\boldsymbol{q}}$. The inclusion goes the other way by the same reason.

This shows the connected components of a set are equivalence classes and partition the set.

A set, $I$ is an interval in $\mathbb{R}$ if and only if whenever $x, y \in I$ then $(x, y) \subseteq I$. The following theorem is about the connected sets in $\mathbb{R}$.

Theorem 10.2.6 A set $C$ in $\mathbb{R}$ is connected if and only if $C$ is an interval.
Proof: Let $C$ be connected. If $C$ consists of a single point, $p$, there is nothing to prove. The interval is just $[p, p]$. Suppose $p<q$ and $p, q \in C$. You need to show $(p, q) \subseteq C$. If

$$
x \in(p, q) \backslash C
$$

let $C \cap(-\infty, x) \equiv A$, and $C \cap(x, \infty) \equiv B$. Then $C=A \cup B$ and the sets $A$ and $B$ separate $C$ contrary to the assumption that $C$ is connected.

Conversely, let $I$ be an interval. Suppose $I$ is separated by $A$ and $B$. Pick $x \in A$ and $y \in B$. Suppose without loss of generality that $x<y$. Now define the set,

$$
S \equiv\{t \in[x, y]:[x, t] \subseteq A\}
$$

and let $l$ be the least upper bound of $S$. Then $l \in \bar{A}$ so $l \notin B$ which implies $l \in A$. But if $l \notin \bar{B}$, then for some $\delta>0,(l, l+\delta) \cap B=\emptyset$ contradicting the definition of $l$ as an upper bound for $S$. Therefore, $l \in \bar{B}$ which implies $l \notin A$ after all, a contradiction. It follows $I$ must be connected.

This yields a generalization of the intermediate value theorem from one variable calculus.

Corollary 10.2.7 Let $E$ be a connected set in a metric space and suppose $f: E \rightarrow \mathbb{R}$ and that $y \in\left(f\left(e_{1}\right), f\left(e_{2}\right)\right)$ where $e_{i} \in E$. Then there exists $e \in E$ such that $f(e)=y$.

Proof: From Theorem 10.2.3, $f(E)$ is a connected subset of $\mathbb{R}$. By Theorem 10.2.6 $f(E)$ must be an interval. In particular, it must contain $y$. This proves the corollary.

The following theorem is a very useful description of the open sets in $\mathbb{R}$.
Theorem 10.2.8 Let $U$ be an open set in $\mathbb{R}$. Then there exist countably many disjoint open sets $\left\{\left(a_{i}, b_{i}\right)\right\}_{i=1}^{\infty}$ such that $U=\cup_{i=1}^{\infty}\left(a_{i}, b_{i}\right)$.

Proof: Let $p \in U$ and let $z \in C_{p}$, the connected component determined by $p$. Since $U$ is open, there exists, $\delta>0$ such that $(z-\delta, z+\delta) \subseteq U$. It follows from Theorem 10.2.2 that

$$
(z-\delta, z+\delta) \subseteq C_{p}
$$

This shows $C_{p}$ is open. By Theorem 10.2.6, this shows $C_{p}$ is an open interval, $(a, b)$ where $a, b \in[-\infty, \infty]$. There are therefore at most countably many of these connected components because each must contain a rational number and the rational numbers are countable. Denote by $\left\{\left(a_{i}, b_{i}\right)\right\}_{i=1}^{\infty}$ the set of these connected components.

Definition 10.2.9 A set $E$ in a metric space is arcwise connected if for any two points, $\boldsymbol{p}, \boldsymbol{q} \in E$, there exists a closed interval, $[a, b]$ and a continuous function, $\gamma:[a, b] \rightarrow E$ such that $\gamma(a)=p$ and $\gamma(b)=q$.

An example of an arcwise connected metric space would be any subset of $\mathbb{R}^{n}$ which is the continuous image of an interval. Arcwise connected is not the same as connected. A well known example is the following.

$$
\begin{equation*}
\left\{\left(x, \sin \frac{1}{x}\right): x \in(0,1]\right\} \cup\{(0, y): y \in[-1,1]\} \tag{10.2}
\end{equation*}
$$

You can verify that this set of points in the normed vector space $\mathbb{R}^{2}$ is not arcwise connected but is connected.

### 10.3 Subspaces Spans And Bases

As shown earlier, $\mathbb{F}^{n}$ is an example of a vector space with field of scalars $\mathbb{F}$. Here is a short review of the major exchange theorem. Here and elsewhere, when it is desired to emphasize that certain things are vectors, bold face will be used. However, sometimes the context makes this sufficiently clear and bold face is not used.

Theorem 10.3.1 If $\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{r}\right) \subseteq \operatorname{span}\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{s}\right) \equiv V$ and $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{r}\right\}$ are linearly independent, then $r \leq s$.

Proof: Suppose $r>s$. Let $E_{p}$ denote a finite list of vectors of $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{s}\right\}$ and let $\left|E_{p}\right|$ denote the number of vectors in the list. Let $F_{p}$ denote the first $p$ vectors in $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{r}\right\}$. In case $p=0, F_{p}$ will denote the empty set. For $0 \leq p \leq s$, let $E_{p}$ have the property

$$
\operatorname{span}\left(F_{p}, E_{p}\right)=V
$$

and $\left|E_{p}\right|$ is as small as possible for this to happen. I claim $\left|E_{p}\right| \leq s-p$ if $E_{p}$ is nonempty.
Here is why. For $p=0$, it is obvious. Suppose true for some $p<s$. Then

$$
\boldsymbol{u}_{p+1} \in \operatorname{span}\left(F_{p}, E_{p}\right)
$$

and so there are constants, $c_{1}, \cdots, c_{p}$ and $d_{1}, \cdots, d_{m}$ where $m \leq s-p$ such that

$$
\boldsymbol{u}_{p+1}=\sum_{i=1}^{p} c_{i} \boldsymbol{u}_{i}+\sum_{j=1}^{m} d_{i} \boldsymbol{z}_{j}
$$

for $\left\{\boldsymbol{z}_{1}, \cdots, \boldsymbol{z}_{m}\right\} \subseteq\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{s}\right\}$.Then not all the $d_{i}$ can equal zero because this would violate the linear independence of the $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{r}\right\}$. Therefore, you can solve for one of
the $\boldsymbol{z}_{k}$ as a linear combination of $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{p+1}\right\}$ and the other $\boldsymbol{z}_{j}$. Thus you can change $F_{p}$ to $F_{p+1}$ and include one fewer vector in $E_{p}$. Thus $\left|E_{p+1}\right| \leq m-1 \leq s-p-1$. This proves the claim.

Therefore, $E_{s}$ is empty and $\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{s}\right)=V$. However, this gives a contradiction because it would require $\boldsymbol{u}_{s+1} \in \operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{s}\right)$ which violates the linear independence of these vectors.

Also recall the following.
Definition 10.3.2 $A$ finite set of vectors, $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{r}\right\}$ is a basis for a vector space $V$ if

$$
\operatorname{span}\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{r}\right)=V
$$

and $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{r}\right\}$ is linearly independent. Thus if $\boldsymbol{v} \in V$ there exist unique scalars, $v_{1}, \cdots, v_{r}$ such that $\boldsymbol{v}=\sum_{i=1}^{r} v_{i} \boldsymbol{x}_{i}$. These scalars are called the components of $\boldsymbol{v}$ with respect to the basis $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{r}\right\}$.

Corollary 10.3.3 Let $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{r}\right\}$ and $\left\{\boldsymbol{y}_{1}, \cdots, \boldsymbol{y}_{s}\right\}$ be two bases ${ }^{1}$ of $\mathbb{F}^{n}$. Then $r=s=n$.
Lemma 10.3.4 Let $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}\right\}$ be a set of vectors. Then $V \equiv \operatorname{span}\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}\right)$ is a subspace.

Definition 10.3.5 Let $V$ be a vector space. Then $\operatorname{dim}(V)$ read as the dimension of $V$ is the number of vectors in a basis.

Of course you should wonder right now whether an arbitrary subspace of a finite dimensional vector space even has a basis. In fact it does and this is in the next theorem. First, here is an interesting lemma which was also presented earlier.

Lemma 10.3.6 Suppose $\boldsymbol{v} \notin \operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}\right)$ and $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}\right\}$ is linearly independent. Then $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}, \boldsymbol{v}\right\}$ is also linearly independent.

Recall that this implies the following theorems also presented earlier.
Theorem 10.3.7 Let $V$ be a nonzero subspace of $Y$ a finite dimensional vector space having dimension $n$. Then $V$ has a basis.

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 10.3.8 Let $V$ be a subspace of $Y$, a finite dimensional vector space of dimension $n$ and let $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}\right\}$ be a linearly independent set of vectors in $V$. Then either it is a basis for $V$ or there exist vectors, $\boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_{s}$ such that $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{r}, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_{s}\right\}$ is a basis for $V$.

Theorem 10.3.9 Let $V$ be a subspace of $Y$, a finite dimensional vector space of dimension $n$ and suppose $\operatorname{span}\left(\boldsymbol{u}_{1} \cdots, \boldsymbol{u}_{p}\right)=V$ where the $\boldsymbol{u}_{i}$ are nonzero vectors. Then there exist vectors, $\left\{\boldsymbol{v}_{1} \cdots, \boldsymbol{v}_{r}\right\}$ such that $\left\{\boldsymbol{v}_{1} \cdots, \boldsymbol{v}_{r}\right\} \subseteq\left\{\boldsymbol{u}_{1} \cdots, \boldsymbol{u}_{p}\right\}$ and $\left\{\boldsymbol{v}_{1} \cdots, \boldsymbol{v}_{r}\right\}$ is a basis for $V$.

[^7]
### 10.4 Inner Product and Normed Linear Spaces

### 10.4.1 The Inner Product in $\mathbb{F}^{n}$

To do calculus, you must understand what you mean by distance. For functions of one variable, the distance was provided by the absolute value of the difference of two numbers. This must be generalized to $\mathbb{F}^{n}$ and to more general situations. This is the most familiar setting for elementary courses. We call it the dot product in calculus and physics but it is a case of something which also works in $\mathbb{C}^{n}$.

Definition 10.4.1 Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^{n}$. Thus $\boldsymbol{x}=\left(x_{1}, \cdots, x_{n}\right)$ where each $x_{k} \in \mathbb{F}$ and a similar formula holding for $\boldsymbol{y}$. Then the inner product of these two vectors is defined to be

$$
\boldsymbol{x} \cdot \boldsymbol{y} \equiv(\boldsymbol{x}, \boldsymbol{y}) \equiv \sum_{j} x_{j} \overline{y_{j}} \equiv x_{1} \overline{y_{1}}+\cdots+x_{n} \overline{y_{n}} .
$$

This is also often denoted by $(\boldsymbol{x}, \boldsymbol{y})$ or as $\langle\boldsymbol{x}, \boldsymbol{y}\rangle$ and is called an inner product. I will use either notation.

Notice how you put the conjugate on the entries of the vector, $\boldsymbol{y}$. It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way ${ }^{2}$. The reason for this is that when you take the inner product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of $\boldsymbol{y}$ in the above definition assures this will take place. Thus

$$
(\boldsymbol{x}, \boldsymbol{x})=\sum_{j} x_{j} \overline{x_{j}}=\sum_{j}\left|x_{j}\right|^{2} \geq 0 .
$$

If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example,

$$
(1+i)^{2}+2^{2}=4+2 i
$$

and this is not a positive number.
The following properties of the inner product follow immediately from the definition and you should verify each of them.

Properties of the inner product:

1. $(\boldsymbol{u}, \boldsymbol{v})=\overline{(\boldsymbol{v}, \boldsymbol{u})}$
2. If $a, b$ are numbers and $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}$ are vectors then $((a \boldsymbol{u}+b \boldsymbol{v}), \boldsymbol{z})=a(\boldsymbol{u}, \boldsymbol{z})+b(\boldsymbol{v}, \boldsymbol{z})$.
3. $(\boldsymbol{u}, \boldsymbol{u}) \geq 0$ and it equals 0 if and only if $\boldsymbol{u}=\mathbf{0}$.

Note this implies $(\boldsymbol{x}, \alpha \boldsymbol{y})=\bar{\alpha}(\boldsymbol{x}, \boldsymbol{y})$ because

$$
(x, \alpha y)=\overline{(\alpha y, x)}=\overline{\alpha(y, x)}=\bar{\alpha}(x, y)
$$

The norm is defined as follows.
Definition 10.4.2 For $\boldsymbol{x} \in \mathbb{F}^{n},|\boldsymbol{x}| \equiv\left(\sum_{k=1}^{n}\left|x_{k}\right|^{2}\right)^{1 / 2}=(\boldsymbol{x}, \boldsymbol{x})^{1 / 2}$

[^8]
### 10.4.2 General Inner Product Spaces

Any time you have a vector space which possesses an inner product, something satisfying the properties $1-3$ above, it is called an inner product space. As usual, $\mathbb{F}$ will mean the field of scalars, either $\mathbb{C}$ or $\mathbb{R}$.

Here is a fundamental inequality called the Cauchy Schwarz inequality which holds in any inner product space. First here is a simple lemma.

Lemma 10.4.3 If $z \in \mathbb{F}$ there exists $\theta \in \mathbb{F}$ such that $\theta z=|z|$ and $|\theta|=1$.
Proof: Let $\theta=1$ if $z=0$ and otherwise, let $\theta=\frac{\bar{z}}{|z|}$. Recall that for $z=x+i y, \bar{z}=x-i y$ and $\bar{z} z=|z|^{2}$. In case $z$ is real, there is no change in the above.

Theorem 10.4.4 (Cauchy Schwarz)Let $H$ be an inner product space. The following inequality holds for $\boldsymbol{x}$ and $\boldsymbol{y} \in H$.

$$
\begin{equation*}
|(\boldsymbol{x}, \boldsymbol{y})| \leq(\boldsymbol{x}, \boldsymbol{x})^{1 / 2}(\boldsymbol{y}, \boldsymbol{y})^{1 / 2} \tag{10.3}
\end{equation*}
$$

Equality holds in this inequality if and only if one vector is a multiple of the other.
Proof: Let $\theta \in \mathbb{F}$ such that $|\theta|=1$ and $\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{y})=|(\boldsymbol{x}, \boldsymbol{y})|$. Consider

$$
p(t) \equiv(\boldsymbol{x}+\bar{\theta} t \boldsymbol{y}, \boldsymbol{x}+t \overline{\boldsymbol{\theta}} \boldsymbol{y})
$$

where $t \in \mathbb{R}$. Then from the above list of properties of the inner product,

$$
\begin{align*}
0 & \leq p(t)=(\boldsymbol{x}, \boldsymbol{x})+t \boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{y})+t \overline{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x})+t^{2}(\boldsymbol{y}, \boldsymbol{y}) \\
& =(\boldsymbol{x}, \boldsymbol{x})+t \boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{y})+t \overline{\boldsymbol{\theta}} \overline{(\boldsymbol{x}, \boldsymbol{y})}+t^{2}(\boldsymbol{y}, \boldsymbol{y}) \\
& =(\boldsymbol{x}, \boldsymbol{x})+2 t \operatorname{Re}(\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{y}))+t^{2}(\boldsymbol{y}, \boldsymbol{y}) \\
& =(\boldsymbol{x}, \boldsymbol{x})+2 t|(\boldsymbol{x}, \boldsymbol{y})|+t^{2}(\boldsymbol{y}, \boldsymbol{y}) \tag{10.4}
\end{align*}
$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(\boldsymbol{y}, \boldsymbol{y})=0$ it must be the case that $|(\boldsymbol{x}, \boldsymbol{y})|=0$ also since otherwise the above inequality would be violated. Therefore, in this case,

$$
|(\boldsymbol{x}, \boldsymbol{y})| \leq(\boldsymbol{x}, \boldsymbol{x})^{1 / 2}(\boldsymbol{y}, \boldsymbol{y})^{1 / 2}
$$

On the other hand, if $(\boldsymbol{y}, \boldsymbol{y}) \neq 0$, then $p(t) \geq 0$ for all $t$ means the graph of $y=p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the $t$ axis or it has no real zeros. From the quadratic formula this happens exactly when

$$
4|(\boldsymbol{x}, \boldsymbol{y})|^{2}-4(\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y}) \leq 0
$$

which is equivalent to 10.3 .
It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 10.3. Conversely, suppose equality does hold. Then this is equivalent to saying $4|(\boldsymbol{x}, \boldsymbol{y})|^{2}-4(\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y})=0$ and so from the quadratic formula, there exists one real zero to $p(t)=0$. Call it $t_{0}$. Then $p\left(t_{0}\right) \equiv\left(\boldsymbol{x}+\overline{\boldsymbol{\theta}} t_{0} \boldsymbol{y}, \boldsymbol{x}+t_{0} \overline{\boldsymbol{\theta}} \boldsymbol{y}\right)=|\boldsymbol{x}+\overline{\boldsymbol{\theta}} t \boldsymbol{y}|^{2}=0$ and so $\boldsymbol{x}=-\bar{\theta} t_{0} \boldsymbol{y}$. This proves the theorem.

Note that in establishing the inequality, I only used part of the above properties of the inner product. It was not necessary to use the one which says that if $(\boldsymbol{x}, \boldsymbol{x})=0$ then $\boldsymbol{x}=\mathbf{0}$.

Now the length of a vector can be defined.

Definition 10.4.5 Let $\boldsymbol{z} \in H$. Then $|\boldsymbol{z}| \equiv(\boldsymbol{z}, \boldsymbol{z})^{1 / 2}$.
Theorem 10.4.6 For length defined in Definition 10.4.5, the following hold.

$$
\begin{gather*}
|\boldsymbol{z}| \geq 0 \text { and }|\boldsymbol{z}|=0 \text { if and only if } \boldsymbol{z}=\mathbf{0}  \tag{10.5}\\
\text { If } \alpha \text { is a scalar, }|\alpha \boldsymbol{z}|=|\alpha||\boldsymbol{z}|  \tag{10.6}\\
|\boldsymbol{z}+\boldsymbol{w}| \leq|\boldsymbol{z}|+|\boldsymbol{w}| . \tag{10.7}
\end{gather*}
$$

Proof: The first two claims are left as exercises. To establish the third,

$$
\begin{aligned}
&|\boldsymbol{z + \boldsymbol { w }}|^{2} \equiv(\boldsymbol{z}+\boldsymbol{w}, \boldsymbol{z}+\boldsymbol{w})=(\boldsymbol{z}, \boldsymbol{z})+(\boldsymbol{w}, \boldsymbol{w})+(\boldsymbol{w}, \boldsymbol{z})+(\boldsymbol{z}, \boldsymbol{w}) \\
&=|\boldsymbol{z}|^{2}+|\boldsymbol{w}|^{2}+2 \operatorname{Re}(\boldsymbol{w}, \boldsymbol{z}) \leq|\boldsymbol{z}|^{2}+|\boldsymbol{w}|^{2}+2|(\boldsymbol{w}, \boldsymbol{z})| \\
& \leq|\boldsymbol{z}|^{2}+|\boldsymbol{w}|^{2}+2|\boldsymbol{w}||\boldsymbol{z}|=(|\boldsymbol{z}|+|\boldsymbol{w}|)^{2}
\end{aligned}
$$

One defines the distance between two vectors $\boldsymbol{x}, \boldsymbol{y}$ in an inner product space as $|\boldsymbol{x}-\boldsymbol{y}|$. This produces a metric in the obvious way: $d(\boldsymbol{x}, \boldsymbol{y}) \equiv|\boldsymbol{x}-\boldsymbol{y}|$.

Not surprisingly we have the following theorem in which $\mathbb{F}$ will be either $\mathbb{R}$ or $\mathbb{C}$.
Theorem 10.4.7 $\mathbb{F}^{n}$ is complete. Also, if $K$ is a nonempty closed and bounded subset of $\mathbb{F}^{n}$, then $K$ is compact. Also, if $f: K \rightarrow \mathbb{R}$, it achieves its maximum and minimum on $K$.

Proof: Recall Example 10.1.23 which established completeness of $\mathbb{F}^{n}$ with the funny norm $\|\boldsymbol{x}\|_{\infty} \equiv \max \left\{\left|x_{i}\right|, i=1,2, \cdots, n\right\}$. However, $\frac{1}{\sqrt{n}}|\boldsymbol{x}| \leq\left|\left|\boldsymbol{x} \|_{\infty} \leq|\boldsymbol{x}|\right.\right.$ and so the Cauchy sequences and limits are exactly the same for the two norms. Thus $\mathbb{F}^{n}$ is complete where the norm is the one just discussed.

Now suppose $K$ is closed and bounded. By the estimate on the norms just given, it is closed and bounded with respect to $\|\cdot\|_{\infty}$ also because a point is a limit point with respect to one norm if and only if it is a limit point with respect to the other. Now if $B(\mathbf{0}, r) \supseteq K$, then $B_{\infty}(\mathbf{0}, r) \supseteq K$ also where this symbol denotes the ball taken with respect to $\|\cdot\|_{\infty}$ rather than $|\cdot|$. Hence $K \subseteq \prod_{j=1}^{n}([-r, r]+[-i r, i r])$. It suffices to verify sequential compactness thanks to Theorem 10.1.38. Letting $\left\{\boldsymbol{x}^{n}\right\} \subseteq K$, it follows that $\operatorname{Re} x_{i}^{n}$ is in $[-r, r]$ and $\operatorname{Im} x_{i}^{n}$ is in $[-r, r]$ and so, taking $2 n$ subsequences, there exists a subsequence still denoted with $n$ such that $\lim _{n \rightarrow \infty} \operatorname{Re} x_{i}^{n}=a_{i} \in[-r, r], \lim _{n \rightarrow \infty} \operatorname{Im} x_{i}^{n}=b_{i}$ for each $i$. Hence $\boldsymbol{x}^{n} \rightarrow \boldsymbol{a}+i \boldsymbol{b} \equiv \boldsymbol{x}$. Now, since $K$ is closed, it follows that $\boldsymbol{x} \in K$ and this shows sequential compactness which is equivalent to compactness.

The last claim is as follows. Let $M \equiv \sup \{f(\boldsymbol{x}): \boldsymbol{x} \in K\}$ and let $\boldsymbol{x}_{n}$ be a maximizing sequence so that $M=\lim _{n \rightarrow \infty} f\left(\boldsymbol{x}_{n}\right)$. By compactness, there is a subsequence $\boldsymbol{x}_{n_{k}} \rightarrow \boldsymbol{x} \in K$. Then by continuity, $M=\lim _{k \rightarrow \infty} f\left(\boldsymbol{x}_{n_{k}}\right)=f(\boldsymbol{x})$. The existence of the minimum is similar.

### 10.4.3 Normed Vector Spaces

The best sort of a norm is one which comes from an inner product, because these norms preserve familiar geometrical ideas. However, any vector space $V$ which has a function $\|\cdot\|$ which maps $V$ to $[0, \infty)$ is called a normed vector space if $\|\cdot\|$ satisfies $10.8-10.10$. That is

$$
\begin{equation*}
\|\boldsymbol{z}\| \geq 0 \text { and }\|\boldsymbol{z}\|=0 \text { if and only if } \boldsymbol{z}=\mathbf{0} \tag{10.8}
\end{equation*}
$$

$$
\begin{align*}
& \text { If } \alpha \text { is a scalar, }\|\alpha \boldsymbol{z}\|=|\alpha|\|\boldsymbol{z}\|  \tag{10.9}\\
& \quad\|\boldsymbol{z}+\boldsymbol{w}\| \leq\|\boldsymbol{z}\|+\|\boldsymbol{w}\| \tag{10.10}
\end{align*}
$$

The last inequality above is called the triangle inequality. Another version of this is

$$
\begin{equation*}
\|\boldsymbol{z}\|-\|\boldsymbol{w}\|\|\leq\| \boldsymbol{z}-\boldsymbol{w} \| \tag{10.11}
\end{equation*}
$$

Note that this shows that $\boldsymbol{x} \rightarrow\|\boldsymbol{x}\|$ is a continuous function. Thus

$$
B(\boldsymbol{z}, r) \equiv\{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{z}\|<r\}
$$

is an open set and

$$
D(\boldsymbol{z}, r) \equiv\{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{z}\| \leq r\}
$$

is a closed set.
To see that 10.11 holds, note

$$
\|\boldsymbol{z}\|=\|\boldsymbol{z}-\boldsymbol{w}+\boldsymbol{w}\| \leq\|\boldsymbol{z}-\boldsymbol{w}\|+\|\boldsymbol{w}\|
$$

which implies

$$
\|\boldsymbol{z}\|-\|\boldsymbol{w}\| \leq\|\boldsymbol{z}-\boldsymbol{w}\|
$$

and now switching $z$ and $\boldsymbol{w}$, yields

$$
\|\boldsymbol{w}\|-\|\boldsymbol{z}\| \leq\|\boldsymbol{z}-\boldsymbol{w}\|
$$

which implies 10.11 .
The distance between $\boldsymbol{x}, \boldsymbol{y}$ is given by

$$
\|\boldsymbol{x}-\boldsymbol{y}\|
$$

This distance satisfies

$$
\begin{gathered}
\|\boldsymbol{x}-\boldsymbol{y}\|=\|\boldsymbol{y}-\boldsymbol{x}\| \\
\|\boldsymbol{x}-\boldsymbol{y}\| \geq 0 \text { and is } \mathbf{0} \text { if and only if } \boldsymbol{x}=\boldsymbol{y} \\
\|\boldsymbol{x}-\boldsymbol{y}\| \leq\|\boldsymbol{x}-\boldsymbol{z}\|+\|\boldsymbol{z}-\boldsymbol{y}\|
\end{gathered}
$$

Thus this yields a metric space, but it has more because it also involves interaction with the algebra of the vector space.

### 10.5 Tietze Extension Theorem

This is an interesting theorem which holds in arbitrary normal topological spaces. In particular it holds in metric space and this is the context in which it will be discussed. First, review Lemma 10.1.14.

Lemma 10.5.1 Let $H, K$ be two nonempty disjoint closed subsets of $X$. Then there exists a continuous function, $g: X \rightarrow[-1 / 3,1 / 3]$ such that $g(H)=-1 / 3, g(K)=1 / 3, g(X) \subseteq$ $[-1 / 3,1 / 3]$.

Proof: Let $f(\boldsymbol{x}) \equiv \frac{\operatorname{dist}(\boldsymbol{x}, H)}{\operatorname{dist}(\boldsymbol{x}, H)+\operatorname{dist}(\boldsymbol{x}, K)}$. The denominator is never equal to zero because if $\operatorname{dist}(\boldsymbol{x}, H)=0$, then $\boldsymbol{x} \in H$ because $H$ is closed. (To see this, pick $\boldsymbol{h}_{k} \in B(\boldsymbol{x}, 1 / k) \cap H$. Then $\boldsymbol{h}_{k} \rightarrow \boldsymbol{x}$ and since $H$ is closed, $\boldsymbol{x} \in H$.) Similarly, if $\operatorname{dist}(\boldsymbol{x}, K)=0$, then $\boldsymbol{x} \in K$ and so the denominator is never zero as claimed. Hence $f$ is continuous and from its definition, $f=0$ on $H$ and $f=1$ on $K$. Now let $g(\boldsymbol{x}) \equiv \frac{2}{3}\left(f(\boldsymbol{x})-\frac{1}{2}\right)$. Then $g$ has the desired properties.

Definition 10.5.2 For $f: M \subseteq X \rightarrow \mathbb{R}$, let $\|f\|_{M} \equiv \sup \{|f(\boldsymbol{x})|: \boldsymbol{x} \in M\}$. This is just notation. I am not claiming this is a norm.

Lemma 10.5.3 Suppose $M$ is a closed set in $X$ and suppose $f: M \rightarrow[-1,1]$ is continuous at every point of $M$. Then there exists a function, $g$ which is defined and continuous on all of $X$ such that $\|f-g\|_{M} \leq \frac{2}{3}, g(X) \subseteq[-1 / 3,1 / 3]$. If $X$ is a normed vector space, and $f$ is odd, meaning that $M$ is symmetric ( $\boldsymbol{x} \in M$ if and only if $-\boldsymbol{x} \in M$ ) and $f(-\boldsymbol{x})=-f(\boldsymbol{x})$. Then we can assume $g$ is also odd.

Proof: Let $H=f^{-1}([-1,-1 / 3]), K=f^{-1}([1 / 3,1])$. Thus $H$ and $K$ are disjoint closed subsets of $M$. Suppose first $H, K$ are both nonempty. Then by Lemma 10.5.1 there exists $g$ such that $g$ is a continuous function defined on all of $X$ and $g(H)=-1 / 3, g(K)=1 / 3$, and $g(X) \subseteq[-1 / 3,1 / 3]$. It follows $\|f-g\|_{M}<2 / 3$. If $H=\emptyset$, then $f$ has all its values in $[-1 / 3,1]$ and so letting $g \equiv 1 / 3$, the desired condition is obtained. If $K=\emptyset$, let $g \equiv-1 / 3$. If both $H, K=\emptyset$, let $g=0$.

When $M$ is symmetric and $f$ is odd, $g(\boldsymbol{x}) \equiv \frac{1}{3} \frac{\operatorname{dist}(\boldsymbol{x}, H)-\operatorname{dist}(\boldsymbol{x}, K)}{\operatorname{dist}(\boldsymbol{x}, H)+\operatorname{dist}(\boldsymbol{x}, K)}$. When $\boldsymbol{x} \in H$ this gives $\frac{1}{3} \frac{-\operatorname{dist}(\boldsymbol{x}, K)}{\operatorname{dist}(\boldsymbol{x}, K)}=-\frac{1}{3}$. Then $\boldsymbol{x} \in K$, this gives $\frac{1}{3} \frac{\operatorname{dist}(\boldsymbol{x}, H)}{\operatorname{dist}(\boldsymbol{x}, H)}=\frac{1}{3}$. Also $g(H)=-1 / 3, f(H) \subseteq$ $[-1,-1 / 3]$ so for $\boldsymbol{x} \in H,|g(\boldsymbol{x})-f(\boldsymbol{x})| \leq \frac{2}{3}$. It is similar for $\boldsymbol{x} \in K$. If $\boldsymbol{x}$ is in neither $H$ nor $K$, then $g(\boldsymbol{x}) \in[-1 / 3,1 / 3]$ and so is $f(\boldsymbol{x})$. Thus $\|f-g\|_{M} \leq \frac{2}{3}$. Now by assumption, since $f$ is odd, $H=-K$. It is clear that $g$ is odd because

$$
\begin{aligned}
g(-\boldsymbol{x}) & =\frac{1}{3} \frac{\operatorname{dist}(-\boldsymbol{x}, H)-\operatorname{dist}(-\boldsymbol{x}, K)}{\operatorname{dist}(-\boldsymbol{x}, H)+\operatorname{dist}(-\boldsymbol{x}, K)}=\frac{1}{3} \frac{\operatorname{dist}(-\boldsymbol{x},-K)-\operatorname{dist}(-\boldsymbol{x},-H)}{\operatorname{dist}(-\boldsymbol{x},-K)+\operatorname{dist}(-\boldsymbol{x},-H)} \\
& =\frac{1}{3} \frac{\operatorname{dist}(\boldsymbol{x}, K)-\operatorname{dist}(\boldsymbol{x}, H)}{\operatorname{dist}(\boldsymbol{x}, K)+\operatorname{dist}(\boldsymbol{x}, H)}=-g(\boldsymbol{x}) .
\end{aligned}
$$

Lemma 10.5.4 Suppose $M$ is a closed set in $X$ and suppose $f: M \rightarrow[-1,1]$ is continuous at every point of $M$. Then there exists a function $g$ which is defined and continuous on all of $X$ such that $g=f$ on $M$ and $g$ has its values in $[-1,1]$. If $X$ is a normed linear space and $f$ is odd, then we can also assume $g$ is odd.

Proof: Using Lemma 10.5 .3 , let $g_{1}$ be such that $g_{1}(X) \subseteq[-1 / 3,1 / 3]$ and $\left\|f-g_{1}\right\|_{M} \leq$ $\frac{2}{3}$. Suppose $g_{1}, \cdots, g_{m}$ have been chosen such that $g_{j}(X) \subseteq[-1 / 3,1 / 3]$ and

$$
\begin{equation*}
\left\|f-\sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} g_{i}\right\|_{M}<\left(\frac{2}{3}\right)^{m} \tag{10.12}
\end{equation*}
$$

This has been done for $m=1$. Then $\left\|\left(\frac{3}{2}\right)^{m}\left(f-\sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} g_{i}\right)\right\|_{M} \leq 1$ and so

$$
\left(\frac{3}{2}\right)^{m}\left(f-\sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} g_{i}\right)
$$

can play the role of $f$ in the first step of the proof. Therefore, there exists $g_{m+1}$ defined and continuous on all of $X$ such that its values are in $[-1 / 3,1 / 3]$ and

$$
\left\|\left(\frac{3}{2}\right)^{m}\left(f-\sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} g_{i}\right)-g_{m+1}\right\|_{M} \leq \frac{2}{3}
$$

Hence

$$
\left\|\left(f-\sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} g_{i}\right)-\left(\frac{2}{3}\right)^{m} g_{m+1}\right\|_{M} \leq\left(\frac{2}{3}\right)^{m+1}
$$

It follows there exists a sequence, $\left\{g_{i}\right\}$ such that each has its values in $[-1 / 3,1 / 3]$ and for every $m 10.12$ holds. Then let $g(\boldsymbol{x}) \equiv \sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i-1} g_{i}(\boldsymbol{x})$. It follows

$$
|g(\boldsymbol{x})| \leq\left|\sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i-1} g_{i}(\boldsymbol{x})\right| \leq \sum_{i=1}^{m}\left(\frac{2}{3}\right)^{i-1} \frac{1}{3} \leq 1
$$

and $\left|\left(\frac{2}{3}\right)^{i-1} g_{i}(\boldsymbol{x})\right| \leq\left(\frac{2}{3}\right)^{i-1} \frac{1}{3}$ so the Weierstrass $M$ test applies and shows convergence is uniform. Therefore $g$ must be continuous by Theorem 10.1.46. The estimate 10.12 implies $f=g$ on $M$. The last claim follows because we can take each $g_{i}$ odd.

The following is the Tietze extension theorem.
Theorem 10.5.5 Let $M$ be a closed nonempty subset of a metric space $X$ and let $f: M \rightarrow$ $[a, b]$ be continuous at every point of $M$. Then there exists a function $g$ continuous on all of $X$ which coincides with $f$ on $M$ such that $g(X) \subseteq[a, b]$. If $[a, b]$ is centered on 0 , and if $X$ is a normed linear space and $f$ is odd, then we can obtain that $g$ is also odd.

Proof: Let $f_{1}(\boldsymbol{x})=1+\frac{2}{b-a}(f(\boldsymbol{x})-b)$. Then $f_{1}$ satisfies the conditions of Lemma 10.5.4 and so there exists $g_{1}: X \rightarrow[-1,1]$ such that $g$ is continuous on $X$ and equals $f_{1}$ on $M$. Let $g(\boldsymbol{x})=\left(g_{1}(\boldsymbol{x})-1\right)\left(\frac{b-a}{2}\right)+b$. This works. The last claim follows from the same arguments which gave Lemma 10.5.4 or the change of variables just given.

Corollary 10.5.6 Let $M$ be a closed nonempty subset of a metric space $X$ and let $f: M \rightarrow$ $[a, b]$ be continuous at every point of $M$. Also let $\|f-g\| \leq \varepsilon$. Then there exists continuous $\hat{f}$ extending $f$ with $\hat{f}(X) \subseteq[a, b]$ and $\hat{g}$ extending $g$ such that $\hat{g}(X) \subseteq[a-\varepsilon, b+\varepsilon]$. Also $\|\hat{f}-\hat{g}\| \leq \varepsilon$.

Proof: Let $\hat{f}$ be the extension of $f$ from the above theorem. Now let $F$ be the extension of $f-g$ with $\|F\| \leq \varepsilon$. Then let $\hat{g}=\hat{f}-F$. Then for $x \in M, \hat{g}(x)=f(x)-(f(x)-g(x))=$ $g(x)$. Thus it extends $g$ and clearly $\hat{g}(X) \subseteq[a-\varepsilon, b+\varepsilon]$.

### 10.5.1 The $p$ Norms

Examples of norms are the $p$ norms on $\mathbb{C}^{n}$. These do not come from an inner product but they are norms just the same.

Definition 10.5.7 Let $\boldsymbol{x} \in \mathbb{C}^{n}$. Then define for $p \geq 1,\|\boldsymbol{x}\|_{p} \equiv\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}$.
The following inequality is called Holder's inequality.

Proposition 10.5.8 For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{n}, \sum_{i=1}^{n}\left|x_{i}\right|\left|y_{i}\right| \leq\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}\left(\sum_{i=1}^{n}\left|y_{i}\right|^{p^{\prime}}\right)^{1 / p^{\prime}}$.
The proof will depend on the following lemma.
Lemma 10.5.9 If $a, b \geq 0$ and $p^{\prime}$ is defined by $\frac{1}{p}+\frac{1}{p^{\prime}}=1$, then $a b \leq \frac{a^{p}}{p}+\frac{b^{p^{\prime}}}{p^{\prime}}$.
Proof of the Proposition: If $\boldsymbol{x}$ or $\boldsymbol{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A=\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}$ and $B=\left(\sum_{i=1}^{n}\left|y_{i}\right|^{p^{\prime}}\right)^{1 / p^{\prime}}$. Then using Lemma 10.5.9,

$$
\begin{aligned}
& \sum_{i=1}^{n} \frac{\left|x_{i}\right|}{A} \frac{\left|y_{i}\right|}{B} \leq \sum_{i=1}^{n}\left[\frac{1}{p}\left(\frac{\left|x_{i}\right|}{A}\right)^{p}+\frac{1}{p^{\prime}}\left(\frac{\left|y_{i}\right|}{B}\right)^{p^{\prime}}\right] \\
= & \frac{1}{p} \frac{1}{A^{p}} \sum_{i=1}^{n}\left|x_{i}\right|^{p}+\frac{1}{p^{\prime}} \frac{1}{B^{p}} \sum_{i=1}^{n}\left|y_{i}\right|^{p^{\prime}}=\frac{1}{p}+\frac{1}{p^{\prime}}=1
\end{aligned}
$$

and so $\sum_{i=1}^{n}\left|x_{i}\right|\left|y_{i}\right| \leq A B=\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}\left(\sum_{i=1}^{n}\left|y_{i}\right|^{p^{\prime}}\right)^{1 / p^{\prime}}$
Theorem 10.5.10 The p norms do indeed satisfy the axioms of a norm.
Proof: It is obvious that $\|\cdot\|_{p}$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $\|\cdot\|$ in place of $\|\cdot\|_{p}$ in what follows. Note also that $\frac{p}{p^{\prime}}=p-1$. Then using the Holder inequality,

$$
\begin{aligned}
& \|\boldsymbol{x}+\boldsymbol{y}\|^{p}=\sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{p} \leq \sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{p-1}\left|x_{i}\right|+\sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{p-1}\left|y_{i}\right| \\
& =\sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{\frac{p}{p^{\prime}}}\left|x_{i}\right|+\sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{\frac{p}{p^{\prime}}}\left|y_{i}\right| \\
& \leq\left(\sum_{i=1}^{n}\left|x_{i}+y_{i}\right|^{p}\right)^{1 / p^{\prime}}\left[\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}+\left(\sum_{i=1}^{n}\left|y_{i}\right|^{p}\right)^{1 / p}\right] \\
& =\|\boldsymbol{x}+\boldsymbol{y}\|^{p / p^{\prime}}\left(\|\boldsymbol{x}\|_{p}+\|\boldsymbol{y}\|_{p}\right)
\end{aligned}
$$

so dividing by $\|\boldsymbol{x}+\boldsymbol{y}\|^{p / p^{\prime}}$, it follows $\|\boldsymbol{x}+\boldsymbol{y}\|^{p}\|\boldsymbol{x}+\boldsymbol{y}\|^{-p / p^{\prime}}=\|\boldsymbol{x}+\boldsymbol{y}\| \leq\|\boldsymbol{x}\|_{p}+$ $\|\boldsymbol{y}\|_{p}\left(p-\frac{p}{p^{\prime}}=p\left(1-\frac{1}{p^{\prime}}\right)=p \frac{1}{p}=1\right)$.

It only remains to prove Lemma 10.5.9.
Proof of the lemma: Let $p^{\prime}=q$ to save on notation and consider the following picture:


$$
a b \leq \int_{0}^{a} t^{p-1} d t+\int_{0}^{b} x^{q-1} d x=\frac{a^{p}}{p}+\frac{b^{q}}{q}
$$

Note equality occurs when $a^{p}=b^{q}$.
Alternate proof of the lemma: First note that if either $a$ or $b$ are zero, then there is nothing to show so we can assume $b, a>0$. Let $b>0$ and let

$$
f(a)=\frac{a^{p}}{p}+\frac{b^{q}}{q}-a b
$$

Then the second derivative of $f$ is positive on $(0, \infty)$ so its graph is convex. Also $f(0)>0$ and $\lim _{a \rightarrow \infty} f(a)=\infty$. Then a short computation shows that there is only one critical point, where $f$ is minimized and this happens when $a$ is such that $a^{p}=b^{q}$. At this point,

$$
f(a)=b^{q}-b^{q / p} b=b^{q}-b^{q-1} b=0
$$

Therefore, $f(a) \geq 0$ for all $a$ and this proves the lemma.
Another example of a very useful norm on $\mathbb{F}^{n}$ is the norm $\|\cdot\|_{\infty}$ defined by

$$
\|\boldsymbol{x}\|_{\infty} \equiv \max \left\{\left|x_{k}\right|: k=1,2, \cdots, n\right\}
$$

You should verify that this satisfies all the axioms of a norm. Here is the triangle inequality.

$$
\begin{aligned}
\|\boldsymbol{x}+\boldsymbol{y}\|_{\infty} & =\max _{k}\left\{\left|x_{k}+y_{k}\right|\right\} \leq \max _{k}\left\{\left|x_{k}\right|+\left|y_{k}\right|\right\} \\
& \leq \max _{k}\left\{\left|x_{k}\right|\right\}+\max _{k}\left\{\left|y_{k}\right|\right\}=\|\boldsymbol{x}\|_{\infty}+\|\boldsymbol{y}\|_{\infty}
\end{aligned}
$$

It turns out that in terms of analysis (limits of sequences, completeness and so forth), it makes absolutely no difference which norm you use. There are however, significant geometric differences. This will be explained later. First is the notion of an orthonormal basis.

### 10.5.2 Orthonormal Bases

Not all bases for an inner product space $H$ are created equal. The best bases are orthonormal.

Definition 10.5.11 Suppose $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{k}\right\}$ is a set of vectors in an inner product space $H$. It is an orthonormal set if

$$
\left(\boldsymbol{v}_{i}, \boldsymbol{v}_{j}\right)=\delta_{i j}=\left\{\begin{array}{l}
1 \text { if } i=j \\
0 \text { if } i \neq j
\end{array}\right.
$$

Every orthonormal set of vectors is automatically linearly independent. Indeed, if

$$
\sum_{k=1}^{n} a_{k} \boldsymbol{v}_{k}=\mathbf{0}
$$

then taking the inner product with $\boldsymbol{v}_{j}$, yields $0=\sum_{k=1}^{n} a_{k}\left(\boldsymbol{v}_{k}, \boldsymbol{v}_{j}\right)=a_{j}$. Thus each $a_{j}=0$. We will use this simple observation whenever convenient.

Proposition 10.5.12 Suppose $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{k}\right\}$ is an orthonormal set of vectors. Then it is linearly independent.

Proof: Suppose $\sum_{i=1}^{k} c_{i} \boldsymbol{v}_{i}=\mathbf{0}$. Then taking inner products with $\boldsymbol{v}_{j}$,

$$
0=\left(\mathbf{0}, \boldsymbol{v}_{j}\right)=\sum_{i} c_{i}\left(\boldsymbol{v}_{i}, \boldsymbol{v}_{j}\right)=\sum_{i} c_{i} \boldsymbol{\delta}_{i j}=c_{j} .
$$

Since $j$ is arbitrary, this shows the set is linearly independent as claimed.
It turns out that if $X$ is any subspace of $H$, then there exists an orthonormal basis for $X$.
Lemma 10.5.13 Let $X$ be a subspace of dimension $n$ whose basis is $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{n}\right\}$. Then there exists an orthonormal basis for $X,\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{n}\right\}$ which has the property that for each $k \leq n, \operatorname{span}\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{k}\right)=\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}\right)$.

Proof: Let $\left\{\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{n}\right\}$ be a basis for $X$. Let $\boldsymbol{u}_{1} \equiv \boldsymbol{x}_{1} /\left|\boldsymbol{x}_{1}\right|$. Therefore, it follows that for $k=1$, span $\left(\boldsymbol{u}_{1}\right)=\operatorname{span}\left(\boldsymbol{x}_{1}\right)$ and $\left\{\boldsymbol{u}_{1}\right\}$ is an orthonormal set. Now suppose for some $k<n, \boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}$ have been chosen such that $\left(\boldsymbol{u}_{j}, \boldsymbol{u}_{l}\right)=\boldsymbol{\delta}_{j l}$ and span $\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{k}\right)=$ $\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}\right)$. Then define

$$
\begin{equation*}
\boldsymbol{u}_{k+1} \equiv \frac{\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{u}_{j}}{\left|\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{u}_{j}\right|}, \tag{10.13}
\end{equation*}
$$

where the denominator is not equal to zero because the $\boldsymbol{x}_{j}$ form a basis and so

$$
\boldsymbol{x}_{k+1} \notin \operatorname{span}\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{k}\right)=\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}\right)
$$

Thus by induction,

$$
\boldsymbol{u}_{k+1} \in \operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}, \boldsymbol{x}_{k+1}\right)=\operatorname{span}\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{k}, \boldsymbol{x}_{k+1}\right) .
$$

Also, $\boldsymbol{x}_{k+1} \in \operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}, \boldsymbol{u}_{k+1}\right)$ which is seen easily by solving 10.13 for $\boldsymbol{x}_{k+1}$ and it follows

$$
\operatorname{span}\left(\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{k}, \boldsymbol{x}_{k+1}\right)=\operatorname{span}\left(\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}, \boldsymbol{u}_{k+1}\right)
$$

If $l \leq k$, then denoting by $C$ the scalar $\left|\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{u}_{j}\right|^{-1}$,

$$
\begin{gathered}
\left(\boldsymbol{u}_{k+1}, \boldsymbol{u}_{l}\right)=C\left(\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{l}\right)-\sum_{j=1}^{k}\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{j}\right)\left(\boldsymbol{u}_{j}, \boldsymbol{u}_{l}\right)\right) \\
=C\left(\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{l}\right)-\sum_{j=1}^{k}\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{\delta}_{l j}\right)=C\left(\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{l}\right)-\left(\boldsymbol{x}_{k+1}, \boldsymbol{u}_{l}\right)\right)=0 .
\end{gathered}
$$

The vectors, $\left\{\boldsymbol{u}_{j}\right\}_{j=1}^{n}$, generated in this way are therefore an orthonormal basis because each vector has unit length.

The process by which these vectors were generated is called the Gram Schmidt process.
The following corollary is obtained from the above process.
Corollary 10.5.14 Let $X$ be a finite dimensional inner product space of dimension $n$ whose basis is $\left\{u_{1}, \cdots, u_{k}, x_{k+1}, \cdots, x_{n}\right\}$. Then if $\left\{u_{1}, \cdots, u_{k}\right\}$ is orthonormal, then the Gram Schmidt process applied to the given list of vectors in order leaves $\left\{u_{1}, \cdots, u_{k}\right\}$ unchanged.

### 10.6 Equivalence Of Norms

As mentioned above, it makes absolutely no difference which norm you decide to use. This holds in general finite dimensional normed spaces and is shown here.

Definition 10.6.1 Let $(V,\|\cdot\|)$ be a normed linear space with basis $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$. For $\boldsymbol{x} \in V$, let its component vector in $\mathbb{F}^{n}$ be $\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ so that $\boldsymbol{x}=\sum_{i} \alpha_{i} \boldsymbol{v}_{i}$. Then define

$$
\theta \boldsymbol{x} \equiv \boldsymbol{\alpha}=\left(\begin{array}{lll}
\alpha_{1} & \cdots & \alpha_{n}
\end{array}\right)^{T}
$$

Thus $\theta$ is well defined, one to one and onto from $V$ to $\mathbb{F}^{n}$. It is also linear and its inverse $\theta^{-1}$ satisfies all the same algebraic properties.

The following fundamental lemma comes from the extreme value theorem for continuous functions defined on a compact set. Let

$$
f(\boldsymbol{\alpha}) \equiv\left\|\sum_{i} \alpha_{i} \boldsymbol{v}_{i}\right\| \equiv\left\|\theta^{-1} \boldsymbol{\alpha}\right\|
$$

Then it is clear that $f$ is a continuous function. This is because $\boldsymbol{\alpha} \rightarrow \sum_{i} \alpha_{i} \boldsymbol{v}_{i}$ is a continuous map into $V$ and from the triangle inequality $\boldsymbol{x} \rightarrow\|\boldsymbol{x}\|$ is continuous as a map from $V$ to $\mathbb{R}$.

Lemma 10.6.2 There exists $\delta>0$ and $\Delta \geq \delta$ such that

$$
\delta=\min \{f(\boldsymbol{\alpha}):|\boldsymbol{\alpha}|=1\}, \Delta=\max \{f(\boldsymbol{\alpha}):|\boldsymbol{\alpha}|=1\}
$$

Also,

$$
\begin{align*}
\boldsymbol{\delta}|\boldsymbol{\alpha}| & \leq\left\|\theta^{-1} \boldsymbol{\alpha}\right\| \leq \Delta|\boldsymbol{\alpha}|  \tag{10.14}\\
\delta|\theta \boldsymbol{v}| & \leq\|\boldsymbol{v}\| \leq \Delta|\theta \boldsymbol{v}| \tag{10.15}
\end{align*}
$$

Proof: These numbers exist thanks to Theorem 10.4.7. It cannot be that $\delta=0$ because if it were, you would have $|\boldsymbol{\alpha}|=1$ but $\sum_{j=1}^{n} \alpha_{k} \boldsymbol{v}_{j}=\mathbf{0}$ which is impossible since $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ is linearly independent. The first of the above inequalities follows from

$$
\delta \leq\left\|\theta^{-1} \frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|}\right\|=f\left(\frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|}\right) \leq \Delta
$$

the second follows from observing that $\theta^{-1} \boldsymbol{\alpha}$ is a generic vector $\boldsymbol{v}$ in $V$.
Now we can draw several conclusions about $(V,\|\cdot\|)$ for $V$ finite dimensional.
Theorem 10.6.3 Let $(V,\|\cdot\|)$ be a finite dimensional normed linear space. Then the compact sets are exactly those which are closed and bounded. Also $(V,\|\cdot\|)$ is complete. If $K$ is a closed and bounded set in $(V,\|\cdot\|)$ and $f: K \rightarrow \mathbb{R}$, then $f$ achieves its maximum and minimum on $K$.

Proof: First note that the inequalities 10.14 and 10.15 show that both $\theta^{-1}$ and $\theta$ are continuous. Thus these take convergent sequences to convergent sequences.

Let $\left\{\boldsymbol{w}_{k}\right\}_{k=1}^{\infty}$ be a Cauchy sequence. Then from 10.15, $\left\{\theta \boldsymbol{w}_{k}\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Thanks to Theorem 10.4.7, it converges to some $\boldsymbol{\beta} \in \mathbb{F}^{n}$. It follows that

$$
\lim _{k \rightarrow \infty} \theta^{-1} \theta \boldsymbol{w}_{k}=\lim _{k \rightarrow \infty} \boldsymbol{w}_{k}=\theta^{-1} \boldsymbol{\beta} \in V
$$

This shows completeness.
Next let $K$ be a closed and bounded set. Let $\left\{\boldsymbol{w}_{k}\right\} \subseteq K$. Then $\left\{\theta \boldsymbol{w}_{k}\right\} \subseteq \theta K$ which is also a closed and bounded set thanks to the inequalities 10.14 and 10.15 . Thus there is a subsequence still denoted with $k$ such that $\theta \boldsymbol{w}_{k} \rightarrow \boldsymbol{\beta} \in \mathbb{F}^{n}$. Then as just done, $\boldsymbol{w}_{k} \rightarrow \boldsymbol{\theta}^{-1} \boldsymbol{\beta}$. Since $K$ is closed, it follows that $\theta^{-1} \beta \in K$.

Finally, why are the only compact sets those which are closed and bounded? Let $K$ be compact. If it is not bounded, then there is a sequence of points of $K,\left\{\boldsymbol{k}^{m}\right\}_{m=1}^{\infty}$ such that $\left\|\boldsymbol{k}^{m}\right\| \geq m$. It follows that it cannot have a convergent subsequence because the points are further apart from each other than $1 / 2$. Hence $K$ is not sequentially compact and consequently it is not compact. It follows that $K$ is bounded. If $K$ is not closed, then there exists a limit point $\boldsymbol{k}$ which is not in $K$. (Recall that closed means it has all its limit points.) By Theorem 10.1.7, there is a sequence of distinct points having no repeats and none equal to $\boldsymbol{k}$ denoted as $\left\{\boldsymbol{k}^{m}\right\}_{m=1}^{\infty}$ such that $\boldsymbol{k}^{m} \rightarrow \boldsymbol{k}$. Then this sequence $\left\{\boldsymbol{k}^{m}\right\}$ fails to have a subsequence which converges to a point of $K$. Hence $K$ is not sequentially compact. Thus, if $K$ is compact then it is closed and bounded.

The last part is identical to the proof in Theorem 10.4.7. You just take a convergent subsequence of a minimizing (maximizing) sequence and exploit continuity.

Next is the theorem which states that any two norms on a finite dimensional vector space are equivalent.

Theorem 10.6.4 Let $\|\cdot\|,\| \| \cdot \| \mid$ be two norms on $V$ a finite dimensional vector space. Then they are equivalent, which means there are constants $0<a<b$ such that for all $\boldsymbol{v}$,

$$
a\|\boldsymbol{v}\| \leq\| \| \boldsymbol{v}\| \| \leq b\|\boldsymbol{v}\|
$$

Proof: In Lemma 10.6 .2 , let $\delta, \Delta$ go with $\|\cdot\|$ and $\hat{\delta}, \hat{\Delta}$ go with $\|\mid \cdot\| \|$. Then using the inequalities of this lemma,

$$
\|\boldsymbol{v}\| \leq \Delta|\theta \boldsymbol{v}| \leq \frac{\Delta}{\hat{\delta}}\| \| \boldsymbol{v}\| \| \leq \frac{\Delta \hat{\Delta}}{\hat{\delta}}|\theta \boldsymbol{v}| \leq \frac{\Delta}{\delta} \hat{\Delta} \hat{\Delta}_{\hat{\delta}}^{\hat{\delta}}\|\boldsymbol{v}\|
$$

and so

$$
\frac{\hat{\delta}}{\Delta}\|\boldsymbol{v}\| \leq\| \| \boldsymbol{v}\left\|\left\lvert\, \leq \frac{\hat{\Delta}}{\delta}\right.\right\| \boldsymbol{v} \|
$$

Thus the norms are equivalent.
It follows right away that the closed and open sets are the same with two different norms. Also, all considerations involving limits are unchanged from one norm to another.

Corollary 10.6.5 Consider the metric spaces $\left(V,\|\cdot\|_{1}\right),\left(V,\|\cdot\|_{2}\right)$ where $V$ has dimension $n$. Then a set is closed or open in one of these if and only if it is respectively closed or open in the other. In other words, the two metric spaces have exactly the same open and closed sets. Also, a set is bounded in one metric space if and only if it is bounded in the other.

Proof: This follows from Theorem 10.1.27, the theorem about the equivalent formulations of continuity. Using this theorem, it follows from Theorem 10.6.4 that the identity $\operatorname{map} I(\boldsymbol{x}) \equiv \boldsymbol{x}$ is continuous. The reason for this is that the inequality of this theorem implies that if $\left\|\boldsymbol{v}^{m}-\boldsymbol{v}\right\|_{1} \rightarrow 0$ then $\left\|\boldsymbol{I} \boldsymbol{v}^{m}-\boldsymbol{I} \boldsymbol{v}\right\|_{2}=\left\|I\left(\boldsymbol{v}^{m}-\boldsymbol{v}\right)\right\|_{2} \rightarrow 0$ and the same holds on switching 1 and 2 in what was just written.

Therefore, the identity map takes open sets to open sets and closed sets to closed sets. In other words, the two metric spaces have the same open sets and the same closed sets.

Suppose $S$ is bounded in $\left(V,\|\cdot\|_{1}\right)$. This means it is contained in $B(0, r)_{1}$ where the subscript of 1 indicates the norm is $\|\cdot\|_{1}$. Let $\delta\|\cdot\|_{1} \leq\|\cdot\|_{2} \leq \Delta\|\cdot\|_{1}$ as described above. Then

$$
S \subseteq B(\mathbf{0}, r)_{1} \subseteq B(\mathbf{0}, \Delta r)_{2}
$$

so $S$ is also bounded in $\left(V,\|\cdot\|_{2}\right)$. Similarly, if $S$ is bounded in $\|\cdot\|_{2}$ then it is bounded in $\|\cdot\|_{1}$.

### 10.7 Norms On $\mathscr{L}(X, Y)$

First here is a definition which applies in all cases, even if $X, Y$ are infinite dimensional.
Definition 10.7.1 Let $X$ and $Y$ be normed linear spaces with norms $\|\cdot\|_{X}$ and $\|\cdot\|_{Y}$ respectively. Then $\mathscr{L}(X, Y)$ denotes the space of linear transformations, called bounded linear transformations, mapping $X$ to $Y$ which have the property that

$$
\|A\| \equiv \sup \left\{\|A x\|_{Y}:\|x\|_{X} \leq 1\right\}<\infty
$$

Then $\|A\|$ is referred to as the operator norm of the bounded linear transformation $A$. We will always assume that if a norm is present the mappings are bounded. However, we show that this boundedness will be automatic in the case of finite dimensions.

It is an easy exercise to verify that $\|\cdot\|$ is a norm on $\mathscr{L}(X, Y)$ and it is always the case that

$$
\|A x\|_{Y} \leq\|A\|\|x\|_{X} .
$$

Furthermore, you should verify that you can replace $\leq 1$ with $=1$ in the definition. Thus

$$
\|A\| \equiv \sup \left\{\|A x\|_{Y}:\|x\|_{X}=1\right\}
$$

In the case that the vector spaces are finite dimensional, the situation becomes very simple.
Lemma 10.7.2 Let $V$ be a finite dimensional vector space with norm $\|\cdot\|_{V}$ and let $W$ be a vector space with norm $\|\cdot\|_{W}$. Then if $A$ is a linear map from $V$ to $W$, then $A$ is continuous and bounded.

Proof: Suppose $\lim _{k \rightarrow \infty} \boldsymbol{v}^{k}=\boldsymbol{v}$ in $V$. Let $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ be a basis and let $\theta$ be the coordinate map of Definition 10.6.1. Then by $10.15, \lim _{k \rightarrow \infty} \theta\left(\boldsymbol{v}^{k}-\boldsymbol{v}\right)=\mathbf{0} \in \mathbb{F}^{n}$. Letting $\boldsymbol{\alpha}^{k}$ and $\boldsymbol{\alpha}$ be $\theta \boldsymbol{v}^{k}$ and $\boldsymbol{\theta} \boldsymbol{v}$ respectively, it follows that $\boldsymbol{\alpha}^{k} \rightarrow \boldsymbol{\alpha}$ and so

$$
A \boldsymbol{v}^{k}=A \sum_{j=1}^{n} \alpha_{j}^{k} \boldsymbol{v}_{j}=\sum_{j=1}^{n} \alpha_{j}^{k} A \boldsymbol{v}_{j}
$$

which converges to $\sum_{k=1}^{n} \alpha_{j} A \boldsymbol{v}_{j}=A \boldsymbol{v}$ as $k \rightarrow \infty$. Thus $A$ is continuous. Then also $\boldsymbol{v} \rightarrow$ $\|A \boldsymbol{v}\|_{W}$ is a continuous function. Now let $D$ be the closed ball of radius 1 in $V$. By Theorem 10.6.3, this set $D$ is compact and so

$$
\max \left\{\|A \boldsymbol{v}\|_{W}:\|\boldsymbol{v}\|_{V} \leq 1\right\} \equiv\|A\|<\infty .
$$

Then we have the following theorem.

Theorem 10.7.3 Let $X$ and $Y$ be finite dimensional normed linear spaces of dimension $n$ and $m$ respectively and denote by $\|\cdot\|$ the norm on either $X$ or $Y$. Then if $A$ is any linear function mapping $X$ to $Y$, then $A \in \mathscr{L}(X, Y)$ and $(\mathscr{L}(X, Y),\|\cdot\|)$ is a complete normed linear space of dimension nm with

$$
\|A x\| \leq\|A\|\|x\|
$$

Also if $A \in \mathscr{L}(X, Y)$ and $B \in \mathscr{L}(Y, Z)$ where $X, Y, Z$ are normed linear spaces,

$$
\|B A\| \leq\|B\|\|A\|
$$

Proof: It is necessary to show the norm defined on linear transformations really is a norm. Again the triangle inequality is the only property which is not obvious. It remains to show this and verify $\|A\|<\infty$. This last follows from the above Lemma 10.7.2. Thus the norm is at least well defined. It remains to verify its properties.

$$
\begin{gathered}
\|A+B\| \equiv \sup \{\|(A+B)(\boldsymbol{x})\|:\|\boldsymbol{x}\| \leq 1\} \\
\leq \sup \{\|A \boldsymbol{x}\|:\|\boldsymbol{x}\| \leq 1\}+\sup \{\|B \boldsymbol{x}\|:\|\boldsymbol{x}\| \leq 1\} \equiv\|A\|+\|B\|
\end{gathered}
$$

Next consider the assertion about the dimension of $\mathscr{L}(X, Y)$. It follows from Theorem 5.1.4. By Theorem 10.6.4 $(\mathscr{L}(X, Y),\|\cdot\|)$ is complete. If $\boldsymbol{x} \neq \mathbf{0}$,

$$
\|A \boldsymbol{x}\| \frac{1}{\|\boldsymbol{x}\|}=\left\|A \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\| \leq\|A\|
$$

Thus $\|A x\| \leq\|A\|\|x\|$.
Consider the last claim.

$$
\|B A\| \equiv \sup _{\|x\| \leq 1}\|B(A(x))\| \leq\|B\| \sup _{\|x\| \leq 1}\|A x\|=\|B\|\|A\|
$$

Note by Theorem 10.6.4 you can define a norm any way desired on any finite dimensional linear space which has the field of scalars $\mathbb{R}$ or $\mathbb{C}$ and any other way of defining a norm on this space yields an equivalent norm. Thus, it doesn't much matter as far as notions of convergence are concerned which norm is used for a finite dimensional space. In particular in the space of $m \times n$ matrices, you can use the operator norm defined above, or some other way of giving this space a norm. A popular choice for a norm is the Frobenius norm.

Definition 10.7.4 Define $A^{*}$ as the transpose of the conjugate of $A$. This is called the adjoint of $A$. Make the space of $m \times n$ matrices into a inner product space by defining

$$
(A, B) \equiv \operatorname{trace}\left(A B^{*}\right) \equiv \sum_{i}\left(A B^{*}\right)_{i i}=\sum_{i} \sum_{j} A_{i j} B_{j i}^{*} \equiv \sum_{i, j} A_{i j} \overline{B_{i j}}
$$

$\|A\| \equiv(A, A)^{1 / 2}$.
This is clearly a norm because, as implied by the notation, $A, B \rightarrow(A, B)$ is an inner product on the space of $m \times n$ matrices. You should verify that this is the case.

### 10.8 Limits Of A Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the limit of a function. The notion of limit of a function makes sense at points $\boldsymbol{x}$, which are limit points of $D(\boldsymbol{f})$ and this concept is defined next. In all that follows $(V,\|\cdot\|)$ and $(W,\|\cdot\|)$ are two normed linear spaces. Recall the definition of limit point first.

Definition 10.8.1 Let $A \subseteq W$ be a set. A point $\boldsymbol{x}$, is a limit point of $A$ if $B(\boldsymbol{x}, r)$ contains infinitely many points of A for every $r>0$.

Definition 10.8.2 Let $\boldsymbol{f}: D(\boldsymbol{f}) \subseteq V \rightarrow W$ be a function and let $\boldsymbol{x}$ be a limit point of $D(\boldsymbol{f})$. Then

$$
\lim _{\boldsymbol{y} \rightarrow \boldsymbol{x}} f(\boldsymbol{y})=\boldsymbol{L}
$$

if and only if the following condition holds. For all $\varepsilon>0$ there exists $\delta>0$ such that if

$$
0<\|\boldsymbol{y}-\boldsymbol{x}\|<\boldsymbol{\delta}, \text { and } \boldsymbol{y} \in D(\boldsymbol{f})
$$

then,

$$
\|\boldsymbol{L}-\boldsymbol{f}(\boldsymbol{y})\|<\boldsymbol{\varepsilon}
$$

Theorem 10.8.3 If $\lim _{\boldsymbol{y} \rightarrow \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y})=\boldsymbol{L}$ and $\lim _{y \rightarrow x} \boldsymbol{f}(\boldsymbol{y})=\boldsymbol{L}_{1}$, then $\boldsymbol{L}=\boldsymbol{L}_{1}$.
Proof: Let $\varepsilon>0$ be given. There exists $\delta>0$ such that if $0<|\boldsymbol{y}-\boldsymbol{x}|<\delta$ and $\boldsymbol{y} \in$ $D(f)$, then

$$
\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}\|<\boldsymbol{\varepsilon},\left\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}_{1}\right\|<\varepsilon .
$$

Pick such a $\boldsymbol{y}$. There exists one because $\boldsymbol{x}$ is a limit point of $D(\boldsymbol{f})$. Then

$$
\left\|\boldsymbol{L}-\boldsymbol{L}_{1}\right\| \leq\|\boldsymbol{L}-\boldsymbol{f}(\boldsymbol{y})\|+\left\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}_{1}\right\|<\varepsilon+\varepsilon=2 \varepsilon .
$$

Since $\varepsilon>0$ was arbitrary, this shows $L=L_{1}$.
As in the case of functions of one variable, one can define $\lim _{\boldsymbol{y} \rightarrow \boldsymbol{x}} f(\boldsymbol{x})= \pm \infty$.
Definition 10.8.4 If $f(\boldsymbol{x}) \in \mathbb{R}, \lim _{\boldsymbol{y} \rightarrow \boldsymbol{x}} f(\boldsymbol{x})=\infty$ iffor every number $l$, there exists $\boldsymbol{\delta}>0$ such that whenever $\|\boldsymbol{y}-\boldsymbol{x}\|<\boldsymbol{\delta}$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $f(\boldsymbol{x})>l . \lim _{\boldsymbol{y} \rightarrow \boldsymbol{x}} f(\boldsymbol{x})=-\infty$ iffor every number $l$, there exists $\delta>0$ such that whenever $\|\boldsymbol{y}-\boldsymbol{x}\|<\boldsymbol{\delta}$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $f(\boldsymbol{x})<l$.

The following theorem is just like the one variable version of calculus.
Theorem 10.8.5 Suppose $\boldsymbol{f}: D(\boldsymbol{f}) \subseteq V \rightarrow \mathbb{F}^{m}$. Then for $\boldsymbol{x}$ a limit point of $D(\boldsymbol{f})$,

$$
\begin{equation*}
\lim _{y \rightarrow x} \boldsymbol{f}(\boldsymbol{y})=\boldsymbol{L} \tag{10.16}
\end{equation*}
$$

if and only if

$$
\begin{equation*}
\lim _{y \rightarrow x} f_{k}(\boldsymbol{y})=L_{k} \tag{10.17}
\end{equation*}
$$

where $\boldsymbol{f}(\boldsymbol{y}) \equiv\left(f_{1}(\boldsymbol{y}), \cdots, f_{p}(\boldsymbol{y})\right)$ and $\boldsymbol{L} \equiv\left(L_{1}, \cdots, L_{p}\right)$.
Suppose here that $f$ has values in $W$, a normed linear space and

$$
\lim _{y \rightarrow x} f(y)=L, \lim _{y \rightarrow \boldsymbol{x}} g(y)=K
$$

where $K, L \in W$. Then if $a, b \in \mathbb{F}$,

$$
\begin{equation*}
\lim _{y \rightarrow x}(a f(y)+b g(y))=a L+b K \tag{10.18}
\end{equation*}
$$

If $W$ is an inner product space,

$$
\begin{equation*}
\lim _{y \rightarrow x}(f, g)(y)=(L, K) \tag{10.19}
\end{equation*}
$$

If $g$ is scalar valued with $\lim _{y \rightarrow x} g(y)=K$,

$$
\begin{equation*}
\lim _{y \rightarrow x} f(y) g(y)=L K \tag{10.20}
\end{equation*}
$$

Also, if $h$ is a continuous function defined near $L$, then

$$
\begin{equation*}
\lim _{y \rightarrow x} h \circ f(y)=h(L) \tag{10.21}
\end{equation*}
$$

Suppose $\lim _{y \rightarrow \boldsymbol{x}} f(y)=L$. If $\|f(y)-b\| \leq r$ for all $y$ sufficiently close to $\boldsymbol{x}$, then $|L-b| \leq r$ also.

Proof: Suppose 10.16. Then letting $\varepsilon>0$ be given there exists $\delta>0$ such that if $0<\|y-x\|<\delta$, it follows

$$
\left|f_{k}(y)-L_{k}\right| \leq\|\boldsymbol{f}(y)-\boldsymbol{L}\|<\varepsilon
$$

which verifies 10.17 .
Now suppose 10.17 holds. Then letting $\varepsilon>0$ be given, there exists $\delta_{k}$ such that if $0<\|y-x\|<\delta_{k}$, then

$$
\left|f_{k}(y)-L_{k}\right|<\varepsilon
$$

Let $0<\delta<\min \left(\delta_{1}, \cdots, \delta_{p}\right)$. Then if $0<\|y-x\|<\delta$, it follows

$$
\|\boldsymbol{f}(y)-\boldsymbol{L}\|_{\infty}<\boldsymbol{\varepsilon}
$$

Any other norm on $\mathbb{F}^{m}$ would work out the same way because the norms are all equivalent.
Each of the remaining assertions follows immediately from the coordinate descriptions of the various expressions and the first part. However, I will give a different argument for these.

The proof of 10.18 is left for you. Now 10.19 is to be verified. Let $\varepsilon>0$ be given. Then by the triangle inequality,

$$
\begin{aligned}
|(f, g)(y)-(L, K)| & \leq|(f, g)(y)-(f(y), K)|+|(f(y), K)-(L, K)| \\
& \leq\|f(y)\|\|g(y)-K\|+\|K\|\|f(y)-L\|
\end{aligned}
$$

There exists $\delta_{1}$ such that if $0<\|y-x\|<\delta_{1}$ and $y \in D(f)$, then

$$
\|f(y)-L\|<1
$$

and so for such $y$, the triangle inequality implies, $\|f(y)\|<1+\|L\|$. Therefore, for $0<$ $\|y-x\|<\delta_{1}$,

$$
\begin{equation*}
|(f, g)(y)-(L, K)| \leq(1+\|K\|+\|L\|)[\|g(y)-K\|+\|f(y)-L\|] . \tag{10.22}
\end{equation*}
$$

Now let $0<\boldsymbol{\delta}_{2}$ be such that if $y \in D(f)$ and $0<\|x-y\|<\boldsymbol{\delta}_{2}$,

$$
\|f(y)-L\|<\frac{\varepsilon}{2(1+\|K\|+\|L\|)},\|g(y)-K\|<\frac{\varepsilon}{2(1+\|K\|+\|L\|)}
$$

Then letting $0<\boldsymbol{\delta} \leq \min \left(\boldsymbol{\delta}_{1}, \boldsymbol{\delta}_{2}\right)$, it follows from 10.22 that

$$
|(f, g)(y)-(L, K)|<\varepsilon
$$

and this proves 10.19 .
The proof of 10.20 is left to you.
Consider 10.21. Since $h$ is continuous near $L$, it follows that for $\varepsilon>0$ given, there exists $\eta>0$ such that if $\|y-L\|<\eta$, then

$$
\|h(y)-h(L)\|<\varepsilon
$$

Now since $\lim _{y \rightarrow x} f(y)=L$, there exists $\delta>0$ such that if $0<\|y-x\|<\delta$, then

$$
\|f(y)-L\|<\eta
$$

Therefore, if $0<\|y-x\|<\delta$,

$$
\|h(f(y))-h(L)\|<\varepsilon
$$

It only remains to verify the last assertion. Assume $\|f(y)-b\| \leq r$. It is required to show that $\|L-b\| \leq r$. If this is not true, then $\|L-b\|>r$. Consider $B(L,\|L-b\|-r)$. Since $L$ is the limit of $f$, it follows $f(y) \in B(L,\|L-b\|-r)$ whenever $y \in D(f)$ is close enough to $x$. Thus, by the triangle inequality,

$$
\|f(y)-L\|<\|L-b\|-r
$$

and so

$$
\begin{aligned}
r & <\|L-b\|-\|f(y)-L\| \leq \mid\|b-L\|-\|f(y)-L\| \| \\
& \leq\|b-f(y)\|
\end{aligned}
$$

a contradiction to the assumption that $\|b-f(y)\| \leq r$.
The relation between continuity and limits is as follows.
Theorem 10.8.6 For $f: D(f) \rightarrow W$ and $x \in D(f)$ a limit point of $D(f), f$ is continuous at $x$ if and only if

$$
\lim _{y \rightarrow x} f(y)=f(x)
$$

Proof: First suppose $f$ is continuous at $x$ a limit point of $D(f)$. Then for every $\varepsilon>0$ there exists $\delta>0$ such that if $\|x-y\|<\delta$ and $y \in D(f)$, then $|f(x)-f(y)|<\varepsilon$. In particular, this holds if $0<\|x-y\|<\delta$ and this is just the definition of the limit. Hence $f(x)=\lim _{y \rightarrow x} f(y)$.

Next suppose $x$ is a limit point of $D(f)$ and $\lim _{y \rightarrow x} f(y)=f(x)$. This means that if $\varepsilon>$ 0 there exists $\delta>0$ such that for $0<\|x-y\|<\delta$ and $y \in D(f)$, it follows $|f(y)-f(x)|<$ $\varepsilon$. However, if $y=x$, then $|f(y)-f(x)|=|f(x)-f(x)|=0$ and so whenever $y \in D(f)$ and $\|x-y\|<\delta$, it follows $|f(x)-f(y)|<\varepsilon$, showing $f$ is continuous at $x$.

Example 10.8.7 Find $\lim _{(x, y) \rightarrow(3,1)}\left(\frac{x^{2}-9}{x-3}, y\right)$.
It is clear that $\lim _{(x, y) \rightarrow(3,1)} \frac{x^{2}-9}{x-3}=6$ and $\lim _{(x, y) \rightarrow(3,1)} y=1$. Therefore, this limit equals $(6,1)$.
Example 10.8.8 Find $\lim _{(x, y) \rightarrow(0,0)} \frac{x y}{x^{2}+y^{2}}$.
First of all, observe the domain of the function is $\mathbb{R}^{2} \backslash\{(0,0)\}$, every point in $\mathbb{R}^{2}$ except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, the limit may not exist. In fact, this is the case here. To see this, take points on the line $y=0$. At these points, the value of the function equals 0 . Now consider points on the line $y=x$ where the value of the function equals $1 / 2$. Since, arbitrarily close to $(0,0)$, there are points where the function equals $1 / 2$ and points where the function has the value 0 , it follows there can be no limit. Just take $\varepsilon=1 / 10$ for example. You cannot be within $1 / 10$ of $1 / 2$ and also within $1 / 10$ of 0 at the same time.

Note it is necessary to rely on the definition of the limit much more than in the case of a function of one variable and there are no easy ways to do limit problems for functions of more than one variable. It is what it is and you will not deal with these concepts without suffering and anguish.

### 10.9 Exercises

1. Consider $C\left([0, T], \mathbb{R}^{n}\right)$ with the norm $\|\boldsymbol{f}\| \equiv \max _{x \in[0, T]}\|\boldsymbol{f}(x)\|_{\infty}$. Explain why the maximum exists. Show this is a complete metric space. Hint: If you have $\left\{\boldsymbol{f}_{m}\right\}$ a Cauchy sequence in $C\left([0, T], \mathbb{R}^{n}\right)$, then for each $x$, you have $\left\{\boldsymbol{f}_{m}(x)\right\}$ a Cauchy sequence in $\mathbb{R}^{n}$ so it converges by completeness of $\mathbb{R}^{n}$. See Example 10.1.23. Thus there exists $\boldsymbol{f}(x) \equiv \lim _{m \rightarrow \infty} \boldsymbol{f}_{m}(x)$. You must show that $\boldsymbol{f}$ is continuous. Consider

$$
\begin{aligned}
\left\|\boldsymbol{f}_{m}(x)-\boldsymbol{f}_{m}(y)\right\| \leq & \left\|\boldsymbol{f}_{m}(x)-\boldsymbol{f}_{n}(x)\right\|+\left\|\boldsymbol{f}_{n}(x)-\boldsymbol{f}_{n}(y)\right\| \\
& +\left\|\boldsymbol{f}_{n}(y)-\boldsymbol{f}_{m}(y)\right\| \\
\leq & 2 \boldsymbol{\varepsilon} / 3+\left\|\boldsymbol{f}_{n}(x)-\boldsymbol{f}_{n}(y)\right\|
\end{aligned}
$$

for $n$ large enough. Now let $m \rightarrow \infty$ to get the same inequality with $f$ on the left. Next use continuity of $f_{n}$. Finally,

$$
\left\|\boldsymbol{f}(x)-\boldsymbol{f}_{n}(x)\right\|=\lim _{m \rightarrow \infty}\left\|\boldsymbol{f}_{m}(x)-\boldsymbol{f}_{n}(x)\right\|
$$

and since a Cauchy sequence, $\left\|\boldsymbol{f}_{m}-\boldsymbol{f}_{n}\right\|<\varepsilon$ whenever $m>n$ for $n$ large enough. Use to show that $\left\|f-f_{n}\right\|_{\infty} \rightarrow 0$.
2. For $f \in C\left([0, T], \mathbb{R}^{n}\right)$, you define the Riemann integral in the usual way using Riemann sums. Alternatively, you can define it as

$$
\int_{0}^{t} f(s) d s=\left(\int_{0}^{t} f_{1}(s) d s, \int_{0}^{t} f_{2}(s) d s, \cdots, \int_{0}^{t} f_{n}(s) d s\right)
$$

Then show that the following limit exists in $\mathbb{R}^{n}$ for each $t \in(0, T)$.

$$
\lim _{h \rightarrow 0} \frac{\int_{0}^{t+h} \boldsymbol{f}(s) d s-\int_{0}^{t} \boldsymbol{f}(s) d s}{h}=\boldsymbol{f}(t)
$$

You should use the fundamental theorem of calculus from one variable calculus and the definition of the norm to verify this. Recall that $\lim _{t \rightarrow s} \boldsymbol{f}(t)=\boldsymbol{l}$ means that for all $\boldsymbol{\varepsilon}>0$, there exists $\delta>0$ such that if $0<|t-s|<\boldsymbol{\delta}$, then $\|\boldsymbol{f}(t)-\boldsymbol{l}\|_{\infty}<\boldsymbol{\varepsilon}$. You have to use the definition of a limit in order to establish that something is a limit.
3. A collection of functions $\mathscr{F}$ of $C\left([0, T], \mathbb{R}^{n}\right)$ is said to be uniformly equicontinuous if for every $\varepsilon>0$ there exists $\delta>0$ such that if $f \in \mathscr{F}$ and $|t-s|<\delta$, then $\|f(t)-\boldsymbol{f}(s)\|_{\infty}<\varepsilon$. Thus the functions are uniformly continuous all at once. The single $\delta$ works for every pair $t, s$ closer together than $\delta$ and for all functions $\boldsymbol{f} \in \mathscr{F}$. As an easy case, suppose there exists $K$ such that for all $f \in \mathscr{F}$,

$$
\|\boldsymbol{f}(t)-\boldsymbol{f}(s)\|_{\infty} \leq K|t-s|
$$

show that $\mathscr{F}$ is uniformly equicontinuous. Now suppose $\mathscr{G}$ is a collection of functions of $C\left([0, T], \mathbb{R}^{n}\right)$ which is bounded. That is,

$$
\|\boldsymbol{f}\|=\max _{t \in[0, T]}\|\boldsymbol{f}(t)\|_{\infty}<M<\infty
$$

for all $f \in \mathscr{G}$. Then let $\mathscr{F}$ denote the functions which are of the form

$$
\boldsymbol{F}(t) \equiv \boldsymbol{y}_{0}+\int_{0}^{t} \boldsymbol{f}(s) d s
$$

where $f \in \mathscr{G}$. Show that $\mathscr{F}$ is uniformly equicontinuous. Hint: This is a really easy problem if you do the right things. Here is the way you should proceed. Remember the triangle inequality from one variable calculus which said that for $a<b$ $\left|\int_{a}^{b} f(s) d s\right| \leq \int_{a}^{b}|f(s)| d s$. Then

$$
\left\|\int_{a}^{b} \boldsymbol{f}(s) d s\right\|_{\infty}=\max _{i}\left|\int_{a}^{b} f_{i}(s) d s\right| \leq \max _{i} \int_{a}^{b}\left|f_{i}(s)\right| d s \leq \int_{a}^{b}\|\boldsymbol{f}(s)\|_{\infty} d s
$$

Reduce to the case just considered using the assumption that these $f$ are bounded.
4. Let $V$ be a vector space with basis $\left\{v_{1}, \cdots, v_{n}\right\}$. For $v \in V$, denote its coordinate vector as $\boldsymbol{v}=\left(\alpha_{1}, \cdots, \alpha_{n}\right)$ where $v=\sum_{k=1}^{n} \alpha_{k} v_{k}$. Now define $\|v\| \equiv\|\boldsymbol{v}\|_{\infty}$. Show that this is a norm on $V$.
5. Let $(X,\|\cdot\|)$ be a normed linear space. A set $A$ is said to be convex if whenever $\boldsymbol{x}, \boldsymbol{y} \in$ $A$ the line segment determined by these points given by $t \boldsymbol{x}+(1-t) \boldsymbol{y}$ for $t \in[0,1]$ is also in $A$. Show that every open or closed ball is convex. Remember a closed ball is $D(\boldsymbol{x}, r) \equiv\{\hat{\boldsymbol{x}}:\|\hat{\boldsymbol{x}}-\boldsymbol{x}\| \leq r\}$ while the open ball is $B(\boldsymbol{x}, r) \equiv\{\hat{\boldsymbol{x}}:\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|<r\}$. This should work just as easily in any normed linear space.
6. A vector $v$ is in the convex hull of a nonempty set $S$ if there are finitely many vectors of $S,\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{m}\right\}$ and nonnegative scalars $\left\{t_{1}, \cdots, t_{m}\right\}$ such that

$$
\boldsymbol{v}=\sum_{k=1}^{m} t_{k} \boldsymbol{v}_{k}, \sum_{k=1}^{m} t_{k}=1
$$

Such a linear combination is called a convex combination. Suppose now that $S \subseteq V$, a vector space of dimension $n$. Show that if $\boldsymbol{v}=\sum_{k=1}^{m} t_{k} \boldsymbol{v}_{k}$ is a vector in the convex
hull for $m>n+1$, then there exist other nonnegative scalars $\left\{t_{k}^{\prime}\right\}$ summing to 1 such that $\boldsymbol{v}=\sum_{k=1}^{m-1} t_{k}^{\prime} \boldsymbol{v}_{k}$. Thus every vector in the convex hull of $S$ can be obtained as a convex combination of at most $n+1$ points of $S$. This incredible result is in Rudin [37]. Convexity is more a geometric property than a topological property. Hint: Consider $L: \mathbb{R}^{m} \rightarrow V \times \mathbb{R}$ defined by

$$
L(\boldsymbol{a}) \equiv\left(\sum_{k=1}^{m} a_{k} \boldsymbol{v}_{k}, \sum_{k=1}^{m} a_{k}\right)
$$

Explain why $\operatorname{ker}(L) \neq\{\mathbf{0}\}$. This will involve observing that $\mathbb{R}^{m}$ has higher dimension that $V \times \mathbb{R}$. Thus $L$ cannot be one to one because one to one functions take linearly independent sets to linearly independent sets and you can't have a linearly independent set with more than $n+1$ vectors in $V \times \mathbb{R}$. Next, letting $\boldsymbol{a} \in \operatorname{ker}(L) \backslash\{\mathbf{0}\}$ and $\lambda \in \mathbb{R}$, note that $\lambda \boldsymbol{a} \in \operatorname{ker}(L)$. Thus for all $\lambda \in \mathbb{R}$,

$$
\boldsymbol{v}=\sum_{k=1}^{m}\left(t_{k}+\lambda a_{k}\right) \boldsymbol{v}_{k}
$$

Now vary $\lambda$ till some $t_{k}+\lambda a_{k}=0$ for some $a_{k} \neq 0$. You can assume each $t_{k}>0$ since otherwise, there is nothing to show. This is a really nice result because it can be used to show that the convex hull of a compact set is also compact. Show this next. This is also Problem 22 but here it is again. This is because it is a really nice result.
7. Show that the usual norm in $\mathbb{F}^{n}$ given by

$$
|x|=(x, x)^{1 / 2}
$$

satisfies the following identities, the first of them being the parallelogram identity and the second being the polarization identity.

$$
\begin{aligned}
|\boldsymbol{x}+\boldsymbol{y}|^{2}+|\boldsymbol{x}-\boldsymbol{y}|^{2} & =2|\boldsymbol{x}|^{2}+2|\boldsymbol{y}|^{2} \\
\operatorname{Re}(\boldsymbol{x}, \boldsymbol{y}) & =\frac{1}{4}\left(|\boldsymbol{x}+\boldsymbol{y}|^{2}-|\boldsymbol{x}-\boldsymbol{y}|^{2}\right)
\end{aligned}
$$

Show that these identities hold in any inner product space, not just $\mathbb{F}^{n}$.
8. Let $K$ be a nonempty closed and convex set in an inner product space $(X,|\cdot|)$ which is complete. For example, $\mathbb{F}^{n}$ or any other finite dimensional inner product space. Let $y \notin K$ and let

$$
\lambda=\inf \{|y-x|: x \in K\}
$$

Let $\left\{x_{n}\right\}$ be a minimizing sequence. That is

$$
\lambda=\lim _{n \rightarrow \infty}\left|y-x_{n}\right|
$$

Explain why such a minimizing sequence exists. Next explain the following using the parallelogram identity in the above problem as follows.

$$
\left|y-\frac{x_{n}+x_{m}}{2}\right|^{2}=\left|\frac{y}{2}-\frac{x_{n}}{2}+\frac{y}{2}-\frac{x_{m}}{2}\right|^{2}
$$

$$
=-\left|\frac{y}{2}-\frac{x_{n}}{2}-\left(\frac{y}{2}-\frac{x_{m}}{2}\right)\right|^{2}+\frac{1}{2}\left|y-x_{n}\right|^{2}+\frac{1}{2}\left|y-x_{m}\right|^{2}
$$

Hence

$$
\begin{aligned}
\left|\frac{x_{m}-x_{n}}{2}\right|^{2} & =-\left|y-\frac{x_{n}+x_{m}}{2}\right|^{2}+\frac{1}{2}\left|y-x_{n}\right|^{2}+\frac{1}{2}\left|y-x_{m}\right|^{2} \\
& \leq-\lambda^{2}+\frac{1}{2}\left|y-x_{n}\right|^{2}+\frac{1}{2}\left|y-x_{m}\right|^{2}
\end{aligned}
$$

Next explain why the right hand side converges to 0 as $m, n \rightarrow \infty$. Thus $\left\{x_{n}\right\}$ is a Cauchy sequence and converges to some $x \in X$. Explain why $x \in K$ and $|x-y|=\lambda$. Thus there exists a closest point in $K$ to $y$. Next show that there is only one closest point. Hint: To do this, suppose there are two $x_{1}, x_{2}$ and consider $\frac{x_{1}+x_{2}}{2}$ using the parallelogram law to show that this average works better than either of the two points which is a contradiction unless they are really the same point. This theorem is of enormous significance.
9. Let $K$ be a closed convex nonempty set in a complete inner product space $(H,|\cdot|)$ (Hilbert space) and let $y \in H$. Denote the closest point to $y$ by $P x$. Show that $P x$ is characterized as being the solution to the following variational inequality

$$
\operatorname{Re}(z-P x, y-P x) \leq 0
$$

for all $z \in K$. Hint: Let $x \in K$. Then, due to convexity, a generic thing in $K$ is of the form $x+t(z-x), t \in[0,1]$ for every $z \in K$. Then

$$
|x+t(z-x)-y|^{2}=|x-y|^{2}+t^{2}|z-x|^{2}-t 2 \operatorname{Re}(z-x, y-x)
$$

If $x=P y$, then the minimum value of this on the left occurs when $t=0$. Function defined on $[0,1]$ has its minimum at $t=0$. What does it say about the derivative of this function at $t=0$ ? Next consider the case that for some $x$ the inequality $\operatorname{Re}(z-x, y-x) \leq 0$. Explain why this shows $x=P y$.
10. Using Problem 9 and Problem 8 show the projection map, $P$ onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is

$$
|P x-P y| \leq|x-y|
$$

11. Suppose, in an inner product space, you know $\operatorname{Re}(x, y)$. Show that you also know $\operatorname{Im}(x, y)$. That is, give a formula for $\operatorname{Im}(x, y)$ in terms of $\operatorname{Re}(x, y)$. Hint:

$$
\begin{aligned}
(x, i y) & =-i(x, y)=-i(\operatorname{Re}(x, y)+i \operatorname{Im}(x, y)) \\
& =-i \operatorname{Re}(x, y)+\operatorname{Im}(x, y)
\end{aligned}
$$

while, by definition,

$$
(x, i y)=\operatorname{Re}(x, i y)+i \operatorname{Im}(x, i y)
$$

Now consider matching real and imaginary parts.
12. Suppose $K$ is a compact subset (If $\mathscr{C}$ is a set of open sets whose union contains $K$, (open cover) then there are finitely many sets of $\mathscr{C}$ whose union contains $K$.) of $(X, d)$ a metric space. Also let $\mathscr{C}$ be an open cover of $K$. Show that there exists $\delta>0$ such that for all $x \in K, B(x, \delta)$ is contained in a single set of $\mathscr{C}$. This number is called a Lebesgue number. Hint: For each $x \in K$, there exists $B\left(x, \delta_{x}\right)$ such that this ball is contained in a set of $\mathscr{C}$. Now consider the balls $\left\{B\left(x, \frac{\delta_{x}}{2}\right)\right\}_{x \in K}$. Finitely many of these cover $K .\left\{B\left(x_{i}, \frac{\delta_{x_{i}}}{2}\right)\right\}_{i=1}^{n}$ Now consider what happens if you let $\delta \leq$ $\min \left\{\frac{\delta_{x_{i}}}{2}, i=1,2, \cdots, n\right\}$. Explain why this works. You might draw a picture to help get the idea.
13. Suppose $\mathscr{C}$ is a set of compact sets (A set is compact if every open cover admits a finite subcover.) in a metric space $(X, d)$ and suppose that the intersection of every finite subset of $\mathscr{C}$ is nonempty. This is called the finite intersection property. Show that $\cap \mathscr{C}$, the intersection of all sets of $\mathscr{C}$ is nonempty. This particular result is enormously important. Hint: You could let $\mathscr{U}$ denote the set $\left\{K^{C}: K \in \mathscr{C}\right\}$. If $\cap \mathscr{C}$ is empty, then its complement is $\cup \mathscr{U}=X$. Picking $K \in \mathscr{C}$, it follows that $\mathscr{U}$ is an open cover of $K$. Therefore, you would need to have $\left\{K_{1}^{C}, \cdots, K_{m}^{C}\right\}$ is a cover of $K$. In other words,

$$
K \subseteq \cup_{i=1}^{m} K_{i}^{C}=\left(\cap_{i=1}^{m} K_{i}\right)^{C}
$$

Now what does this say about the intersection of $K$ with these $K_{i}$.
14. Show that if $f$ is continuous and defined on a compact set $K$ in a metric space, then it is uniformly continuous. Continuous means continuous at every point. Uniformly continuous means: For every $\varepsilon>0$ there exists $\delta>0$ such that if $d(x, y)<\delta$, then $d(f(x), f(y))<\varepsilon$. The difference is that $\delta$ does not depend on $x$. Hint: Use the existence of the Lebesgue number in Problem 12 to prove continuity on a compact set $K$ implies uniform continuity on this set. Hint: Consider

$$
\mathscr{C} \equiv\left\{f^{-1}(B(f(x), \varepsilon / 2)): x \in X\right\} .
$$

This is an open cover of $X$. Let $\delta$ be a Lebesgue number for this open cover. Suppose $d(x, \hat{x})<\delta$. Then both $x, \hat{x}$ are in $B(x, \delta)$ and so both are in $f^{-1}\left(B\left(f(\bar{x}), \frac{\varepsilon}{2}\right)\right)$. Hence $\rho(f(x), f(\bar{x}))<\frac{\varepsilon}{2}$ and $\rho(f(\hat{x}), f(\bar{x}))<\frac{\varepsilon}{2}$. Now consider the triangle inequality. Recall the usual definition of continuity. In metric space it is as follows: For $(D, d),(Y, \rho)$ metric spaces, $f: D \rightarrow Y$ is continuous at $x \in D$ means that for all $\varepsilon>0$ there exists $\delta>0$ such that if $d(y, x)<\delta$, then $\rho(f(x), f(y))<\varepsilon$. Continuity on $D$ means continuity at every point of $D$.
15. The definition of compactness is that a set $K$ is compact if and only if every open cover (collection of open sets whose union contains $K$ ) has a finite subset which is also an open cover. Show that this is equivalent to saying that every open cover consisting of balls has a finite subset which is also an open cover.
16. A set $K$ in a metric space is said to be sequentially compact if whenever $\left\{x_{n}\right\}$ is a sequence in $K$, there exists a subsequence which converges to a point of $K$. Show that if $K$ is compact, then it is sequentially compact. Hint: Explain why if $x \in K$, then there exist an open set $B_{x}$ containing $x$ which has $x_{k}$ for only finitely many values of $k$. Then use compactness. This was shown in the chapter, but do your own proof of this part of it.
17. Show that $f: D \rightarrow Y$ is continuous at $x \in D$ where $(D, d),(Y, \rho)$ are metric spaces if and only if whenever $x_{n} \rightarrow x$ in $D$, it follows that $f\left(x_{n}\right) \rightarrow f(x)$. Recall the usual definition of continuity. $f$ is continuous at $x$ means that for all $\varepsilon>0$ there exists $\delta>0$ such that if $d(y, x)<\delta$, then $\rho(f(x), f(y))<\varepsilon$. Continuity on $D$ means continuity at every point of $D$. This is in the chapter, but go through the proof and write it down in your own words.
18. Give an easier proof of the result of Problem 14. Hint: If $f$ is not uniformly continuous, then there exists $\boldsymbol{\varepsilon}>0$ and $x_{n}, y_{n}, d\left(x_{n}, y_{n}\right)<\frac{1}{n}$ but $d\left(f\left(x_{n}\right), f\left(y_{n}\right)\right) \geq \varepsilon$. Now use sequential compactness of $K$ to get a contradiction.
19. This problem will reveal the best kept secret in undergraduate mathematics, the definition of the derivative of a function of $n$ variables. Let $\|\cdot\|_{V}$ be a norm on $V$ and also denote by $\|\cdot\|_{W}$ a norm on $W$. Write $\|\cdot\|$ for both to save notation. Let $U \subseteq V$ be an open set. Let $f: U \mapsto W$ be a function having values in $W$. Then $f$ is differentiable at $\boldsymbol{x} \in U$ means that there exists $A \in \mathscr{L}(V, W)$ such that for every $\varepsilon>0$, there exists a $\delta>0$ such that whenever $0<\|\boldsymbol{v}\|<\boldsymbol{\delta}$, it follows that

$$
\frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})-A \boldsymbol{v}\|}{\|\boldsymbol{v}\|}<\varepsilon
$$

Stated more simply,

$$
\lim _{\|\boldsymbol{v}\| \rightarrow 0} \frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})-A \boldsymbol{v}\|}{\|\boldsymbol{v}\|}=0
$$

Show that $A$ is unique. It is written as $D \boldsymbol{f}(\boldsymbol{x})=A$. This is what is meant by the derivative of $\boldsymbol{f}$. If $V=\mathbb{R}^{n}$, and $W=\mathbb{R}^{m}$, show that with respect to the usual bases, the matrix of $D \boldsymbol{f}(\boldsymbol{x})$ is an $m \times n$ matrix whose $k^{t h}$ column is $\frac{\partial \boldsymbol{f}}{\partial x_{k}}$.
20. Let $V, W$ be finite dimensional normed linear spaces and let $\left\{A_{n}\right\}_{n=1}^{\infty}$ be a sequence of linear transformations in $\mathscr{L}(V, W)$ such that $\sup _{n}\left\|A_{n}\right\|<\infty$. Show that there exists a subsequence $\left\{A_{n_{k}}\right\}$ and $A \in \mathscr{L}(V, W)$ such that $\lim _{n \rightarrow \infty}\left\|A_{n}-A\right\|=0$.
21. Given an example of a sequence $\left\{A_{k}\right\} \subseteq \mathscr{L}(V, V)$ such that the minimum polynomial of each $A_{k}$ has degree $n=\operatorname{dim}(V)$ but $\left\|A_{k}-A\right\| \rightarrow 0$ and the minimum polynomial of $A$ has degree less than $n$. Hint: You might want to think in terms of the Jordan form.
22. Let $U_{n}, n=1,2, \ldots$ be an open set in a complete metric space $(X, d)$. Suppose that $U_{n}$ is also dense so that $\bar{U}_{n}=X$. Show that $\cap_{n=1}^{\infty} U_{n}$ is dense. Hint: Start with $p \in X$ and form $B_{p}$ a ball containing $p$. There exists a point $x_{1}$ of $U_{1}$ in $B_{p}$. Now let $\bar{B}_{x_{1}} \subseteq B_{p} \cap U_{1}$ and let $B_{x_{1}}$ have diameter no more than $1 / 2^{1}$. Iterate this. This is called Bair's theorem.

## Chapter 11

## Limits of Vectors and Matrices

### 11.1 Regular Markov Matrices

The existence of the Jordan form is the basis for the proof of limit theorems for certain kinds of matrices called Markov matrices.

Definition 11.1.1 An $n \times n$ matrix $A=\left(a_{i j}\right)$, is a Markov matrix if $a_{i j} \geq 0$ for all $i, j$ and

$$
\sum_{i} a_{i j}=1 .
$$

It may also be called a stochastic matrix or a transition matrix. A Markov or stochastic matrix is called regular if some power of $A$ has all entries strictly positive. A vector $\boldsymbol{v} \in \mathbb{R}^{n}$, is a steady state if $A \boldsymbol{v}=\boldsymbol{v}$.

Lemma 11.1.2 The property of being a stochastic matrix is preserved by taking products. It is also true if the sum is of the form $\sum_{j} a_{i j}=1$.

Proof: Suppose the sum over a row equals 1 for $A$ and $B$. Then letting the entries be denoted by $\left(a_{i j}\right)$ and $\left(b_{i j}\right)$ respectively and the entries of $A B$ by $\left(c_{i j}\right)$,

$$
\sum_{i} c_{i j}=\sum_{i} \sum_{k} a_{i k} b_{k j}=\sum_{k} \sum_{i} a_{i k} b_{k j}=\sum_{k} b_{k j}=1
$$

It is obvious that when the product is taken, if each $a_{i j}, b_{i j} \geq 0$, then the same will be true of sums of products of these numbers. Similar reasoning works for the assumption that $\sum_{j} a_{i j}=1$.

The following theorem is convenient for showing the existence of limits.
Theorem 11.1.3 Let A be a real $p \times p$ matrix having the properties

1. $a_{i j} \geq 0$
2. Either $\sum_{i=1}^{p} a_{i j}=1$ or $\sum_{j=1}^{p} a_{i j}=1$.
3. The distinct eigenvalues of $A$ are $\left\{1, \lambda_{2}, \ldots, \lambda_{m}\right\}$ where each $\left|\lambda_{j}\right|<1$.

Then $\lim _{n \rightarrow \infty} A^{n}=A_{\infty}$ exists in the sense that $\lim _{n \rightarrow \infty} a_{i j}^{n}=a_{i j}^{\infty}$, the $i j^{\text {th }}$ entry $A_{\infty}$. Here $a_{i j}^{n}$ denotes the $i j$ th entry of $A^{n}$. Also, if $\lambda=1$ has algebraic multiplicity $r$, then the Jordan block corresponding to $\lambda=1$ is just the $r \times r$ identity.

Proof. By the existence of the Jordan form for $A$, it follows that there exists an invertible matrix $P$ such that

$$
P^{-1} A P=\left(\begin{array}{cccc}
I+N & & & \\
& J_{r_{2}}\left(\lambda_{2}\right) & & \\
& & \ddots & \\
& & & J_{r_{m}}\left(\lambda_{m}\right)
\end{array}\right)=J
$$

where $I$ is $r \times r$ for $r$ the multiplicity of the eigenvalue 1 and $N$ is a nilpotent matrix for which $N^{r}=0$. I will show that because of Condition $2, N=0$.

First of all,

$$
J_{r_{i}}\left(\lambda_{i}\right)=\lambda_{i} I+N_{i}
$$

where $N_{i}$ satisfies $N_{i}^{r_{i}}=0$ for some $r_{i}>0$. It is clear that $N_{i}\left(\lambda_{i} I\right)=\left(\lambda_{i} I\right) N$ and so

$$
\left(J_{r_{i}}\left(\lambda_{i}\right)\right)^{n}=\sum_{k=0}^{n}\binom{n}{k} N^{k} \lambda_{i}^{n-k}=\sum_{k=0}^{r_{i}}\binom{n}{k} N^{k} \lambda_{i}^{n-k}
$$

which converges to 0 due to the assumption that $\left|\lambda_{i}\right|<1$. There are finitely many terms and a typical one is a matrix whose entries are no larger than an expression of the form

$$
\left|\lambda_{i}\right|^{n-k} C_{k} n(n-1) \cdots(n-k+1) \leq C_{k}\left|\lambda_{i}\right|^{n-k} n^{k}
$$

which converges to 0 because, by the root test, the series $\sum_{n=1}^{\infty}\left|\lambda_{i}\right|^{n-k} n^{k}$ converges. Thus for each $i=2, \ldots, p$,

$$
\lim _{n \rightarrow \infty}\left(J_{r_{i}}\left(\lambda_{i}\right)\right)^{n}=0
$$

By Condition 2, if $a_{i j}^{n}$ denotes the $i j^{t h}$ entry of $A^{n}$, then either

$$
\sum_{i=1}^{p} a_{i j}^{n}=1 \text { or } \sum_{j=1}^{p} a_{i j}^{n}=1, a_{i j}^{n} \geq 0
$$

This follows from Lemma 11.1.2. It is obvious each $a_{i j}^{n} \geq 0$, and so the entries of $A^{n}$ must be bounded independent of $n$.

It follows easily from

$$
\overbrace{P^{-1} A P P^{-1} A P P^{-1} A P \cdots P^{-1} A P}^{n \text { times }}=P^{-1} A^{n} P
$$

that

$$
\begin{equation*}
P^{-1} A^{n} P=J^{n} \tag{11.1}
\end{equation*}
$$

Hence $J^{n}$ must also have bounded entries as $n \rightarrow \infty$. However, this requirement is incompatible with an assumption that $N \neq 0$.

If $N \neq 0$, then $N^{s} \neq 0$ but $N^{s+1}=0$ for some $1 \leq s \leq r$. Then

$$
(I+N)^{n}=I+\sum_{k=1}^{s}\binom{n}{k} N^{k}
$$

One of the entries of $N^{s}$ is nonzero by the definition of $s$. Let this entry be $n_{i j}^{s}$. Then this implies that one of the entries of $(I+N)^{n}$ is of the form $\binom{n}{s} n_{i j}^{s}$. This entry dominates the $i j^{\text {th }}$ entries of $\binom{n}{k} N^{k}$ for all $k<s$ because

$$
\lim _{n \rightarrow \infty}\binom{n}{s} /\binom{n}{k}=\infty
$$

Therefore, the entries of $(I+N)^{n}$ cannot all be bounded. From block multiplication,

$$
P^{-1} A^{n} P=\left(\begin{array}{llll}
(I+N)^{n} & & & \\
& \left(J_{r_{2}}\left(\lambda_{2}\right)\right)^{n} & & \\
& & \ddots & \\
& & & \left(J_{r_{m}}\left(\lambda_{m}\right)\right)^{n}
\end{array}\right)
$$

and this is a contradiction because entries are bounded on the left and unbounded on the right.

Since $N=0$, the above equation implies $\lim _{n \rightarrow \infty} A^{n}$ exists and equals

$$
P\left(\begin{array}{cccc}
I & & & \\
& 0 & & \\
& & \ddots & \\
& & & 0
\end{array}\right) P^{-1} \square
$$

Are there examples which will cause the eigenvalue condition of this theorem to hold? The following lemma gives such a condition. It turns out that if $a_{i j}>0$, not just $\geq 0$, then the eigenvalue condition of the above theorem is valid.

Lemma 11.1.4 Suppose $A=\left(a_{i j}\right)$ is a stochastic matrix. Then $\lambda=1$ is an eigenvalue. If $a_{i j}>0$ for all $i, j$, then if $\mu$ is an eigenvalue of $A$, either $|\mu|<1$ or $\mu=1$.

Proof: First consider the claim that 1 is an eigenvalue. By definition,

$$
\sum_{i} 1 a_{i j}=1
$$

and so $A^{T} \boldsymbol{v}=\boldsymbol{v}$ where $\boldsymbol{v}=\left(\begin{array}{lll}1 & \cdots & 1\end{array}\right)^{T}$. Since $A, A^{T}$ have the same eigenvalues, this shows 1 is an eigenvalue of $A$. Suppose then that $\mu$ is an eigenvalue. Is $|\mu|<1$ or $\mu=1$ ? Let $\boldsymbol{v}$ be an eigenvector for $A^{T}$ and let $\left|v_{i}\right|$ be the largest of the $\left|v_{j}\right|$.

$$
\mu v_{i}=\sum_{j} a_{j i} v_{j}
$$

and now multiply both sides by $\bar{\mu} \overline{v_{i}}$ to obtain

$$
\begin{aligned}
|\mu|^{2}\left|v_{i}\right|^{2} & =\sum_{j} a_{j i} v_{j} \bar{\mu} \overline{v_{i}}=\sum_{j} a_{j i} \operatorname{Re}\left(v_{j} \bar{\mu} \overline{v_{i}}\right) \\
& \leq \sum_{j} a_{j i}\left|v_{i}\right|^{2}|\mu|=|\mu|\left|v_{i}\right|^{2}
\end{aligned}
$$

Therefore, $|\mu| \leq 1$. If $|\mu|=1$, then equality must hold in the above, and so $v_{j} \overline{v_{i}} \bar{\mu}$ must be real and nonnegative for each $j$. In particular, this holds for $j=i$ which shows $\bar{\mu}$ is real and nonnegative. Thus, in this case, $\mu=1$ because $\bar{\mu}=\mu$ is nonnegative and equal to 1 . The only other case is where $|\mu|<1$.

The next lemma is sort of a conservation result. It says the sign and sum of entries of a vector are preserved when multiplying by a Markov matrix.

Lemma 11.1.5 Let $A$ be any Markov matrix and let $\boldsymbol{v}$ be a vector having all its components non negative with $\sum_{i} v_{i}=c$. Then if $\boldsymbol{w}=A \boldsymbol{v}$, it follows that $w_{i} \geq 0$ for all $i$ and $\sum_{i} w_{i}=c$.

Proof: From the definition of $\boldsymbol{w}$,

$$
w_{i} \equiv \sum_{j} a_{i j} v_{j} \geq 0
$$

Also

$$
\sum_{i} w_{i}=\sum_{i} \sum_{j} a_{i j} v_{j}=\sum_{j} \sum_{i} a_{i j} v_{j}=\sum_{j} v_{j}=c .
$$

The following theorem about limits is now easy to obtain.
Theorem 11.1.6 Suppose $A$ is a Markov matrix in which $a_{i j}>0$ for all $i, j$ and suppose $\boldsymbol{w}$ is a vector. Then for each $i$,

$$
\lim _{k \rightarrow \infty}\left(A^{k} \boldsymbol{w}\right)_{i}=v_{i}
$$

where $A \boldsymbol{v}=\boldsymbol{v}$. In words, $A^{k} \boldsymbol{w}$ always converges to a steady state. In addition to this, if the vector $\boldsymbol{w}$ satisfies $w_{i} \geq 0$ for all $i$ and $\sum_{i} w_{i}=c$, then the vector $\boldsymbol{v}$ will also satisfy the conditions, $v_{i} \geq 0, \sum_{i} v_{i}=c$.

Proof: By Lemma 11.1.4, since each $a_{i j}>0$, the eigenvalues are either 1 or have absolute value less than 1. Therefore, the claimed limit exists by Theorem 11.1.3. The assertion that the components are nonnegative and sum to $c$ follows from Lemma 11.1.5. That $A v=v$ follows from

$$
\boldsymbol{v}=\lim _{n \rightarrow \infty} A^{n} \boldsymbol{w}=\lim _{n \rightarrow \infty} A^{n+1} \boldsymbol{w}=A \lim _{n \rightarrow \infty} A^{n} \boldsymbol{w}=A \boldsymbol{v}
$$

It is not hard to generalize the conclusion of this theorem to regular Markov processes which are those having some power with all positive entries.
Corollary 11.1.7 Suppose $A$ is a regular Markov matrix, one for which the entries of $A^{k}$ are all positive for some $k$, and suppose $\boldsymbol{w}$ is a vector. Then for each $i$,

$$
\lim _{n \rightarrow \infty}\left(A^{n} \boldsymbol{w}\right)_{i}=v_{i}
$$

where $A \boldsymbol{v}=\boldsymbol{v}$. In words, $A^{n} \boldsymbol{w}$ always converges to a steady state. In addition to this, if the vector $\boldsymbol{w}$ satisfies $w_{i} \geq 0$ for all $i$ and $\sum_{i} w_{i}=c$, Then the vector $\boldsymbol{v}$ will also satisfy the conditions $v_{i} \geq 0, \sum_{i} v_{i}=c$.

Proof: Let the entries of $A^{k}$ be all positive for some $k$. Now suppose that $a_{i j} \geq 0$ for all $i, j$ and $A=\left(a_{i j}\right)$ is a Markov matrix. Then if $B=\left(b_{i j}\right)$ is a Markov matrix with $b_{i j}>0$ for all $i j$, it follows that $B A$ is a Markov matrix which has strictly positive entries. This is because the $i j^{\text {th }}$ entry of $B A$ is

$$
\sum_{k} b_{i k} a_{k j}>0
$$

Thus, from Lemma 11.1.4, $A^{k}$ has eigenvalues $\left\{1, \lambda_{1}, \cdots, \lambda_{r}\right\},\left|\lambda_{r}\right|<1$. The same must be true of $A$. If $A \boldsymbol{x}=\mu \boldsymbol{x}$ for $\boldsymbol{x} \neq \mathbf{0}$ and $\mu \neq 1$, Then $A^{k} \boldsymbol{x}=\mu^{k} \boldsymbol{x}$ and so either $\mu^{k}=1$ or $|\mu|<$ 1. If $\mu^{k}=1$, then $|\mu|=1$ and the eigenvalues of $A^{k+1}$ are either 1 or have absolute value less than 1 because $A^{k+1}$ has all postive entries thanks to Lemma 11.1.4. Thus $\mu^{k+1}=1$ and so

$$
1=\mu^{k+1}=\mu \mu^{k}=\mu
$$

By Theorem 11.1.3, $\lim _{n \rightarrow \infty} A^{n} \boldsymbol{w}$ exists. The rest follows as in Theorem 11.1.6.

### 11.2 Migration Matrices

Definition 11.2.1 Let $n$ locations be denoted by the numbers $1,2, \cdots, n$. Also suppose it is the case that each year $a_{i j}$ denotes the proportion of residents in location $j$ which move to location $i$. Also suppose no one escapes or emigrates from without these $n$ locations. This last assumption requires $\sum_{i} a_{i j}=1$. Thus $\left(a_{i j}\right)$ is a Markov matrix referred to as a migration matrix.

If $\boldsymbol{v}=\left(x_{1}, \cdots, x_{n}\right)^{T}$ where $x_{i}$ is the population of location $i$ at a given instant, you obtain the population of location $i$ one year later by computing $\sum_{j} a_{i j} x_{j}=(A \boldsymbol{v})_{i}$. Therefore, the population of location $i$ after $k$ years is $\left(A^{k} \boldsymbol{v}\right)_{i}$. Furthermore, Corollary 11.1.7 can be used to predict in the case where $A$ is regular what the long time population will be for the given locations.

As an example of the above, consider the case where $n=3$ and the migration matrix is of the form

$$
\left(\begin{array}{ccc}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & .2 & .9
\end{array}\right)
$$

Now

$$
\left(\begin{array}{lll}
.6 & 0 & .1 \\
.2 & .8 & 0 \\
.2 & .2 & .9
\end{array}\right)^{2}=\left(\begin{array}{lll}
.38 & .02 & .15 \\
.28 & .64 & .02 \\
.34 & .34 & .83
\end{array}\right)
$$

and so the Markov matrix is regular. Therefore, $\left(A^{k} v\right)_{i}$ will converge to the $i^{t h}$ component of a steady state. It follows the steady state can be obtained from solving the system

$$
\begin{gathered}
.6 x+.1 z=x \\
.2 x+.8 y=y \\
.2 x+.2 y+.9 z=z
\end{gathered}
$$

along with the stipulation that the sum of $x, y$, and $z$ must equal the constant value present at the beginning of the process. The solution to this system is

$$
\{y=x, z=4 x, x=x\}
$$

If the total population at the beginning is 150,000 , then you solve the following system

$$
y=x, z=4 x, x+y+z=150000
$$

whose solution is easily seen to be $\{x=25000, z=100000, y=25000\}$. Thus, after a long time there would be about four times as many people in the third location as in either of the other two.

### 11.3 Absorbing States

There is a different kind of Markov process containing so called absorbing states which result in transition matrices which are not regular. However, Theorem 11.1.3 may still apply. One such example is the Gambler's ruin problem. There is a total amount of money
denoted by $b$. The Gambler starts with an amount $j>0$ and gambles till he either loses everything or gains everything. He does this by playing a game in which he wins with probability $p$ and loses with probability $q$. When he wins, the amount of money he has increases by 1 and when he loses, the amount of money he has decreases by 1 . Thus the states are the integers from 0 to $b$. Let $p_{i j}$ denote the probability that the gambler has $i$ at the end of a game given that he had $j$ at the beginning. Let $p_{i j}^{n}$ denote the probability that the gambler has $i$ after $n$ games given that he had $j$ initially. Thus

$$
p_{i j}^{n+1}=\sum_{k} p_{i k} p_{k j}^{n}
$$

and so $p_{i j}^{n}$ is the $i j^{t h}$ entry of $P^{n}$ where $P$ is the transition matrix. The above description indicates that this transition probability matrix is of the form

$$
P=\left(\begin{array}{cccccc}
1 & q & 0 & 0 & \cdots & 0  \tag{11.2}\\
0 & 0 & q & & & 0 \\
0 & p & 0 & \ddots & & \vdots \\
\vdots & & \ddots & \ddots & q & 0 \\
0 & & & p & 0 & 0 \\
0 & 0 & \cdots & 0 & p & 1
\end{array}\right)
$$

The absorbing states are 0 and $b$. In the first, the gambler has lost everything and hence has nothing else to gamble, so the process stops. In the second, he has won everything and there is nothing else to gain, so again the process stops.

Consider the eigenvalues of this matrix which is a piece of the above transition matrix.
Lemma 11.3.1 Let $p, q>0$ and $p+q=1$. Then the eigenvalues of

$$
A \equiv\left(\begin{array}{cccc}
0 & q & & 0 \\
p & 0 & \ddots & \\
& \ddots & \ddots & q \\
0 & & p & 0
\end{array}\right)
$$

have absolute value less than 1.
Proof: By Gerschgorin's theorem, (See Page 139.) if $\lambda$ is an eigenvalue, then $|\lambda| \leq 1$. Alternatively, you note that $\sum_{i} A_{i j} \leq 1$. If $\lambda$ is an eigenvalue of $A$ then it is also one for $A^{T}$ and if $A^{T} \boldsymbol{x}=\lambda \boldsymbol{x}$ where $\left|x_{i}\right|$ is the largest of the $\left|x_{j}\right|$,

$$
\sum_{j} A_{j i} x_{j}=\lambda x_{i}, \quad|\lambda|\left|x_{i}\right| \leq \sum_{j} A_{j i}\left|x_{j}\right| \leq\left|x_{i}\right| \text { so }|\lambda| \leq 1 .
$$

Now suppose $\boldsymbol{v}$ is an eigenvector for $\lambda$. Then

$$
A \boldsymbol{v}=\left(\begin{array}{c}
q v_{2} \\
p v_{1}+q v_{3} \\
\vdots \\
p v_{n-2}+q v_{n} \\
p v_{n-1}
\end{array}\right)=\lambda\left(\begin{array}{c}
v_{1} \\
v_{2} \\
\vdots \\
v_{n-1} \\
v_{n}
\end{array}\right) .
$$

Suppose $|\lambda|=1$. Let $v_{k}$ be the first nonzero entry. Then $v_{k-1}=0$ and so

$$
q v_{k+1}=\lambda v_{k}
$$

which implies $\left|v_{k+1}\right|>\left|v_{k}\right|$. Thus $\left|v_{k+1}\right| \geq\left|v_{k}\right|$. Then consider the next term. From the above equations and what was just shown,

$$
\left|v_{k+1}\right|=\left|p v_{k}+q v_{k+1}\right| \leq p\left|v_{k}\right|+q\left|v_{k+2}\right| \leq p\left|v_{k+1}\right|+q\left|v_{k+2}\right|
$$

and so

$$
q\left|v_{k+1}\right| \leq q\left|v_{k+2}\right|
$$

Continuing this way, it follows that the sequence $\left\{\left|v_{j}\right|\right\}_{j=k}^{n}$ must be increasing. Specifically, if $\left\{\left|v_{j}\right|\right\}_{j=k}^{m}$ is increasing for some $m>k$, then

$$
p\left|v_{m-1}\right|+q\left|v_{m}\right| \geq\left|p v_{m-2}+q v_{m}\right|=\left|\lambda v_{m-1}\right|=\left|v_{m-1}\right|
$$

and so $q\left|v_{m}\right| \geq q\left|v_{m-1}\right|$. Hence $\left|v_{n}\right| \geq\left|v_{n-1}\right|>0$. However, this is contradicted by the the last line which states that $p\left|v_{n-1}\right|=\left|v_{n}\right|$ which requires that $\left|v_{n-1}\right|>\left|v_{n}\right|$, a contradiction. Therefore, it must be that $|\lambda|<1$.

Now consider the eigenvalues of 11.2. For $P$ given there,

$$
P-\lambda I=\left(\begin{array}{ccccc}
1-\lambda & q & 0 & \cdots & 0 \\
0 & -\lambda & \ddots & & 0 \\
0 & p & \ddots & q & \vdots \\
\vdots & & \ddots & -\lambda & 0 \\
0 & \cdots & 0 & p & 1-\lambda
\end{array}\right)
$$

and so, expanding the determinant of the matrix along the first column and then along the last column yields

$$
(1-\lambda)^{2} \operatorname{det}\left(\begin{array}{cccc}
-\lambda & q & & \\
p & \ddots & \ddots & \\
& \ddots & -\lambda & q \\
& & p & -\lambda
\end{array}\right)
$$

The roots of the polynomial after $(1-\lambda)^{2}$ have absolute value less than 1 because they are just the eigenvalues of a matrix of the sort in Lemma 11.3.1. It follows that the conditions of Theorem 11.1.3 apply and therefore, $\lim _{n \rightarrow \infty} P^{n}$ exists.

Of course, the above transition matrix, models many other kinds of problems. It is called a Markov process with two absorbing states, sometimes a random walk with two aborbing states.

It is interesting to find the probability that the gambler loses all his money. This is given by $\lim _{n \rightarrow \infty} p_{0 j}^{n}$. From the transition matrix for the gambler's ruin problem, it follows that

$$
\begin{aligned}
& p_{0 j}^{n}=\sum_{k} p_{0 k}^{n-1} p_{k j}=q p_{0(j-1)}^{n-1}+p p_{0(j+1)}^{n-1} \text { for } j \in[1, b-1] \\
& p_{00}^{n}=1, \text { and } p_{0 b}^{n}=0
\end{aligned}
$$

Assume here that $p \neq q$. Now it was shown above that $\lim _{n \rightarrow \infty} p_{0 j}^{n}$ exists. Denote by $P_{j}$ this limit. Then the above becomes much simpler if written as

$$
\begin{align*}
P_{j} & =q P_{j-1}+p P_{j+1} \text { for } j \in[1, b-1]  \tag{11.3}\\
P_{0} & =1 \text { and } P_{b}=0 \tag{11.4}
\end{align*}
$$

It is only required to find a solution to the above difference equation with boundary conditions. To do this, look for a solution in the form $P_{j}=r^{j}$ and use the difference equation with boundary conditions to find the correct values of $r$. Thus you need

$$
r^{j}=q r^{j-1}+p r^{j+1}
$$

and so to find $r$ you need to have $p r^{2}-r+q=0$, and so the solutions for $r$ are $r=$

$$
\frac{1}{2 p}(1+\sqrt{1-4 p q}), \frac{1}{2 p}(1-\sqrt{1-4 p q})
$$

Now

$$
\sqrt{1-4 p q}=\sqrt{1-4 p(1-p)}=\sqrt{1-4 p+4 p^{2}}=1-2 p
$$

Thus the two values of $r$ simplify to

$$
\frac{1}{2 p}(1+1-2 p)=\frac{q}{p}, \frac{1}{2 p}(1-(1-2 p))=1
$$

Therefore, for any choice of $C_{i}, i=1,2$,

$$
C_{1}+C_{2}\left(\frac{q}{p}\right)^{j}
$$

will solve the difference equation. Now choose $C_{1}, C_{2}$ to satisfy the boundary conditions 11.4. Thus you need to have

$$
C_{1}+C_{2}=1, C_{1}+C_{2}\left(\frac{q}{p}\right)^{b}=0
$$

It follows that

$$
C_{2}=\frac{p^{b}}{p^{b}-q^{b}}, \quad C_{1}=\frac{q^{b}}{q^{b}-p^{b}}
$$

Thus $P_{j}=$

$$
\frac{q^{b}}{q^{b}-p^{b}}+\frac{p^{b}}{p^{b}-q^{b}}\left(\frac{q}{p}\right)^{j}=\frac{q^{b}}{q^{b}-p^{b}}-\frac{p^{b-j} q^{j}}{q^{b}-p^{b}}=\frac{q^{j}\left(q^{b-j}-p^{b-j}\right)}{q^{b}-p^{b}}
$$

To find the solution in the case of a fair game, one could take the $\lim _{p \rightarrow 1 / 2}$ of the above solution. Taking this limit, you get

$$
P_{j}=\frac{b-j}{b}
$$

You could also verify directly in the case where $p=q=1 / 2$ in 11.3 and 11.4 that $P_{j}=1$ and $P_{j}=j$ are two solutions to the difference equation and proceeding as before.

### 11.4 Positive Matrices

Earlier theorems about Markov matrices were presented. These were matrices in which all the entries were nonnegative and either the columns or the rows added to 1. It turns out that many of the theorems presented can be generalized to positive matrices. When this is done, the resulting theory is mainly due to Perron and Frobenius. I will give an introduction to this theory here following Karlin and Taylor [27].

Definition 11.4.1 For A a matrix or vector, the notation, $A \gg 0$ will mean every entry of $A$ is positive. By $A>0$ is meant that every entry is nonnegative and at least one is positive. By $A \geq 0$ is meant that every entry is nonnegative. Thus the matrix or vector consisting only of zeros is $\geq 0$. An expression like $A \gg B$ will mean $A-B \gg 0$ with similar modifications for $>$ and $\geq$.

For the sake of this section only, define the following for $\boldsymbol{x}=\left(x_{1}, \cdots, x_{n}\right)^{T}$, a vector.

$$
|\boldsymbol{x}| \equiv\left(\left|x_{1}\right|, \cdots,\left|x_{n}\right|\right)^{T}
$$

Thus $|\boldsymbol{x}|$ is the vector which results by replacing each entry of $\boldsymbol{x}$ with its absolute value ${ }^{1}$. Also define for $\boldsymbol{x} \in \mathbb{C}^{n}$,

$$
\|\boldsymbol{x}\|_{1} \equiv \sum_{k}\left|x_{k}\right| .
$$

Lemma 11.4.2 Let $A \gg 0$ and let $\boldsymbol{x}>\mathbf{0}$. Then $A \boldsymbol{x} \gg \boldsymbol{0}$.
Proof: $(A \boldsymbol{x})_{i}=\sum_{j} A_{i j} x_{j}>0$ because all the $A_{i j}>0$ and at least one $x_{j}>0$.
Lemma 11.4.3 Let $A \gg 0$. Define

$$
S \equiv\{\boldsymbol{\lambda}: A \boldsymbol{x}>\boldsymbol{\lambda} \boldsymbol{x} \text { for some } \boldsymbol{x} \gg \boldsymbol{0}\}
$$

and let

$$
K \equiv\left\{\boldsymbol{x} \geq \mathbf{0} \text { such that }\|\boldsymbol{x}\|_{1}=1\right\}
$$

Now define

$$
S_{1} \equiv\{\boldsymbol{\lambda}: A \boldsymbol{x} \geq \lambda \boldsymbol{x} \text { for some } \boldsymbol{x} \in K\}
$$

Then

$$
\sup (S)=\sup \left(S_{1}\right)
$$

Proof: Let $\lambda \in S$. Then there exists $\boldsymbol{x} \gg \boldsymbol{0}$ such that $A \boldsymbol{x}>\boldsymbol{\lambda} \boldsymbol{x}$. Consider the unit vector $\boldsymbol{y} \equiv \boldsymbol{x} /\|\boldsymbol{x}\|_{1}$. Then $\|\boldsymbol{y}\|_{1}=1$ and $A \boldsymbol{y}>\lambda \boldsymbol{y}$. Therefore, $\lambda \in S_{1}$ and so $S \subseteq S_{1}$. Therefore, $\sup (S) \leq \sup \left(S_{1}\right)$.

Now let $\lambda \in S_{1}$. Then there exists $\boldsymbol{x} \geq \mathbf{0}$ such that $\|\boldsymbol{x}\|_{1}=1$ so $\boldsymbol{x}>\boldsymbol{0}$ and $A \boldsymbol{x}>\lambda \boldsymbol{x}$. Letting $\boldsymbol{y} \equiv A \boldsymbol{x}$, it follows from Lemma 11.4.2 that $A \boldsymbol{y} \gg \boldsymbol{\lambda} \boldsymbol{y}$ and $\boldsymbol{y} \gg \boldsymbol{0}$. Thus $\lambda \in S$ and so $S_{1} \subseteq S$ which shows that $\sup \left(S_{1}\right) \leq \sup (S)$.

This lemma is significant because the set, $\left\{\boldsymbol{x} \geq \mathbf{0}\right.$ such that $\left.\|\boldsymbol{x}\|_{1}=1\right\} \equiv K$ is a compact set in $\mathbb{R}^{n}$. Define

$$
\begin{equation*}
\lambda_{0} \equiv \sup (S)=\sup \left(S_{1}\right) \tag{11.5}
\end{equation*}
$$

The following theorem is due to Perron.

[^9]Theorem 11.4.4 Let $A \gg 0$ be an $n \times n$ matrix and let $\lambda_{0}$ be given in 11.5. Then

1. $\lambda_{0}>0$ and there exists $x_{0} \gg 0$ such that $A x_{0}=\lambda_{0} x_{0}$ so $\lambda_{0}$ is an eigenvalue for $A$.
2. If $A \boldsymbol{x}=\mu \boldsymbol{x}$ where $\boldsymbol{x} \neq \mathbf{0}$, and $\mu \neq \lambda_{0}$. Then $|\mu|<\lambda_{0}$.
3. The eigenspace for $\lambda_{0}$ has dimension 1 .

Proof: To see $\lambda_{0}>0$, consider the vector, $e \equiv(1, \cdots, 1)^{T}$. Then

$$
(A \boldsymbol{e})_{i}=\sum_{j} A_{i j}>0
$$

and so $\lambda_{0}$ is at least as large as

$$
\min _{i} \sum_{j} A_{i j}
$$

Let $\left\{\lambda_{k}\right\}$ be an increasing sequence of numbers from $S_{1}$ converging to $\lambda_{0}$. Letting $\boldsymbol{x}_{k}$ be the vector from $K$ which occurs in the definition of $S_{1}$, these vectors are in a compact set. Therefore, there exists a subsequence, still denoted by $\boldsymbol{x}_{k}$ such that $\boldsymbol{x}_{k} \rightarrow \boldsymbol{x}_{0} \in K$ and $\lambda_{k} \rightarrow \lambda_{0}$. Then passing to the limit,

$$
A x_{0} \geq \lambda_{0} x_{0}, \boldsymbol{x}_{0}>\mathbf{0}
$$

If $A \boldsymbol{x}_{0}>\lambda_{0} \boldsymbol{x}_{0}$, then letting $\boldsymbol{y} \equiv A \boldsymbol{x}_{0}$, it follows from Lemma 11.4.2 that $A \boldsymbol{y} \gg \lambda_{0} \boldsymbol{y}$ and $\boldsymbol{y} \gg \boldsymbol{0}$. But this contradicts the definition of $\lambda_{0}$ as the supremum of the elements of $S$ because since $A \boldsymbol{y} \gg \lambda_{0} \boldsymbol{y}$, it follows $A \boldsymbol{y} \gg\left(\lambda_{0}+\boldsymbol{\varepsilon}\right) \boldsymbol{y}$ for $\varepsilon$ a small positive number. Therefore, $A \boldsymbol{x}_{0}=\lambda_{0} \boldsymbol{x}_{0}$. It remains to verify that $\boldsymbol{x}_{0} \gg \boldsymbol{0}$. But this follows immediately from

$$
0<\sum_{j} A_{i j} x_{0 j}=\left(A x_{0}\right)_{i}=\lambda_{0} x_{0 i}
$$

This proves 1.
Next suppose $A \boldsymbol{x}=\mu \boldsymbol{x}$ and $\boldsymbol{x} \neq \mathbf{0}$ and $\mu \neq \lambda_{0}$. Then $|A \boldsymbol{x}|=|\mu||\boldsymbol{x}|$. But this implies $A|\boldsymbol{x}| \geq|\boldsymbol{\mu}||\boldsymbol{x}|$. (See the above abominable definition of $|\boldsymbol{x}|$.)

Case 1: $|\boldsymbol{x}| \neq \boldsymbol{x}$ and $|\boldsymbol{x}| \neq-\boldsymbol{x}$.
In this case, $A|\boldsymbol{x}|>|A \boldsymbol{x}|=|\mu||\boldsymbol{x}|$ and letting $\boldsymbol{y}=A|\boldsymbol{x}|$, it follows $\boldsymbol{y} \gg \boldsymbol{0}$ and $A \boldsymbol{y} \gg|\mu| \boldsymbol{y}$ which shows $A \boldsymbol{y} \gg(|\mu|+\boldsymbol{\varepsilon}) \boldsymbol{y}$ for sufficiently small positive $\varepsilon$ and verifies $|\mu|<\lambda_{0}$.

Case 2: $|\boldsymbol{x}|=\boldsymbol{x}$ or $|\boldsymbol{x}|=-\boldsymbol{x}$
In this case, the entries of $\boldsymbol{x}$ are all real and have the same sign. Therefore, $A|\boldsymbol{x}|=$ $|A \boldsymbol{x}|=|\mu||\boldsymbol{x}|$. Now let $\boldsymbol{y} \equiv|\boldsymbol{x}| /\|\boldsymbol{x}\|_{1}$. Then $A \boldsymbol{y}=|\mu| \boldsymbol{y}$ and so $|\boldsymbol{\mu}| \in S_{1}$ showing that $|\mu| \leq \lambda_{0}$. But also, the fact the entries of $\boldsymbol{x}$ all have the same sign shows $\mu=|\mu|$ and so $\mu \in S_{1}$. Since $\mu \neq \lambda_{0}$, it must be that $\mu=|\mu|<\lambda_{0}$. This proves 2 .

It remains to verify 3. Suppose then that $A \boldsymbol{y}=\lambda_{0} \boldsymbol{y}$ and for all scalars $\alpha, \alpha \boldsymbol{x}_{0} \neq \boldsymbol{y}$. Then

$$
A \operatorname{Re} \boldsymbol{y}=\lambda_{0} \operatorname{Re} \boldsymbol{y}, A \operatorname{Im} \boldsymbol{y}=\lambda_{0} \operatorname{Im} \boldsymbol{y}
$$

If $\operatorname{Re} \boldsymbol{y}=\alpha_{1} \boldsymbol{x}_{0}$ and $\operatorname{Im} \boldsymbol{y}=\alpha_{2} \boldsymbol{x}_{0}$ for real numbers, $\alpha_{i}$, then $\boldsymbol{y}=\left(\alpha_{1}+i \alpha_{2}\right) \boldsymbol{x}_{0}$ and it is assumed this does not happen. Therefore, either

$$
t \operatorname{Re} \boldsymbol{y} \neq \boldsymbol{x}_{0} \text { for all } t \in \mathbb{R}
$$

or

$$
t \operatorname{Im} \boldsymbol{y} \neq \boldsymbol{x}_{0} \text { for all } t \in \mathbb{R}
$$

Assume the first holds. Then varying $t \in \mathbb{R}$, there exists a value of $t$ such that $\boldsymbol{x}_{0}+t \operatorname{Re} \boldsymbol{y}>\mathbf{0}$ but it is not the case that $\boldsymbol{x}_{0}+t \operatorname{Re} \boldsymbol{y} \gg 0$. Then $A\left(\boldsymbol{x}_{0}+t \operatorname{Re} \boldsymbol{y}\right) \gg 0$ by Lemma 11.4.2. But this implies $\lambda_{0}\left(\boldsymbol{x}_{0}+t \operatorname{Re} \boldsymbol{y}\right) \gg 0$ which is a contradiction. Hence there exist real numbers, $\alpha_{1}$ and $\alpha_{2}$ such that $\operatorname{Re} \boldsymbol{y}=\alpha_{1} \boldsymbol{x}_{0}$ and $\operatorname{Im} \boldsymbol{y}=\alpha_{2} \boldsymbol{x}_{0}$ showing that $\boldsymbol{y}=\left(\alpha_{1}+i \alpha_{2}\right) \boldsymbol{x}_{0}$. This proves 3 .

It is possible to obtain a simple corollary to the above theorem.
Corollary 11.4.5 If $A>0$ and $A^{m} \gg 0$ for some $m \in \mathbb{N}$, then all the conclusions of the above theorem hold.

Proof: There exists $\mu_{0}>0$ such that $A^{m} \boldsymbol{y}_{0}=\mu_{0} \boldsymbol{y}_{0}$ for $\boldsymbol{y}_{0} \gg 0$ by Theorem 11.4.4 and

$$
\mu_{0}=\sup \left\{\mu: A^{m} \boldsymbol{x} \geq \mu \boldsymbol{x} \text { for some } \boldsymbol{x} \in K\right\} .
$$

Let $\lambda_{0}^{m}=\mu_{0}$. Then

$$
\left(A-\lambda_{0} I\right)\left(A^{m-1}+\lambda_{0} A^{m-2}+\cdots+\lambda_{0}^{m-1} I\right) \boldsymbol{y}_{0}=\left(A^{m}-\lambda_{0}^{m} I\right) \boldsymbol{y}_{0}=\mathbf{0}
$$

and so letting $\boldsymbol{x}_{0} \equiv\left(A^{m-1}+\lambda_{0} A^{m-2}+\cdots+\lambda_{0}^{m-1} I\right) \boldsymbol{y}_{0}$, it follows $\boldsymbol{x}_{0} \gg 0$ and $A \boldsymbol{x}_{0}=$ $\lambda_{0} x_{0}$.

Suppose now that $A \boldsymbol{x}=\mu \boldsymbol{x}$ for $\boldsymbol{x} \neq \mathbf{0}$ and $\mu \neq \lambda_{0}$. Suppose $|\mu| \geq \lambda_{0}$. Multiplying both sides by $A$, it follows $A^{m} \boldsymbol{x}=\mu^{m} \boldsymbol{x}$ and $\left|\mu^{m}\right|=|\mu|^{m} \geq \lambda_{0}^{m}=\mu_{0}$ and so from Theorem 11.4.4, since $\left|\mu^{m}\right| \geq \mu_{0}$, and $\mu^{m}$ is an eigenvalue of $A^{m}$, it follows that $\mu^{m}=\mu_{0}$. But by Theorem 11.4.4 again, this implies $\boldsymbol{x}=c \boldsymbol{y}_{0}$ for some scalar, $c$ and hence $A \boldsymbol{y}_{0}=\mu \boldsymbol{y}_{0}$. Since $\boldsymbol{y}_{0} \gg \boldsymbol{0}$, it follows $\mu \geq 0$ and so $\mu=\lambda_{0}$, a contradiction. Therefore, $|\mu|<\lambda_{0}$.

Finally, if $A \boldsymbol{x}=\lambda_{0} \boldsymbol{x}$, then $A^{m} \boldsymbol{x}=\lambda_{0}^{m} \boldsymbol{x}$ and so $\boldsymbol{x}=c \boldsymbol{y}_{0}$ for some scalar, $c$. Consequently,

$$
\begin{aligned}
\left(A^{m-1}+\lambda_{0} A^{m-2}+\cdots+\lambda_{0}^{m-1} I\right) \boldsymbol{x} & =c\left(A^{m-1}+\lambda_{0} A^{m-2}+\cdots+\lambda_{0}^{m-1} I\right) \boldsymbol{y}_{0} \\
& =c \boldsymbol{x}_{0}
\end{aligned}
$$

Hence

$$
m \lambda_{0}^{m-1} \boldsymbol{x}=c \boldsymbol{x}_{0}
$$

which shows the dimension of the eigenspace for $\lambda_{0}$ is one.
The following corollary is an extremely interesting convergence result involving the powers of positive matrices.

Corollary 11.4.6 Let $A>0$ and $A^{m} \gg 0$ for some $m \in \mathbb{N}$. Then for $\lambda_{0}$ given in 11.5, there exists a rank one matrix $P$ such that $\lim _{m \rightarrow \infty}\left\|\left(\frac{A}{\lambda_{0}}\right)^{m}-P\right\|=0$.

Proof: Considering $A^{T}$, and the fact that $A$ and $A^{T}$ have the same eigenvalues, Corollary 11.4.5 implies the existence of a vector, $\boldsymbol{v} \gg \boldsymbol{0}$ such that

$$
A^{T} \boldsymbol{v}=\lambda_{0} \boldsymbol{v}
$$

Also let $\boldsymbol{x}_{0}$ denote the vector such that $A \boldsymbol{x}_{0}=\lambda_{0} \boldsymbol{x}_{0}$ with $\boldsymbol{x}_{0} \gg \boldsymbol{0}$. First note that $\boldsymbol{x}_{0}^{T} \boldsymbol{v}>0$ because both these vectors have all entries positive. Therefore, $\boldsymbol{v}$ may be scaled such that

$$
\begin{equation*}
\boldsymbol{v}^{T} \boldsymbol{x}_{0}=\boldsymbol{x}_{0}^{T} \boldsymbol{v}=1 \tag{11.6}
\end{equation*}
$$

Define

$$
P \equiv \boldsymbol{x}_{0} \boldsymbol{v}^{T}
$$

Thanks to 11.6,

$$
\begin{equation*}
\frac{A}{\lambda_{0}} P=\boldsymbol{x}_{0} \boldsymbol{v}^{T}=P, P\left(\frac{A}{\lambda_{0}}\right)=\boldsymbol{x}_{0} \boldsymbol{v}^{T}\left(\frac{A}{\lambda_{0}}\right)=\boldsymbol{x}_{0} \boldsymbol{v}^{T}=P \tag{11.7}
\end{equation*}
$$

and

$$
\begin{equation*}
P^{2}=\boldsymbol{x}_{0} \boldsymbol{v}^{T} \boldsymbol{x}_{0} \boldsymbol{v}^{T}=\boldsymbol{v}^{T} \boldsymbol{x}_{0}=P \tag{11.8}
\end{equation*}
$$

Therefore,

$$
\begin{aligned}
\left(\frac{A}{\lambda_{0}}-P\right)^{2} & =\left(\frac{A}{\lambda_{0}}\right)^{2}-2\left(\frac{A}{\lambda_{0}}\right) P+P^{2} \\
& =\left(\frac{A}{\lambda_{0}}\right)^{2}-P
\end{aligned}
$$

Continuing this way, using 11.7 repeatedly, it follows

$$
\begin{equation*}
\left(\left(\frac{A}{\lambda_{0}}\right)-P\right)^{m}=\left(\frac{A}{\lambda_{0}}\right)^{m}-P \tag{11.9}
\end{equation*}
$$

The eigenvalues of $\left(\frac{A}{\lambda_{0}}\right)-P$ are of interest because it is powers of this matrix which determine the convergence of $\left(\frac{A}{\lambda_{0}}\right)^{m}$ to $P$. Therefore, let $\mu$ be a nonzero eigenvalue of this matrix. Thus

$$
\begin{equation*}
\left(\left(\frac{A}{\lambda_{0}}\right)-P\right) \boldsymbol{x}=\mu \boldsymbol{x} \tag{11.10}
\end{equation*}
$$

for $\boldsymbol{x} \neq \mathbf{0}$, and $\mu \neq 0$. Applying $P$ to both sides and using the second formula of 11.7 yields

$$
\mathbf{0}=(P-P) \boldsymbol{x}=\left(P\left(\frac{A}{\lambda_{0}}\right)-P^{2}\right) \boldsymbol{x}=\mu P x
$$

But since $P \boldsymbol{x}=\mathbf{0}$, it follows from 11.10 that

$$
A x=\lambda_{0} \mu x
$$

which implies $\lambda_{0} \mu$ is an eigenvalue of $A$. Therefore, by Corollary 11.4.5 it follows that either $\lambda_{0} \mu=\lambda_{0}$ in which case $\mu=1$, or $\lambda_{0}|\mu|<\lambda_{0}$ which implies $|\mu|<1$. But if $\mu=1$, then $\boldsymbol{x}$ is a multiple of $\boldsymbol{x}_{0}$ and 11.10 would yield

$$
\left(\left(\frac{A}{\lambda_{0}}\right)-P\right) x_{0}=x_{0}
$$

which says $x_{0}-x_{0} v^{T} x_{0}=x_{0}$ and so by $11.6, x_{0}=\mathbf{0}$ contrary to the property that $x_{0} \gg$ $\mathbf{0}$. Therefore, $|\mu|<1$ and so this has shown that the absolute values of all eigenvalues of $\left(\frac{A}{\lambda_{0}}\right)-P$ are less than 1. By Gelfand's theorem, Theorem 14.2.4, it follows

$$
\left\|\left(\left(\frac{A}{\lambda_{0}}\right)-P\right)^{m}\right\|^{1 / m}<r<1
$$

whenever $m$ is large enough. Now by 11.9 this yields

$$
\left\|\left(\frac{A}{\lambda_{0}}\right)^{m}-P\right\|=\left\|\left(\left(\frac{A}{\lambda_{0}}\right)-P\right)^{m}\right\| \leq r^{m}
$$

whenever $m$ is large enough. It follows

$$
\lim _{m \rightarrow \infty}\left\|\left(\frac{A}{\lambda_{0}}\right)^{m}-P\right\|=0
$$

as claimed.
What about the case when $A>0$ but maybe it is not the case that $A \gg 0$ ? As before,

$$
K \equiv\left\{\boldsymbol{x} \geq \mathbf{0} \text { such that }\|\boldsymbol{x}\|_{1}=1\right\}
$$

Now define

$$
S_{1} \equiv\{\lambda: A x \geq \lambda x \text { for some } x \in K\}
$$

and

$$
\begin{equation*}
\lambda_{0} \equiv \sup \left(S_{1}\right) \tag{11.11}
\end{equation*}
$$

Theorem 11.4.7 Let $A>0$ and let $\lambda_{0}$ be defined in 11.11. Then there exists $\boldsymbol{x}_{0}>\mathbf{0}$ such that $A x_{0}=\lambda_{0} x_{0}$.

Proof: Let $E$ consist of the matrix which has a one in every entry. Then from Theorem 11.4.4 it follows there exists $\boldsymbol{x}_{\delta} \gg \boldsymbol{0},\left\|\boldsymbol{x}_{\delta}\right\|_{1}=1$, such that $(A+\delta E) \boldsymbol{x}_{\delta}=\lambda_{0 \delta} \boldsymbol{x}_{\delta}$ where

$$
\lambda_{0 \delta} \equiv \sup \{\lambda:(A+\delta E) \boldsymbol{x} \geq \lambda \boldsymbol{x} \text { for some } \boldsymbol{x} \in K\}
$$

Now if $\alpha<\delta$

$$
\begin{gathered}
\{\boldsymbol{\lambda}:(A+\alpha E) \boldsymbol{x} \geq \boldsymbol{\lambda} \boldsymbol{x} \text { for some } \boldsymbol{x} \in K\} \subseteq \\
\{\boldsymbol{\lambda}:(A+\delta E) \boldsymbol{x} \geq \boldsymbol{\lambda} \boldsymbol{x} \text { for some } \boldsymbol{x} \in K\}
\end{gathered}
$$

and so $\lambda_{0 \delta} \geq \lambda_{0 \alpha}$ because $\lambda_{0 \delta}$ is the sup of the second set and $\lambda_{0 \alpha}$ is the sup of the first. It follows the limit, $\lambda_{1} \equiv \lim _{\delta \rightarrow 0+} \lambda_{0 \delta}$ exists. Taking a subsequence and using the compactness of $K$, there exists a subsequence, still denoted by $\delta$ such that as $\delta \rightarrow 0, \boldsymbol{x}_{\delta} \rightarrow$ $\boldsymbol{x} \in K$. Therefore,

$$
A x=\lambda_{1} x
$$

and so, in particular, $A x \geq \lambda_{1} x$ and so $\lambda_{1} \leq \lambda_{0}$. But also, if $\lambda \leq \lambda_{0}$,

$$
\boldsymbol{\lambda} \boldsymbol{x} \leq A \boldsymbol{x}<(A+\delta E) \boldsymbol{x}
$$

showing that $\lambda_{0 \delta} \geq \lambda$ for all such $\lambda$. But then $\lambda_{0 \delta} \geq \lambda_{0}$ also. Hence $\lambda_{1} \geq \lambda_{0}$, showing these two numbers are the same. Hence $A \boldsymbol{x}=\lambda_{0} \boldsymbol{x}$.

If $A^{m} \gg 0$ for some $m$ and $A>0$, it follows that the dimension of the eigenspace for $\lambda_{0}$ is one and that the absolute value of every other eigenvalue of $A$ is less than $\lambda_{0}$. If it is only assumed that $A>0$, not necessarily $\gg 0$, this is no longer true. However, there is something which is very interesting which can be said. First here is an interesting lemma.

Lemma 11.4.8 Let $M$ be a matrix of the form

$$
M=\left(\begin{array}{cc}
A & 0 \\
B & C
\end{array}\right)
$$

or

$$
M=\left(\begin{array}{cc}
A & B \\
0 & C
\end{array}\right)
$$

where $A$ is an $r \times r$ matrix and $C$ is an $(n-r) \times(n-r)$ matrix. Then it follows that $\operatorname{det}(M)=\operatorname{det}(A) \operatorname{det}(B)$ and $\sigma(M)=\sigma(A) \cup \sigma(C)$.

Proof: To verify the claim about the determinants, note

$$
\left(\begin{array}{ll}
A & 0 \\
B & C
\end{array}\right)=\left(\begin{array}{cc}
A & 0 \\
0 & I
\end{array}\right)\left(\begin{array}{ll}
I & 0 \\
B & C
\end{array}\right)
$$

Therefore,

$$
\operatorname{det}\left(\begin{array}{cc}
A & 0 \\
B & C
\end{array}\right)=\operatorname{det}\left(\begin{array}{cc}
A & 0 \\
0 & I
\end{array}\right) \operatorname{det}\left(\begin{array}{cc}
I & 0 \\
B & C
\end{array}\right) .
$$

But it is clear from the method of Laplace expansion that

$$
\operatorname{det}\left(\begin{array}{cc}
A & 0 \\
0 & I
\end{array}\right)=\operatorname{det} A
$$

and from the multilinear properties of the determinant and row operations that

$$
\operatorname{det}\left(\begin{array}{cc}
I & 0 \\
B & C
\end{array}\right)=\operatorname{det}\left(\begin{array}{cc}
I & 0 \\
0 & C
\end{array}\right)=\operatorname{det} C
$$

The case where $M$ is upper block triangular is similar.
This immediately implies $\sigma(M)=\sigma(A) \cup \sigma(C)$.
Theorem 11.4.9 Let $A>0$ and let $\lambda_{0}$ be given in 11.11. If $\lambda$ is an eigenvalue for $A$ such that $|\lambda|=\lambda_{0}$, then $\lambda / \lambda_{0}$ is a root of unity. Thus $\left(\lambda / \lambda_{0}\right)^{m}=1$ for some $m \in \mathbb{N}$.

Proof: Applying Theorem 11.4 .7 to $A^{T}$, there exists $\boldsymbol{v}>\mathbf{0}$ such that $A^{T} \boldsymbol{v}=\lambda_{0} \boldsymbol{v}$. In the first part of the argument it is assumed $\boldsymbol{v} \gg \boldsymbol{0}$. Now suppose $A \boldsymbol{x}=\boldsymbol{\lambda} \boldsymbol{x}, \boldsymbol{x} \neq \mathbf{0}$ and that $|\lambda|=\lambda_{0}$. Then

$$
A|\boldsymbol{x}| \geq|\lambda||\boldsymbol{x}|=\lambda_{0}|\boldsymbol{x}|
$$

and it follows that if $A|\boldsymbol{x}|>|\boldsymbol{\lambda}||\boldsymbol{x}|$, then since $\boldsymbol{v} \gg \boldsymbol{0}$,

$$
\lambda_{0}(\boldsymbol{v},|\boldsymbol{x}|)<(\boldsymbol{v}, A|\boldsymbol{x}|)=\left(A^{T} \boldsymbol{v},|\boldsymbol{x}|\right)=\lambda_{0}(\boldsymbol{v},|\boldsymbol{x}|),
$$

a contradiction. Therefore,

$$
\begin{equation*}
A|x|=\lambda_{0}|x| \tag{11.12}
\end{equation*}
$$

It follows that

$$
\left|\sum_{j} A_{i j} x_{j}\right|=\lambda_{0}\left|x_{i}\right|=\sum_{j} A_{i j}\left|x_{j}\right|
$$

and so the complex numbers,

$$
A_{i j} x_{j}, A_{i k} x_{k}
$$

must have the same argument for every $k, j$ because equality holds in the triangle inequality. Therefore, there exists a complex number, $\mu_{i}$ such that

$$
\begin{equation*}
A_{i j} x_{j}=\mu_{i} A_{i j}\left|x_{j}\right| \tag{11.13}
\end{equation*}
$$

and so, letting $r \in \mathbb{N}$,

$$
A_{i j} x_{j} \mu_{j}^{r}=\mu_{i} A_{i j}\left|x_{j}\right| \mu_{j}^{r}
$$

Summing on $j$ yields

$$
\begin{equation*}
\sum_{j} A_{i j} x_{j} \mu_{j}^{r}=\mu_{i} \sum_{j} A_{i j}\left|x_{j}\right| \mu_{j}^{r} \tag{11.14}
\end{equation*}
$$

Also, summing 11.13 on $j$ and using that $\lambda$ is an eigenvalue for $\boldsymbol{x}$, it follows from 11.12 that

$$
\begin{equation*}
\lambda x_{i}=\sum_{j} A_{i j} x_{j}=\mu_{i} \sum_{j} A_{i j}\left|x_{j}\right|=\mu_{i} \lambda_{0}\left|x_{i}\right| \tag{11.15}
\end{equation*}
$$

From 11.14 and 11.15,

$$
\begin{aligned}
& \sum_{j} A_{i j} x_{j} \mu_{j}^{r}=\mu_{i} \sum_{j} A_{i j}\left|x_{j}\right| \mu_{j}^{r}=\mu_{i} \sum_{j} A_{i j} \overbrace{\mu_{j}\left|x_{j}\right|}^{\text {see } 11.15} \mu_{j}^{r-1} \\
& =\mu_{i} \sum_{j} A_{i j}\left(\frac{\lambda}{\lambda_{0}}\right) x_{j} \mu_{j}^{r-1}=\mu_{i}\left(\frac{\lambda}{\lambda_{0}}\right) \sum_{j} A_{i j} x_{j} \mu_{j}^{r-1}
\end{aligned}
$$

Now from 11.14 with $r$ replaced by $r-1$, this equals

$$
\mu_{i}^{2}\left(\frac{\lambda}{\lambda_{0}}\right) \sum_{j} A_{i j}\left|x_{j}\right| \mu_{j}^{r-1}=\mu_{i}^{2}\left(\frac{\lambda}{\lambda_{0}}\right) \sum_{j} A_{i j} \mu_{j}\left|x_{j}\right| \mu_{j}^{r-2}=\mu_{i}^{2}\left(\frac{\lambda}{\lambda_{0}}\right)^{2} \sum_{j} A_{i j} x_{j} \mu_{j}^{r-2}
$$

Continuing this way,

$$
\sum_{j} A_{i j} x_{j} \mu_{j}^{r}=\mu_{i}^{k}\left(\frac{\lambda}{\lambda_{0}}\right)^{k} \sum_{j} A_{i j} x_{j} \mu_{j}^{r-k}
$$

and eventually, this shows

$$
\sum_{j} A_{i j} x_{j} \mu_{j}^{r}=\mu_{i}^{r}\left(\frac{\lambda}{\lambda_{0}}\right)^{r} \sum_{j} A_{i j} x_{j}=\left(\frac{\lambda}{\lambda_{0}}\right)^{r} \lambda\left(x_{i} \mu_{i}^{r}\right)
$$

and this says $\left(\frac{\lambda}{\lambda_{0}}\right)^{r+1}$ is an eigenvalue for $\left(\frac{A}{\lambda_{0}}\right)$ with the eigenvector being

$$
\left(x_{1} \mu_{1}^{r}, \cdots, x_{n} \mu_{n}^{r}\right)^{T}
$$

Now recall that $r \in \mathbb{N}$ was arbitrary and so this has shown that

$$
\left(\frac{\lambda}{\lambda_{0}}\right)^{2},\left(\frac{\lambda}{\lambda_{0}}\right)^{3},\left(\frac{\lambda}{\lambda_{0}}\right)^{4}, \cdots
$$

are each eigenvalues of $\left(\frac{A}{\lambda_{0}}\right)$ which has only finitely many and hence this sequence must repeat. Therefore, $\left(\frac{\lambda}{\lambda_{0}}\right)$ is a root of unity as claimed. This proves the theorem in the case that $\boldsymbol{v} \gg \mathbf{0}$.

Now it is necessary to consider the case where $\boldsymbol{v}>\mathbf{0}$ but it is not the case that $\boldsymbol{v} \gg \mathbf{0}$. Then in this case, there exists a permutation matrix $P$ such that

$$
P \boldsymbol{v}=\left(\begin{array}{c}
v_{1} \\
\vdots \\
v_{r} \\
0 \\
\vdots \\
0
\end{array}\right) \equiv\binom{\boldsymbol{u}}{\mathbf{0}} \equiv \boldsymbol{v}_{1}
$$

Then $\lambda_{0} \boldsymbol{v}=A^{T} \boldsymbol{v}=A^{T} P \boldsymbol{v}_{1}$. Therefore, $\lambda_{0} \boldsymbol{v}_{1}=P A^{T} P \boldsymbol{v}_{1}=G \boldsymbol{v}_{1}$. Now $P^{2}=I$ because it is a permutation matrix. Therefore, the matrix $G \equiv P A^{T} P$ and $A$ are similar. Consequently, they have the same eigenvalues and it suffices from now on to consider the matrix $G$ rather than $A$. Then

$$
\lambda_{0}\binom{\boldsymbol{u}}{\mathbf{0}}=\left(\begin{array}{ll}
M_{1} & M_{2} \\
M_{3} & M_{4}
\end{array}\right)\binom{\boldsymbol{u}}{\mathbf{0}}
$$

where $M_{1}$ is $r \times r$ and $M_{4}$ is $(n-r) \times(n-r)$. It follows from block multiplication and the assumption that $A$ and hence $G$ are $>0$ that

$$
G=\left(\begin{array}{cc}
A^{\prime} & B \\
0 & C
\end{array}\right)
$$

Now let $\lambda$ be an eigenvalue of $G$ such that $|\lambda|=\lambda_{0}$. Then from Lemma 11.4.8, either $\lambda \in \sigma\left(A^{\prime}\right)$ or $\lambda \in \sigma(C)$. Suppose without loss of generality that $\lambda \in \sigma\left(A^{\prime}\right)$. Since $A^{\prime}>0$ it has a largest positive eigenvalue $\lambda_{0}^{\prime}$ which is obtained from 11.11. Thus $\lambda_{0}^{\prime} \leq \lambda_{0}$ but $\lambda$ being an eigenvalue of $A^{\prime}$, has its absolute value bounded by $\lambda_{0}^{\prime}$ and so $\lambda_{0}=|\lambda| \leq \lambda_{0}^{\prime} \leq \lambda_{0}$ showing that $\lambda_{0} \in \sigma\left(A^{\prime}\right)$. Now if there exists $\boldsymbol{v} \gg \boldsymbol{0}$ such that $A^{T T} \boldsymbol{v}=\lambda_{0} \boldsymbol{v}$, then the first part of this proof applies to the matrix $A$ and so $\left(\lambda / \lambda_{0}\right)$ is a root of unity. If such a vector, $\boldsymbol{v}$ does not exist, then let $A^{\prime}$ play the role of $A$ in the above argument and reduce to the consideration of

$$
G^{\prime} \equiv\left(\begin{array}{cc}
A^{\prime \prime} & B^{\prime} \\
0 & C^{\prime}
\end{array}\right)
$$

where $G^{\prime}$ is similar to $A^{\prime}$ and $\lambda, \lambda_{0} \in \sigma\left(A^{\prime \prime}\right)$. Stop if $A^{\prime \prime T} \boldsymbol{v}=\lambda_{0} \boldsymbol{v}$ for some $\boldsymbol{v} \gg \boldsymbol{0}$. Otherwise, decompose $A^{\prime \prime}$ similar to the above and add another prime. Continuing this way you must eventually obtain the situation where $\left(A^{\prime \cdots \prime}\right)^{T} \boldsymbol{v}=\boldsymbol{\lambda}_{0} \boldsymbol{v}$ for some $\boldsymbol{v} \gg \boldsymbol{0}$. Indeed, this happens no later than when $A^{\prime \cdots /}$ is a $1 \times 1$ matrix.

### 11.5 Functions Of Matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

$$
\begin{equation*}
f(\lambda)=\sum_{n=0}^{\infty} a_{n} \lambda^{n} \tag{11.16}
\end{equation*}
$$

for all $|\lambda|<R$. There is a formula for $f(A) \equiv \sum_{n=0}^{\infty} a_{n} A^{n}$ which makes sense whenever $\rho(A)<R$. Thus you can speak of $\sin (A)$ or $e^{A}$ for $A$ an $n \times n$ matrix. To begin with, define

$$
f_{P}(\lambda) \equiv \sum_{n=0}^{P} a_{n} \lambda^{n}
$$

so for $k<P$

$$
\begin{equation*}
f_{P}^{(k)}(\lambda)=\sum_{n=k}^{P} a_{n} n \cdots(n-k+1) \lambda^{n-k}=\sum_{n=k}^{P} a_{n}\binom{n}{k} k!\lambda^{n-k} . \tag{11.17}
\end{equation*}
$$

Thus

$$
\begin{equation*}
\frac{f_{P}^{(k)}(\lambda)}{k!}=\sum_{n=k}^{P} a_{n}\binom{n}{k} \lambda^{n-k} \tag{11.18}
\end{equation*}
$$

To begin with consider $f\left(J_{m}(\lambda)\right)$ where $J_{m}(\lambda)$ is an $m \times m$ Jordan block. Thus $J_{m}(\lambda)=$ $D+N$ where $N^{m}=0$ and $N$ commutes with $D$. Therefore, letting $P>m$

$$
\begin{align*}
\sum_{n=0}^{P} a_{n} J_{m}(\lambda)^{n} & =\sum_{n=0}^{P} a_{n} \sum_{k=0}^{n}\binom{n}{k} D^{n-k} N^{k}=\sum_{k=0}^{P} \sum_{n=k}^{P} a_{n}\binom{n}{k} D^{n-k} N^{k} \\
& =\sum_{k=0}^{m-1} N^{k} \sum_{n=k}^{P} a_{n}\binom{n}{k} D^{n-k} \tag{11.19}
\end{align*}
$$

From 11.18 this equals

$$
\begin{equation*}
\sum_{k=0}^{m-1} N^{k} \operatorname{diag}\left(\frac{f_{P}^{(k)}(\lambda)}{k!}, \cdots, \frac{f_{P}^{(k)}(\lambda)}{k!}\right) \tag{11.20}
\end{equation*}
$$

where for $k=0, \cdots, m-1$, define $\operatorname{diag}_{k}\left(a_{1}, \cdots, a_{m-k}\right)$ the $m \times m$ matrix which equals zero everywhere except on the $k^{\text {th }}$ super diagonal where this diagonal is filled with the numbers, $\left\{a_{1}, \cdots, a_{m-k}\right\}$ from the upper left to the lower right. With no subscript, it is just the diagonal matrices having the indicated entries. Thus in $4 \times 4$ matrices, $\operatorname{diag}_{2}(1,2)$ would be the matrix

$$
\left(\begin{array}{llll}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Then from 11.20 and 11.17,

$$
\sum_{n=0}^{P} a_{n} J_{m}(\lambda)^{n}=\sum_{k=0}^{m-1} \operatorname{diag}_{k}\left(\frac{f_{P}^{(k)}(\lambda)}{k!}, \cdots, \frac{f_{P}^{(k)}(\lambda)}{k!}\right)
$$

Therefore, $\sum_{n=0}^{P} a_{n} J_{m}(\lambda)^{n}=$

$$
\left(\begin{array}{ccccc}
f_{P}(\lambda) & \frac{f_{P}^{\prime}(\lambda)}{1!} & \frac{f_{P}^{(2)}(\lambda)}{2!} & \cdots & \frac{f_{P}^{(m-1)}(\lambda)}{(m-1)!}  \tag{11.21}\\
& f_{P}(\lambda) & \frac{f_{P}^{\prime}(\lambda)}{1!} & \ddots & \vdots \\
& & f_{P}(\lambda) & \ddots & \frac{f_{P}^{(2)}(\lambda)}{2!} \\
& & & \ddots & \frac{f_{P}^{\prime}(\lambda)}{1!} \\
0 & & & & f_{P}(\lambda)
\end{array}\right)
$$

Now let $A$ be an $n \times n$ matrix with $\rho(A)<R$ where $R$ is given above. Then the Jordan form of $A$ is of the form

$$
J=\left(\begin{array}{cccc}
J_{1} & & & 0  \tag{11.22}\\
& J_{2} & & \\
& & \ddots & \\
0 & & & J_{r}
\end{array}\right)
$$

where $J_{k}=J_{m_{k}}\left(\lambda_{k}\right)$ is an $m_{k} \times m_{k}$ Jordan block and $A=S^{-1} J S$. Then, letting $P>m_{k}$ for all $k$,

$$
\sum_{n=0}^{P} a_{n} A^{n}=S^{-1} \sum_{n=0}^{P} a_{n} J^{n} S
$$

and because of block multiplication of matrices,

$$
\sum_{n=0}^{P} a_{n} J^{n}=\left(\begin{array}{cccc}
\sum_{n=0}^{P} a_{n} J_{1}^{n} & & & 0 \\
& \ddots & & \\
& & \ddots & \\
0 & & & \sum_{n=0}^{P} a_{n} J_{r}^{n}
\end{array}\right)
$$

and from $11.21 \sum_{n=0}^{P} a_{n} J_{k}^{n}$ converges as $P \rightarrow \infty$ to the $m_{k} \times m_{k}$ matrix

$$
\left(\begin{array}{ccccc}
f\left(\lambda_{k}\right) & \frac{f^{\prime}\left(\lambda_{k}\right)}{1!} & \frac{f^{(2)}\left(\lambda_{k}\right)}{2!} & \cdots & \frac{f^{(m-1)}\left(\lambda_{k}\right)}{\left(m_{k}-1\right)!}  \tag{11.23}\\
0 & f\left(\lambda_{k}\right) & \frac{f^{\prime}\left(\lambda_{k}\right)}{1!} & \ddots & \vdots \\
0 & 0 & f\left(\lambda_{k}\right) & \ddots & \frac{f^{(2)}\left(\lambda_{k}\right)}{2!} \\
\vdots & & \ddots & \ddots & \frac{f^{\prime}\left(\lambda_{k}\right)}{1!} \\
0 & 0 & \cdots & 0 & f\left(\lambda_{k}\right)
\end{array}\right)
$$

There is no convergence problem because $|\lambda|<R$ for all $\lambda \in \sigma(A)$. This has proved the following theorem.

Theorem 11.5.1 Let $f$ be given by 11.16 and suppose $\rho(A)<R$ where $R$ is the radius of convergence of the power series in 11.16. Then the series,

$$
\begin{equation*}
\sum_{k=0}^{\infty} a_{n} A^{n} \tag{11.24}
\end{equation*}
$$

converges in the space $\mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{n}\right)$ with respect to any of the norms on this space and furthermore,

$$
\sum_{k=0}^{\infty} a_{n} A^{n}=S^{-1}\left(\begin{array}{cccc}
\sum_{n=0}^{\infty} a_{n} J_{1}^{n} & & & 0 \\
& \ddots & & \\
& & \ddots & \\
0 & & & \sum_{n=0}^{\infty} a_{n} J_{r}^{n}
\end{array}\right) S
$$

where $\sum_{n=0}^{\infty} a_{n} J_{k}^{n}$ is an $m_{k} \times m_{k}$ matrix of the form given in 11.23 where $A=S^{-1} J S$ and the Jordan form of $A, J$ is given by 11.22. Therefore, you can define $f(A)$ by the series in 11.24.

Here is a simple example.
Example 11.5.2 Find $\sin (A)$ where $A=\left(\begin{array}{cccc}4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4\end{array}\right)$.
In this case, the Jordan canonical form of the matrix is not too hard to find.

$$
\begin{aligned}
& \left(\begin{array}{cccc}
4 & 1 & -1 & 1 \\
1 & 1 & 0 & -1 \\
0 & -1 & 1 & -1 \\
-1 & 2 & 1 & 4
\end{array}\right)=\left(\begin{array}{cccc}
2 & 0 & -2 & -1 \\
1 & -4 & -2 & -1 \\
0 & 0 & -2 & 1 \\
-1 & 4 & 4 & 2
\end{array}\right) . \\
& \left(\begin{array}{llll}
4 & 0 & 0 & 0 \\
0 & 2 & 1 & 0 \\
0 & 0 & 2 & 1 \\
0 & 0 & 0 & 2
\end{array}\right)\left(\begin{array}{cccc}
\frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\
0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{array}\right) .
\end{aligned}
$$

Then from the above theorem $\sin (J)$ is given by

$$
\sin \left(\begin{array}{cccc}
4 & 0 & 0 & 0 \\
0 & 2 & 1 & 0 \\
0 & 0 & 2 & 1 \\
0 & 0 & 0 & 2
\end{array}\right)=\left(\begin{array}{cccc}
\sin 4 & 0 & 0 & 0 \\
0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\
0 & 0 & \sin 2 & \cos 2 \\
0 & 0 & 0 & \sin 2
\end{array}\right) .
$$

Therefore, $\sin (A)=$

$$
\left(\begin{array}{cccc}
2 & 0 & -2 & -1 \\
1 & -4 & -2 & -1 \\
0 & 0 & -2 & 1 \\
-1 & 4 & 4 & 2
\end{array}\right)\left(\begin{array}{cccc}
\sin 4 & 0 & 0 & 0 \\
0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\
0 & 0 & \sin 2 & \cos 2 \\
0 & 0 & 0 & \sin 2
\end{array}\right)\left(\begin{array}{cccc}
\frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\
0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{array}\right)
$$

$=M$ where the columns of $M$ are as follows from left to right,

$$
\begin{aligned}
& \left(\begin{array}{c}
\sin 4 \\
\frac{1}{2} \sin 4-\frac{1}{2} \sin 2 \\
0 \\
-\frac{1}{2} \sin 4+\frac{1}{2} \sin 2
\end{array}\right),\left(\begin{array}{c}
\sin 4-\sin 2-\cos 2 \\
\frac{1}{2} \sin 4+\frac{3}{2} \sin 2-2 \cos 2 \\
-\cos 2 \\
-\frac{1}{2} \sin 4-\frac{1}{2} \sin 2+3 \cos 2
\end{array}\right),\left(\begin{array}{c}
-\cos 2 \\
\sin 2 \\
\sin 2-\cos 2 \\
\cos 2-\sin 2
\end{array}\right) \\
& \left(\begin{array}{c}
\sin 4-\sin 2-\cos 2 \\
\frac{1}{2} \sin 4+\frac{1}{2} \sin 2-2 \cos 2 \\
-\cos 2 \\
-\frac{1}{2} \sin 4+\frac{1}{2} \sin 2+3 \cos 2
\end{array}\right)
\end{aligned}
$$

Perhaps this isn't the first thing you would think of. Of course the ability to get this nice closed form description of $\sin (A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

The following corollary is known as the spectral mapping theorem.
Corollary 11.5.3 Let $A$ be an $n \times n$ matrix and let $\rho(A)<R$ where for $|\lambda|<R, f(\lambda)=$ $\sum_{n=0}^{\infty} a_{n} \lambda^{n}$.Then $f(A)$ is also an $n \times n$ matrix and furthermore, $\sigma(f(A))=f(\sigma(A))$. Thus the eigenvalues of $f(A)$ are exactly the numbers $f(\lambda)$ where $\lambda$ is an eigenvalue of $A$. Furthermore, the algebraic multiplicity of $f(\lambda)$ coincides with the algebraic multiplicity of $\lambda$.

All of these things can be generalized to linear transformations defined on infinite dimensional spaces and when this is done the main tool is the Dunford integral along with the methods of complex analysis. It is good to see it done for finite dimensional situations first because it gives an idea of what is possible.

### 11.6 Exercises

1. Suppose the migration matrix for three locations is

$$
\left(\begin{array}{ccc}
.5 & 0 & .3 \\
.3 & .8 & 0 \\
.2 & .2 & .7
\end{array}\right)
$$

Find a comparison for the populations in the three locations after a long time.
2. Show that if $\sum_{i} a_{i j}=1$, then if $A=\left(a_{i j}\right)$, then the sum of the entries of $A \boldsymbol{v}$ equals the sum of the entries of $\boldsymbol{v}$. Thus it does not matter whether $a_{i j} \geq 0$ for this to be so.
3. If $A$ satisfies the conditions of the above problem, can it be concluded that $\lim _{n \rightarrow \infty} A^{n}$ exists?
4. Give an example of a non regular Markov matrix which has an eigenvalue equal to -1 .
5. Show that when a Markov matrix is non defective, all of the above theory can be proved very easily. In particular, prove the theorem about the existence of $\lim _{n \rightarrow \infty} A^{n}$ if the eigenvalues are either 1 or have absolute value less than 1.
6. Find a formula for $A^{n}$ where

$$
A=\left(\begin{array}{cccc}
\frac{5}{2} & -\frac{1}{2} & 0 & -1 \\
5 & 0 & 0 & -4 \\
\frac{7}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{5}{2} \\
\frac{7}{2} & -\frac{1}{2} & 0 & -2
\end{array}\right)
$$

Does $\lim _{n \rightarrow \infty} A^{n}$ exist? Note that all the rows sum to 1 . Hint: This matrix is similar to a diagonal matrix. The eigenvalues are $1,-1, \frac{1}{2}, \frac{1}{2}$.
7. Find a formula for $A^{n}$ where

$$
A=\left(\begin{array}{cccc}
2 & -\frac{1}{2} & \frac{1}{2} & -1 \\
4 & 0 & 1 & -4 \\
\frac{5}{2} & -\frac{1}{2} & 1 & -2 \\
3 & -\frac{1}{2} & \frac{1}{2} & -2
\end{array}\right)
$$

Note that the rows sum to 1 in this matrix also. Hint: This matrix is not similar to a diagonal matrix but you can find the Jordan form and consider this in order to obtain a formula for this product. The eigenvalues are $1,-1, \frac{1}{2}, \frac{1}{2}$.
8. Find $\lim _{n \rightarrow \infty} A^{n}$ if it exists for the matrix

$$
A=\left(\begin{array}{cccc}
\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 \\
-\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{3}{2} & 0 \\
\frac{3}{2} & \frac{3}{2} & \frac{3}{2} & 1
\end{array}\right)
$$

The eigenvalues are $\frac{1}{2}, 1,1,1$.
9. Give an example of a matrix $A$ which has eigenvalues which are either equal to $1,-1$, or have absolute value strictly less than 1 but which has the property that $\lim _{n \rightarrow \infty} A^{n}$ does not exist.
10. If $A$ is an $n \times n$ matrix such that all the eigenvalues have absolute value less than 1 , show $\lim _{n \rightarrow \infty} A^{n}=0$.
11. Find an example of a $3 \times 3$ matrix $A$ such that $\lim _{n \rightarrow \infty} A^{n}$ does not exist but $\lim _{r \rightarrow \infty} A^{5 r}$ does exist.
12. If $A$ is a Markov matrix and $B$ is similar to $A$, does it follow that $B$ is also a Markov matrix?
13. In Theorem 11.1.3 suppose everything is unchanged except that you assume either $\sum_{j} a_{i j} \leq 1$ or $\sum_{i} a_{i j} \leq 1$. Would the same conclusion be valid? What if you don't insist that each $a_{i j} \geq 0$ ? Would the conclusion hold in this case?
14. Let $V$ be an $n$ dimensional vector space and let $\boldsymbol{x} \in V$ and $\boldsymbol{x} \neq \mathbf{0}$. Consider

$$
\beta_{x} \equiv x, A x, \cdots, A^{m-1} x
$$

where

$$
A^{m} \boldsymbol{x} \in \operatorname{span}\left(\boldsymbol{x}, A \boldsymbol{x}, \cdots, A^{m-1} \boldsymbol{x}\right)
$$

and $m$ is the smallest such that the above inclusion in the span takes place. Show that $\left\{\boldsymbol{x}, A \boldsymbol{x}, \cdots, A^{m-1} \boldsymbol{x}\right\}$ must be linearly independent. Next suppose $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ is a basis for $V$. Consider $\beta_{\boldsymbol{v}_{i}}$ as just discussed, having length $m_{i}$. Thus $A^{m_{i}} \boldsymbol{v}_{i}$ is a linearly combination of $\boldsymbol{v}_{i}, A \boldsymbol{v}_{i}, \cdots, A^{m-1} \boldsymbol{v}_{i}$ for $m$ as small as possible. Let $p_{\boldsymbol{v}_{i}}(\boldsymbol{\lambda})$ be the monic polynomial which expresses this linear combination. Thus $p_{\boldsymbol{v}_{i}}(A) \boldsymbol{v}_{i}=0$ and the degree of $p_{v_{i}}(\lambda)$ is as small as possible for this to take place. Show that the minimum polynomial for $A$ must be the monic polynomial which is the least common multiple of these polynomials $p_{\boldsymbol{v}_{i}}(\boldsymbol{\lambda})$.

## Chapter 12

## Inner Product Spaces, Least Squares

In this chapter is a more complete discussion of important theorems for inner product spaces. These results are presented for inner product spaces, the typical example being $\mathbb{C}^{n}$ or $\mathbb{R}^{n}$. The extra generality is used because most of the ideas have a straight forward generalization to something called a Hilbert space which is just a complete inner product space. First is a major result about projections.

### 12.1 Orthogonal Projections

Recall that any finite dimensional normed linear space is complete. The following definition includes the case where the norm comes from an inner product.

Definition 12.1.1 Let $(H,(\cdot, \cdot))$ be a complete inner product space. This means the norm comes from an inner product as described on Page 263, $|v| \equiv(v, v)^{1 / 2}$. Such a space is called a Hilbert space

As shown earlier, if $H$ is finite dimensional, then it is a Hilbert space automatically. The following is the definition of a convex set. This is a set with the property that the line segment between any two points in the set is in the set.

Definition 12.1.2 A nonempty subset $K$ of a vector space is said to be convex if whenever $x, y \in K$ and $t \in[0,1]$, it follows that $t x+(1-t) y \in K$.

Theorem 12.1.3 Let $K$ be a closed and convex nonempty subset of a Hilbert space and let $y \in H$. Also let

$$
\lambda \equiv \inf \{|x-y|: x \in K\}
$$

Then if $\left\{x_{n}\right\} \subseteq K$ is a sequence such that $\lim _{n \rightarrow \infty}\left|x_{n}-y\right|=\lambda$, then it follows that $\left\{x_{n}\right\}$ is a Cauchy sequence and $\lim _{n \rightarrow \infty} x_{n}=x \in K$ with $|x-y|=\lambda$. Also if $|x-y|=\lambda=|\hat{x}-y|$, then $\hat{x}=x$.

Proof: Recall the parallelogram identity valid in any innner product space:

$$
|x+y|^{2}+|x-y|^{2}=2|x|^{2}+2|y|^{2}
$$

First consider the claim about uniqueness. Letting $x, \hat{x}$ be as given,

$$
\begin{aligned}
\left|\frac{x+\hat{x}}{2}-y\right|^{2}+\left|\frac{x-\hat{x}}{2}\right|^{2} & =\left|\frac{x-y}{2}+\frac{\hat{x}-y}{2}\right|^{2}+\left|\frac{x-\hat{x}}{2}\right|^{2} \\
& =2\left|\frac{x-y}{2}\right|^{2}+2\left|\frac{\hat{x}-y}{2}\right|^{2}=\lambda^{2}
\end{aligned}
$$

Since $\frac{x+\hat{x}}{2} \in K$ due to convexity, this is a contradiction unless $x=\hat{x}$ since it shows that $\frac{x+\hat{x}}{2}$ is closer to $y$ than $\lambda$.

Now consider the minimizing sequence. From the same computation just given,

$$
\begin{aligned}
\left|\frac{x_{n}+x_{m}}{2}-y\right|^{2}+\left|\frac{x_{n}-x_{m}}{2}\right|^{2} & =2\left|\frac{x_{n}-y}{2}\right|^{2}+2\left|\frac{x_{m}-y}{2}\right|^{2} \\
& =\frac{1}{2}\left|x_{n}-y\right|^{2}+\frac{1}{2}\left|x_{m}-y\right|^{2}
\end{aligned}
$$

since $\frac{x_{n}+x_{m}}{2} \in K$,

$$
\left|\frac{x_{n}-x_{m}}{2}\right|^{2} \leq \frac{1}{2}\left|x_{n}-y\right|^{2}+\frac{1}{2}\left|x_{m}-y\right|^{2}-\lambda^{2}
$$

and as $n, m \rightarrow \infty$, the right side converges to 0 by definition. Hence $\left\{x_{n}\right\}$ is a Cauchy sequence as claimed. By completeness, it converges to some $x \in H$. Since $K$ is closed, it follows that $x \in K$. Then from the triangle inequality,

$$
\lim _{n \rightarrow \infty}\left|y-x_{n}\right|=|y-x|=\lambda .
$$

In the above theorem, denote by Py the vector $x \in K$ closest to $y$. It turns out there is an easy way to characterize $P y$. For a given $z \in K$, one can consider the function $t \rightarrow$ $|x+t(z-x)-y|^{2}$ for $x \in K$. By properties of the inner product, this is

$$
t \rightarrow|x-y|^{2}+2 t \operatorname{Re}(z-x, x-y)+t^{2}|z-x|^{2}
$$

according to whether $\operatorname{Re}(z-x, x-y) \geq 0$. Thus elementary considerations yield the two possibilities shown in the graph. Either this function is increasing on $[0,1]$ or it is not. In the case $\operatorname{Re}(z-x, x-y)<0$ the graph shows that $x \neq P y$ because there is a positive value of $t$ such that the function is less than $|x-y|^{2}$ and in case $\operatorname{Re}(z-x, x-y) \geq 0$, we obtain $x=P y$ if this is always true for any $z \in K$. Note that by convexity, $x+t(z-x) \in K$ for all $t \in[0,1]$ since it equals $(1-t) x+t z$.



Theorem 12.1.4 Let $x \in K$ and $y \in H$. Then there exists a closest point of $K$ to $y$ denoted by Py. Then $x=$ Py if and only if

$$
\begin{equation*}
\operatorname{Re}(z-x, y-x) \leq 0 \tag{12.1}
\end{equation*}
$$

for all $z \in K$.


Proof: First suppose 12.1 so $\operatorname{Re}(z-x, x-y) \geq 0$. Then for arbitrary $z \in K$,

$$
|x+t(z-x)-y|^{2}=|x-y|^{2}+2 t \operatorname{Re}(z-x, x-y)+t^{2}|z-x|^{2}
$$

and is an increasing function on $[0,1]$. Thus it has its minimum at $t=0$. In particular $|x-y|^{2} \leq|z-y|^{2}$. Since $z$ is arbitrary, this shows $x=P y$.

Next suppose $x=P y$. Then for arbitrary $z \in K$, the minimum of

$$
t \rightarrow|x+t(z-x)-y|^{2}
$$

occurs when $t=0$ since $x+t(z-x) \in K$. This will not happen unless

$$
\operatorname{Re}(z-x, x-y) \geq 0
$$

because if this is less than 0 , the minimum of that function will take place for some positive $t$. Thus 12.1 holds.

Every subspace is a closed and convex set. Note that Re and Im are real linear maps from $\mathbb{C}$ to $\mathbb{R}$.

$$
\operatorname{Re}(x+i y) \equiv x, \operatorname{Im}(x+i y) \equiv y
$$

That is, for $a, b$ real scalars and $z, w$ complex numbers,

$$
\begin{aligned}
& \operatorname{Re}(a z+b w)=a \operatorname{Re}(z)+b \operatorname{Re}(w) \\
& \operatorname{Im}(a z+b w)=a \operatorname{Im}(z)+b \operatorname{Im}(w)
\end{aligned}
$$

This assertions follow directly from the definitions of complex arithmetic and will be used without any mention whenever convenient. The next proposition will be very useful in what follows.

Proposition 12.1.5 If $W$ is a subspace. Then $\operatorname{Re}(z, w) \leq 0$ for all $w \in W$, if and only if $(z, w)=0$ for all $w \in W$.

Proof: $\Rightarrow$ First of all, $\operatorname{Re}(z,-w)=-\operatorname{Re}(z, w)$ so if $\operatorname{Re}(z, w) \leq 0$ for all $w \in W$, then for each $w \in W$,

$$
0 \geq \operatorname{Re}(z,-w)=-\operatorname{Re}(z, w) \geq 0
$$

Thus

$$
\operatorname{Re}(z, w)=-\operatorname{Re}(z,-w)=0
$$

Now also

$$
\begin{aligned}
(z, i w) & =\operatorname{Re}(z, i w)+i \operatorname{Im}(z, w)=-i(z, w) \\
& =-i[\operatorname{Re}(z, w)+i \operatorname{Im}(z, w)] \\
& =-i \operatorname{Re}(z, w)+\operatorname{Im}(z, w)
\end{aligned}
$$

and so $\operatorname{Im}(z, w)=(z, i w)$. Therefore, if $\operatorname{Re}(z, w) \leq 0$ for all $w \in W$, then $\operatorname{Re}(z, w)=0$ for all $w$ and hence $\operatorname{Im}(z, w)=0$ for all $w \in W$ and so $(z, w)=0$ for all $w \in W$.
$\Leftarrow$ Conversely, if $(z, w)=0$ for all $w \in W$, then obviously $\operatorname{Re}(z, w)=0$ for all $w \in W$.
Next is a fundamental result used in inner product spaces. It is called the Gram Schmidt process.

Lemma 12.1.6 Let $\left\{v_{1}, \cdots, v_{n}\right\}$ be a linearly independent subset of an inner product space $H$. Then there exists orthonormal vectors $\left\{u_{1}, \cdots, u_{n}\right\}$ which have the property that for each $k \leq n, \operatorname{span}\left(v_{1}, \cdots, v_{k}\right)=\operatorname{span}\left(u_{1}, \cdots, u_{k}\right)$.

Proof: Let $u_{1} \equiv v_{1} /\left|v_{1}\right|$. Thus for $k=1, \operatorname{span}\left(u_{1}\right)=\operatorname{span}\left(v_{1}\right)$ and $\left\{u_{1}\right\}$ is an orthonormal set. Now suppose for some $k<n, u_{1}, \cdots, u_{k}$ have been chosen such that $\left(u_{j}, u_{l}\right)=\delta_{j l}$ and $\operatorname{span}\left(v_{1}, \cdots, v_{k}\right)=\operatorname{span}\left(u_{1}, \cdots, u_{k}\right)$. Then define

$$
\begin{equation*}
u_{k+1} \equiv \frac{v_{k+1}-\sum_{j=1}^{k}\left(v_{k+1}, u_{j}\right) u_{j}}{\left|v_{k+1}-\sum_{j=1}^{k}\left(v_{k+1}, u_{j}\right) u_{j}\right|} \tag{12.2}
\end{equation*}
$$

where the denominator is not equal to zero because the $v_{j}$ form a basis, and so

$$
v_{k+1} \notin \operatorname{span}\left(v_{1}, \cdots, v_{k}\right)=\operatorname{span}\left(u_{1}, \cdots, u_{k}\right)
$$

Thus by induction,

$$
u_{k+1} \in \operatorname{span}\left(u_{1}, \cdots, u_{k}, v_{k+1}\right)=\operatorname{span}\left(v_{1}, \cdots, v_{k}, v_{k+1}\right)
$$

Also, $v_{k+1} \in \operatorname{span}\left(u_{1}, \cdots, u_{k}, u_{k+1}\right)$ which is seen easily by solving 10.13 for $v_{k+1}$, and it follows

$$
\operatorname{span}\left(v_{1}, \cdots, v_{k}, v_{k+1}\right)=\operatorname{span}\left(u_{1}, \cdots, u_{k}, u_{k+1}\right) .
$$

If $l \leq k$,

$$
\begin{gathered}
\left(u_{k+1}, u_{l}\right)=C\left(\left(v_{k+1}, u_{l}\right)-\sum_{j=1}^{k}\left(v_{k+1}, u_{j}\right)\left(u_{j}, u_{l}\right)\right)= \\
C\left(\left(v_{k+1}, u_{l}\right)-\sum_{j=1}^{k}\left(v_{k+1}, u_{j}\right) \delta_{l j}\right)=C\left(\left(v_{k+1}, u_{l}\right)-\left(v_{k+1}, u_{l}\right)\right)=0
\end{gathered}
$$

The vectors, $\left\{u_{j}\right\}_{j=1}^{n}$, generated in this way are therefore orthonormal because each vector has unit length.

Theorem 12.1.7 Let $K$ be a nonempty closed subspace of $H$ a Hilbert space. Let $y \in H$. Then $x=P y$, the closest point in $K$ to $y$ if and only if

$$
(y-x, w)=0
$$

for all $w \in K$. If $K$ is a finite dimensional subspace of $H$ then by Lemma 12.1.6 it has an orthonormal basis $\left\{u_{1}, \cdots, u_{n}\right\}$. Then $P y=\sum_{k=1}^{n}\left(y, u_{k}\right) u_{k}$. In particular, if $y \in K$, then $y=P y=\sum_{k=1}^{n}\left(y, u_{k}\right) u_{k}$

Proof: From Theorem 12.1.4, $x=P y, x \in K$ if and only if for all $z \in K$,

$$
\operatorname{Re}(y-x, z-x) \leq 0
$$

However, if $w \in K$, let $z=x+w$ and this shows that $x=P y$ if and only if for all $w \in K$,

$$
\operatorname{Re}(y-x, w) \leq 0
$$

From Proposition 12.1.5 above, this happens if and only if $(y-x, w)=0$.
It only remains to verify the orthogonality condition for the vector claimed to be the closest point.

$$
\left(y-\sum_{k=1}^{n}\left(y, u_{k}\right) u_{k}, u_{j}\right)=\left(y, u_{j}\right)-\sum_{k=1}^{n}\left(y, u_{k}\right)\left(u_{k}, u_{j}\right)=\left(y, u_{j}\right)-\left(y, u_{j}\right)=0
$$

and so, from the first part, Py is indeed given by the claimed formula.
Because of this theorem, $P y$ is called the orthogonal projection.
What if $H$ is not complete but $K$ is a finite dimensional subspace? Is it still the case that you can obtain a projection?

Proposition 12.1.8 Let $H$ be an inner product space, not necessarily complete and let $K$ be a finite dimensional subspace. Then if $u \in H$, a point $z \in K$ is the closest point to $u$ if and only if $(u-z, w)=0$ for all $w \in K$. Furthermore, there exists a closest point and it is given by $\sum_{i=1}^{n}\left(u, e_{i}\right) e_{i}$ where $\left\{e_{1}, \ldots, e_{n}\right\}$ is an orthonormal basis for $K$.

Proof: Suppose $z$ is the closest point to $u$ in $K$. Then if $w \in K,|u-(z+t w)|^{2}$ has a minimum at $t=0$. However, the function of $t$ has a derivative. The function of $t$ equals $|u-z|^{2}-2 t \operatorname{Re}(u-z, w)+t^{2}|w|^{2}$ and so its derivative is $-2 \operatorname{Re}(u-z, w)+t|w|^{2}$ and when $t=0$ this is to be zero so $\operatorname{Re}(u-z, w)=0$ for all $w \in K$. Now $(u-z, w)=$ $\operatorname{Re}(u-z, w)+i \operatorname{Im}(u-z, w)$ and so $(u-z, i w)=\operatorname{Re}(u-z, i w)+i \operatorname{Im}(u-z, i w)$ which implies that $-i(u-z, w)=-i \operatorname{Re}(u-z, w)+\operatorname{Im}(u-z, w)=\operatorname{Re}(u-z, i w)+i \operatorname{Im}(u-z, i w)$ so $\operatorname{Im}(u-z, w)=\operatorname{Re}(u-z, i w)$ and this shows that $\operatorname{Im}(u-z, w)=0=\operatorname{Re}(u-z, w)$ so $(u-z, w)=0$.

Next suppose $(u-z, w)=0$ for all $w \in K$. Then $|u-w|^{2}=|u-z+z-w|^{2}=|u-z|^{2}+$ $|z-w|^{2}$ because $2 \operatorname{Re}(u-z, z-w)=0$ and so it follows that $|u-z|^{2} \leq|u-w|^{2}$ for all $w \in K$.

It remains to verify that $\sum_{i=1}^{n}\left(e_{i}, u\right) e_{i}$ is as close as possible. From what was just shown, it suffices to verify that $\left(u-\sum_{i=1}^{n}\left(u, e_{i}\right) e_{i}, e_{k}\right)=0$ for all $e_{k}$. However, this is just $\left(u, e_{k}\right)-$ $\sum_{i}\left(u, e_{i}\right)\left(e_{i}, e_{k}\right)=\left(u, e_{k}\right)-\left(u, e_{k}\right)=0$.

Example 12.1.9 Consider $X$ equal to the continuous functions defined on $[-\pi, \pi]$ and let the inner product be given by

$$
\int_{-\pi}^{\pi} f(x) \overline{g(x)} d x
$$

It is left to the reader to verify that this is an inner product. Letting $e_{k}$ be the function $x \rightarrow \frac{1}{\sqrt{2 \pi}} e^{i k x}$, define $M \equiv \operatorname{span}\left(\left\{e_{k}\right\}_{k=-n}^{n}\right)$. Then you can verify that

$$
\left(e_{k}, e_{m}\right)=\int_{-\pi}^{\pi}\left(\frac{1}{\sqrt{2 \pi}} e^{-i k x}\right)\left(\frac{1}{\sqrt{2 \pi}} \bar{e}^{\overline{m i x}}\right) d x=\frac{1}{2 \pi} \int_{-\pi}^{\pi} e^{i(m-k) x}=\delta_{k m}
$$

then for a given function $f \in X$, the function from $M$ which is closest to $f$ in this inner product norm is $g=\sum_{k=-n}^{n}\left(f, e_{k}\right) e_{k}$ In this case $\left(f, e_{k}\right)=\frac{1}{\sqrt{2 \pi}} \int_{-\pi}^{\pi} f(x) e^{i k x} d x$. These are the Fourier coefficients. The above is the $n^{\text {th }}$ partial sum of the Fourier series.

To show how this kind of thing approximates a given function, let $f(x)=x^{2}$. Let $M=\operatorname{span}\left(\left\{\frac{1}{\sqrt{2 \pi}} e^{-i k x}\right\}_{k=-3}^{3}\right)$. Then, doing the computations, you find the closest point is of the form

$$
\begin{aligned}
& \frac{1}{3} \sqrt{2} \pi^{\frac{5}{2}}\left(\frac{1}{\sqrt{2 \pi}}\right)+\sum_{k=1}^{3}\left(\frac{(-1)^{k} 2}{k^{2}}\right) \sqrt{2} \sqrt{\pi} \frac{1}{\sqrt{2 \pi}} e^{-i k x} \\
& +\sum_{k=1}^{3}\left(\frac{(-1)^{k} 2}{k^{2}}\right) \sqrt{2} \sqrt{\pi} \frac{1}{\sqrt{2 \pi}} e^{i k x}
\end{aligned}
$$

and now simplify to get

$$
\frac{1}{3} \pi^{2}+\sum_{k=1}^{3}(-1)^{k}\left(\frac{4}{k^{2}}\right) \cos k x
$$

Then a graph of this along with the graph of $y=x^{2}$ is given below. In this graph, the dashed graph is of $y=x^{2}$ and the solid line is the graph of the above Fourier series approximation. If we had taken the partial sum up to $n$ much bigger, it would have
been very hard to distinguish between the graph of the partial sum
of the Fourier series and the graph of the function it is approximating.
This is in contrast to approximation by Taylor series in which you only the function on the entire interval.

### 12.2 Formula for Distance to a Subspace

Let $V$ be a finite dimensional subspace of a real inner product space $H$, for the sake of convenience, and suppose a basis for $V$ is $\left\{v_{1}, \ldots, v_{n}\right\}$. Thus this is a closed subspace. Then each point of $H$ has a closest point in $V$ thanks to Proposition 12.1.8. I want a convenient formula for the distance to $V$.

Definition 12.2.1 If $G_{i j} \equiv\left(v_{i}, v_{j}\right)$ where $\left\{v_{1}, \ldots, v_{n}\right\}$ are vectors, then $G$ is called the Grammian matrix, also the metric tensor. This matrix will also be denoted as $G\left(v_{1}, \ldots, v_{n}\right)$ to indicate the vectors used in defining $G$. Thus, it is an $n \times n$ matrix.

Proposition 12.2.2 $\left\{v_{1}, \ldots, v_{n}\right\}$ is linearly independent, if and only if $G\left(v_{1}, \ldots, v_{n}\right)$ is invertible.

Proof: If $G$ is invertible, then if $\sum_{i=1}^{n} x^{i} v_{i}=0, \sum_{i}\left(v_{j}, v_{i}\right) x^{i}=0$ and so $G \boldsymbol{x}=\mathbf{0}$ which can only hapen if $\boldsymbol{x}=\mathbf{0}$ because $G$ is invertible.

If $G$ is not invertible, then for some $\boldsymbol{x} \neq 0, \sum_{j} G_{i j} x^{j}=\sum_{j}\left(v_{i}, v_{j}\right) x^{j}=0$ for each $i$. However, this requires that $\left(\sum_{j} v_{j} x^{j}, v_{i}\right)=0$ for each $v_{i}$ and so $\sum_{j} v_{j} x^{j}=0$ where $\boldsymbol{x} \neq \mathbf{0}$ so $\left\{v_{1}, \ldots, v_{n}\right\}$ is not linearly independent.Thus, $G$ is invertible if and only if $\left\{v_{1}, \ldots, v_{n}\right\}$ is independent.

Let $V \equiv \operatorname{span}\left(v_{1}, \ldots, v_{n}\right)$ where these spanning vectors constitute a linearly independent set. Suppose $u \in H$. I want to find a convenient formula for the distance between $u$ and $V$. From Theorem 12.1.7, $P u \equiv z$, the projection of $u$ onto $V$ which is the closest point of $V$ to $u$, is defined by $\left(u-z, v_{i}\right)=0$ for all $v_{i}$ or equivalently $(u-z, v)=0$ for all $v \in V$. Thus, for $d$ the distance from $u$ to $V$,

$$
\begin{equation*}
|u|^{2}=|u-z|^{2}+|z|^{2}=d^{2}+|z|^{2},\left(u, v_{i}\right)=\left(z, v_{i}\right) \text { for each } i \tag{12.3}
\end{equation*}
$$

Let $z=\sum_{i=1}^{n} z^{i} v_{i}$. Then in the above,

$$
\left(u, v_{i}\right)=\left(z, v_{i}\right)=\left(\sum_{j=1}^{n} z^{j} v_{j}, v_{i}\right)=\sum_{j=1}^{n} G_{i j} z^{j}
$$

Letting $\boldsymbol{z} \equiv\left(z^{1}, \ldots, z^{n}\right)^{T}$ and $\boldsymbol{y} \equiv\left(\left(u, v_{1}\right),\left(u, v_{2}\right), \ldots,\left(u, v_{3}\right)\right)$,

$$
\begin{equation*}
G\left(v_{1}, \ldots, v_{n}\right) \boldsymbol{z}=\boldsymbol{y}, \boldsymbol{z}^{T} G\left(v_{1}, \ldots, v_{n}\right)=\boldsymbol{y}^{T} \tag{12.4}
\end{equation*}
$$

From 12.3 and 12.4,

$$
|u|^{2}=d^{2}+\sum_{i, j} G_{i j} z^{i} z^{j}=d^{2}+\boldsymbol{z}^{T} G\left(v_{1}, \ldots, v_{n}\right) \boldsymbol{z}=d^{2}+\boldsymbol{y}^{T} \boldsymbol{z}
$$

Then from 12.3 and 12.4,

$$
\left(\begin{array}{cc}
G\left(v_{1}, \ldots, v_{n}\right) & \mathbf{0} \\
\boldsymbol{y}^{T} & 1
\end{array}\right)\binom{\boldsymbol{z}}{d^{2}}=\binom{\boldsymbol{y}}{|u|^{2}}
$$

By Cramer's rule,

$$
d^{2}=\frac{\operatorname{det}\left(\begin{array}{cc}
G\left(v_{1}, \ldots, v_{n}\right) & \boldsymbol{y} \\
\boldsymbol{y}^{T} & |u|^{2}
\end{array}\right)}{\operatorname{det}\left(G\left(v_{1}, \ldots, v_{n}\right)\right)} \equiv \frac{\operatorname{det}\left(G\left(v_{1}, \ldots, v_{n}, u\right)\right)}{\operatorname{det}\left(G\left(v_{1}, \ldots, v_{n}\right)\right)}
$$

This proves the interesting approximation theorem.
Theorem 12.2.3 Suppose $\left\{v_{1}, \ldots, v_{n}\right\}$ is a linearly independent set of vectors in $H$ an inner product space. Then if $u \in H$, and $d$ is the distance to $V \equiv \operatorname{span}\left(v_{1}, \ldots, v_{n}\right)$, then $d^{2}=$ $\frac{\operatorname{det}\left(G\left(v_{1}, \ldots, v_{n}, u\right)\right)}{\operatorname{det}\left(G\left(v_{1}, \ldots, v_{n}\right)\right)}$.

### 12.3 Riesz Representation Theorem, Adjoint Map

The next theorem is one of the most important results in the theory of inner product spaces. It is called the Riesz representation theorem.

Theorem 12.3.1 Let $f \in \mathscr{L}(H, \mathbb{F})$ where $H$ is a Hilbert space and $f$ is continuous. Recall that in finite dimensions, this is automatic. Then there exists a unique $z \in H$ such that for all $x \in H, f(x)=(x, z)$.

Proof: First I will verify uniqueness. Suppose $z_{j}$ works for $j=1,2$. Then for all $x \in H$,

$$
0=f(x)-f(x)=\left(x, z_{1}-z_{2}\right)
$$

and so $z_{1}=z_{2}$.
If $f(H)=0$, let $z=0$ and this works. Otherwise, let $u \notin f^{-1}(0)$ which is a closed subspace of $H$. Let $w=u-P u \neq 0$. Then

$$
f(f(w) x-f(x) w)=f(w) f(x)-f(x) f(w)=0
$$

and so from Theorem 12.1.7,

$$
0=(f(w) x-f(x) w, w)=f(w)(x, w)-f(x)(w, w)
$$

It follows that for all $x, f(x)=\left(x, \frac{\overline{f(w)} w}{|w|^{2}}\right)$.
This leads to the following important definition.

Corollary 12.3.2 Let $A \in \mathscr{L}(X, Y)$ where $X$ and $Y$ are two inner product spaces of finite dimension or else Hilbert spaces. Then there exists a unique $A^{*} \in \mathscr{L}(Y, X)$, the bounded linear transformations, such that

$$
\begin{equation*}
(A x, y)_{Y}=\left(x, A^{*} y\right)_{X} \tag{12.5}
\end{equation*}
$$

for all $x \in X$ and $y \in Y$. The following formula holds

$$
(\alpha A+\beta B)^{*}=\bar{\alpha} A^{*}+\bar{\beta} B^{*}
$$

Also, $\left(A^{*}\right)^{*}=A$.
Proof: Let $f_{y} \in \mathscr{L}(X, \mathbb{F})$ be defined as

$$
f_{y}(x) \equiv(A x, y)_{Y}
$$

This is linear and

$$
\left|f_{y}(x)\right|=\left|(A x, y)_{Y}\right| \leq|A x||y| \leq(||A|||y|)|x|
$$

Then by the Riesz representation theorem, there exists a unique element of $X, A^{*}(y)$ such that

$$
(A x, y)_{Y}=\left(x, A^{*}(y)\right)_{X}
$$

It only remains to verify that $A^{*}$ is linear. Let $a$ and $b$ be scalars. Then for all $x \in X$,

$$
\begin{gathered}
\left(x, A^{*}\left(a y_{1}+b y_{2}\right)\right)_{X} \equiv\left(A x,\left(a y_{1}+b y_{2}\right)\right)_{Y} \\
\equiv \bar{a}\left(A x, y_{1}\right)+\bar{b}\left(A x, y_{2}\right) \equiv \\
\bar{a}\left(x, A^{*}\left(y_{1}\right)\right)+\bar{b}\left(x, A^{*}\left(y_{2}\right)\right)=\left(x, a A^{*}\left(y_{1}\right)+b A^{*}\left(y_{2}\right)\right) .
\end{gathered}
$$

Since this holds for every $x$, it follows

$$
A^{*}\left(a y_{1}+b y_{2}\right)=a A^{*}\left(y_{1}\right)+b A^{*}\left(y_{2}\right)
$$

which shows $A^{*}$ is linear as claimed.
Consider the last assertion that * is conjugate linear.

$$
\begin{aligned}
& \left(x,(\alpha A+\beta B)^{*} y\right) \equiv((\alpha A+\beta B) x, y) \\
= & \alpha(A x, y)+\beta(B x, y)=\alpha\left(x, A^{*} y\right)+\beta\left(x, B^{*} y\right) \\
= & \left(x, \bar{\alpha} A^{*} y\right)+\left(x, \bar{\beta} A^{*} y\right)=\left(x,\left(\bar{\alpha} A^{*}+\bar{\beta} A^{*}\right) y\right) .
\end{aligned}
$$

Since $x$ is arbitrary,

$$
(\alpha A+\beta B)^{*} y=\left(\bar{\alpha} A^{*}+\bar{\beta} A^{*}\right) y
$$

and since this is true for all $y$,

$$
(\alpha A+\beta B)^{*}=\bar{\alpha} A^{*}+\bar{\beta} A^{*}
$$

Finally, $\left(A^{*} x, y\right)=\overline{\left(y, A^{*} x\right)}=\overline{(A y, x)}=(x, A y)$ while $\left(A^{*} x, y\right)=\left(x,\left(A^{*}\right)^{*} y\right)$ and so for all $x$,

$$
\left(x,\left(A^{*}\right)^{*} y-A y\right)=0
$$

and so $\left(A^{*}\right)^{*}=A$.

Definition 12.3.3 The linear map, $A^{*}$ is called the adjoint of $A$. In the case when $A: X \rightarrow X$ and $A=A^{*}, A$ is called a self adjoint map. Such a map is also called Hermitian.

Theorem 12.3.4 Let $M$ be an $m \times n$ matrix. Then $M^{*}=(\bar{M})^{T}$ in words, the transpose of the conjugate of $M$ is equal to the adjoint.

Proof: Using the definition of the inner product in $\mathbb{C}^{n}$,

$$
(M \boldsymbol{x}, \boldsymbol{y})=\left(\boldsymbol{x}, M^{*} \boldsymbol{y}\right) \equiv \sum_{i} x_{i} \overline{\sum_{j}\left(M^{*}\right)_{i j} y_{j}}=\sum_{i, j} \overline{\left(M^{*}\right)_{i j}} \overline{y_{j}} x_{i} .
$$

Also $(\boldsymbol{M x}, \boldsymbol{y})=\sum_{j} \sum_{i} M_{j i} \overline{y_{j}} x_{i}$. Since $\boldsymbol{x}, \boldsymbol{y}$ are arbitrary vectors, it follows that $M_{j i}=\overline{\left(M^{*}\right)_{i j}}$ and so, taking conjugates of both sides, $M_{i j}^{*}=\overline{M_{j i}}$

Some linear transformations preserve distance. Something special can be asserted about these which is in the next lemma.

Lemma 12.3.5 Suppose $R \in \mathscr{L}(X, Y)$ where $X, Y$ are inner product spaces and $R$ preserves distances. Then $R^{*} R=I$.

Proof: Since $R$ preserves distances, $|R u|=|u|$ for every $u$. Let $u, v$ be arbitrary vectors in $X$

$$
\begin{aligned}
|u+v|^{2} & =|u|^{2}+|v|^{2}+2 \operatorname{Re}(u, v) \\
|R u+R v|^{2} & =|R u|^{2}+|R v|^{2}+2 \operatorname{Re}(R u, R v) \\
& =|u|^{2}+|v|^{2}+2 \operatorname{Re}\left(R^{*} R u, v\right)
\end{aligned}
$$

Thus $\operatorname{Re}\left(R^{*} R u-u, v\right)=0$ for all $v$ and so by Proposition 12.1.5, $\left(R^{*} R u-u, v\right)=0$ for all $v$ and so $R^{*} R u=u$ for all $u$ which implies $R^{*} R=I$.

The next theorem is interesting. You have a $p$ dimensional subspace of $\mathbb{F}^{n}$ where $\mathbb{F}=\mathbb{R}$ or $\mathbb{C}$. Of course this might be "slanted". However, there is a linear transformation $Q$ which preserves distances which maps this subspace to $\mathbb{F}^{p}$.

Theorem 12.3.6 Suppose $V$ is a subspace of $\mathbb{F}^{n}$ having dimension $p \leq n$. Then there exists $a Q \in \mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{n}\right)$ such that

$$
Q V \subseteq \operatorname{span}\left(\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{p}\right)
$$

and $|Q \boldsymbol{x}|=|\boldsymbol{x}|$ for all $\boldsymbol{x}$. Also

$$
Q^{*} Q=Q Q^{*}=I
$$

Proof: By Lemma 12.1.6 there exists an orthonormal basis for $V,\left\{\boldsymbol{v}_{i}\right\}_{i=1}^{p}$. By using the Gram Schmidt process this may be extended to an orthonormal basis of the whole space $\mathbb{F}^{n}$,

$$
\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{p}, \boldsymbol{v}_{p+1}, \cdots, \boldsymbol{v}_{n}\right\}
$$

Now define $Q \in \mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{n}\right)$ by $Q\left(\boldsymbol{v}_{i}\right) \equiv \boldsymbol{e}_{i}$ and extend linearly. If $\sum_{i=1}^{n} x_{i} \boldsymbol{v}_{i}$ is an arbitrary element of $\mathbb{F}^{n}$,

$$
\left|Q\left(\sum_{i=1}^{n} x_{i} \boldsymbol{v}_{i}\right)\right|^{2}=\left|\sum_{i=1}^{n} x_{i} \boldsymbol{e}_{i}\right|^{2}=\sum_{i=1}^{n}\left|x_{i}\right|^{2}=\left|\sum_{i=1}^{n} x_{i} \boldsymbol{v}_{i}\right|^{2}
$$

Thus $Q$ preserves lengths and so, by Lemma 12.3.5, it follows that $Q^{*} Q=I$. Also, this shows that $Q$ maps $V$ onto $V$ and so a generic element of $V$ is of the form $Q x$. Now

$$
\left|Q^{*} Q x\right|^{2}=\left(Q^{*} Q x, Q^{*} Q x\right)=(Q x, Q \overbrace{Q^{*} Q}^{=I})=(Q x, Q x)=|Q x|^{2}
$$

showing that $Q^{*}$ also preserves lengths. Hence it is also the case that $Q Q^{*}=I$ because from the definition of the adjoint, $\left(Q^{*}\right)^{*}=Q$.

Definition 12.3.7 If $U \in \mathscr{L}(X, X)$ for $X$ an inner product space, then $U$ is called unitary if $U^{*} U=U U^{*}=I$.

Note that it is actually shown that $Q V=\operatorname{span}\left(\boldsymbol{e}_{1}, \cdots, \boldsymbol{e}_{p}\right)$ and that in case $p=n$ one obtains that a linear transformation which maps an orthonormal basis to an orthonormal basis is unitary. Unitary matrices are also characterized by preserving length. More generally

Corollary 12.3.8 Suppose $U \in \mathscr{L}(X, X)$ where $X$ is an inner product space. Then $U$ is unitary if and only if $|U x|=|x|$ for all $x$ so it preserves distance.

Proof: $\Rightarrow$ If $U$ is unitary, then $|U x|^{2}=(U x, U x)=\left(U^{*} U x, x\right)=(x, x)=|x|^{2}$.
$\Leftarrow$ If $|U x|=|x|$ for all $x$ then by Lemma 12.3.5, $U^{*} U=I$. Thus $U$ is onto since it is one to one and so a generic element of $X$ is $U x$. Note how this would fail if you had $U \in \mathscr{L}(X, Y)$ where the dimension of $Y$ is larger than the dimension of $X$. Then as above,

$$
\left|U^{*} U x\right|^{2}=\left(U^{*} U x, U^{*} U x\right)=\left(U x, U U^{*} U x\right)=(U x, U x)=|U x|^{2}
$$

Thus also $U U^{*}=I$ because $U^{*}$ preserves distances and $\left(U^{*}\right)^{*}=U$ from the definition.
Now here is an important result on factorization of an $m \times n$ matrix. It is called a $Q R$ factorization.

Theorem 12.3.9 Let $A$ be an $m \times n$ complex matrix. Then there exists a unitary $Q$ and $R$, all zero below the main diagonal $\left(R_{i j}=0\right.$ if $\left.i>j\right)$ such that $A=Q R$.

Proof: This is obvious if $m=1$.

$$
\left(\begin{array}{lll}
a_{1} & \cdots & a_{n}
\end{array}\right)=(1)\left(\begin{array}{lll}
a_{1} & \cdots & a_{n}
\end{array}\right)
$$

Suppose true for $m-1$ and let

$$
A=\left(\begin{array}{lll}
\boldsymbol{a}_{1} & \cdots & \boldsymbol{a}_{n}
\end{array}\right), A \text { is } m \times n
$$

Using Theorem 12.3.6, there exists $Q_{1}$ a unitary matrix such that $Q_{1}\left(a_{1} /\left|\boldsymbol{a}_{1}\right|\right)=\boldsymbol{e}_{1}$ in case $\boldsymbol{a}_{1} \neq \mathbf{0}$. Thus $Q_{1} \boldsymbol{a}_{1}=\left|\boldsymbol{a}_{1}\right| \boldsymbol{e}_{1}$. If $\boldsymbol{a}_{1}=\mathbf{0}$, let $Q_{1}=I$. Thus

$$
Q_{1} A=\left(\begin{array}{ll}
a & \boldsymbol{b}^{T} \\
\mathbf{0} & A_{1}
\end{array}\right)
$$

where $A_{1}$ is $(m-1) \times(n-1)$. If $n=1$, this obtains

$$
Q_{1} A=\binom{a}{\mathbf{0}}, A=Q_{1}^{*}\binom{a}{\mathbf{0}}, \text { let } Q=Q_{1}^{*}
$$

That which is desired is obtained. So assume $n>1$. By induction, there exists $Q_{2}^{\prime}$ an $(m-1) \times(n-1)$ unitary matrix such that $Q_{2}^{\prime} A_{1}=R^{\prime}, R_{i j}^{\prime}=0$ if $i>j$. Then

$$
\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & Q_{2}^{\prime}
\end{array}\right) Q_{1} A=\left(\begin{array}{cc}
a & \boldsymbol{b}^{T} \\
\mathbf{0} & R^{\prime}
\end{array}\right)=R
$$

Since the product of unitary matrices is unitary, there exists $Q$ unitary such that $Q^{*} A=R$ and so $A=Q R$.

### 12.4 Least Squares

A common problem in experimental work is to find a straight line which approximates as well as possible a collection of points in the plane $\left\{\left(x_{i}, y_{i}\right)\right\}_{i=1}^{p}$. The usual way of dealing with these problems is by the method of least squares and it turns out that all these sorts of approximation problems can be reduced to $A \boldsymbol{x}=\boldsymbol{b}$ where the problem is to find the best $\boldsymbol{x}$ for solving this equation even when there is no solution.

Lemma 12.4.1 Let $V$ and $W$ be finite dimensional inner product spaces and let $A: V \rightarrow W$ be linear. For each $y \in W$ there exists $x \in V$ such that

$$
|A x-y| \leq\left|A x_{1}-y\right|
$$

for all $x_{1} \in V$. Also, $x \in V$ is a solution to this minimization problem if and only if $x$ is a solution to the equation, $A^{*} A x=A^{*} y$.

Proof: By Theorem 12.1.7 on Page 310 there exists a point, $A x_{0}$, in the finite dimensional subspace, $A(V)$, of $W$ such that for all $x \in V,|A x-y|^{2} \geq\left|A x_{0}-y\right|^{2}$. Also, from this theorem, this happens if and only if $A x_{0}-y$ is perpendicular to every $A x \in A(V)$. Therefore, the solution is characterized by $\left(A x_{0}-y, A x\right)=0$ for all $x \in V$ which is the same as saying $\left(A^{*} A x_{0}-A^{*} y, x\right)=0$ for all $x \in V$. In other words the solution is obtained by solving $A^{*} A x_{0}=A^{*} y$ for $x_{0}$.

Consider the problem of finding the least squares regression line in statistics. Suppose you have given points in the plane, $\left\{\left(x_{i}, y_{i}\right)\right\}_{i=1}^{n}$ and you would like to find constants $m$ and $b$ such that the line $y=m x+b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find $m, b$ such that you do the best you can to solve the system

$$
\left(\begin{array}{c}
y_{1} \\
\vdots \\
y_{n}
\end{array}\right)=\left(\begin{array}{cc}
x_{1} & 1 \\
\vdots & \vdots \\
x_{n} & 1
\end{array}\right)\binom{m}{b}
$$

which is of the form $\boldsymbol{y}=A \boldsymbol{x}$. In other words try to make $\left|A\binom{m}{b}-\left(\begin{array}{c}y_{1} \\ \vdots \\ y_{n}\end{array}\right)\right|^{2}$ as small as possible. According to what was just shown, it is desired to solve the following for $m$ and $b$.

$$
A^{*} A\binom{m}{b}=A^{*}\left(\begin{array}{c}
y_{1} \\
\vdots \\
y_{n}
\end{array}\right)
$$

Since $A^{*}=A^{T}$ in this case,

$$
\left(\begin{array}{cc}
\sum_{i=1}^{n} x_{i}^{2} & \sum_{i=1}^{n} x_{i} \\
\sum_{i=1}^{n} x_{i} & n
\end{array}\right)\binom{m}{b}=\binom{\sum_{i=1}^{n} x_{i} y_{i}}{\sum_{i=1}^{n} y_{i}}
$$

Solving this system of equations for $m$ and $b$,

$$
m=\frac{-\left(\sum_{i=1}^{n} x_{i}\right)\left(\sum_{i=1}^{n} y_{i}\right)+\left(\sum_{i=1}^{n} x_{i} y_{i}\right) n}{\left(\sum_{i=1}^{n} x_{i}^{2}\right) n-\left(\sum_{i=1}^{n} x_{i}\right)^{2}}
$$

and

$$
b=\frac{-\left(\sum_{i=1}^{n} x_{i}\right) \sum_{i=1}^{n} x_{i} y_{i}+\left(\sum_{i=1}^{n} y_{i}\right) \sum_{i=1}^{n} x_{i}^{2}}{\left(\sum_{i=1}^{n} x_{i}^{2}\right) n-\left(\sum_{i=1}^{n} x_{i}\right)^{2}}
$$

One could clearly do a least squares fit for curves of the form $y=a x^{2}+b x+c$ in the same way. In this case you solve as well as possible for $a, b$, and $c$ the system

$$
\left(\begin{array}{ccc}
x_{1}^{2} & x_{1} & 1 \\
\vdots & \vdots & \vdots \\
x_{n}^{2} & x_{n} & 1
\end{array}\right)\left(\begin{array}{l}
a \\
b \\
c
\end{array}\right)=\left(\begin{array}{c}
y_{1} \\
\vdots \\
y_{n}
\end{array}\right)
$$

using the same techniques.

### 12.5 Fredholm Alternative

The best context in which to study the Fredholm alternative is in inner product spaces. This is done here.

Definition 12.5.1 Let $S$ be a subset of an inner product space, $X$. Define

$$
S^{\perp} \equiv\{x \in X:(x, s)=0 \text { for all } s \in S\}
$$

The following theorem also follows from the above lemma. It is sometimes called the Fredholm alternative.

Theorem 12.5.2 Let $A: V \rightarrow W$ where $A$ is linear and $V$ and $W$ are inner product spaces. Then $A(V)=\operatorname{ker}\left(A^{*}\right)^{\perp}$.

Proof: Let $y=A x$ so $y \in A(V)$. Then if $A^{*} z=0$,

$$
(y, z)=(A x, z)=\left(x, A^{*} z\right)=0
$$

showing that $y \in \operatorname{ker}\left(A^{*}\right)^{\perp}$. Thus $A(V) \subseteq \operatorname{ker}\left(A^{*}\right)^{\perp}$.
Now suppose $y \in \operatorname{ker}\left(A^{*}\right)^{\perp}$. Does there exists $x$ such that $A x=y$ ? Since this might not be immediately clear, take the least squares solution to the problem. Thus let $x$ be a solution to $A^{*} A x=A^{*} y$. It follows $A^{*}(y-A x)=0$ and so $y-A x \in \operatorname{ker}\left(A^{*}\right)$ which implies from the assumption about $y$ that $(y-A x, y)=0$. Also, since $A x$ is the closest point to $y$ in $A(V)$, Theorem 12.1.7 on Page 310 implies that $\left(y-A x, A x_{1}\right)=0$ for all $x_{1} \in V$. In particular this is true for $x_{1}=x$ and so $0=(y-A x, y)-\overbrace{(y-A x, A x)}^{=0}=|y-A x|^{2}$, showing that $y=A x$. Thus $A(V) \supseteq \operatorname{ker}\left(A^{*}\right)^{\perp}$.

Corollary 12.5.3 Let $A, V$, and $W$ be as described above. If the only solution to $A^{*} y=0$ is $y=0$, then $A$ is onto $W$.

Proof: If the only solution to $A^{*} y=0$ is $y=0$, then $\operatorname{ker}\left(A^{*}\right)=\{0\}$ and so every vector from $W$ is contained in $\operatorname{ker}\left(A^{*}\right)^{\perp}$ and by the above theorem, this shows $A(V)=W$.

### 12.6 The Determinant and Volume

The determinant is the essential algebraic tool which provides a way to give a unified treatment of the concept of $p$ dimensional volume. Here is the definition of what is meant by such a thing. In what follows, $X$ will be typically some $\mathbb{R}^{m}$.

Definition 12.6.1 Let $u_{1}, \cdots, u_{p}$ be vectors in some inner product space $X$. The parallelepiped determined by these vectors will be denoted by $P\left(u_{1}, \cdots, u_{p}\right)$ and it is defined as

$$
P\left(u_{1}, \cdots, u_{p}\right) \equiv\left\{\sum_{j=1}^{p} s_{j} u_{j}: s_{j} \in[0,1]\right\}=U Q, Q=[0,1]^{p}
$$

The volume of this parallelepiped is defined as

$$
\text { volume of } P\left(u_{1}, \cdots, u_{p}\right) \equiv v\left(P\left(u_{1}, \cdots, u_{p}\right)\right) \equiv(\operatorname{det}(G))^{1 / 2} .
$$

where $G_{i j}=u_{i} \cdot u_{j}$. This $G$ is called the metric tensor, sometimes the Grammian matrix. Let $G\left(u_{1}, \cdots, u_{p}\right)$ denote the metric tensor determined by $u_{1}, \cdots, u_{p}$. The vectors $u_{i}$ are dependent, if and only if the $p$ dimensional volume just defined gives 0 . That is, if and only if $\operatorname{det}\left(G\left(u_{1}, \cdots, u_{p}\right)\right)=0$.

The last assertion follows from Proposition 12.2.2.
I am going to show that this is the only reasonable definition of volume for such a parallelepiped if you desire to preserve Euclidean ideas of distance and volume. Here is a

picture which shows the relation between $P\left(u_{1}, \cdots, u_{p-1}\right)$ and $P\left(u_{1}, \cdots, u_{p}\right)$.
In particular, if you have a parallelepiped $P\left(u_{1}, \cdots, u_{p-1}\right)$, then by adding another vector $u$ not in the span of the $\left\{u_{1}, \cdots, u_{p-1}\right\}$ you would want the $p$ dimensional volume of $P\left(u_{1}, \cdots, u_{p-1}, u\right)$ to equal the distance from $u$ to the subspace spanned by $u_{1}, \cdots, u_{p-1}$ multiplied by the $p-1$ dimensional volume of $P\left(u_{1}, \cdots, u_{p-1}\right), v\left(P\left(u_{1}, \cdots, u_{p-1}\right)\right)$. Thus, from Theorem 12.2.3, assuming $P\left(u_{1}, \cdots, u_{p-1}\right) \neq 0$ so that $\operatorname{det} G\left(u_{1}, \cdots, u_{p-1}\right) \neq 0$ you would want the $p$ dimensional volume to satisfy

$$
\begin{aligned}
v\left(P\left(u_{1}, \cdots, u_{p-1}, u\right)\right)^{2} & =\frac{\operatorname{det}\left(G\left(u_{1}, \cdots, u_{p-1}, u\right)\right)}{\operatorname{det}\left(G\left(u_{1}, \cdots, u_{p-1}\right)\right)} v\left(P\left(u_{1}, \cdots, u_{p-1}\right)\right)^{2} \\
& =\operatorname{det}\left(G\left(u_{1}, \cdots, u_{p-1}, u\right)\right)
\end{aligned}
$$

and so it follows that this is the only geometrically reasonable definition of the volume of a parallelepiped if the one dimensional volume is $\operatorname{det}(G(v))^{1 / 2}$, the Euclidean length. Clearly if $v=0$, this gives what the volume should be, 0 . If $v \neq 0$, then $P(v)$ is just a line of the form $0+t v: t \in[0,1]$ and the endpoints would be 0 and $v$. We would want the one dimensional volume of this line to be its length. But if length is to be defined in terms of the Pythagorean theorem, this length is just $(v, v)^{1 / 2}=\operatorname{det}(G(v))^{1 / 2}$. Therefore, the above is the only reasonable definition of Euclidean volume.

### 12.7 Finding an Orthogonal Basis

The Gram Schmidt process described above gives a way to generate an orthogonal set of vectors from a linearly independent set. Is there a convenient way to do this? Probably not. However, if you have access to a computer algebra system there might be a way which could help. In the following lemma, $v_{i}$ will be a vector and it is assumed that $v_{i}, i=1, \ldots, n$ are linearly independent.

Lemma 12.7.1 Let $\left\{v_{1}, \ldots, v_{n}\right\}$ be linearly independent and consider the following formal determinant:

$$
\operatorname{det}\left(\begin{array}{ccccc}
\left(v_{1}, v_{1}\right) & \left(v_{1}, v_{2}\right) & \cdots & \left(v_{1}, v_{n-1}\right) & v_{1} \\
\left(v_{2}, v_{1}\right) & \left(v_{2}, v_{2}\right) & \cdots & \left(v_{2}, v_{n-1}\right) & v_{2} \\
\vdots & \vdots & & \vdots & \vdots \\
\left(v_{n-1}, v_{1}\right) & \left(v_{n-1}, v_{2}\right) & \cdots & \left(v_{n-1}, v_{n-1}\right) & v_{n-1} \\
\left(v_{n}, v_{1}\right) & \left(v_{n}, v_{2}\right) & \cdots & \left(v_{n}, v_{n-1}\right) & v_{n}
\end{array}\right)
$$

Then the vector which results from expanding this determinant formally is perpendicular to each of $v_{1}, \ldots, v_{n-1}$.

Proof: It is of the form $\sum_{i=1}^{n} v_{i} C_{i}$ where $C_{i}$ is a suitable $(n-1) \times(n-1)$ determinant. Thus the inner product of this with $v_{k}$ for $k \leq n-1$ is the expansion of a determinant which has two equal columns. However, the inner product with $v_{n}$ will be the Grammian of $\left\{v_{1}, \ldots, v_{n}\right\}$ which is not zero since these vectors $v_{i}$ are independent, this by Proposition 12.2.2

Example 12.7.2 The vectors $1, x, x^{2}, x^{3}$ are linearly independent on $[0,1]$, the vector space being the continuous functions defined on $[0,1]$. You might show this. An inner product is given by $\int_{0}^{1} f(x) g(x) d x$. Find an orthogonal basis for $\operatorname{span}\left(1, x, x^{2}, x^{3}\right)$.

You could use the above lemma. $u_{1}(x)=1$. Now I will assemble the formal determinants as given above.

$$
\operatorname{det}\left(\begin{array}{cc}
1 & 1 \\
\frac{1}{2} & x
\end{array}\right), \operatorname{det}\left(\begin{array}{ccc}
1 & \frac{1}{2} & 1 \\
\frac{1}{2} & \frac{1}{3} & x \\
\frac{1}{3} & \frac{1}{4} & x^{2}
\end{array}\right), \operatorname{det}\left(\begin{array}{cccc}
1 & \frac{1}{2} & \frac{1}{3} & 1 \\
\frac{1}{2} & \frac{1}{3} & \frac{1}{4} & x \\
\frac{1}{3} & \frac{1}{4} & \frac{1}{5} & x^{2} \\
\frac{1}{4} & \frac{1}{5} & \frac{1}{6} & x^{3}
\end{array}\right)
$$

Now the orthogonal basis is obtained from evaluating these determinants and adding 1 to the list. Thus an orthonormal basis is
$\left\{1, x-\frac{1}{2}, \frac{1}{12} x^{2}-\frac{1}{12} x+\frac{1}{72}, \frac{1}{2160} x^{3}-\frac{1}{1440} x^{2}+\frac{1}{3600} x-\frac{1}{43200}\right\}$

Is this horrible? Yes it is. However, if you have a computer algebra system do it for you, it isn't so bad. For example, to get the last term, you just do

$$
\left(\begin{array}{c}
1 \\
x \\
x^{2} \\
x^{3}
\end{array}\right)\left(\begin{array}{lll}
1 & x & x^{2}
\end{array}\right)=\left(\begin{array}{ccc}
1 & x & x^{2} \\
x & x^{2} & x^{3} \\
x^{2} & x^{3} & x^{4} \\
x^{3} & x^{4} & x^{5}
\end{array}\right)
$$

Then you do the following.

$$
\int_{0}^{1}\left(\begin{array}{ccc}
1 & x & x^{2} \\
x & x^{2} & x^{3} \\
x^{2} & x^{3} & x^{4} \\
x^{3} & x^{4} & x^{5}
\end{array}\right) d x=\left(\begin{array}{ccc}
1 & \frac{1}{2} & \frac{1}{3} \\
\frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\
\frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\
\frac{1}{4} & \frac{1}{5} & \frac{1}{6}
\end{array}\right)
$$

You could get Matlab to do it for you. Then you add in the last column which consists of the original vectors. If you wanted an orthonormal basis, you could divide each vector by its magnitude. This was only painless because I let the computer do all the tedious busy work. However, I think it has independent interest because it gives a formula for a vector which will be orthogonal to a given set of linearly independent vectors.

### 12.8 Exercises

1. Find the best solution to the system

$$
\begin{gathered}
x+2 y=6 \\
2 x-y=5 \\
3 x+2 y=0
\end{gathered}
$$

2. Find an orthonormal basis for $\mathbb{R}^{3},\left\{\boldsymbol{w}_{1}, \boldsymbol{w}_{2}, \boldsymbol{w}_{3}\right\}$ given that $\boldsymbol{w}_{1}$ is a multiple of the vector ( $1,1,2$ ).
3. Suppose $A=A^{T}$ is a symmetric real $n \times n$ matrix which has all positive eigenvalues. Define

$$
(\boldsymbol{x}, \boldsymbol{y}) \equiv(A \boldsymbol{x}, \boldsymbol{y})
$$

Show this is an inner product on $\mathbb{R}^{n}$. What does the Cauchy Schwarz inequality say in this case?
4. Let $\|\boldsymbol{x}\|_{\infty} \equiv \max \left\{\left|x_{j}\right|: j=1,2, \cdots, n\right\}$. Show this is a norm on $\mathbb{C}^{n}$. Here

$$
\boldsymbol{x}=\left(\begin{array}{lll}
x_{1} & \cdots & x_{n}
\end{array}\right)^{T}
$$

Show $\|x\|_{\infty} \leq|\boldsymbol{x}| \equiv(\boldsymbol{x}, \boldsymbol{x})^{1 / 2}$ where the above is the usual inner product on $\mathbb{C}^{n}$.
5. Let $\|\boldsymbol{x}\|_{1} \equiv \sum_{j=1}^{n}\left|x_{j}\right|$. Show this is a norm on $\mathbb{C}^{n}$. Here $\boldsymbol{x}=\left(\begin{array}{lll}x_{1} & \cdots & x_{n}\end{array}\right)^{T}$. Show $\|\boldsymbol{x}\|_{1} \geq|\boldsymbol{x}| \equiv(\boldsymbol{x}, \boldsymbol{x})^{1 / 2}$. where the above is the usual inner product on $\mathbb{C}^{n}$. Show there cannot exist an inner product such that this norm comes from the inner product as described above for inner product spaces.
6. Show that if $\|\cdot\|$ is any norm on any vector space, then $\|\|x\|-\| y\|\|\leq\| x-y\|$.
7. Relax the assumptions in the axioms for the inner product. Change the axiom about $(x, x) \geq 0$ and equals 0 if and only if $x=0$ to simply read $(x, x) \geq 0$. Show the Cauchy Schwarz inequality still holds in the following form. $|(x, y)| \leq(x, x)^{1 / 2}(y, y)^{1 / 2}$.
8. Let $H$ be an inner product space and let $\left\{u_{k}\right\}_{k=1}^{n}$ be an orthonormal basis for $H$. Show

$$
(x, y)=\sum_{k=1}^{n}\left(x, u_{k}\right) \overline{\left(y, u_{k}\right)}
$$

9. Let the vector space $V$ consist of real polynomials of degree no larger than 3. Thus a typical vector is a polynomial of the form $a+b x+c x^{2}+d x^{3}$. For $p, q \in V$ define the inner product, $(p, q) \equiv \int_{0}^{1} p(x) q(x) d x$. Show this is indeed an inner product. Then state the Cauchy Schwarz inequality in terms of this inner product. Show $\left\{1, x, x^{2}, x^{3}\right\}$ is a basis for $V$. Finally, find an orthonormal basis for $V$. This is an example of some orthonormal polynomials.
10. Let $P_{n}$ denote the polynomials of degree no larger than $n-1$ which are defined on an interval $[a, b]$. Let $\left\{x_{1}, \cdots, x_{n}\right\}$ be $n$ distinct points in $[a, b]$. Now define for $p, q \in P_{n}$,

$$
(p, q) \equiv \sum_{j=1}^{n} p\left(x_{j}\right) \overline{q\left(x_{j}\right)}
$$

Show this yields an inner product on $P_{n}$. Hint: Most of the axioms are obvious. The one which says $(p, p)=0$ if and only if $p=0$ is the only interesting one. To verify this one, note that a nonzero polynomial of degree no more than $n-1$ has at most $n-1$ zeros.
11. Let $C([0,1])$ denote the vector space of continuous real valued functions defined on $[0,1]$. Let the inner product be given as

$$
(f, g) \equiv \int_{0}^{1} f(x) g(x) d x
$$

Show this is an inner product. Also let $V$ be the subspace described in Problem 9. Using the result of this problem, find the vector in $V$ which is closest to $x^{4}$.
12. A regular Sturm Liouville problem involves the differential equation, for an unknown function of $x$ which is denoted here by $y$,

$$
\left(p(x) y^{\prime}\right)^{\prime}+(\lambda q(x)+r(x)) y=0, x \in[a, b]
$$

and it is assumed that $p(t), q(t)>0$ for any $t \in[a, b]$ and also there are boundary conditions,

$$
\begin{aligned}
& C_{1} y(a)+C_{2} y^{\prime}(a)=0 \\
& C_{3} y(b)+C_{4} y^{\prime}(b)=0
\end{aligned}
$$

where

$$
C_{1}^{2}+C_{2}^{2}>0, \text { and } C_{3}^{2}+C_{4}^{2}>0
$$

There is an immense theory connected to these important problems. The constant, $\lambda$ is called an eigenvalue. Show that if $y$ is a solution to the above problem corresponding to $\lambda=\lambda_{1}$ and if $z$ is a solution corresponding to $\lambda=\lambda_{2} \neq \lambda_{1}$, then

$$
\begin{equation*}
\int_{a}^{b} q(x) y(x) z(x) d x=0 \tag{12.6}
\end{equation*}
$$

and this defines an inner product. Hint: Do something like this:

$$
\begin{aligned}
& \left(p(x) y^{\prime}\right)^{\prime} z+\left(\lambda_{1} q(x)+r(x)\right) y z=0 \\
& \left(p(x) z^{\prime}\right)^{\prime} y+\left(\lambda_{2} q(x)+r(x)\right) z y=0 .
\end{aligned}
$$

Now subtract and either use integration by parts or show

$$
\left(p(x) y^{\prime}\right)^{\prime} z-\left(p(x) z^{\prime}\right)^{\prime} y=\left(\left(p(x) y^{\prime}\right) z-\left(p(x) z^{\prime}\right) y\right)^{\prime}
$$

and then integrate. Use the boundary conditions to show that

$$
y^{\prime}(a) z(a)-z^{\prime}(a) y(a)=0
$$

and $y^{\prime}(b) z(b)-z^{\prime}(b) y(b)=0$. The formula, 12.6 is called an orthogonality relation. It turns out there are typically infinitely many eigenvalues and it is interesting to write given functions as an infinite series of these "eigenfunctions".
13. Consider the continuous functions defined on $[0, \pi], C([0, \pi])$. Show that the expres$\operatorname{sion}(f, g) \equiv \int_{0}^{\pi} f g d x$ is an inner product on this vector space. Show the functions

$$
\left\{\sqrt{\frac{2}{\pi}} \sin (n x)\right\}_{n=1}^{\infty}
$$

are an orthonormal set. What does this mean about the dimension of the vector space $C([0, \pi])$ ? Now let

$$
V_{N}=\operatorname{span}\left(\sqrt{\frac{2}{\pi}} \sin (x), \cdots, \sqrt{\frac{2}{\pi}} \sin (N x)\right)
$$

For $f \in C([0, \pi])$ find a formula for the vector in $V_{N}$ which is closest to $f$ with respect to the norm determined from the above inner product. This is called the $N^{t h}$ partial sum of the Fourier series of $f$. An important problem is to determine whether and in what way this Fourier series converges to the function $f$. The norm which comes from this inner product is sometimes called the mean square norm.
14. Consider the subspace $V \equiv \operatorname{ker}(A)$ where

$$
A=\left(\begin{array}{cccc}
1 & 4 & -1 & -1 \\
2 & 1 & 2 & 3 \\
4 & 9 & 0 & 1 \\
5 & 6 & 3 & 4
\end{array}\right)
$$

Find an orthonormal basis for $V$. Hint: You might first find a basis and then use the Gram Schmidt procedure.
15. The Gram Schmidt process starts with a basis for a subspace $\left\{v_{1}, \cdots, v_{n}\right\}$ and produces an orthonormal basis for the same subspace $\left\{u_{1}, \cdots, u_{n}\right\}$ such that

$$
\operatorname{span}\left(v_{1}, \cdots, v_{k}\right)=\operatorname{span}\left(u_{1}, \cdots, u_{k}\right)
$$

for each $k$. Show that in the case of $\mathbb{R}^{m}$ the $Q R$ factorization does the same thing. Specifically, if $A=\left(\begin{array}{lll}\boldsymbol{v}_{1} & \cdots & \boldsymbol{v}_{n}\end{array}\right)$ and if $A=Q R \equiv\left(\begin{array}{lll}\boldsymbol{q}_{1} & \cdots & \boldsymbol{q}_{n}\end{array}\right) R$ then the vectors $\left\{\boldsymbol{q}_{1}, \cdots, \boldsymbol{q}_{n}\right\}$ is an orthonormal set of vectors and for each $k$,

$$
\operatorname{span}\left(\boldsymbol{q}_{1}, \cdots, \boldsymbol{q}_{k}\right)=\operatorname{span}\left(\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{k}\right)
$$

16. Verify the parallelogram identify for any inner product space,

$$
|x+y|^{2}+|x-y|^{2}=2|x|^{2}+2|y|^{2} .
$$

Why is it called the parallelogram identity?
17. Let $H$ be an inner product space and let $K \subseteq H$ be a nonempty convex subset. This means that if $k_{1}, k_{2} \in K$, then the line segment consisting of points of the form

$$
t k_{1}+(1-t) k_{2} \text { for } t \in[0,1]
$$

is also contained in $K$. Suppose for each $x \in H$, there exists $P x$ defined to be a point of $K$ closest to $x$. Show that $P x$ is unique so that $P$ actually is a map. Hint: Suppose $z_{1}$ and $z_{2}$ both work as closest points. Consider the midpoint, $\left(z_{1}+z_{2}\right) / 2$ and use the parallelogram identity of Problem 16 in an auspicious manner.
18. In the situation of Problem 17 suppose $K$ is a closed convex subset and that $H$ is complete. This means every Cauchy sequence converges. Recall a sequence $\left\{k_{n}\right\}$ is a Cauchy sequence if for every $\varepsilon>0$ there exists $N_{\varepsilon}$ such that whenever $m, n>N_{\varepsilon}$, it follows $\left|k_{m}-k_{n}\right|<\varepsilon$. Let $\left\{k_{n}\right\}$ be a sequence of points of $K$ such that

$$
\lim _{n \rightarrow \infty}\left|x-k_{n}\right|=\inf \{|x-k|: k \in K\}
$$

This is called a minimizing sequence. Show there exists a unique $k \in K$ such that $\lim _{n \rightarrow \infty}\left|k_{n}-k\right|$ and that $k=P x$. That is, there exists a well defined projection map onto the convex subset of $H$. Hint: Use the parallelogram identity in an auspicious manner to show $\left\{k_{n}\right\}$ is a Cauchy sequence which must therefore converge. Since $K$ is closed it follows this will converge to something in $K$ which is the desired vector.
19. Let $H$ be an inner product space which is also complete and let $P$ denote the projection map onto a convex closed subset, $K$. Show this projection map is characterized by the inequality $\operatorname{Re}(k-P x, x-P x) \leq 0$ for all $k \in K$. That is, a point $z \in K$ equals $P x$ if and only if the above variational inequality holds. This is what that inequality is called. This is because $k$ is allowed to vary and the inequality continues to hold for all $k \in K$.
20. Using Problem 19 and Problems 17-18 show the projection map, $P$ onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is $|P x-P y| \leq$ $|x-y|$
21. Give an example of two vectors in $\mathbb{R}^{4}$ or $\mathbb{R}^{3} \boldsymbol{x}, \boldsymbol{y}$ and a subspace $V$ such that $\boldsymbol{x} \cdot \boldsymbol{y}=0$ but $P \boldsymbol{x} \cdot P \boldsymbol{y} \neq 0$ where $P$ denotes the projection map which sends $\boldsymbol{x}$ to its closest point on $V$.
22. Suppose you are given the data, $(1,2),(2,4),(3,8),(0,0)$. Find the linear regression line using the formulas derived above. Then graph the given data along with your regression line.
23. Generalize the least squares procedure to the situation in which data is given and you desire to fit it with an expression of the form $y=a f(x)+b g(x)+c$ where the problem would be to find $a, b$ and $c$ in order to minimize the error. Could this be generalized to higher dimensions? How about more functions?
24. Let $A \in \mathscr{L}(X, Y)$ where $X$ and $Y$ are finite dimensional vector spaces with the dimension of $X$ equal to $n$. Define $\operatorname{rank}(A) \equiv \operatorname{dim}(A(X))$ and nullity $(A) \equiv \operatorname{dim}(\operatorname{ker}(A))$. Show that $\operatorname{nullity}(A)+\operatorname{rank}(A)=\operatorname{dim}(X)$. Hint: Let $\left\{x_{i}\right\}_{i=1}^{r}$ be a basis for $\operatorname{ker}(A)$ and let $\left\{x_{i}\right\}_{i=1}^{r} \cup\left\{y_{i}\right\}_{i=1}^{n-r}$ be a basis for $X$. Then show that $\left\{A y_{i}\right\}_{i=1}^{n-r}$ is linearly independent and spans $A X$.
25. Let $A$ be an $m \times n$ matrix. Show the column rank of $A$ equals the column rank of $A^{*} A$. Next verify column rank of $A^{*} A$ is no larger than column rank of $A^{*}$. Next justify the following inequality to conclude the column rank of $A$ equals the column rank of $A^{*}$.

$$
\begin{gathered}
\operatorname{rank}(A)=\operatorname{rank}\left(A^{*} A\right) \leq \operatorname{rank}\left(A^{*}\right) \leq \\
=\operatorname{rank}\left(A A^{*}\right) \leq \operatorname{rank}(A)
\end{gathered}
$$

Hint: Start with an orthonormal basis, $\left\{A \boldsymbol{x}_{j}\right\}_{j=1}^{r}$ of $A\left(\mathbb{F}^{n}\right)$ and verify $\left\{A^{*} A \boldsymbol{x}_{j}\right\}_{j=1}^{r}$ is a basis for $A^{*} A\left(\mathbb{F}^{n}\right)$.
26. Let $A$ be a real $m \times n$ matrix and let $A=Q R$ be the $Q R$ factorization with $Q$ orthogonal and $R$ upper triangular. Show that there exists a solution $\boldsymbol{x}$ to the equation

$$
R^{T} R \boldsymbol{x}=R^{T} Q^{T} \boldsymbol{b}
$$

and that this solution is also a least squares solution defined above such that $A^{T} A \boldsymbol{x}=$ $A^{T} \boldsymbol{b}$.
27. Here are three vectors in $\mathbb{R}^{4}:(1,2,0,3)^{T},(2,1,-3,2)^{T},(0,0,1,2)^{T}$. Find the three dimensional volume of the parallelepiped determined by these three vectors.
28. Here are two vectors in $\mathbb{R}^{4}:(1,2,0,3)^{T},(2,1,-3,2)^{T}$. Find the volume of the parallelepiped determined by these two vectors.
29. Here are three vectors in $\mathbb{R}^{2}:(1,2)^{T},(2,1)^{T},(0,1)^{T}$. Find the three dimensional volume of the parallelepiped determined by these three vectors. Recall that from the above theorem, this should equal 0 .
30. Find the equation of the plane through the three points

$$
(1,2,3),(2,-3,1),(1,1,7) .
$$

31. Let $T$ map a vector space $V$ to itself. Explain why $T$ is one to one if and only if $T$ is onto. It is in the text, but do it again in your own words.
32. $\uparrow$ Let all matrices be complex with complex field of scalars and let $A$ be an $n \times n$ matrix and $B$ a $m \times m$ matrix while $X$ will be an $n \times m$ matrix. The problem is to consider solutions to Sylvester's equation. Solve the following equation for $X$

$$
A X-X B=C
$$

where $C$ is an arbitrary $n \times m$ matrix. Show there exists a unique solution if and only if $\sigma(A) \cap \sigma(B)=\emptyset$. Hint: If $q(\lambda)$ is a polynomial, show first that if $A X-X B=0$, then $q(A) X-X q(B)=0$. Next define the linear map $T$ which maps the $n \times m$ matrices to the $n \times m$ matrices as follows. $T X \equiv A X-X B$. Show that the only solution to $T X=0$ is $X=0$ so that $T$ is one to one if and only if $\sigma(A) \cap \sigma(B)=\emptyset$. Do this by using the first part for $q(\lambda)$ the characteristic polynomial for $B$ and then use the Cayley Hamilton theorem. Explain why $q(A)^{-1}$ exists if and only if the condition $\sigma(A) \cap \sigma(B)=\emptyset$.
33. What is the geometric significance of the Binet Cauchy theorem, Theorem 8.4.5?
34. Let $U, H$ be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let $A$ be a linear map from $U$ to $H$. Thus $A U$ is a subspace of $H$. For $\boldsymbol{g} \in A U$, define $A^{-1} \boldsymbol{g}$ to be the unique element of $\{\boldsymbol{x}: A \boldsymbol{x}=\boldsymbol{g}\}$ which is closest to $\mathbf{0}$. Then define $(\boldsymbol{h}, \boldsymbol{g})_{A U} \equiv\left(A^{-1} \boldsymbol{g}, A^{-1} \boldsymbol{h}\right)_{U}$. Show that this is a well defined inner product. Let $U, H$ be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let $A$ be a linear map from $U$ to $H$. Thus $A U$ is a subspace of $H$. For $\boldsymbol{g} \in A U$, define $A^{-1} \boldsymbol{g}$ to be the unique element of $\{\boldsymbol{x}: A \boldsymbol{x}=\boldsymbol{g}\}$ which is closest to 0 . Then define $(\boldsymbol{h}, \boldsymbol{g})_{A U} \equiv\left(A^{-1} \boldsymbol{g}, A^{-1} \boldsymbol{h}\right)_{U}$. Show that this is a well defined inner product and that if $A$ is one to one, then $\|\boldsymbol{h}\|_{A U}=\left\|A^{-1} \boldsymbol{h}\right\|_{U}$ and $\|A \boldsymbol{x}\|_{A U}=\|\boldsymbol{x}\|_{U}$.
35. For $f$ a piecewise continuous function,

$$
S_{n} f(x)=\frac{1}{2 \pi} \sum_{k=-n}^{n} e^{i k x}\left(\int_{-\pi}^{\pi} f(y) e^{-i k y} d y\right)
$$

where $S_{n} f(x)$ denotes the $n^{\text {th }}$ partial sum of the Fourier series. Recall that this Fourier series was of the form

$$
\sum_{k=-n}^{n} a_{n} \frac{1}{\sqrt{2 \pi}} e^{i k x}, a_{n} \equiv \frac{1}{\sqrt{2 \pi}} \int_{-\pi}^{\pi} f(y) e^{-i k y} d y
$$

Show this can be written in the form

$$
S_{n} f(x)=\int_{-\pi}^{\pi} f(y) D_{n}(x-y) d y
$$

where

$$
D_{n}(t)=\frac{1}{2 \pi} \sum_{k=-n}^{n} e^{i k t}
$$

This is called the Dirichlet kernel. Show that

$$
D_{n}(t)=\frac{1}{2 \pi} \frac{\sin (n+(1 / 2)) t}{\sin (t / 2)}
$$

For $V$ the vector space of piecewise continuous functions, define $S_{n}: V \mapsto V$ by

$$
S_{n} f(x)=\int_{-\pi}^{\pi} f(y) D_{n}(x-y) d y
$$

Show that $S_{n}$ is a linear transformation. (In fact, $S_{n} f$ is not just piecewise continuous but infinitely differentiable. Why?) Explain why $\int_{-\pi}^{\pi} D_{n}(t) d t=1$. Hint: To obtain the formula, do the following.

$$
\begin{aligned}
e^{i(t / 2)} D_{n}(t) & =\frac{1}{2 \pi} \sum_{k=-n}^{n} e^{i(k+(1 / 2)) t} \\
e^{i(-t / 2)} D_{n}(t) & =\frac{1}{2 \pi} \sum_{k=-n}^{n} e^{i(k-(1 / 2)) t}
\end{aligned}
$$

Change the variable of summation in the bottom sum and then subtract and solve for $D_{n}(t)$.
36. $\uparrow$ Let $V$ be an inner product space and let $U$ be a finite dimensional subspace with an orthonormal basis $\left\{u_{i}\right\}_{i=1}^{n}$. If $\boldsymbol{y} \in V$, show

$$
|y|^{2} \geq \sum_{k=1}^{n}\left|\left\langle y, u_{k}\right\rangle\right|^{2}
$$

Let $\left\{\boldsymbol{u}_{k}\right\}_{k=1}^{\infty}$ be an orthonormal set of vectors of $V$. Explain why $\lim _{k \rightarrow \infty}\left\langle y, u_{k}\right\rangle=0$. When applied to functions, this is a special case of the Riemann Lebesgue lemma.
37. $\uparrow$ Let $f$ be any piecewise continuous real function which is bounded on $[-\pi, \pi]$. Show, using the above problem, that

$$
\lim _{n \rightarrow \infty} \int_{-\pi}^{\pi} f(t) \sin (n t) d t=\lim _{n \rightarrow \infty} \int_{-\pi}^{\pi} f(t) \cos (n t) d t=0
$$

38. $\uparrow^{*}$ Let $f$ be a function which is defined on $(-\pi, \pi]$. The $2 \pi$ periodic extension is given by the formula $f(x+2 \pi)=f(x)$. In the rest of this problem, $f$ will refer to this $2 \pi$ periodic extension. Assume that $f$ is piecewise continuous, bounded, and also that the following limits exist for this $2 \pi$ extension.

$$
\lim _{y \rightarrow 0+} \frac{f(x+y)-f(x+)}{y}, \lim _{y \rightarrow 0+} \frac{f(x-y)-f(x+)}{y}
$$

Here it is assumed that $f(x+) \equiv \lim _{h \rightarrow 0+} f(x+h), \quad f(x-) \equiv \lim _{h \rightarrow 0+} f(x-h)$. both exist at every point. The above conditions rule out functions where the slope taken from either side becomes infinite. Actually, you don't need anything about these quotients being bounded. It is enough to have what is called a Dini condition which is a bound involving a Holder condition and it gives the quotients in $L^{1}$ but this kind of thing involves more analysis. The above result is still very interesting. Justify
the following assertions and eventually conclude that under these very reasonable conditions (more general ones are possible.)

$$
\lim _{n \rightarrow \infty} S_{n} f(x)=(f(x+)+f(x-)) / 2
$$

the mid point of the jump. In words, the Fourier series converges to the midpoint of the jump of the function.

$$
S_{n} f(x)=\int_{-\pi}^{\pi} f(y) D_{n}(x-y) d y=\int_{-\pi}^{\pi} f(x-y) D_{n}(y) d y
$$

You just change variables and then use $2 \pi$ periodicity to get this.

$$
\begin{aligned}
& \left|S_{n} f(x)-\frac{f(x+)+f(x-)}{2}\right| \\
= & \left|\int_{-\pi}^{\pi}\left(f(x-y)-\frac{f(x+)+f(x-)}{2}\right) D_{n}(y) d y\right| \\
= & \mid \int_{0}^{\pi} f(x-y) D_{n}(y) d y+\int_{0}^{\pi} f(x+y) D_{n}(y) d y \\
& -\int_{0}^{\pi}(f(x+)+f(x-)) D_{n}(y) d y \mid \\
\leq & \left|\int_{0}^{\pi}(f(x-y)-f(x-)) D_{n}(y) d y\right|+\left|\int_{0}^{\pi}(f(x+y)-f(x+)) D_{n}(y) d y\right|
\end{aligned}
$$

Now apply some trig. identities and use the result of Problem 37 to conclude that both of these terms must converge to 0 .

## Chapter 13

## Matrices and the Inner Product

### 13.1 Schur's Theorem, Hermitian Matrices

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Schur's theorem and it is the most important theorem in the spectral theory of matrices. The important result which makes this theorem possible is the Gram Schmidt procedure of Lemma 10.5.13.

Definition 13.1.1 An $n \times n$ matrix $U$, is unitary if $U U^{*}=I=U^{*} U$ where $U^{*}$ is defined to be the transpose of the conjugate of $U$. Thus $\overline{U_{i j}}=U_{j i}^{*}$. Note that every real orthogonal, meaning $Q^{T} Q=I$, matrix is unitary. For $A$ any matrix, $A^{*}$, just defined as the conjugate of the transpose, is called the adjoint. As shown above, this is also defined by

$$
(A \boldsymbol{x}, \boldsymbol{y})=\left(\boldsymbol{x}, A^{*} \boldsymbol{y}\right)
$$

Note that if $U=\left(\begin{array}{lll}\boldsymbol{v}_{1} & \cdots & \boldsymbol{v}_{n}\end{array}\right)$ where the $\boldsymbol{v}_{k}$ are orthonormal vectors in $\mathbb{C}^{n}$, then $U$ is unitary. This follows because the $i j^{\text {th }}$ entry of $U^{*} U$ is $\overline{\boldsymbol{v}_{i}^{T}} \boldsymbol{v}_{j}=\delta_{i j}$ since the $\boldsymbol{v}_{i}$ are assumed orthonormal.

Lemma 13.1.2 The following holds. $(A B)^{*}=B^{*} A^{*}$.
Proof: Using the definition in terms of inner products,

$$
\left(\boldsymbol{x},(A B)^{*} \boldsymbol{y}\right)=(A B \boldsymbol{x}, \boldsymbol{y})=\left(B \boldsymbol{x}, A^{*} \boldsymbol{y}\right)=\left(\boldsymbol{x}, B^{*} A^{*} \boldsymbol{y}\right)
$$

and so, since $\boldsymbol{x}$ is arbitrary, $(A B)^{*} \boldsymbol{y}=B^{*} A^{*} \boldsymbol{y}$ which shows the result since $\boldsymbol{y}$ is arbitrary.

Theorem 13.1.3 Let $A$ be an $n \times n$ matrix. Then there exists a unitary matrix $U$ such that

$$
\begin{equation*}
U^{*} A U=T \tag{13.1}
\end{equation*}
$$

where $T$ is an upper triangular matrix having the eigenvalues of $A$ on the main diagonal listed according to multiplicity as roots of the characteristic equation. If A is a real matrix having all real eigenvalues, then $U$ can be chosen to be an orthogonal real matrix.

Proof: The theorem is clearly true if $A$ is a $1 \times 1$ matrix. Just let $U=1$, the $1 \times 1$ matrix which has entry 1 . Suppose it is true for $(n-1) \times(n-1)$ matrices, $n \geq 2$ and let $A$ be an $n \times n$ matrix. Then let $\boldsymbol{v}_{1}$ be a unit eigenvector for $A$. Then there exists $\lambda_{1}$ such that

$$
A \boldsymbol{v}_{1}=\lambda_{1} \boldsymbol{v}_{1},\left|\boldsymbol{v}_{1}\right|=1
$$

Extend $\left\{\boldsymbol{v}_{1}\right\}$ to a basis and then use the Gram - Schmidt process or Theorem 12.3.6 to obtain $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$, an orthonormal basis of $\mathbb{C}^{n}$. Let $U_{0}$ be a matrix whose $i^{\text {th }}$ column is $\boldsymbol{v}_{i}$ so that $U_{0}$ is unitary. Consider $U_{0}^{*} A U_{0}$

$$
U_{0}^{*} A U_{0}=\left(\begin{array}{c}
\boldsymbol{v}_{1}^{*} \\
\vdots \\
\boldsymbol{v}_{n}^{*}
\end{array}\right)\left(\begin{array}{lll}
A \boldsymbol{v}_{1} & \cdots & A \boldsymbol{v}_{n}
\end{array}\right)=\left(\begin{array}{c}
\boldsymbol{v}_{1}^{*} \\
\vdots \\
\boldsymbol{v}_{n}^{*}
\end{array}\right)\left(\begin{array}{lll}
\lambda_{1} \boldsymbol{v}_{1} & \cdots & A \boldsymbol{v}_{n}
\end{array}\right)
$$

Thus $U_{0}^{*} A U_{0}$ is of the form

$$
\left(\begin{array}{cc}
\lambda_{1} & a \\
0 & A_{1}
\end{array}\right)
$$

where $A_{1}$ is an $n-1 \times n-1$ matrix. Now by induction, there exists an $(n-1) \times(n-1)$ unitary matrix $\widetilde{U}_{1}$ such that $\widetilde{U}_{1}^{*} A_{1} \widetilde{U}_{1}=T_{n-1}$, an upper triangular matrix. Consider

$$
U_{1} \equiv\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & \widetilde{U}_{1}
\end{array}\right)
$$

Then

$$
U_{1}^{*} U_{1}=\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & \widetilde{U}_{1}^{*}
\end{array}\right)\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & \widetilde{U}_{1}
\end{array}\right)=\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & I_{n-1}
\end{array}\right)
$$

Also

$$
\begin{aligned}
U_{1}^{*} U_{0}^{*} A U_{0} U_{1} & =\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & \widetilde{U}_{1}^{*}
\end{array}\right)\left(\begin{array}{cc}
\lambda_{1} & * \\
\mathbf{0} & A_{1}
\end{array}\right)\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & \widetilde{U}_{1}
\end{array}\right) \\
& =\left(\begin{array}{cc}
\lambda_{1} & * \\
\mathbf{0} & T_{n-1}
\end{array}\right) \equiv T
\end{aligned}
$$

where $T$ is upper triangular. Then let $U=U_{0} U_{1}$. It is clear that this is unitary because both matrices preserve distance. Therefore, so does the product and hence $U$. Alternatively,

$$
I=U_{0} U_{1} U_{1}^{*} U_{0}^{*}=\left(U_{0} U_{1}\right)\left(U_{0} U_{1}\right)^{*}
$$

and so, it follows that $A$ is similar to $T$ and that $U_{0} U_{1}$ is unitary. Hence $A$ and $T$ have the same characteristic polynomials, and therefore the same eigenvalues listed according to multiplicity as roots of the characteristic equation. These are the diagonal entries of $T$ listed with multiplicity and so this proves the main conclusion of the theorem. In case $A$ is real with all real eigenvalues, the above argument can be repeated word for word using only the real dot product to show that $U$ can be taken to be real and orthogonal.

As a simple consequence of the above theorem, here is an interesting lemma.
Lemma 13.1.4 Let A be of the form

$$
A=\left(\begin{array}{ccc}
P_{1} & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & P_{s}
\end{array}\right)
$$

where $P_{k}$ is an $m_{k} \times m_{k}$ matrix. Then

$$
\operatorname{det}(A)=\prod_{k} \operatorname{det}\left(P_{k}\right)
$$

Proof: Let $U_{k}$ be an $m_{k} \times m_{k}$ unitary matrix such that

$$
U_{k}^{*} P_{k} U_{k}=T_{k}
$$

where $T_{k}$ is upper triangular. Then letting $U$ denote the block diagonal matrix, having the $U_{i}$ as the blocks on the diagonal,

$$
U=\left(\begin{array}{ccc}
U_{1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & U_{s}
\end{array}\right), U^{*}=\left(\begin{array}{ccc}
U_{1}^{*} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & U_{s}^{*}
\end{array}\right)
$$

and

$$
\left(\begin{array}{ccc}
U_{1}^{*} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & U_{s}^{*}
\end{array}\right)\left(\begin{array}{ccc}
P_{1} & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & P_{s}
\end{array}\right)\left(\begin{array}{ccc}
U_{1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & U_{s}
\end{array}\right)=\left(\begin{array}{ccc}
T_{1} & \cdots & * \\
\vdots & \ddots & \vdots \\
0 & \cdots & T_{s}
\end{array}\right)
$$

and so

$$
\operatorname{det}(A)=\prod_{k} \operatorname{det}\left(T_{k}\right)=\prod_{k} \operatorname{det}\left(P_{k}\right)
$$

Definition 13.1.5 An $n \times n$ matrix $A$ is called Hermitian if $A=A^{*}$. Thus a real symmetric ( $A=A^{T}$ ) matrix is Hermitian.

The following is the major result about Hermitian matrices. It says that any Hermitian matrix is similar to a diagonal matrix. We say it is unitarily similar because the matrix $U$ in the following theorem which gives the similarity transformation is a unitary matrix.

Theorem 13.1.6 If $A$ is an $n \times n$ Hermitian matrix, there exists a unitary matrix $U$ such that

$$
\begin{equation*}
U^{*} A U=D \tag{13.2}
\end{equation*}
$$

where $D$ is a real diagonal matrix. That is, $D$ has nonzero entries only on the main diagonal and these are real. Furthermore, the columns of $U$ are an orthonormal basis of eigenvectors for $\mathbb{C}^{n}$. If $A$ is real and symmetric, then $U$ can be assumed to be a real orthogonal matrix and the columns of $U$ form an orthonormal basis for $\mathbb{R}^{n}$.

Proof: From Schur's theorem above, there exists $U$ unitary (real and orthogonal if $A$ is real) such that

$$
U^{*} A U=T
$$

where $T$ is an upper triangular matrix. Then from Lemma 13.1.2

$$
T^{*}=\left(U^{*} A U\right)^{*}=U^{*} A^{*} U=U^{*} A U=T
$$

Thus $T=T^{*}$ and $T$ is upper triangular. This can only happen if $T$ is really a diagonal matrix having real entries on the main diagonal. (If $i \neq j$, one of $T_{i j}$ or $T_{j i}$ equals zero. But $T_{i j}=\overline{T_{j i}}$ and so they are both zero. Also $T_{i i}=\overline{T_{i i}}$.)

Finally, let

$$
U=\left(\begin{array}{llll}
\boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \cdots & \boldsymbol{u}_{n}
\end{array}\right)
$$

where the $\boldsymbol{u}_{i}$ denote the columns of $U$ and

$$
D=\left(\begin{array}{lll}
\lambda_{1} & & 0 \\
& \ddots & \\
0 & & \lambda_{n}
\end{array}\right)
$$

The equation, $U^{*} A U=D$ implies

$$
\begin{aligned}
A U & =\left(\begin{array}{llll}
A \boldsymbol{u}_{1} & A \boldsymbol{u}_{2} & \cdots & A \boldsymbol{u}_{n}
\end{array}\right) \\
& =U D=\left(\begin{array}{llll}
\lambda_{1} \boldsymbol{u}_{1} & \lambda_{2} \boldsymbol{u}_{2} & \cdots & \lambda_{n} \boldsymbol{u}_{n}
\end{array}\right)
\end{aligned}
$$

where the entries denote the columns of $A U$ and $U D$ respectively. Therefore, $A \boldsymbol{u}_{i}=\lambda_{i} \boldsymbol{u}_{i}$ and since the matrix is unitary, the $i j^{t h}$ entry of $U^{*} U$ equals $\delta_{i j}$ and so

$$
\delta_{i j}=\overline{\boldsymbol{u}}_{i}^{T} \boldsymbol{u}_{j}=\overline{\boldsymbol{u}_{i}^{T} \overline{\boldsymbol{u}}_{j}}=\overline{\boldsymbol{u}_{i} \cdot \boldsymbol{u}_{j}} .
$$

This proves the corollary because it shows the vectors $\left\{\boldsymbol{u}_{i}\right\}$ form an orthonormal basis. In case $A$ is real and symmetric, simply ignore all complex conjugations in the above argument.

This theorem is particularly nice because the diagonal entries are all real. What of a matrix which is unitarily similar to a diagonal matrix without assuming the diagonal entries are real? That is, $A$ is an $n \times n$ matrix with

$$
U^{*} A U=D
$$

Then this requires

$$
U^{*} A^{*} U=D^{*}
$$

and so since the two diagonal matrices commute,

$$
\begin{aligned}
A A^{*} & =U D U^{*} U D^{*} U^{*}=U D D^{*} U^{*}=U D^{*} D U^{*} \\
& =U D^{*} U^{*} U D U^{*}=A^{*} A
\end{aligned}
$$

The following definition describes these matrices.
Definition 13.1.7 An $n \times n$ matrix is normal means: $A^{*} A=A A^{*}$.
We just showed that if $A$ is unitarily similar to a diagonal matrix, then it is normal. The converse is also true. This involves the following lemma.

Lemma 13.1.8 If $T$ is upper triangular and normal, then $T$ is a diagonal matrix. If $A$ is normal and $U$ is unitary, then $U^{*} A U$ is also normal.

Proof: This is obviously true if $T$ is $1 \times 1$. In fact, it can't help being diagonal in this case. Suppose then that the lemma is true for $(n-1) \times(n-1)$ matrices and let $T$ be an upper triangular normal $n \times n$ matrix. Thus $T$ is of the form

$$
T=\left(\begin{array}{cc}
t_{11} & \boldsymbol{a}^{*} \\
\mathbf{0} & T_{1}
\end{array}\right), T^{*}=\left(\begin{array}{cc}
\overline{t_{11}} & \mathbf{0}^{T} \\
\boldsymbol{a} & T_{1}^{*}
\end{array}\right)
$$

Then

$$
\begin{aligned}
& T T^{*}=\left(\begin{array}{cc}
t_{11} & \boldsymbol{a}^{*} \\
\mathbf{0} & T_{1}
\end{array}\right)\left(\begin{array}{cc}
\overline{t_{11}} & \mathbf{0}^{T} \\
\boldsymbol{a} & T_{1}^{*}
\end{array}\right)=\left(\begin{array}{cc}
\left|t_{11}\right|^{2}+\boldsymbol{a}^{*} \boldsymbol{a} & \boldsymbol{a}^{*} T_{1}^{*} \\
T_{1} \boldsymbol{a} & T_{1} T_{1}^{*}
\end{array}\right) \\
& T^{*} T=\left(\begin{array}{cc}
\overline{t_{11}} & \mathbf{0}^{T} \\
\boldsymbol{a} & T_{1}^{*}
\end{array}\right)\left(\begin{array}{cc}
t_{11} & \boldsymbol{a}^{*} \\
\mathbf{0} & T_{1}
\end{array}\right)=\left(\begin{array}{cc}
\left|t_{11}\right|^{2} & \overline{t_{11}} \boldsymbol{a}^{*} \\
\boldsymbol{a} t_{11} & \boldsymbol{a} \boldsymbol{a}^{*}+T_{1}^{*} T_{1}
\end{array}\right)
\end{aligned}
$$

Since these two matrices are equal, it follows $\boldsymbol{a}=\mathbf{0}$. But now it follows that $T_{1}^{*} T_{1}=T_{1} T_{1}^{*}$ and so by induction $T_{1}$ is a diagonal matrix $D_{1}$. Therefore,

$$
T=\left(\begin{array}{cc}
t_{11} & \mathbf{0}^{T} \\
\mathbf{0} & D_{1}
\end{array}\right)
$$

a diagonal matrix.
As to the last claim, let $A$ be normal. Then

$$
\begin{aligned}
\left(U^{*} A U\right)^{*}\left(U^{*} A U\right) & =U^{*} A^{*} U U^{*} A U=U^{*} A^{*} A U \\
& =U^{*} A A^{*} U=U^{*} A U U^{*} A^{*} U \\
& =\left(U^{*} A U\right)\left(U^{*} A U\right)^{*}
\end{aligned}
$$

Theorem 13.1.9 An $n \times n$ matrix is unitarily similar to a diagonal matrix if and only if it is normal.

Proof: It was already shown above that if $A$ is similar to a diagonal matrix then it is normal. Suppose now that $A$ is normal. By Schur's theorem, there is a unitary matrix $U$ such that

$$
U^{*} A U=T
$$

where $T$ is upper triangular. By Lemma 13.1.8, $T$ is normal and, since it is upper triangular, it is a diagonal matrix.

### 13.2 Quadratic Forms

Definition 13.2.1 A quadratic form in three dimensions is an expression of the form

$$
\left(\begin{array}{lll}
x & y & z
\end{array}\right) A\left(\begin{array}{l}
x  \tag{13.3}\\
y \\
z
\end{array}\right)
$$

where $A$ is a $3 \times 3$ symmetric matrix. In higher dimensions the idea is the same except you use a larger symmetric matrix in place of $A$. In two dimensions $A$ is a $2 \times 2$ matrix.

For example, consider

$$
\left(\begin{array}{lll}
x & y & z
\end{array}\right)\left(\begin{array}{ccc}
3 & -4 & 1  \tag{13.4}\\
-4 & 0 & -4 \\
1 & -4 & 3
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)
$$

which equals $3 x^{2}-8 x y+2 x z-8 y z+3 z^{2}$. This is very awkward because of the mixed terms such as $-8 x y$. The idea is to pick different axes such that if $x, y, z$ are taken with respect to these axes, the quadratic form is much simpler. In other words, look for new variables, $x^{\prime}, y^{\prime}$, and $z^{\prime}$ and a unitary matrix $U$ such that

$$
U\left(\begin{array}{l}
x^{\prime}  \tag{13.5}\\
y^{\prime} \\
z^{\prime}
\end{array}\right)=\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)
$$

and if you write the quadratic form in terms of the primed variables, there will be no mixed terms. Any symmetric real matrix is Hermitian and is therefore normal. From Corollary 13.1.6, it follows there exists a real unitary matrix $U$, (an orthogonal matrix) such that $U^{T} A U=D$ a diagonal matrix. Thus in the quadratic form, 13.3

$$
\begin{aligned}
\left(\begin{array}{lll}
x & y & z
\end{array}\right) A\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right) & =\left(\begin{array}{lll}
x^{\prime} & y^{\prime} & z^{\prime}
\end{array}\right) U^{T} A U\left(\begin{array}{l}
x^{\prime} \\
y^{\prime} \\
z^{\prime}
\end{array}\right) \\
& =\left(\begin{array}{lll}
x^{\prime} & y^{\prime} & z^{\prime}
\end{array}\right) D\left(\begin{array}{l}
x^{\prime} \\
y^{\prime} \\
z^{\prime}
\end{array}\right)
\end{aligned}
$$

and in terms of these new variables, the quadratic form becomes

$$
\lambda_{1}\left(x^{\prime}\right)^{2}+\lambda_{2}\left(y^{\prime}\right)^{2}+\lambda_{3}\left(z^{\prime}\right)^{2}
$$

where $D=\operatorname{diag}\left(\lambda_{1}, \lambda_{2}, \lambda_{3}\right)$. Similar considerations apply equally well in any other dimension. For the given example,

$$
\begin{aligned}
& \left(\begin{array}{ccc}
-\frac{1}{2} \sqrt{2} & 0 & \frac{1}{2} \sqrt{2} \\
\frac{1}{6} \sqrt{6} & \frac{1}{3} \sqrt{6} & \frac{1}{6} \sqrt{6} \\
\frac{1}{3} \sqrt{3} & -\frac{1}{3} \sqrt{3} & \frac{1}{3} \sqrt{3}
\end{array}\right)\left(\begin{array}{ccc}
3 & -4 & 1 \\
-4 & 0 & -4 \\
1 & -4 & 3
\end{array}\right) . \\
& \left(\begin{array}{ccc}
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\
0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}}
\end{array}\right)=\left(\begin{array}{ccc}
2 & 0 & 0 \\
0 & -4 & 0 \\
0 & 0 & 8
\end{array}\right)
\end{aligned}
$$

and so if the new variables are given by

$$
\left(\begin{array}{ccc}
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\
0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}}
\end{array}\right)\left(\begin{array}{c}
x^{\prime} \\
y^{\prime} \\
z^{\prime}
\end{array}\right)=\left(\begin{array}{c}
x \\
y \\
z
\end{array}\right)
$$

it follows that in terms of the new variables the quadratic form is $2\left(x^{\prime}\right)^{2}-4\left(y^{\prime}\right)^{2}+8\left(z^{\prime}\right)^{2}$. You can work other examples the same way.

### 13.3 The Estimation Of Eigenvalues

There are ways to estimate the eigenvalues for matrices. The most famous is known as Gerschgorin's theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

Theorem 13.3.1 Let A be an $n \times n$ matrix. Consider the $n$ Gerschgorin discs defined as

$$
D_{i} \equiv\left\{\lambda \in \mathbb{C}:\left|\lambda-a_{i i}\right| \leq \sum_{j \neq i}\left|a_{i j}\right|\right\} .
$$

Then every eigenvalue is contained in some Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the $i^{\text {th }}$ row which are off the main diagonal and form the disc centered at $a_{i i}$ having this radius. The union of these discs contains $\sigma(A)$.

Proof: Suppose $A \boldsymbol{x}=\boldsymbol{\lambda} \boldsymbol{x}$ where $\boldsymbol{x} \neq \mathbf{0}$. Then for $A=\left(a_{i j}\right)$, let $\left|x_{k}\right| \geq\left|x_{j}\right|$ for all $x_{j}$. Thus $\left|x_{k}\right| \neq 0$.

$$
\sum_{j \neq k} a_{k j} x_{j}=\left(\lambda-a_{k k}\right) x_{k} .
$$

Then

$$
\left|x_{k}\right| \sum_{j \neq k}\left|a_{k j}\right| \geq \sum_{j \neq k}\left|a_{k j}\right|\left|x_{j}\right| \geq\left|\sum_{j \neq k} a_{k j} x_{j}\right|=\left|\lambda-a_{i i}\right|\left|x_{k}\right|
$$

Now dividing by $\left|x_{k}\right|$, it follows $\lambda$ is contained in the $k^{t h}$ Gerschgorin disc.
Example 13.3.2 Here is a matrix. Estimate its eigenvalues.

$$
\left(\begin{array}{lll}
2 & 1 & 1 \\
3 & 5 & 0 \\
0 & 1 & 9
\end{array}\right)
$$

According to Gerschgorin's theorem the eigenvalues are contained in the disks

$$
\begin{gathered}
D_{1}=\{\lambda \in \mathbb{C}:|\lambda-2| \leq 2\}, D_{2}=\{\lambda \in \mathbb{C}:|\lambda-5| \leq 3\} \\
D_{3}=\{\lambda \in \mathbb{C}:|\lambda-9| \leq 1\}
\end{gathered}
$$

It is important to observe that these disks are in the complex plane. In general this is the case. If you want to find eigenvalues they will be complex numbers.


So what are the values of the eigenvalues? In this case they are real. You can compute them by graphing the characteristic polynomial, $\lambda^{3}-16 \lambda^{2}+70 \lambda-66$ and then zooming in on the zeros. If you do this you find the solution is $\{\lambda=1.2953\},\{\lambda=5.5905\}$, $\{\lambda=9.1142\}$. Of course these are only approximations and so this information is useless for finding eigenvectors. However, in many applications, it is the size of the eigenvalues which is important and so these numerical values would be helpful for such applications. In this case, you might think there is no real reason for Gerschgorin's theorem. Why not just compute the characteristic equation and graph and zoom? This is fine up to a point, but what if the matrix was huge? Then it might be hard to find the characteristic polynomial. Remember the difficulties in expanding a big matrix along a row or column. Also, what if the eigenvalues were complex? You don't see these by following this procedure. However, Gerschgorin's theorem will at least estimate them.

### 13.4 Advanced Theorems

More can be said but this requires some theory from complex variables ${ }^{1}$. The following is a fundamental theorem about counting zeros.

Theorem 13.4.1 Let $U$ be a region and let $\gamma:[a, b] \rightarrow U$ be closed, continuous, bounded variation, and the winding number, $n(\gamma, z)=0$ for all $z \notin U$. Suppose also that $f$ is analytic on $U$ having zeros $a_{1}, \cdots, a_{m}$ where the zeros are repeated according to multiplicity, and suppose that none of these zeros are on $\gamma([a, b])$. Then

$$
\frac{1}{2 \pi i} \int_{\gamma} \frac{f^{\prime}(z)}{f(z)} d z=\sum_{k=1}^{m} n\left(\gamma, a_{k}\right)
$$

Proof: It is given that $f(z)=\prod_{j=1}^{m}\left(z-a_{j}\right) g(z)$ where $g(z) \neq 0$ on $U$. Hence using the product rule,

$$
\frac{f^{\prime}(z)}{f(z)}=\sum_{j=1}^{m} \frac{1}{z-a_{j}}+\frac{g^{\prime}(z)}{g(z)}
$$

where $\frac{g^{\prime}(z)}{g(z)}$ is analytic on $U$ and so

$$
\frac{1}{2 \pi i} \int_{\gamma} \frac{f^{\prime}(z)}{f(z)} d z=\sum_{j=1}^{m} n\left(\gamma, a_{j}\right)+\frac{1}{2 \pi i} \int_{\gamma} \frac{g^{\prime}(z)}{g(z)} d z=\sum_{j=1}^{m} n\left(\gamma, a_{j}\right)
$$

Now let $A$ be an $n \times n$ matrix. Recall that the eigenvalues of $A$ are given by the zeros of the polynomial, $p_{A}(z)=\operatorname{det}(z I-A)$ where $I$ is the $n \times n$ identity. You can argue that small changes in $A$ will produce small changes in $p_{A}(z)$ and $p_{A}^{\prime}(z)$. Let $\gamma_{k}$ denote a very small closed circle which winds around $z_{k}$, one of the eigenvalues of $A$, in the counter clockwise direction so that $n\left(\gamma_{k}, z_{k}\right)=1$. This circle is to enclose only $z_{k}$ and is to have no other eigenvalue on it. Then apply Theorem 13.4.1. According to this theorem

$$
\frac{1}{2 \pi i} \int_{\gamma} \frac{p_{A}^{\prime}(z)}{p_{A}(z)} d z
$$

is always an integer equal to the multiplicity of $z_{k}$ as a root of $p_{A}(t)$. Therefore, small changes in $A$ result in no change to the above contour integral because it must be an integer and small changes in $A$ result in small changes in the integral. Therefore whenever $B$ is close enough to $A$, the two matrices have the same number of zeros inside $\gamma_{k}$, the zeros being counted according to multiplicity. By making the radius of the small circle equal to $\varepsilon$ where $\varepsilon$ is less than the minimum distance between any two distinct eigenvalues of $A$, this shows that if $B$ is close enough to $A$, every eigenvalue of $B$ is closer than $\varepsilon$ to some eigenvalue of $A$.

Theorem 13.4.2 If $\lambda$ is an eigenvalue of $A$, then if all the entries of $B$ are close enough to the corresponding entries of $A$, some eigenvalue of $B$ will be within $\varepsilon$ of $\lambda$.

Consider the situation that $A(t)$ is an $n \times n$ matrix and that $t \rightarrow A(t)$ is continuous for $t \in[0,1]$.

[^10]Lemma 13.4.3 Let $\lambda(t) \in \sigma(A(t))$ for $t<1$ and let $\Sigma_{t}=\cup_{s \geq t} \sigma(A(s))$. Also let $K_{t}$ be the connected component of $\lambda(t)$ in $\Sigma_{t}$. Then there exists $\eta>0$ such that $K_{t} \cap \sigma(A(s)) \neq \emptyset$ for all $s \in[t, t+\eta]$.

Proof: Denote by $D(\lambda(t), \delta)$ the disc centered at $\lambda(t)$ having radius $\delta>0$, with other occurrences of this notation being defined similarly. Thus

$$
D(\lambda(t), \delta) \equiv\{z \in \mathbb{C}:|\lambda(t)-z| \leq \delta\}
$$

Suppose $\delta>0$ is small enough that $\lambda(t)$ is the only element of $\sigma(A(t))$ contained in $D(\lambda(t), \delta)$ and that $p_{A(t)}$ has no zeroes on the boundary of this disc. Then by continuity, and the above discussion and theorem, there exists $\eta>0, t+\eta<1$, such that for $s \in[t, t+\eta], p_{A(s)}$ also has no zeroes on the boundary of this disc and $A(s)$ has the same number of eigenvalues, counted according to multiplicity, in the disc as $A(t)$. Thus $\sigma(A(s)) \cap D(\lambda(t), \delta) \neq \emptyset$ for all $s \in[t, t+\eta]$. Now let

$$
H=\bigcup_{s \in[t, t+\eta]} \sigma(A(s)) \cap D(\lambda(t), \delta)
$$

It will be shown that $H$ is connected. Suppose not. Then $H=P \cup Q$ where $P, Q$ are separated and $\lambda(t) \in P$. Let $s_{0} \equiv \inf \{s: \lambda(s) \in Q$ for some $\lambda(s) \in \sigma(A(s))\}$. There exists $\lambda\left(s_{0}\right) \in$ $\sigma\left(A\left(s_{0}\right)\right) \cap D(\lambda(t), \delta)$. If $\lambda\left(s_{0}\right) \notin Q$, then from the above discussion there are $\lambda(s) \in$ $\sigma(A(s)) \cap Q$ for $s>s_{0}$ arbitrarily close to $\lambda\left(s_{0}\right)$. Therefore, $\lambda\left(s_{0}\right) \in Q$ which shows that $s_{0}>t$ because $\lambda(t)$ is the only element of $\sigma(A(t))$ in $D(\lambda(t), \delta)$ and $\lambda(t) \in P$. Now let $s_{n} \uparrow s_{0}$. Then $\lambda\left(s_{n}\right) \in P$ for any $\lambda\left(s_{n}\right) \in \sigma\left(A\left(s_{n}\right)\right) \cap D(\lambda(t), \delta)$ and also it follows from the above discussion that for some choice of $s_{n} \rightarrow s_{0}, \lambda\left(s_{n}\right) \rightarrow \lambda\left(s_{0}\right)$ which contradicts $P$ and $Q$ separated and nonempty. Since $P$ is nonempty, this shows $Q=\emptyset$. Therefore, $H$ is connected as claimed. But $K_{t} \supseteq H$ and so $K_{t} \cap \sigma(A(s)) \neq \emptyset$ for all $s \in[t, t+\eta]$.

Theorem 13.4.4 Suppose $A(t)$ is an $n \times n$ matrix and that $t \rightarrow A(t)$ is continuous for $t \in[0,1]$. Let $\lambda(0) \in \sigma(A(0))$ and define $\Sigma \equiv \cup_{t \in[0,1]} \sigma(A(t))$. Let $K_{\lambda(0)}=K_{0}$ denote the connected component of $\lambda(0)$ in $\Sigma$. Then $K_{0} \cap \sigma(A(t)) \neq \emptyset$ for all $t \in[0,1]$.

## Proof: Let

$$
S \equiv\left\{t \in[0,1]: K_{0} \cap \sigma(A(s)) \neq \emptyset \text { for all } s \in[0, t]\right\}
$$

Then $0 \in S$. Let $t_{0}=\sup (S)$. Say $\sigma\left(A\left(t_{0}\right)\right)=\lambda_{1}\left(t_{0}\right), \cdots, \lambda_{r}\left(t_{0}\right)$.
Claim: At least one of these is a limit point of $K_{0}$ and consequently must be in $K_{0}$ which shows that $S$ has a last point. Why is this claim true? Let $s_{n} \uparrow t_{0}$ so $s_{n} \in S$. Now let the discs, $D\left(\lambda_{i}\left(t_{0}\right), \boldsymbol{\delta}\right), i=1, \cdots, r$ be disjoint with $p_{A\left(t_{0}\right)}$ having no zeroes on $\gamma_{i}$ the boundary of $D\left(\lambda_{i}\left(t_{0}\right), \delta\right)$. Then for $n$ large enough it follows from Theorem 13.4.1 and the discussion following it that $\sigma\left(A\left(s_{n}\right)\right)$ is contained in $\cup_{i=1}^{r} D\left(\lambda_{i}\left(t_{0}\right), \delta\right)$. It follows that $K_{0} \cap\left(\sigma\left(A\left(t_{0}\right)\right)+D(0, \delta)\right) \neq \emptyset$ for all $\delta$ small enough. This requires at least one of the $\lambda_{i}\left(t_{0}\right)$ to be in $\overline{K_{0}}$. Therefore, $t_{0} \in S$ and $S$ has a last point.

Now by Lemma 13.4.3, if $t_{0}<1$, then $K_{0} \cup K_{t}$ would be a strictly larger connected set containing $\lambda(0)$. (The reason this would be strictly larger is that $K_{0} \cap \sigma(A(s))=\emptyset$ for some $s \in(t, t+\eta)$ while $K_{t} \cap \sigma(A(s)) \neq \emptyset$ for all $s \in[t, t+\eta]$.) Therefore, $t_{0}=1$.

Corollary 13.4.5 Suppose one of the Gerschgorin discs, $D_{i}$ is disjoint from the union of the others. Then $D_{i}$ contains an eigenvalue of A. Also, if there are $n$ disjoint Gerschgorin discs, then each one contains an eigenvalue of $A$.

Proof: Denote by $A(t)$ the matrix $\left(a_{i j}^{t}\right)$ where if $i \neq j, a_{i j}^{t}=t a_{i j}$ and $a_{i i}^{t}=a_{i i}$. Thus to get $A(t)$ multiply all non diagonal terms by $t$. Let $t \in[0,1]$. Then $A(0)=\operatorname{diag}\left(a_{11}, \cdots, a_{n n}\right)$ and $A(1)=A$. Furthermore, the map, $t \rightarrow A(t)$ is continuous. Denote by $D_{j}^{t}$ the Gerschgorin disc obtained from the $j^{t h}$ row for the matrix $A(t)$. Then it is clear that $D_{j}^{t} \subseteq D_{j}$ the $j^{t h}$ Gerschgorin disc for $A$. It follows $a_{i i}$ is the eigenvalue for $A(0)$ which is contained in the disc, consisting of the single point $a_{i i}$ which is contained in $D_{i}$. Letting $K$ be the connected component in $\Sigma$ for $\Sigma$ defined in Theorem 13.4.4 which is determined by $a_{i i}$, Gerschgorin's theorem implies that $K \cap \sigma(A(t)) \subseteq \cup_{j=1}^{n} D_{j}^{t} \subseteq \cup_{j=1}^{n} D_{j}=D_{i} \cup\left(\cup_{j \neq i} D_{j}\right)$ and also, since $K$ is connected, there are not points of $K$ in both $D_{i}$ and $\left(\cup_{j \neq i} D_{j}\right)$. Since at least one point of $K$ is in $D_{i},\left(a_{i i}\right)$, it follows all of $K$ must be contained in $D_{i}$. Now by Theorem 13.4.4 this shows there are points of $K \cap \sigma(A)$ in $D_{i}$. The last assertion follows immediately.

This can be improved even more. This involves the following lemma.
Lemma 13.4.6 In the situation of Theorem 13.4.4 suppose $\lambda(0)=K_{0} \cap \sigma(A(0))$ and that $\lambda(0)$ is a simple root of the characteristic equation of $A(0)$. Then for all $t \in[0,1]$,

$$
\sigma(A(t)) \cap K_{0}=\lambda(t)
$$

where $\lambda(t)$ is a simple root of the characteristic equation of $A(t)$.
Proof: Let

$$
S \equiv\left\{t \in[0,1]: K_{0} \cap \sigma(A(s))=\lambda(s), \text { a simple eigenvalue for all } s \in[0, t]\right\}
$$

Then $0 \in S$ so it is nonempty. Let $t_{0}=\sup (S)$ and suppose $\lambda_{1} \neq \lambda_{2}$ are two elements of $\sigma\left(A\left(t_{0}\right)\right) \cap K_{0}$. Then choosing $\eta>0$ small enough, and letting $D_{i}$ be disjoint discs containing $\lambda_{i}$ respectively, similar arguments to those of Lemma 13.4.3 can be used to conclude

$$
H_{i} \equiv \cup_{s \in\left[t_{0}-\eta, t_{0}\right]} \sigma(A(s)) \cap D_{i}
$$

is a connected and nonempty set for $i=1,2$ which would require that $H_{i} \subseteq K_{0}$. But then there would be two different eigenvalues of $A(s)$ contained in $K_{0}$, contrary to the definition of $t_{0}$. Therefore, there is at most one eigenvalue $\lambda\left(t_{0}\right) \in K_{0} \cap \sigma\left(A\left(t_{0}\right)\right)$. Could it be a repeated root of the characteristic equation? Suppose $\lambda\left(t_{0}\right)$ is a repeated root of the characteristic equation. As before, choose a small disc, $D$ centered at $\lambda\left(t_{0}\right)$ and $\eta$ small enough that

$$
H \equiv \cup_{s \in\left[t_{0}-\eta, t_{0}\right]} \sigma(A(s)) \cap D
$$

is a nonempty connected set containing either multiple eigenvalues of $A(s)$ or else a single repeated root to the characteristic equation of $A(s)$. But since $H$ is connected and contains $\lambda\left(t_{0}\right)$ it must be contained in $K_{0}$ which contradicts the condition for $s \in S$ for all these $s \in\left[t_{0}-\eta, t_{0}\right]$. Therefore, $t_{0} \in S$ as hoped. If $t_{0}<1$, there exists a small disc centered at $\lambda\left(t_{0}\right)$ and $\eta>0$ such that for all $s \in\left[t_{0}, t_{0}+\eta\right], A(s)$ has only simple eigenvalues in $D$ and the only eigenvalues of $A(s)$ which could be in $K_{0}$ are in $D$. (This last assertion follows from noting that $\lambda\left(t_{0}\right)$ is the only eigenvalue of $A\left(t_{0}\right)$ in $K_{0}$ and so the others are at a positive distance from $K_{0}$. For $s$ close enough to $t_{0}$, the eigenvalues of $A(s)$ are either close to these eigenvalues of $A\left(t_{0}\right)$ at a positive distance from $K_{0}$ or they are close to the eigenvalue $\lambda\left(t_{0}\right)$ in which case it can be assumed they are in $D$.) But this shows that $t_{0}$ is not really an upper bound to $S$. Therefore, $t_{0}=1$ and the lemma is proved.

With this lemma, the conclusion of the above corollary can be sharpened.

Corollary 13.4.7 Suppose one of the Gerschgorin discs, $D_{i}$ is disjoint from the union of the others. Then $D_{i}$ contains exactly one eigenvalue of $A$ and this eigenvalue is a simple root to the characteristic polynomial of $A$.

Proof: In the proof of Corollary 13.4.5, note that $a_{i i}$ is a simple root of $A(0)$ since otherwise the $i^{\text {th }}$ Gerschgorin disc would not be disjoint from the others. Also, $K$, the connected component determined by $a_{i i}$ must be contained in $D_{i}$ because it is connected and by Gerschgorin's theorem above, $K \cap \sigma(A(t))$ must be contained in the union of the Gerschgorin discs. Since all the other eigenvalues of $A(0)$, the $a_{j j}$, are outside $D_{i}$, it follows that $K \cap \sigma(A(0))=a_{i i}$. Therefore, by Lemma 13.4.6, $K \cap \sigma(A(1))=K \cap \sigma(A)$ consists of a single simple eigenvalue.

## Example 13.4.8 Consider the matrix

$$
\left(\begin{array}{lll}
5 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 0
\end{array}\right)
$$

The Gerschgorin discs are $D(5,1), D(1,2)$, and $D(0,1)$. Observe $D(5,1)$ is disjoint from the other discs. Therefore, there should be an eigenvalue in $D(5,1)$. The actual eigenvalues are not easy to find. They are the roots of the characteristic equation, $t^{3}-6 t^{2}+$ $3 t+5=0$. The numerical values of these are $-.66966,1.4231$, and 5.24655 , verifying the predictions of Gerschgorin's theorem.

### 13.5 Exercises

1. Explain why it is typically impossible to compute the upper triangular matrix whose existence is guaranteed by Schur's theorem.
2. Now recall the $Q R$ factorization of Theorem 12.3.9 on Page 316. The $Q R$ algorithm is a technique which does compute the upper triangular matrix in Schur's theorem sometimes. There is much more to the $Q R$ algorithm than will be presented here. In fact, what I am about to show you is not the way it is done in practice. One first obtains what is called a Hessenburg matrix for which the algorithm will work better. However, the idea is as follows. Start with $A$ an $n \times n$ matrix having real eigenvalues. Form $A=Q R$ where $Q$ is orthogonal and $R$ is upper triangular. (Right triangular.) This can be done using the technique of Theorem 12.3.9 using Householder matrices. Next take $A_{1} \equiv R Q$. Show that $A=Q A_{1} Q^{T}$. In other words these two matrices, $A, A_{1}$ are similar. Explain why they have the same eigenvalues. Continue by letting $A_{1}$ play the role of $A$. Thus the algorithm is of the form $A_{n}=Q R_{n}$ and $A_{n+1}=R_{n+1} Q$. Explain why $A=Q_{n} A_{n} Q_{n}^{T}$ for some $Q_{n}$ orthogonal. Thus $A_{n}$ is a sequence of matrices each similar to $A$. The remarkable thing is that often these matrices converge to an upper triangular matrix $T$ and $A=Q T Q^{T}$ for some orthogonal matrix, the limit of the $Q_{n}$ where the limit means the entries converge. Then the process computes the upper triangular Schur form of the matrix $A$. Thus the eigenvalues of $A$ appear on the diagonal of $T$. You will see approximately what these are as the process continues.
3. $\uparrow$ Try the $Q R$ algorithm on

$$
\left(\begin{array}{cc}
-1 & -2 \\
6 & 6
\end{array}\right)
$$

which has eigenvalues 3 and 2. I suggest you use a computer algebra system to do the computations.
4. $\uparrow$ Now try the $Q R$ algorithm on

$$
\left(\begin{array}{cc}
0 & -1 \\
2 & 0
\end{array}\right)
$$

Show that the algorithm cannot converge for this example. Hint: Try a few iterations of the algorithm. Use a computer algebra system if you like.
5. $\uparrow$ Show the two matrices $A \equiv\left(\begin{array}{cc}0 & -1 \\ 4 & 0\end{array}\right)$ and $B \equiv\left(\begin{array}{cc}0 & -2 \\ 2 & 0\end{array}\right)$ are similar; that is there exists a matrix $S$ such that $A=S^{-1} B S$ but there is no orthogonal matrix $Q$ such that $Q^{T} B Q=A$. Show the $Q R$ algorithm does converge for the matrix $B$ although it fails to do so for $A$.
6. Let $F$ be an $m \times n$ matrix. Show that $F^{*} F$ has all real eigenvalues and furthermore, they are all nonnegative.
7. If $A$ is a real $n \times n$ matrix and $\lambda$ is a complex eigenvalue $\lambda=a+i b, b \neq 0$, of $A$ having eigenvector $\boldsymbol{z}+i \boldsymbol{w}$, show that $\boldsymbol{w} \neq \mathbf{0}$.
8. Suppose $A=Q^{T} D Q$ where $Q$ is an orthogonal matrix and all the matrices are real. Also $D$ is a diagonal matrix. Show that $A$ must be symmetric.
9. Suppose $A$ is an $n \times n$ matrix and there exists a unitary matrix $U$ such that

$$
A=U^{*} D U
$$

where $D$ is a diagonal matrix. Explain why $A$ must be normal.
10. If $A$ is Hermitian, show that $\operatorname{det}(A)$ must be real.
11. Show that every unitary matrix preserves distance. That is, if $U$ is unitary,

$$
|U \boldsymbol{x}|=|\boldsymbol{x}| .
$$

12. Show that if a matrix does preserve distances, then it must be unitary.
13. $\uparrow$ Show that a complex normal matrix $A$ is unitary if and only if its eigenvalues have magnitude equal to 1 .
14. Suppose $A$ is an $n \times n$ matrix which is diagonally dominant. Recall this means

$$
\sum_{j \neq i}\left|a_{i j}\right|<\left|a_{i i}\right|
$$

show $A^{-1}$ must exist.
15. Give some disks in the complex plane whose union contains all the eigenvalues of the matrix

$$
\left(\begin{array}{ccc}
1+2 i & 4 & 2 \\
0 & i & 3 \\
5 & 6 & 7
\end{array}\right)
$$

16. Show a square matrix is invertible if and only if it has no zero eigenvalues.
17. Using Schur's theorem, show the trace of an $n \times n$ matrix equals the sum of the eigenvalues and the determinant of an $n \times n$ matrix is the product of the eigenvalues.
18. Using Schur's theorem, show that if $A$ is any complex $n \times n$ matrix having eigenvalues $\left\{\lambda_{i}\right\}$ listed according to multiplicity, then $\sum_{i, j}\left|A_{i j}\right|^{2} \geq \sum_{i=1}^{n}\left|\lambda_{i}\right|^{2}$. Show that equality holds if and only if $A$ is normal. This inequality is called Schur's inequality. [33]
19. Here is a matrix.

$$
\left(\begin{array}{cccc}
1234 & 6 & 5 & 3 \\
0 & -654 & 9 & 123 \\
98 & 123 & 10,000 & 11 \\
56 & 78 & 98 & 400
\end{array}\right)
$$

I know this matrix has an inverse before doing any computations. How do I know?
20. Show the critical points of the following function are

$$
(0,-3,0),(2,-3,0), \text { and }\left(1,-3,-\frac{1}{3}\right)
$$

and classify them as local minima, local maxima or saddle points.
$f(x, y, z)=-\frac{3}{2} x^{4}+6 x^{3}-6 x^{2}+z x^{2}-2 z x-2 y^{2}-12 y-18-\frac{3}{2} z^{2}$.
21. Here is a function of three variables.

$$
f(x, y, z)=13 x^{2}+2 x y+8 x z+13 y^{2}+8 y z+10 z^{2}
$$

change the variables so that in the new variables there are no mixed terms, terms involving $x y, y z$ etc. Two eigenvalues are 12 and 18.
22. Here is a function of three variables.

$$
f(x, y, z)=2 x^{2}-4 x+2+9 y x-9 y-3 z x+3 z+5 y^{2}-9 z y-7 z^{2}
$$

change the variables so that in the new variables there are no mixed terms, terms involving $x y, y z$ etc. The eigenvalues of the matrix which you will work with are $-\frac{17}{2}, \frac{19}{2},-1$.
23. Here is a function of three variables.

$$
f(x, y, z)=-x^{2}+2 x y+2 x z-y^{2}+2 y z-z^{2}+x
$$

change the variables so that in the new variables there are no mixed terms, terms involving $x y, y z$ etc.
24. Show the critical points of the function,

$$
f(x, y, z)=-2 y x^{2}-6 y x-4 z x^{2}-12 z x+y^{2}+2 y z
$$

are points of the form,

$$
(x, y, z)=\left(t, 2 t^{2}+6 t,-t^{2}-3 t\right)
$$

for $t \in \mathbb{R}$ and classify them as local minima, local maxima or saddle points.
25. Show the critical points of the function

$$
f(x, y, z)=\frac{1}{2} x^{4}-4 x^{3}+8 x^{2}-3 z x^{2}+12 z x+2 y^{2}+4 y+2+\frac{1}{2} z^{2} .
$$

are $(0,-1,0),(4,-1,0)$, and $(2,-1,-12)$ and classify them as local minima, local maxima or saddle points.
26. Let $f(x, y)=3 x^{4}-24 x^{2}+48-y x^{2}+4 y$. Find and classify the critical points using the second derivative test.
27. Let $f(x, y)=3 x^{4}-5 x^{2}+2-y^{2} x^{2}+y^{2}$. Find and classify the critical points using the second derivative test.
28. Let $f(x, y)=5 x^{4}-7 x^{2}-2-3 y^{2} x^{2}+11 y^{2}-4 y^{4}$. Find and classify the critical points using the second derivative test.
29. Let $f(x, y, z)=-2 x^{4}-3 y x^{2}+3 x^{2}+5 x^{2} z+3 y^{2}-6 y+3-3 z y+3 z+z^{2}$. Find and classify the critical points using the second derivative test.
30. Let $f(x, y, z)=3 y x^{2}-3 x^{2}-x^{2} z-y^{2}+2 y-1+3 z y-3 z-3 z^{2}$. Find and classify the critical points using the second derivative test.
31. Let $Q$ be orthogonal. Find the possible values of $\operatorname{det}(Q)$.
32. Let $U$ be unitary. Find the possible values of $\operatorname{det}(U)$.
33. If a matrix is nonzero can it have only zero for eigenvalues?
34. A matrix $A$ is called nilpotent if $A^{k}=0$ for some positive integer $k$. Suppose $A$ is a nilpotent matrix. Show it has only 0 for an eigenvalue.
35. If $A$ is a nonzero nilpotent matrix, show it must be defective.
36. Suppose $A$ is a nondefective $n \times n$ matrix and its eigenvalues are all either 0 or 1 . Show $A^{2}=A$. Could you say anything interesting if the eigenvalues were all either 0,1, or -1 ? By DeMoivre's theorem, an $n^{\text {th }}$ root of unity is of the form

$$
\left(\cos \left(\frac{2 k \pi}{n}\right)+i \sin \left(\frac{2 k \pi}{n}\right)\right)
$$

Could you generalize the sort of thing just described to get $A^{n}=A$ ? Hint: Since $A$ is nondefective, there exists $S$ such that $S^{-1} A S=D$ where $D$ is a diagonal matrix.
37. This and the following problems will present most of a differential equations course. Most of the explanations are given. You fill in any details needed. To begin with, consider the scalar initial value problem

$$
y^{\prime}=a y, y\left(t_{0}\right)=y_{0}
$$

When $a$ is real, show the unique solution to this problem is $y=y_{0} e^{a\left(t-t_{0}\right)}$. Next suppose

$$
\begin{equation*}
y^{\prime}=(a+i b) y, y\left(t_{0}\right)=y_{0} \tag{13.6}
\end{equation*}
$$

where $y(t)=u(t)+i v(t)$. Show there exists a unique solution and it is given by $y(t)=$

$$
\begin{equation*}
y_{0} e^{a\left(t-t_{0}\right)}\left(\cos b\left(t-t_{0}\right)+i \sin b\left(t-t_{0}\right)\right) \equiv e^{(a+i b)\left(t-t_{0}\right)} y_{0} \tag{13.7}
\end{equation*}
$$

Next show that for $a$ real or complex there exists a unique solution to the initial value problem

$$
y^{\prime}=a y+f, y\left(t_{0}\right)=y_{0}
$$

and it is given by

$$
y(t)=e^{a\left(t-t_{0}\right)} y_{0}+e^{a t} \int_{t_{0}}^{t} e^{-a s} f(s) d s
$$

Hint: For the first part write as $y^{\prime}-a y=0$ and multiply both sides by $e^{-a t}$. Then explain why you get

$$
\frac{d}{d t}\left(e^{-a t} y(t)\right)=0, y\left(t_{0}\right)=0
$$

Now you finish the argument. To show uniqueness in the second part, suppose

$$
y^{\prime}=(a+i b) y, y\left(t_{0}\right)=0
$$

and verify this requires $y(t)=0$. To do this, note

$$
\bar{y}^{\prime}=(a-i b) \bar{y}, \bar{y}\left(t_{0}\right)=0
$$

and that $|y|^{2}\left(t_{0}\right)=0$ and

$$
\begin{gathered}
\frac{d}{d t}|y(t)|^{2}=y^{\prime}(t) \bar{y}(t)+\bar{y}^{\prime}(t) y(t) \\
=(a+i b) y(t) \bar{y}(t)+(a-i b) \bar{y}(t) y(t)=2 a|y(t)|^{2} .
\end{gathered}
$$

Thus from the first part $|y(t)|^{2}=0 e^{-2 a t}=0$. Finally observe by a simple computation that 13.6 is solved by 13.7. For the last part, write the equation as

$$
y^{\prime}-a y=f
$$

and multiply both sides by $e^{-a t}$ and then integrate from $t_{0}$ to $t$ using the initial condition.
38. Now consider $A$ an $n \times n$ matrix. By Schur's theorem there exists unitary $Q$ such that

$$
Q^{-1} A Q=T
$$

where $T$ is upper triangular. Now consider the first order initial value problem

$$
\boldsymbol{x}^{\prime}=A \boldsymbol{x}, \boldsymbol{x}\left(t_{0}\right)=\boldsymbol{x}_{0} .
$$

Show there exists a unique solution to this first order system. Hint: Let $\boldsymbol{y}=Q^{-1} \boldsymbol{x}$ and so the system becomes

$$
\begin{equation*}
\boldsymbol{y}^{\prime}=T \boldsymbol{y}, \boldsymbol{y}\left(t_{0}\right)=Q^{-1} \boldsymbol{x}_{0} \tag{13.8}
\end{equation*}
$$

Now letting $\boldsymbol{y}=\left(y_{1}, \cdots, y_{n}\right)^{T}$, the bottom equation becomes

$$
y_{n}^{\prime}=t_{n n} y_{n}, y_{n}\left(t_{0}\right)=\left(Q^{-1} x_{0}\right)_{n}
$$

Then use the solution you get in this to get the solution to the initial value problem which occurs one level up, namely

$$
y_{n-1}^{\prime}=t_{(n-1)(n-1)} y_{n-1}+t_{(n-1) n} y_{n}, y_{n-1}\left(t_{0}\right)=\left(Q^{-1} x_{0}\right)_{n-1}
$$

Continue doing this to obtain a unique solution to 13.8 .
39. Now suppose $\Phi(t)$ is an $n \times n$ matrix of the form

$$
\Phi(t)=\left(\begin{array}{lll}
x_{1}(t) & \cdots & x_{n}(t) \tag{13.9}
\end{array}\right)
$$

where

$$
\boldsymbol{x}_{k}^{\prime}(t)=A \boldsymbol{x}_{k}(t)
$$

Explain why

$$
\Phi^{\prime}(t)=A \Phi(t)
$$

if and only if $\Phi(t)$ is given in the form of 13.9. Also explain why if $\boldsymbol{c} \in \mathbb{F}^{n}, \boldsymbol{y}(t) \equiv$ $\Phi(t) \boldsymbol{c}$ solves the equation $\boldsymbol{y}^{\prime}(t)=A \boldsymbol{y}(t)$.
40. In the above problem, consider the question whether all solutions to

$$
\begin{equation*}
\boldsymbol{x}^{\prime}=A \boldsymbol{x} \tag{13.10}
\end{equation*}
$$

are obtained in the form $\Phi(t) c$ for some choice of $c \in \mathbb{F}^{n}$. In other words, is the general solution to this equation $\Phi(t) c$ for $c \in \mathbb{F}^{n}$ ? Prove the following theorem using linear algebra.

Theorem 13.5.1 Suppose $\Phi(t)$ is an $n \times n$ matrix which satisfies $\Phi^{\prime}(t)=A \Phi(t)$. Then the general solution to 13.10 is $\Phi(t) c$ if and only if $\Phi(t)^{-1}$ exists for some $t$. Furthermore, if $\Phi^{\prime}(t)=A \Phi(t)$, then either $\Phi(t)^{-1}$ exists for all $t$ or $\Phi(t)^{-1}$ never exists for any $t$.
( $\operatorname{det}(\Phi(t))$ is called the Wronskian and this theorem is sometimes called the Wronskian alternative.)
Hint: Suppose first the general solution is of the form $\Phi(t) c$ where $c$ is an arbitrary constant vector in $\mathbb{F}^{n}$. You need to verify $\Phi(t)^{-1}$ exists for some $t$. In fact, show $\Phi(t)^{-1}$ exists for every $t$. Suppose then that $\Phi\left(t_{0}\right)^{-1}$ does not exist. Explain why
there exists $\boldsymbol{c} \in \mathbb{F}^{n}$ such that there is no solution $\boldsymbol{x}$ to the equation $\boldsymbol{c}=\Phi\left(t_{0}\right) \boldsymbol{x}$. By the existence part of Problem 38 there exists a solution to

$$
\boldsymbol{x}^{\prime}=A \boldsymbol{x}, \boldsymbol{x}\left(t_{0}\right)=\boldsymbol{c}
$$

but this cannot be in the form $\Phi(t) c$. Thus for every $t, \Phi(t)^{-1}$ exists. Next suppose for some $t_{0}, \Phi\left(t_{0}\right)^{-1}$ exists. Let $\boldsymbol{z}^{\prime}=A \boldsymbol{z}$ and choose $\boldsymbol{c}$ such that

$$
z\left(t_{0}\right)=\Phi\left(t_{0}\right) \boldsymbol{c}
$$

Then both $\boldsymbol{z}(t), \Phi(t) \boldsymbol{c}$ solve

$$
\boldsymbol{x}^{\prime}=A \boldsymbol{x}, \boldsymbol{x}\left(t_{0}\right)=\boldsymbol{z}\left(t_{0}\right)
$$

Apply uniqueness to conclude $\boldsymbol{z}=\Phi(t) \boldsymbol{c}$. Finally, consider that $\Phi(t) \boldsymbol{c}$ for $\boldsymbol{c} \in \mathbb{F}^{n}$ either is the general solution or it is not the general solution. If it is, then $\Phi(t)^{-1}$ exists for all $t$. If it is not, then $\Phi(t)^{-1}$ cannot exist for any $t$ from what was just shown.
41. Let $\Phi^{\prime}(t)=A \Phi(t)$. Then $\Phi(t)$ is called a fundamental matrix if $\Phi(t)^{-1}$ exists for all $t$. Show there exists a unique solution to the equation

$$
\begin{equation*}
\boldsymbol{x}^{\prime}=A \boldsymbol{x}+\boldsymbol{f}, \boldsymbol{x}\left(t_{0}\right)=\boldsymbol{x}_{0} \tag{13.11}
\end{equation*}
$$

and it is given by the formula

$$
\boldsymbol{x}(t)=\Phi(t) \Phi\left(t_{0}\right)^{-1} \boldsymbol{x}_{0}+\Phi(t) \int_{t_{0}}^{t} \Phi(s)^{-1} \boldsymbol{f}(s) d s
$$

Now these few problems have done virtually everything of significance in an entire undergraduate differential equations course, illustrating the superiority of linear algebra. The above formula is called the variation of constants formula.
Hint: Uniquenss is easy. If $\boldsymbol{x}_{1}, \boldsymbol{x}_{2}$ are two solutions then let $\boldsymbol{u}(t)=\boldsymbol{x}_{1}(t)-\boldsymbol{x}_{2}(t)$ and argue $\boldsymbol{u}^{\prime}=A \boldsymbol{u}, \boldsymbol{u}\left(t_{0}\right)=\mathbf{0}$. Then use Problem 38. To verify there exists a solution, you could just differentiate the above formula using the fundamental theorem of calculus and verify it works. Another way is to assume the solution in the form

$$
\boldsymbol{x}(t)=\Phi(t) \boldsymbol{c}(t)
$$

and find $\boldsymbol{c}(t)$ to make it all work out. This is called the method of variation of parameters.
42. Show there exists a special $\Phi$ such that $\Phi^{\prime}(t)=A \Phi(t), \Phi(0)=I$, and suppose $\Phi(t)^{-1}$ exists for all $t$. Show using uniqueness that

$$
\Phi(-t)=\Phi(t)^{-1}
$$

and that for all $t, s \in \mathbb{R}$

$$
\Phi(t+s)=\Phi(t) \Phi(s)
$$

Explain why with this special $\Phi$, the solution to 13.11 can be written as

$$
\boldsymbol{x}(t)=\Phi\left(t-t_{0}\right) x_{0}+\int_{t_{0}}^{t} \Phi(t-s) \boldsymbol{f}(s) d s
$$

Hint: Let $\Phi(t)$ be such that the $j^{t h}$ column is $\boldsymbol{x}_{j}(t)$ where

$$
\boldsymbol{x}_{j}^{\prime}=A \boldsymbol{x}_{j}, \boldsymbol{x}_{j}(0)=\boldsymbol{e}_{j}
$$

Use uniqueness as required.
43. You can see more on this problem and the next one in the latest version of Horn and Johnson, [25]. Two $n \times n$ matrices $A, B$ are said to be congruent if there is an invertible $P$ such that

$$
B=P A P^{*}
$$

Let $A$ be a Hermitian matrix. Thus it has all real eigenvalues. Let $n_{+}$be the number of positive eigenvalues, $n_{-}$, the number of negative eigenvalues and $n_{0}$ the number of zero eigenvalues. For $k$ a positive integer, let $I_{k}$ denote the $k \times k$ identity matrix and $O_{k}$ the $k \times k$ zero matrix. Then the inertia matrix of $A$ is the following block diagonal $n \times n$ matrix.

$$
\left(\begin{array}{ccc}
I_{n_{+}} & & \\
& I_{n_{-}} & \\
& & O_{n_{0}}
\end{array}\right)
$$

Show that $A$ is congruent to its inertia matrix. Next show that congruence is an equivalence relation on the set of Hermitian matrices. Finally, show that if two Hermitian matrices have the same inertia matrix, then they must be congruent. Hint: First recall that there is a unitary matrix, $U$ such that

$$
U^{*} A U=\left(\begin{array}{ccc}
D_{n_{+}} & & \\
& D_{n_{-}} & \\
& & O_{n_{0}}
\end{array}\right)
$$

where the $D_{n_{+}}$is a diagonal matrix having the positive eigenvalues of $A, D_{n_{-}}$being defined similarly. Now let $\left|D_{n_{-}}\right|$denote the diagonal matrix which replaces each entry of $D_{n_{-}}$with its absolute value. Consider the two diagonal matrices

$$
D=D^{*}=\left(\begin{array}{ccc}
D_{n_{+}}^{-1 / 2} & & \\
& \left|D_{n_{-}}\right|^{-1 / 2} & \\
& & I_{n_{0}}
\end{array}\right)
$$

Now consider $D^{*} U^{*} A U D$.
44. Show that if $A, B$ are two congruent Hermitian matrices, then they have the same inertia matrix. Hint: Let $A=S B S^{*}$ where $S$ is invertible. Show that $A, B$ have the same rank and this implies that they are each unitarily similar to a diagonal matrix which has the same number of zero entries on the main diagonal. Therefore, letting $V_{A}$ be the span of the eigenvectors associated with positive eigenvalues of $A$ and $V_{B}$ being defined similarly, it suffices to show that these have the same dimensions. Show that $(A \boldsymbol{x}, \boldsymbol{x})>0$ for all $\boldsymbol{x} \in V_{A}$. Next consider $S^{*} V_{A}$. For $\boldsymbol{x} \in V_{A}$, explain why

$$
\begin{aligned}
\left(B S^{*} \boldsymbol{x}, S^{*} \boldsymbol{x}\right) & =\left(S^{-1} A\left(S^{*}\right)^{-1} S^{*} \boldsymbol{x}, S^{*} \boldsymbol{x}\right) \\
& =\left(S^{-1} A \boldsymbol{x}, S^{*} \boldsymbol{x}\right)=\left(A \boldsymbol{x},\left(S^{-1}\right)^{*} S^{*} \boldsymbol{x}\right)=(A \boldsymbol{x}, \boldsymbol{x})>0
\end{aligned}
$$

Next explain why this shows that $S^{*} V_{A}$ is a subspace of $V_{B}$ and so the dimension of $V_{B}$ is at least as large as the dimension of $V_{A}$. Hence there are at least as many positive eigenvalues for $B$ as there are for $A$. Switching $A, B$ you can turn the inequality around. Thus the two have the same inertia matrix.
45. Let $A$ be an $m \times n$ matrix. Then if you unraveled it, you could consider it as a vector in $\mathbb{C}^{n m}$. The Frobenius inner product on the vector space of $m \times n$ matrices is defined as

$$
(A, B) \equiv \operatorname{trace}\left(A B^{*}\right)
$$

Show that this really does satisfy the axioms of an inner product space and that it also amounts to nothing more than considering $m \times n$ matrices as vectors in $\mathbb{C}^{n m}$.
46. $\uparrow$ Consider the $n \times n$ unitary matrices. Show that whenever $U$ is such a matrix, it follows that

$$
|U|_{\mathbb{C}^{n n}}=\sqrt{n}
$$

Next explain why if $\left\{U_{k}\right\}$ is any sequence of unitary matrices, there exists a subsequence $\left\{U_{k_{m}}\right\}_{m=1}^{\infty}$ such that $\lim _{m \rightarrow \infty} U_{k_{m}}=U$ where $U$ is unitary. Here the limit takes place in the sense that the entries of $U_{k_{m}}$ converge to the corresponding entries of $U$.
47. $\uparrow$ Let $A, B$ be two $n \times n$ matrices. Denote by $\sigma(A)$ the set of eigenvalues of $A$. Define

$$
\operatorname{dist}(\sigma(A), \sigma(B))=\max _{\lambda \in \sigma(A)} \min \{|\lambda-\mu|: \mu \in \sigma(B)\}
$$

Explain why $\operatorname{dist}(\sigma(A), \sigma(B))$ is small if and only if every eigenvalue of $A$ is close to some eigenvalue of $B$. Now prove the following theorem using the above problem and Schur's theorem. This theorem says roughly that if $A$ is close to $B$ then the eigenvalues of $A$ are close to those of $B$ in the sense that every eigenvalue of $A$ is close to an eigenvalue of $B$. This is a very important observation when you try to approximate eigenvalues using the $Q R$ algorithm.

Theorem 13.5.2 Suppose $\lim _{k \rightarrow \infty} A_{k}=A$. Then

$$
\lim _{k \rightarrow \infty} \operatorname{dist}\left(\sigma\left(A_{k}\right), \sigma(A)\right)=0
$$

48. Let $A=\left(\begin{array}{ll}a & b \\ c & d\end{array}\right)$ be a $2 \times 2$ matrix which is not a multiple of the identity. Show that $A$ is similar to a $2 \times 2$ matrix which has at least one diagonal entry equal to 0 . Hint: First note that there exists a vector $\boldsymbol{a}$ such that $A \boldsymbol{a}$ is not a multiple of $\boldsymbol{a}$. Then consider

$$
B=\left(\begin{array}{ll}
\boldsymbol{a} & A \boldsymbol{a}
\end{array}\right)^{-1} A\left(\begin{array}{ll}
\boldsymbol{a} & A \boldsymbol{a}
\end{array}\right)
$$

Show $B$ has a zero on the main diagonal.
49. $\uparrow$ Let $A$ be a complex $n \times n$ matrix which has trace equal to 0 . Show that $A$ is similar to a matrix which has all zeros on the main diagonal. Hint: Use Problem 39 on Page 99 to argue that you can say that a given matrix is similar to one which has the diagonal entries permuted in any order desired. Then use the above problem and block multiplication to show that if the $A$ has $k$ nonzero entries, then it is similar to a
matrix which has $k-1$ nonzero entries. Finally, when $A$ is similar to one which has at most one nonzero entry, this one must also be zero because of the condition on the trace.
50. $\uparrow$ An $n \times n$ matrix $X$ is a commutator if there are $n \times n$ matrices $A, B$ such that $X=$ $A B-B A$. Show that the trace of any commutator is 0 . Next show that if a complex matrix $X$ has trace equal to 0 , then it is in fact a commutator. Hint: Use the above problem to show that it suffices to consider $X$ having all zero entries on the main diagonal. Then define

$$
A=\left(\begin{array}{cccc}
1 & & & 0 \\
& 2 & & \\
& & \ddots & \\
0 & & & n
\end{array}\right), B_{i j}=\left\{\begin{array}{c}
\frac{X_{i j}}{i-j} \text { if } i \neq j \\
0 \text { if } i=j
\end{array}\right.
$$

### 13.6 Cauchy's Interlacing Theorem, Eigenvalues

Recall that every Hermitian matrix has all real eigenvalues. The Cauchy interlacing theorem compares the location of the eigenvalues of a Hermitian matrix with the eigenvalues of a principal submatrix. It is an extremely interesting theorem.

Theorem 13.6.1 Let $A$ be a Hermitian $n \times n$ matrix and let

$$
A=\left(\begin{array}{ll}
a & y^{*} \\
\boldsymbol{y} & B
\end{array}\right)
$$

where $B$ is $(n-1) \times(n-1)$. Let the eigenvalues of $B$ be $\mu_{1} \leq \mu_{2} \leq \cdots \leq \mu_{n-1}$. Then if the eigenvalues of $A$ are $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$, it follows that $\lambda_{1} \leq \mu_{1} \leq \lambda_{2} \leq \mu_{2} \leq \cdots \leq$ $\mu_{n-1} \leq \lambda_{n}$.

Proof: First note that $B$ is Hermitian because

$$
A^{*}=\left(\begin{array}{ll}
\bar{a} & \boldsymbol{y}^{*} \\
\boldsymbol{y} & B^{*}
\end{array}\right)=A=\left(\begin{array}{cc}
a & \boldsymbol{y}^{*} \\
\boldsymbol{y} & B
\end{array}\right)
$$

It is easiest to consider the case where strict inequality holds for the eigenvalues for $B$ so first is an outline of reducing to this case.

There exists $U$ unitary, depending on $B$ such that $U^{*} B U=D$ where

$$
D=\left(\begin{array}{ccc}
\mu_{1} & & 0 \\
& \ddots & \\
0 & & \mu_{n-1}
\end{array}\right)
$$

Now let $\left\{\varepsilon_{k}\right\}$ be a decreasing sequence of very small positive numbers converging to 0 and let $B_{k}$ be defined by

$$
U^{*} B_{k} U=D_{k}, \quad D_{k} \equiv\left(\begin{array}{cccc}
\mu_{1}+\varepsilon_{k} & & & 0 \\
& \mu_{2}+2 \varepsilon_{k} & & \\
& & \ddots & \\
0 & & & \mu_{n-1}+(n-1) \varepsilon_{k}
\end{array}\right)
$$

where $U$ is the above unitary matrix. Thus the eigenvalues of $B_{k}, \hat{\mu}_{1}<\cdots<\hat{\mu}_{n-1}$ are strictly increasing and $\hat{\mu}_{j} \equiv \mu_{j}+j \varepsilon_{k}$. Let $A_{k}$ be given by

$$
A_{k}=\left(\begin{array}{cc}
a & \boldsymbol{y}^{*} \\
\boldsymbol{y} & B_{k}
\end{array}\right)
$$

Then

$$
\begin{aligned}
& \left(\begin{array}{cc}
1 & \mathbf{0}^{*} \\
\mathbf{0} & U^{*}
\end{array}\right) A_{k}\left(\begin{array}{cc}
1 & \mathbf{0}^{*} \\
\mathbf{0} & U
\end{array}\right) \\
= & \left(\begin{array}{cc}
1 & \mathbf{0}^{*} \\
\mathbf{0} & U^{*}
\end{array}\right)\left(\begin{array}{cc}
a & \boldsymbol{y}^{*} \\
\boldsymbol{y} & B_{k}
\end{array}\right)\left(\begin{array}{cc}
1 & \mathbf{0}^{*} \\
\mathbf{0} & U
\end{array}\right) \\
= & \left(\begin{array}{cc}
a & \boldsymbol{y}^{*} \\
U^{*} \boldsymbol{y} & U^{*} B_{k}
\end{array}\right)\left(\begin{array}{cc}
1 & \mathbf{0}^{*} \\
\mathbf{0} & U
\end{array}\right)=\left(\begin{array}{cc}
a & \boldsymbol{y}^{*} U \\
U^{*} \boldsymbol{y} & D_{k}
\end{array}\right)
\end{aligned}
$$

We can replace $\boldsymbol{y}$ in the statement of the theorem with $\boldsymbol{y}_{k}$ such that $\lim _{k \rightarrow \infty} \boldsymbol{y}_{k}=\boldsymbol{y}$ but $\boldsymbol{z}_{k} \equiv U^{*} \boldsymbol{y}_{k}$ has the property that each component of $\boldsymbol{z}_{k}$ is nonzero. This will probably take place automatically but if not, make the change. This makes a change in $A_{k}$ but still $\lim _{k \rightarrow \infty} A_{k}=A$. The main part of this argument which follows has to do with fixed $k$.

Expanding $\operatorname{det}\left(\lambda I-A_{k}\right)$ along the top row, the characteristic polynomial for $A_{k}$ is then

$$
\begin{equation*}
q(\lambda)=(\lambda-a) \prod_{i=1}^{n-1}\left(\lambda-\hat{\mu}_{i}\right)-\sum_{i=2}^{n-1}\left|z_{i}\right|^{2}\left(\lambda-\hat{\mu}_{1}\right) \cdots\left(\widehat{\lambda-\hat{\mu}_{i}}\right) \cdots\left(\lambda-\hat{\mu}_{n-1}\right) \tag{13.12}
\end{equation*}
$$

where $\left(\widehat{\lambda-\hat{\mu}_{i}}\right)$ indicates that this factor is omitted from the product $\prod_{i=1}^{n-1}\left(\lambda-\hat{\mu}_{i}\right)$. To see why this is so, consider the case where $B_{k}$ is $3 \times 3$. In this case, you would have

$$
\left(\begin{array}{cc}
1 & \mathbf{0}^{T} \\
\mathbf{0} & U^{*}
\end{array}\right)\left(\lambda I-A_{k}\right)\left(\begin{array}{cc}
1 & \mathbf{0}^{T} \\
\mathbf{0} & U
\end{array}\right)=\left(\begin{array}{cccc}
\lambda-a & \bar{z}_{1} & \bar{z}_{2} & \bar{z}_{3} \\
z_{1} & \lambda-\hat{\mu}_{1} & 0 & 0 \\
z_{2} & 0 & \lambda-\hat{\mu}_{2} & 0 \\
z_{3} & 0 & 0 & \lambda-\hat{\mu}_{3}
\end{array}\right)
$$

In general, you would have an $n \times n$ matrix on the right with the same appearance. Then expanding as indicated, the determinant is

$$
\begin{array}{r}
(\lambda-a) \prod_{i=1}^{3}\left(\lambda-\hat{\mu}_{i}\right)-\bar{z}_{1} \operatorname{det}\left(\begin{array}{ccc}
z_{1} & 0 & 0 \\
z_{2} & \lambda-\hat{\mu}_{2} & 0 \\
z_{3} & 0 & \lambda-\hat{\mu}_{3}
\end{array}\right) \\
+\bar{z}_{2} \operatorname{det}\left(\begin{array}{ccc}
z_{1} & \lambda-\hat{\mu}_{1} & 0 \\
z_{2} & 0 & 0 \\
z_{3} & 0 & \lambda-\hat{\mu}_{3}
\end{array}\right)-\bar{z}_{3} \operatorname{det}\left(\begin{array}{ccc}
z_{1} & \lambda-\hat{\mu}_{1} & 0 \\
z_{2} & 0 & \lambda-\hat{\mu}_{2} \\
z_{3} & 0 & 0
\end{array}\right) \\
=(\lambda-a) \prod_{i=1}^{3}\left(\lambda-\hat{\mu}_{i}\right)-\binom{\left|z_{1}\right|^{2}\left(\lambda-\hat{\mu}_{2}\right)\left(\lambda-\hat{\mu}_{3}\right)+\left|z_{2}\right|^{2}\left(\lambda-\hat{\mu}_{1}\right)\left(\lambda-\hat{\mu}_{3}\right)}{+\left|z_{3}\right|^{2}\left(\lambda-\hat{\mu}_{1}\right)\left(\lambda-\hat{\mu}_{2}\right)}
\end{array}
$$

Notice how, when you expand the $3 \times 3$ determinants along the first column, you have only one non-zero term and the sign is adjusted to give the above claim. Clearly, it works the same for any size matrix. Since the $\hat{\mu}_{i}$ are strictly increasing in $i$, it follows from 13.12 that $q\left(\hat{\mu}_{i}\right) q\left(\hat{\mu}_{i+1}\right) \leq 0$. However, since each $\left|z_{i}\right| \neq 0$, none of the $q\left(\hat{\mu}_{i}\right)$ can equal 0 and so $q\left(\hat{\mu}_{i}\right) q\left(\hat{\mu}_{i+1}\right)<0$. Hence, from the intermediate value theorem of calculus, there is a root of $q(\lambda)$ in each of the disjoint open intervals $\left(\hat{\mu}_{i}, \hat{\mu}_{i+1}\right)$. There are $n-2$ of these intervals and so this accounts for $n-2$ roots of $q(\lambda)$.

$$
q(\lambda)=(\lambda-a) \prod_{i=1}^{n-1}\left(\lambda-\hat{\mu}_{i}\right)-\sum_{i=2}^{n-1}\left|z_{i}\right|^{2}\left(\lambda-\hat{\mu}_{1}\right) \cdots\left(\widehat{\lambda-\hat{\mu}_{i}}\right) \cdots\left(\lambda-\hat{\mu}_{n-1}\right)
$$

What of $q\left(\hat{\mu}_{1}\right)$ ? Its sign is the same as $(-1)^{n-3}$ and also $q\left(\hat{\mu}_{n-1}\right)<0$. Therefore, there is a root to $q(\lambda)$ which is larger than $\hat{\mu}_{n-1}$. Indeed, $\lim _{\lambda \rightarrow \infty} q(\lambda)=\infty$ so there exists a root of $q(\lambda)$ strictly larger than $\hat{\mu}_{n-1}$. This accounts for $n-1$ roots of $q(\lambda)$. Now consider $q\left(\hat{\mu}_{1}\right)$. Suppose first that $n$ is odd. Then you have $q\left(\hat{\mu}_{1}\right)>0$. Hence, there is a root of $q(\lambda)$ which is no larger than $\hat{\mu}_{1}$ because in this case, $\lim _{\lambda \rightarrow-\infty} q(\lambda)=-\infty$. If $n$ is even, then $q\left(\hat{\mu}_{1}\right)<0$ and so there is a root of $q(\lambda)$ which is smaller than $\hat{\mu}_{1}$ because in this case, $\lim _{\lambda \rightarrow-\infty} q(\lambda)=\infty$. This accounts for all roots of $q(\lambda)$. Hence, if the roots of $q(\lambda)$ are $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$, it follows that

$$
\lambda_{1}<\hat{\mu}_{1}<\lambda_{2}<\hat{\mu}_{2}<\cdots<\hat{\mu}_{n-1}<\lambda_{n}
$$

To get the complete result, simply take the limit as $k \rightarrow \infty$. Then $\lim _{k \rightarrow \infty} \hat{\mu}_{k}=\mu_{k}$ and $A_{k} \rightarrow A$ and so the eigenvalues of $A_{k}$ converge to the corresponding eigenvalues of $A$ (See Problem 47 on Page 347), and so, passing to the limit, gives the desired result in which it may be necessary to replace $<$ with $\leq$.

Definition 13.6.2 Let A be an $n \times n$ matrix. An $(n-r) \times(n-r)$ matrix is called a principal submatrix of $A$ if it is obtained by deleting from $A$ the rows $i_{1}, i_{2}, \cdots, i_{r}$ and the columns $i_{1}, i_{2}, \cdots, i_{r}$.

Now the Cauchy interlacing theorem is really the following corollary.
Corollary 13.6.3 Let $A$ be an $n \times n$ Hermitian matrix and let $B$ be an $(n-1) \times(n-1)$ principal submatrix. Then the interlacing inequality holds $\lambda_{1} \leq \mu_{1} \leq \lambda_{2} \leq \mu_{2} \leq \cdots \leq$ $\mu_{n-1} \leq \lambda_{n}$ where the $\mu_{i}$ are the eigenvalues of $B$ listed in increasing order and the $\lambda_{i}$ are the eigenvalues of $A$ listed in increasing order.

Proof: Suppose $B$ is obtained from $A$ by deleting the $i^{\text {th }}$ row and the $i^{t h}$ column. Then let $P$ be the permutation matrix which switches the $i^{\text {th }}$ row with the first row. It is an orthogonal matrix and so its inverse is its transpose. The transpose switches the $i^{\text {th }}$ column with the first column. See Problem 40 on Page 99. Thus $P A P^{T}=\left(\begin{array}{cc}a & y^{*} \\ \boldsymbol{y} & B\end{array}\right)$ and it follows that the result of the multiplication is indeed as shown, a Hermitian matrix because $P, P^{T}$ are orthogonal matrices. Now the conclusion of the corollary follows from Theorem 13.6.1.

### 13.7 The Right Polar Factorization

The right polar factorization involves writing a matrix as a product of two other matrices, one which preserves distances and the other which stretches and distorts. This is of fundamental significance in geometric measure theory and also in continuum mechanics. Not surprisingly the stress should depend on the part which stretches and distorts. See [18].

First here are some lemmas which review and add to many of the topics discussed so far about adjoints and orthonormal sets and such things.

Lemma 13.7.1 Let A be a Hermitian matrix such that all its eigenvalues are nonnegative. Then there exists a Hermitian matrix $A^{1 / 2}$ such that $A^{1 / 2}$ has all nonnegative eigenvalues and $\left(A^{1 / 2}\right)^{2}=A$.

Proof: Since $A$ is Hermitian, there exists a diagonal matrix $D$ having all real nonnegative entries and a unitary matrix $U$ such that $A=U^{*} D U$. Then denote by $D^{1 / 2}$ the matrix which is obtained by replacing each diagonal entry of $D$ with its square root. Thus $D^{1 / 2} D^{1 / 2}=D$. Then define

$$
A^{1 / 2} \equiv U^{*} D^{1 / 2} U
$$

Then

$$
\left(A^{1 / 2}\right)^{2}=U^{*} D^{1 / 2} U U^{*} D^{1 / 2} U=U^{*} D U=A
$$

Since $D^{1 / 2}$ is real,

$$
\left(U^{*} D^{1 / 2} U\right)^{*}=U^{*}\left(D^{1 / 2}\right)^{*}\left(U^{*}\right)^{*}=U^{*} D^{1 / 2} U
$$

so $A^{1 / 2}$ is Hermitian.
In fact this square root is unique. This is shown a little later after the main result of this section.

Next it is helpful to recall the Gram Schmidt algorithm and observe a certain property stated in the next lemma.

Lemma 13.7.2 Suppose $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_{p}\right\}$ is a linearly independent set of vectors such that $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}\right\}$ is an orthonormal set of vectors. Then when the Gram Schmidt process is applied to the vectors in the given order, it will not change any of the $\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}$.

Proof: Let $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{p}\right\}$ be the orthonormal set delivered by the Gram Schmidt process. Then $\boldsymbol{u}_{1}=\boldsymbol{w}_{1}$ because by definition, $\boldsymbol{u}_{1} \equiv \boldsymbol{w}_{1} /\left|\boldsymbol{w}_{1}\right|=\boldsymbol{w}_{1}$. Now suppose $\boldsymbol{u}_{j}=\boldsymbol{w}_{j}$ for all $j \leq k \leq r$. Then if $k<r$, consider the definition of $\boldsymbol{u}_{k+1}$.

$$
\boldsymbol{u}_{k+1} \equiv \frac{\boldsymbol{w}_{k+1}-\sum_{j=1}^{k+1}\left(\boldsymbol{w}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{u}_{j}}{\left|\boldsymbol{w}_{k+1}-\sum_{j=1}^{k+1}\left(\boldsymbol{w}_{k+1}, \boldsymbol{u}_{j}\right) \boldsymbol{u}_{j}\right|}
$$

By induction, $\boldsymbol{u}_{j}=\boldsymbol{w}_{j}$ and so this reduces to $\boldsymbol{w}_{k+1} /\left|\boldsymbol{w}_{k+1}\right|=\boldsymbol{w}_{k+1}$ since $\left|\boldsymbol{w}_{k+1}\right|=1$.
This lemma immediately implies the following lemma.
Lemma 13.7.3 Let $V$ be a subspace of dimension $p$ and let $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}\right\}$ be an orthonormal set of vectors in $V$. Then this orthonormal set of vectors may be extended to an orthonormal basis for $V$,

$$
\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_{p}\right\}
$$

Proof: First extend the given linearly independent set $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}\right\}$ to a basis for $V$ and then apply the Gram Schmidt theorem to the resulting basis. Since $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}\right\}$ is orthonormal it follows from Lemma 13.7.2 the result is of the desired form, an orthonormal basis extending $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{r}\right\}$.

Recall Lemma 12.3.5 which is about preserving distances. It is restated here in the case of an $m \times n$ matrix.

Lemma 13.7.4 Suppose $R$ is an $m \times n$ matrix with $m \geq n$ and $R$ preserves distances. Then $R^{*} R=I$.

With this preparation, here is the big theorem about the right polar factorization.
Theorem 13.7.5 Let $F$ be an $m \times n$ matrix where $m \geq n$. Then there exists a Hermitian $n \times$ $n$ matrix $U$ which has all nonnegative eigenvalues and an $m \times n$ matrix $R$ which preserves distances and satisfies $R^{*} R=I$ such that $F=R U$.

Proof: Consider $F^{*} F$. This is a Hermitian matrix because

$$
\left(F^{*} F\right)^{*}=F^{*}\left(F^{*}\right)^{*}=F^{*} F
$$

Also the eigenvalues of the $n \times n$ matrix $F^{*} F$ are all nonnegative. This is because if $\boldsymbol{x}$ is an eigenvalue,

$$
\lambda(\boldsymbol{x}, \boldsymbol{x})=\left(F^{*} F \boldsymbol{x}, \boldsymbol{x}\right)=(\boldsymbol{F} \boldsymbol{x}, F \boldsymbol{x}) \geq 0 .
$$

Therefore, by Lemma 13.7.1, there exists an $n \times n$ Hermitian matrix $U$ having all nonnegative eigenvalues such that

$$
U^{2}=F^{*} F
$$

Consider the subspace $U\left(\mathbb{F}^{n}\right)$. Let $\left\{U \boldsymbol{x}_{1}, \cdots, U \boldsymbol{x}_{r}\right\}$ be an orthonormal basis for

$$
U\left(\mathbb{F}^{n}\right) \subseteq \mathbb{F}^{n}
$$

Note that $U\left(\mathbb{F}^{n}\right)$ might not be all of $\mathbb{F}^{n}$. Using Lemma 13.7.3, extend to an orthonormal basis for all of $\mathbb{F}^{n}$,

$$
\left\{U \boldsymbol{x}_{1}, \cdots, U \boldsymbol{x}_{r}, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_{n}\right\} .
$$

Next observe that $\left\{F \boldsymbol{x}_{1}, \cdots, F \boldsymbol{x}_{r}\right\}$ is also an orthonormal set of vectors in $\mathbb{F}^{m}$. This is because

$$
\begin{aligned}
\left(F \boldsymbol{x}_{k}, F \boldsymbol{x}_{j}\right) & =\left(F^{*} F \boldsymbol{x}_{k}, \boldsymbol{x}_{j}\right)=\left(U^{2} \boldsymbol{x}_{k}, \boldsymbol{x}_{j}\right) \\
& =\left(U \boldsymbol{x}_{k}, U^{*} \boldsymbol{x}_{j}\right)=\left(U \boldsymbol{x}_{k}, U \boldsymbol{x}_{j}\right)=\delta_{j k}
\end{aligned}
$$

Therefore, from Lemma 13.7.3 again, this orthonormal set of vectors can be extended to an orthonormal basis for $\mathbb{F}^{m}$,

$$
\left\{F \boldsymbol{x}_{1}, \cdots, F \boldsymbol{x}_{r}, \boldsymbol{z}_{r+1}, \cdots, \boldsymbol{z}_{m}\right\}
$$

Thus there are at least as many $\boldsymbol{z}_{k}$ as there are $\boldsymbol{y}_{j}$ because $m \geq n$. Now for $\boldsymbol{x} \in \mathbb{F}^{n}$, since

$$
\left\{U \boldsymbol{x}_{1}, \cdots, U \boldsymbol{x}_{r}, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_{n}\right\}
$$

is an orthonormal basis for $\mathbb{F}^{n}$, there exist unique scalars,

$$
c_{1} \cdots, c_{r}, d_{r+1}, \cdots, d_{n}
$$

such that

$$
\boldsymbol{x}=\sum_{k=1}^{r} c_{k} U \boldsymbol{x}_{k}+\sum_{k=r+1}^{n} d_{k} \boldsymbol{y}_{k}
$$

Define

$$
\begin{equation*}
R \boldsymbol{x} \equiv \sum_{k=1}^{r} c_{k} F \boldsymbol{x}_{k}+\sum_{k=r+1}^{n} d_{k} \boldsymbol{z}_{k} \tag{13.13}
\end{equation*}
$$

Thus, since $\left\{F \boldsymbol{x}_{1}, \cdots, F \boldsymbol{x}_{r}, \boldsymbol{z}_{r+1}, \cdots, \boldsymbol{z}_{n}\right\}$ is orthonormal,

$$
|R \boldsymbol{x}|^{2}=\sum_{k=1}^{r}\left|c_{k}\right|^{2}+\sum_{k=r+1}^{n}\left|d_{k}\right|^{2}=|\boldsymbol{x}|^{2}
$$

and so it follows from Corollary 12.3 .8 or Lemma 13.7.4 that $R^{*} R=I$. Then also there exist scalars $b_{k}$ such that

$$
\begin{equation*}
U \boldsymbol{x}=\sum_{k=1}^{r} b_{k} U \boldsymbol{x}_{k} \tag{13.14}
\end{equation*}
$$

and so from 13.13,

$$
R U \boldsymbol{x}=\sum_{k=1}^{r} b_{k} F \boldsymbol{x}_{k}=F\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}\right)
$$

Is $F\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}\right)=F(\boldsymbol{x})$ ? Using 13.14,

$$
\begin{aligned}
& \left(F\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}\right)-F(\boldsymbol{x}), F\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}\right)-F(\boldsymbol{x})\right) \\
& =\left(\left(F^{*} F\right)\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right),\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right)\right) \\
& =\left(U^{2}\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right),\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right)\right) \\
& =\left(U\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right), U\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}-\boldsymbol{x}\right)\right) \\
& =\left(\sum_{k=1}^{r} b_{k} U \boldsymbol{x}_{k}-U \boldsymbol{x}, \sum_{k=1}^{r} b_{k} U \boldsymbol{x}_{k}-U \boldsymbol{x}\right)=0
\end{aligned}
$$

Therefore, $F\left(\sum_{k=1}^{r} b_{k} \boldsymbol{x}_{k}\right)=F(\boldsymbol{x})$ and this shows $R U \boldsymbol{x}=F \boldsymbol{x}$.
Note that $U^{2}$ is completely determined by $F$ because $F^{*} F=U R^{*} R U=U^{2}$. In fact, $U$ is also uniquely determined. This will be shown later in Theorem 13.8.1. First is an easy corollary of this theorem.

Corollary 13.7.6 Let $F$ be $m \times n$ and suppose $n \geq m$. Then there exists a Hermitian $U$ and and $R$, such that

$$
F=U R, R R^{*}=I
$$

Proof: Recall that $L^{* *}=L$ and $(M L)^{*}=L^{*} M^{*}$. Now apply Theorem 13.7.5 to $F^{*}$. Thus, $F^{*}=R^{*} U$ where $R^{*}$ and $U$ satisfy the conditions of that theorem. In particular $R^{*}$ preserves distances. Then $F=U R$ and $R R^{*}=R^{* *} R^{*}=I$.

### 13.8 The Square Root

Now here is a uniqueness and existence theorem for the square root. It follows from this theorem that $U$ in the above right polar decomposition of Theorem 13.7.5 is unique.

Theorem 13.8.1 Let $A$ be a self adjoint and nonnegative $n \times n$ matrix (all eigenvalues are nonnegative). Then there exists a unique self adjoint nonnegative matrix $B$ such that $B^{2}=A$.

Proof: Suppose $B^{2}=A$ where $B$ is such a Hermitian square root for $A$ with nonnegative eigenvalues. Then by Theorem 13.1.6, $B$ has an orthonormal basis for $\mathbb{F}^{n}$ of eigenvectors $\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{n}\right\}$.

$$
B \boldsymbol{u}_{i}=\mu_{i} \boldsymbol{u}_{i}
$$

Thus

$$
B=\sum_{i} \mu_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}
$$

because both linear transformations agree on the orthonormal basis. But this implies that

$$
A \boldsymbol{u}_{i}=B^{2} \boldsymbol{u}_{i}=\mu_{i}^{2} \boldsymbol{u}_{i}
$$

Thus these are also an orthonormal basis of eigenvectors for $A$. Hence, letting $\lambda_{i}=\mu_{i}^{2}$

$$
A=\sum_{i} \lambda_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}, B=\sum_{i} \lambda_{i}^{1 / 2} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}
$$

Let $p(\lambda)$ be a polynomial such that $p\left(\lambda_{i}\right)=\lambda_{i}^{1 / 2}$. Say $p(\lambda)=a_{0}+a_{1} \lambda \cdots+a_{p} \lambda^{p}$. Then

$$
\begin{align*}
A^{m}= & \left(\sum_{i} \lambda_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}\right)^{m}=\sum_{i_{1}, \cdots, i_{m}} \lambda_{i_{1}} \boldsymbol{u}_{i_{1}} \boldsymbol{u}_{i_{1}}^{*} \lambda_{i_{2}} \boldsymbol{u}_{i_{2}} \boldsymbol{u}_{i_{2}}^{*} \cdots \lambda_{i_{m}} \boldsymbol{u}_{i_{m}} \boldsymbol{u}_{i_{m}}^{*}  \tag{13.15}\\
= & \sum_{i_{1}, \cdots, i_{m}} \lambda_{i_{1}} \lambda_{i_{2}} \cdots \lambda_{i_{m}} \boldsymbol{u}_{i_{1}} \boldsymbol{u}_{i_{1}}^{*} \boldsymbol{u}_{i_{2}} \boldsymbol{u}_{i_{2}}^{*} \cdots \boldsymbol{u}_{i_{m}} \boldsymbol{u}_{i_{m}}^{*} \\
= & \sum_{i_{1}, \cdots, i_{m}} \lambda_{i_{1}} \lambda_{i_{2}} \cdots \lambda_{i_{m}} \boldsymbol{u}_{i_{1}} \boldsymbol{u}_{i_{m}}^{*} \delta_{i_{1} i_{2}} \delta_{i_{2} i_{3}} \cdots \delta_{i_{m-1} i_{m}} \\
= & \sum_{i_{1}, \cdots, i_{m-1}} \lambda_{i_{1}} \lambda_{i_{2}} \cdots \lambda_{i_{m-1}}^{2} \boldsymbol{u}_{i_{1}} \boldsymbol{u}_{i_{m-1}}^{*} \delta_{i_{1} i_{2}} \delta_{i_{2} i_{3}} \cdots \delta_{i_{m-2} i_{m-1}} \\
& \vdots \\
= & \sum_{i_{1}} \lambda_{i_{1}}^{m} \boldsymbol{u}_{i_{1}} u_{i_{1}}^{*}=\sum_{i} \lambda_{i}^{m} \boldsymbol{u}_{i} u_{i}^{*} \tag{13.16}
\end{align*}
$$

Therefore,

$$
\begin{gather*}
p(A)=a_{0} I+a_{1} A \cdots+a_{p} A^{p} \\
=a_{0} \sum_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}+a_{1} \sum_{i} \lambda_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}+\cdots+a_{p} \sum_{i} \lambda_{i}^{p} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*} \\
=\sum_{i} p\left(\lambda_{i}\right) \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}=\sum_{i} \lambda_{i}^{1 / 2} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}=B \tag{13.17}
\end{gather*}
$$

and so $B$ commutes with every matrix which commutes with $A$. To see this, suppose $C A=$ $A C$, then

$$
B C=p(A) C=C p(A)=B
$$

This shows that if $B$ is such a square root, then it commutes with every matrix $C$ which commutes with $A$. It also shows, by a repeat of the argument 13.15-13.16 that $B^{2}=A$.

Could there be another such Hermitian square root which has all nonnegative eigenvalues? It was just shown that any such square root commutes with every matrix which commutes with $A$. Suppose $B_{1}$ is another square root which is self adjoint, and has nonnegative eignevalues. Since both $B, B_{1}$ are nonnegative,

$$
\begin{align*}
& \left(B\left(B-B_{1}\right) x,\left(B-B_{1}\right) x\right) \geq 0, \\
& \left(B_{1}\left(B-B_{1}\right) x,\left(B-B_{1}\right) x\right) \geq 0 \tag{13.18}
\end{align*}
$$

Now, adding these together, and using the fact that the two commute because they both commute with every matrix which commutes with $A$,

$$
\begin{gathered}
\left(\left(B+B_{1}\right)\left(B-B_{1}\right) \boldsymbol{x},\left(B-B_{1}\right) \boldsymbol{x}\right) \geq 0 \\
\left(\left(B^{2}-B_{1}^{2}\right) \boldsymbol{x},\left(B-B_{1}\right) \boldsymbol{x}\right)=\left((A-A) \boldsymbol{x},\left(B-B_{1}\right) \boldsymbol{x}\right)=0
\end{gathered}
$$

It follows that both inner products in 13.18 equal 0 . Next use the existence part shown above to take the square root of $B$ and $B_{1}$ which is denoted by $\sqrt{B}, \sqrt{B_{1}}$ respectively. Then

$$
\begin{aligned}
& 0=\left(\sqrt{B}\left(B-B_{1}\right) x, \sqrt{B}\left(B-B_{1}\right) x\right) \\
& 0=\left(\sqrt{B_{1}}\left(B-B_{1}\right) x, \sqrt{B_{1}}\left(B-B_{1}\right) x\right)
\end{aligned}
$$

which implies $\sqrt{B}\left(B-B_{1}\right) \boldsymbol{x}=\sqrt{B_{1}}\left(B-B_{1}\right) \boldsymbol{x}=0$. Thus also,

$$
B\left(B-B_{1}\right) x=B_{1}\left(B-B_{1}\right) x=0
$$

Hence

$$
0=\left(B\left(B-B_{1}\right) \boldsymbol{x}-B_{1}\left(B-B_{1}\right) \boldsymbol{x}, \boldsymbol{x}\right)=\left(\left(B-B_{1}\right) \boldsymbol{x},\left(B-B_{1}\right) \boldsymbol{x}\right)
$$

and so, since $\boldsymbol{x}$ is arbitrary, $B_{1}=B$.
Corollary 13.8.2 The $U$ in Theorem 13.7.5 is unique.

### 13.9 An Application To Statistics

A random vector is a function $\boldsymbol{X}: \Omega \rightarrow \mathbb{R}^{p}$ where $\Omega$ is a probability space. This means that there exists a $\sigma$ algebra of measurable sets $\mathscr{F}$ and a probability measure $P: \mathscr{F} \rightarrow[0,1]$. In practice, people often don't worry too much about the underlying probability space and instead pay more attention to the distribution measure of the random variable. For $E$ a suitable subset of $\mathbb{R}^{p}$, this measure gives the probability that $\boldsymbol{X}$ has values in $E$. There are often excellent reasons for believing that a random vector is normally distributed. This means that the probability that $\boldsymbol{X}$ has values in a set $E$ is given by

$$
\int_{E} \frac{1}{(2 \pi)^{p / 2} \operatorname{det}(\Sigma)^{1 / 2}} \exp \left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{*} \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})\right) d \boldsymbol{x}
$$

The expression in the integral is called the normal probability density function. There are two parameters, $\boldsymbol{m}$ and $\Sigma$ where $\boldsymbol{m}$ is called the mean and $\Sigma$ is called the covariance matrix. It is a symmetric matrix which has all real eigenvalues which are all positive. While it may be reasonable to assume this is the distribution, in general, you won't know $\boldsymbol{m}$ and $\Sigma$ and in order to use this formula to predict anything, you would need to know these quantities. I am following a nice discussion given in Wikipedia which makes use of the existence of square roots.

What people do to estimate $\boldsymbol{m}$, and $\Sigma$ is to take $n$ independent observations $\boldsymbol{x}_{1}, \cdots, \boldsymbol{x}_{n}$ and try to predict what $\boldsymbol{m}$ and $\Sigma$ should be based on these observations. One criterion used for making this determination is the method of maximum likelihood. In this method, you seek to choose the two parameters in such a way as to maximize the likelihood which is given as

$$
\prod_{i=1}^{n} \frac{1}{\operatorname{det}(\Sigma)^{1 / 2}} \exp \left(-\frac{1}{2}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\right)
$$

For convenience the term $(2 \pi)^{p / 2}$ was ignored. Maximizing the above is equivalent to maximizing the $\ln$ of the above. So taking $\ln$,

$$
\frac{n}{2} \ln \left(\operatorname{det}\left(\Sigma^{-1}\right)\right)-\frac{1}{2} \sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)
$$

Note that the above is a function of the entries of $\boldsymbol{m}$. Take the partial derivative with respect to $m_{l}$. Since the matrix $\Sigma^{-1}$ is symmetric this implies

$$
\sum_{i=1}^{n} \sum_{r}\left(x_{i r}-m_{r}\right) \Sigma_{r l}^{-1}=0 \text { each } l
$$

Written in terms of vectors,

$$
\sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}=\mathbf{0}
$$

and so, multiplying by $\Sigma$ on the right and then taking adjoints, this yields

$$
\sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)=\mathbf{0}, n \boldsymbol{m}=\sum_{i=1}^{n} \boldsymbol{x}_{i}, \boldsymbol{m}=\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{i} \equiv \overline{\boldsymbol{x}} .
$$

Now that $\boldsymbol{m}$ is determined, it remains to find the best estimate for $\Sigma$.

$$
\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)
$$

is a scalar, so since trace $(A B)=\operatorname{trace}(B A)$,

$$
\begin{aligned}
\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right) & =\operatorname{trace}\left(\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\right) \\
& =\operatorname{trace}\left(\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\right)
\end{aligned}
$$

Therefore, the thing to maximize is

$$
\begin{aligned}
& n \ln \left(\operatorname{det}\left(\Sigma^{-1}\right)\right)-\sum_{i=1}^{n} \operatorname{trace}\left(\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*} \Sigma^{-1}\right) \\
= & n \ln \left(\operatorname{det}\left(\Sigma^{-1}\right)\right)-\operatorname{trace}(\overbrace{\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*}\right)}^{S} \Sigma^{-1})
\end{aligned}
$$

We assume that $S$ has rank $p$. Thus it is a self adjoint matrix which has all positive eigenvalues. Therefore, from the property of the trace, trace $(A B)=\operatorname{trace}(B A)$, the thing to maximize is

$$
n \ln \left(\operatorname{det}\left(\Sigma^{-1}\right)\right)-\operatorname{trace}\left(S^{1 / 2} \Sigma^{-1} S^{1 / 2}\right)
$$

Now let $B=S^{1 / 2} \Sigma^{-1} S^{1 / 2}$. Then $B$ is positive and self adjoint also and so there exists $U$ unitary such that $B=U^{*} D U$ where $D$ is the diagonal matrix having the positive scalars $\lambda_{1}, \cdots, \lambda_{p}$ down the main diagonal. Solving for $\Sigma^{-1}$ in terms of $B$, this yields

$$
S^{-1 / 2} B S^{-1 / 2}=\Sigma^{-1}
$$

and so

$$
\begin{aligned}
\ln \left(\operatorname{det}\left(\Sigma^{-1}\right)\right) & =\ln \left(\operatorname{det}\left(S^{-1 / 2}\right) \operatorname{det}(B) \operatorname{det}\left(S^{-1 / 2}\right)\right) \\
& =\ln \left(\operatorname{det}\left(S^{-1}\right)\right)+\ln (\operatorname{det}(B))
\end{aligned}
$$

which yields

$$
C(S)+n \ln (\operatorname{det}(B))-\operatorname{trace}(B)
$$

as the thing to maximize. Of course this yields

$$
\begin{aligned}
& C(S)+n \ln \left(\prod_{i=1}^{p} \lambda_{i}\right)-\sum_{i=1}^{p} \lambda_{i} \\
= & C(S)+n \sum_{i=1}^{p} \ln \left(\lambda_{i}\right)-\sum_{i=1}^{p} \lambda_{i}
\end{aligned}
$$

as the quantity to be maximized. To do this, take $\partial / \partial \lambda_{k}$ and set equal to 0 . This yields $\lambda_{k}=n$. Therefore, from the above, $B=U^{*} n I U=n I$. Also from the above,

$$
B^{-1}=\frac{1}{n} I=S^{-1 / 2} \Sigma S^{-1 / 2}
$$

and so

$$
\Sigma=\frac{1}{n} S=\frac{1}{n} \sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*}
$$

This has shown that the maximum likelihood estimates are

$$
\boldsymbol{m}=\overline{\boldsymbol{x}} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{i}, \Sigma=\frac{1}{n} \sum_{i=1}^{n}\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{m}\right)^{*}
$$

### 13.10 Simultaneous Diagonalization

Recall the following definition of what it means for a matrix to be diagonalizable.
Definition 13.10.1 Let $A$ be an $n \times n$ matrix. It is said to be diagonalizable if there exists an invertible matrix $S$ such that

$$
S^{-1} A S=D
$$

where $D$ is a diagonal matrix.

Also, here is a useful observation.
Observation 13.10.2 If $A$ is an $n \times n$ matrix and $A S=S D$ for $D$ a diagonal matrix, then each column of $S$ is an eigenvector or else it is the zero vector. This follows from observing that for $s_{k}$ the $k^{\text {th }}$ column of $S$ and from the way we multiply matrices,

$$
A s_{k}=\lambda_{k} s_{k}
$$

It is sometimes interesting to consider the problem of finding a single similarity transformation which will diagonalize all the matrices in some set.

Lemma 13.10.3 Let $A$ be an $n \times n$ matrix and let $B$ be an $m \times m$ matrix. Denote by $C$ the matrix

$$
C \equiv\left(\begin{array}{cc}
A & 0 \\
0 & B
\end{array}\right)
$$

Then $C$ is diagonalizable if and only if both $A$ and $B$ are diagonalizable.
Proof: Suppose $S_{A}^{-1} A S_{A}=D_{A}$ and $S_{B}^{-1} B S_{B}=D_{B}$ where $D_{A}$ and $D_{B}$ are diagonal matrices. You should use block multiplication to verify that $S \equiv\left(\begin{array}{cc}S_{A} & 0 \\ 0 & S_{B}\end{array}\right)$ is such that $S^{-1} C S=D_{C}$, a diagonal matrix.

Consider the converse that $C$ is diagonalizable. It is necessary to show that $A$ has a basis of eigenvectors for $\mathbb{F}^{n}$ and that $B$ has a basis of eigenvectors in $\mathbb{F}^{m}$. Thus $S$ has columns $s_{i}$. Suppose $C$ is diagonalized by $S=\left(\begin{array}{lll}s_{1} & \cdots & s_{n+m}\end{array}\right)$. For each of these columns, write in the form

$$
s_{i}=\binom{\boldsymbol{x}_{i}}{\boldsymbol{y}_{i}}
$$

where $\boldsymbol{x}_{i} \in \mathbb{F}^{n}$ and where $\boldsymbol{y}_{i} \in \mathbb{F}^{m}$. The result is

$$
S=\left(\begin{array}{ll}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{array}\right)
$$

where $S_{11}$ is an $n \times n$ matrix and $S_{22}$ is an $m \times m$ matrix. Then there is a diagonal matrix, $D_{1}$ being $n \times n$ and $D_{2} m \times m$ such that

$$
D=\operatorname{diag}\left(\lambda_{1}, \cdots, \lambda_{n+m}\right)=\left(\begin{array}{cc}
D_{1} & 0 \\
0 & D_{2}
\end{array}\right)
$$

such that

$$
\begin{aligned}
& \left(\begin{array}{cc}
A & 0 \\
0 & B
\end{array}\right)\left(\begin{array}{ll}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{array}\right) \\
= & \left(\begin{array}{ll}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{array}\right)\left(\begin{array}{cc}
D_{1} & 0 \\
0 & D_{2}
\end{array}\right)
\end{aligned}
$$

Hence by block multiplication,

$$
\left(\begin{array}{ll}
A S_{11} & A S_{12} \\
B S_{21} & B S_{22}
\end{array}\right)=\left(\begin{array}{ll}
S_{11} D_{1} & S_{12} D_{2} \\
S_{21} D_{1} & S_{22} D_{2}
\end{array}\right)
$$

Thus,

$$
\begin{aligned}
& A S_{11}=S_{11} D_{1}, B S_{22}=S_{22} D_{2} \\
& B S_{21}=S_{21} D_{1}, A S_{12}=S_{12} D_{2}
\end{aligned}
$$

It follows each of the $x_{i}$ is an eigenvector of $A$ or else is the zero vector and that each of the $\boldsymbol{y}_{i}$ is an eigenvector of $B$ or is the zero vector. If there are $n$ linearly independent $\boldsymbol{x}_{i}$, then $A$ is diagonalizable by Theorem 6.4.3 on Page 6.4.3.

The row rank of the top half of $S$, the matrix $\left(\begin{array}{ccc}\boldsymbol{x}_{1} & \cdots & x_{n+m}\end{array}\right)$ must be $n$ because if this is not so, the row rank of $S$ would be less than $n+m$ which would mean $S^{-1}$ does not exist. Therefore, since the column rank equals the row rank, this top half of $S$ has column rank equal to $n$ and this means there are $n$ linearly independent eigenvectors of $A$ implying that $A$ is diagonalizable. Similar reasoning applies to $B$ by considering the bottom half of $S$.

Note that once you know that each of $A, B$ are diagonalizable, you can then use the specific method used in the first part to accomplish the diagonalization.

The following corollary follows from the same type of argument as the above.
Corollary 13.10.4 Let $A_{k}$ be an $n_{k} \times n_{k}$ matrix and let $C$ denote the block diagonal

$$
\left(\sum_{k=1}^{r} n_{k}\right) \times\left(\sum_{k=1}^{r} n_{k}\right)
$$

matrix given below.

$$
C \equiv\left(\begin{array}{ccc}
A_{1} & & 0 \\
& \ddots & \\
0 & & A_{r}
\end{array}\right)
$$

Then $C$ is diagonalizable if and only if each $A_{k}$ is diagonalizable.
Definition 13.10.5 A set, $\mathscr{F}$ of $n \times n$ matrices is said to be simultaneously diagonalizable if and only if there exists a single invertible matrix $S$ such that for every $A \in \mathscr{F}, S^{-1} A S=D_{A}$ where $D_{A}$ is a diagonal matrix. $\mathscr{F}$ is a commuting family of matrices if whenever $A, B \in \mathscr{F}$, $A B=B A$.

Lemma 13.10.6 If $\mathscr{F}$ is a set of $n \times n$ matrices which is simultaneously diagonalizable, then $\mathscr{F}$ is a commuting family of matrices.

Proof: Let $A, B \in \mathscr{F}$ and let $S$ be a matrix which has the property that $S^{-1} A S$ is a diagonal matrix for all $A \in \mathscr{F}$. Then $S^{-1} A S=D_{A}$ and $S^{-1} B S=D_{B}$ where $D_{A}$ and $D_{B}$ are diagonal matrices. Since diagonal matrices commute,

$$
\begin{aligned}
A B & =S D_{A} S^{-1} S D_{B} S^{-1}=S D_{A} D_{B} S^{-1} \\
& =S D_{B} D_{A} S^{-1}=S D_{B} S^{-1} S D_{A} S^{-1}=B A
\end{aligned}
$$

Lemma 13.10.7 Let $D$ be a diagonal matrix of the form

$$
D \equiv\left(\begin{array}{cccc}
\lambda_{1} I_{n_{1}} & 0 & \cdots & 0  \tag{13.19}\\
0 & \lambda_{2} I_{n_{2}} & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_{r} I_{n_{r}}
\end{array}\right)
$$

where $I_{n_{i}}$ denotes the $n_{i} \times n_{i}$ identity matrix and $\lambda_{i} \neq \lambda_{j}$ for $i \neq j$ and suppose $B$ is a matrix which commutes with $D$. Then $B$ is a block diagonal matrix of the form

$$
B=\left(\begin{array}{cccc}
B_{1} & 0 & \cdots & 0  \tag{13.20}\\
0 & B_{2} & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & B_{r}
\end{array}\right)
$$

where $B_{i}$ is an $n_{i} \times n_{i}$ matrix.
Proof: Let $B=\left(B_{i j}\right)$ where $B_{i i}=B_{i}$ a block matrix as above in 13.20. Since it commutes with $D$,

$$
\begin{aligned}
& \left(\begin{array}{cccc}
B_{11} & B_{12} & \cdots & B_{1 r} \\
B_{21} & B_{22} & \ddots & B_{2 r} \\
\vdots & \ddots & \ddots & \vdots \\
B_{r 1} & B_{r 2} & \cdots & B_{r r}
\end{array}\right)\left(\begin{array}{cccc}
\lambda_{1} I_{n_{1}} & 0 & \cdots & 0 \\
0 & \lambda_{2} I_{n_{2}} & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_{r} I_{n_{r}}
\end{array}\right) \\
& =\left(\begin{array}{cccc}
\lambda_{1} I_{n_{1}} & 0 & \cdots & 0 \\
0 & \lambda_{2} I_{n_{2}} & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_{r} I_{n_{r}}
\end{array}\right)\left(\begin{array}{cccc}
B_{11} & B_{12} & \cdots & B_{1 r} \\
B_{21} & B_{22} & \ddots & B_{2 r} \\
\vdots & \ddots & \ddots & \vdots \\
B_{r 1} & B_{r 2} & \cdots & B_{r r}
\end{array}\right)
\end{aligned}
$$

Thus

$$
\lambda_{j} B_{i j}=\lambda_{i} B_{i j}
$$

Therefore, if $i \neq j, B_{i j}=0$. Hence $B$ as the form which is claimed.
Lemma 13.10.8 Let $\mathscr{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathscr{F}$ is diagonalizable. Then $\mathscr{F}$ is simultaneously diagonalizable.

$$
\text { commuting }+ \text { diagonalizable } \Rightarrow \text { simultaneously diagonalizable }
$$

Proof: First note that if every matrix in $\mathscr{F}$ has only one eigenvalue, there is nothing to prove. This is because for $A$ such a matrix,

$$
S^{-1} A S=\lambda I
$$

and so

$$
A=\lambda I
$$

Thus all the matrices in $\mathscr{F}$ are diagonal matrices and you could pick any $S$ to diagonalize them all. Therefore, without loss of generality, assume some matrix in $\mathscr{F}$ has more than one eigenvalue.

The significant part of the lemma is proved by induction on $n$. If $n=1$, there is nothing to prove because all the $1 \times 1$ matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n-1$ where $n \geq 2$ and let $\mathscr{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathscr{F}$ which has more than one eigenvalue and let $S$ be an invertible matrix such that $S^{-1} A S=D$ where $D$ is of the form given in 13.19. By permuting the columns of $S$ there is no loss of generality in assuming $D$ has this form. Now denote by $\widetilde{\mathscr{F}}$ the collection of matrices, $\left\{S^{-1} C S: C \in \mathscr{F}\right\}$. Note $\widetilde{\mathscr{F}}$ features the single matrix $S$.

It follows easily that $\widetilde{\mathscr{F}}$ is also a commuting family of diagonalizable matrices. Indeed,

$$
\left(S^{-1} C S\right)\left(S^{-1} \hat{C} S\right)=S^{-1} C \hat{C} S=S^{-1} \hat{C} C S=\left(S^{-1} \hat{C} S\right)\left(S^{-1} C S\right)
$$

so the matrices commute. Now if $M$ is a matrix in $\widetilde{\mathscr{F}}$, then $S^{-1} C S=M$ where $C \in \mathscr{F}$ and so

$$
C=S M S^{-1}
$$

By assumption, there exists $T$ such that $T^{-1} C T=D$ and so

$$
D=T^{-1} C T=T^{-1} S M S^{-1} T=\left(S^{-1} T\right)^{-1} M S^{-1} T
$$

showing that $M$ is also diagonalizable.
By Lemma 13.10.7 every $B \in \widetilde{\mathscr{F}}$ is a block diagonal matrix of the form given in 13.20 because each of these commutes with $D$ described above as $S^{-1} A S$ and so by block multiplication, the diagonal blocks $B_{i}, \hat{B}_{i}$ corresponding respectively to $B, \hat{B} \in \widetilde{\mathscr{F}}$ commute.

By Corollary 13.10.4 each of these blocks is diagonalizable. This is because $B$ is known to be so. Therefore, by induction, since all the blocks are no larger than $n-1 \times n-1$, thanks to the assumption that $A$ has more than one eigenvalue, there exist invertible $n_{i} \times n_{i}$ matrices, $T_{i}$ such that $T_{i}^{-1} B_{i} T_{i}$ is a diagonal matrix whenever $B_{i}$ is one of the matrices making up the block diagonal of any $B \in \widetilde{\mathscr{F}}$. It follows that for $T$ defined by

$$
T \equiv\left(\begin{array}{cccc}
T_{1} & 0 & \cdots & 0 \\
0 & T_{2} & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & T_{r}
\end{array}\right)
$$

then $T^{-1} B T=$ a diagonal matrix for every $B \in \widetilde{\mathscr{F}}$ including $D$. Consider $S T$. It follows that for all $C \in \mathscr{F}$,

$$
T^{-1} \overbrace{S^{-1} C S}^{\text {something in } \widetilde{\mathscr{F}}} T=(S T)^{-1} C(S T)=\text { a diagonal matrix. }
$$

Theorem 13.10.9 Let $\mathscr{F}$ denote a family of matrices which are diagonalizable. Then $\mathscr{F}$ is simultaneously diagonalizable if and only if $\mathscr{F}$ is a commuting family.

Proof: If $\mathscr{F}$ is a commuting family, it follows from Lemma 13.10.8 that it is simultaneously diagonalizable. If it is simultaneously diagonalizable, then it follows from Lemma 13.10.6 that it is a commuting family.

This is really a remarkable theorem. Recall that if $S^{-1} A S=D$ a diagonal matrix, then the columns of $S$ are a basis of eigenvectors. Hence this says that when you have a commuting family of non defective matrices, then they have the same eigenvectors. This shows how remarkable it is when a set of matrices commutes.

### 13.11 Fractional Powers

The main result is the following theorem.
Theorem 13.11.1 Let $A$ be a self adjoint and nonnegative $n \times n$ matrix (all eigenvalues are nonnegative) and let $k$ be a positive integer. Then there exists a unique self adjoint nonnegative matrix $B$ such that $B^{k}=A$.

Proof: By Theorem 13.1.6, there exists an orthonormal basis of eigenvectors of $A$, say $\left\{v_{i}\right\}_{i=1}^{n}$ such that $A v_{i}=\lambda_{i} v_{i}$ with each $\lambda_{i}$ real. In particular, there exists a unitary matrix $U$ such that

$$
U^{*} A U=D, A=U D U^{*}
$$

where $D$ has nonnegative diagonal entries. Define $B$ in the obvious way.

$$
B \equiv U D^{1 / k} U^{*}
$$

Then it is clear that $B$ is self adjoint and nonnegative. Also it is clear that $B^{k}=A$. What of uniqueness? Let $p(t)$ be a polynomial whose graph contains the ordered pairs $\left(\lambda_{i}, \lambda_{i}^{1 / k}\right)$ where the $\lambda_{i}$ are the diagonal entries of $D$, the eigenvalues of $A$. Then

$$
p(A)=U P(D) U^{*}=U D^{1 / k} U^{*} \equiv B
$$

Suppose then that $C^{k}=A$ and $C$ is also self adjoint and nonnegative.

$$
C B=C p(A)=C p\left(C^{k}\right)=p\left(C^{k}\right) C=p(A) C=B C
$$

and so $\{B, C\}$ is a commuting family of non defective matrices. By Theorem 13.10.9 this family of matrices is simultaneously diagonalizable. Hence there exists a single $S$ such that

$$
S^{-1} B S=D_{B}, \quad S^{-1} C S=D_{C}
$$

Where $D_{C}, D_{B}$ denote diagonal matrices. Hence, raising to the power $k$, it follows that

$$
A=B^{k}=S D_{B}^{k} S^{-1}, A=C^{k}=S D_{C}^{k} S^{-1}
$$

Hence

$$
S D_{B}^{k} S^{-1}=S D_{C}^{k} S^{-1}
$$

and so $D_{B}^{k}=D_{C}^{k}$. Since the entries of the two diagonal matrices are nonnegative, this implies $D_{B}=D_{C}$ and so $S^{-1} B S=S^{-1} C S$ which shows $B=C$.

A similar result holds for a general finite dimensional inner product space. See Problem 21 in the exercises.

### 13.12 Roots of Positive Linear Maps

In this section, $H$ will be a Hilbert space, real or complex, and $T$ will denote an operator which satisfies the following definition. This will be a more general result than the above because it will hold for infinite dimensional spaces.

Definition 13.12.1 Let $T$ satisfy $T=T^{*}$ (Hermitian) and for all $x \in H$,

$$
\begin{equation*}
(T x, x) \geq 0 \tag{13.21}
\end{equation*}
$$

Such an operator is referred to as positive and self adjoint. It is probably better to refer to such an operator as "nonnegative" since the possibility that $T x=0$ for some $x \neq 0$ is not being excluded. Instead of "self adjoint" you can also use the term, Hermitian. To save on notation, write $T \geq 0$ to mean $T$ is positive, satisfying 13.21 . When we say $A \leq B$ this means $B-A \geq 0$.

A useful theorem about the existence of roots of positive self adjoint operators is presented. This proof is very elementary. I found it in [28] for square roots.

### 13.12.1 The Product of Positive Self Adjoint Operators

With the above definition here is a fundamental result about positive self adjoint operators.

Proposition 13.12.2 Let $S, T$ be positive and self adjoint such that $S T=T S$. Then $S T$ is also positive and self adjoint.

Proof: It is obvious that $S T$ is self adjoint.

$$
(S T x, y)=(T S x, y)=(S x, T y)=(x, S T y)
$$

The only problem is to show that $S T$ is positive. The idea is to write $S=S_{n+1}+\sum_{k=0}^{n} S_{k}^{2}$ where $S_{0}=S$ and the operators $S_{k}$ are self adjoint. This is because if you have $\left(T S^{2} x, x\right)$, where everything commutes, this equals $(S T S x, x)=(T S x, S x) \geq 0$. Thus it will be possible to deal with the terms of the sum which are squared. First assume $(S x, x) \leq(x, x)$ so $S \leq I$.

Define a sequence recursively as follows.

$$
\begin{equation*}
S_{n+1}=S_{n}-S_{n}^{2}, S \equiv S_{0} \tag{13.22}
\end{equation*}
$$

Then $\sum_{k=0}^{n} S_{k}^{2}=\sum_{k=0}^{n}\left(S_{k}-S_{k+1}\right)=S-S_{n+1}, S=S_{n+1}+\sum_{k=0}^{n} S_{k}^{2}$. Now $S_{0} \geq 0$ by assumption. Assume $S_{n} \geq 0$. Then

$$
S_{n+1}=S_{n}-S_{n}^{2}=\left(I-S_{n}\right) S_{n}\left(S_{n}+\left(I-S_{n}\right)\right)=S_{n}^{2}\left(I-S_{n}\right)+\left(I-S_{n}\right)^{2} S_{n}
$$

It follows that $S_{n+1} \geq 0$ because clearly those two terms on the end are positive. Therefore,

$$
(S x, x)=\left(S_{n+1} x, x\right)+\sum_{k=0}^{n}\left(S_{k}^{2} x, x\right) \geq \sum_{k=0}^{n}\left\|S_{k} x\right\|^{2},(S x, x) \geq \sum_{k=0}^{\infty}\left\|S_{k} x\right\|^{2}
$$

also and so $\lim _{k \rightarrow \infty}\left\|S_{k} x\right\|=0$.

$$
T S x=T S_{n+1} x+\sum_{k=0}^{n} T S_{k}^{2} x
$$

$$
(T S x, x)=\left(S_{n+1} x, T x\right)+\sum_{k=0}^{n}\left(T S_{k}^{2} x, x\right)=\left(S_{n+1} x, T x\right)+\sum_{k=0}^{n}\left(T S_{k} x, S_{k} x\right)
$$

so passing to a limit as $n \rightarrow \infty,(T S x, x)=0+\lim \sup _{n \rightarrow \infty} \sum_{S=0}^{n}\left(T S_{k} x, S_{k} x\right) \geq 0$.
Thus if $S \leq I$, the theorem is proved. If $S$ is general, $\frac{S}{\|S\|} \leq I$ and in this case, it follows that $\left(T \frac{S}{\|S\|} x, x\right)=\left(\frac{S}{\|S\|} T x, x\right) \geq 0$ and so $(S T x, x) \geq 0$.

The proposition is like the familiar statement about real numbers which says that when you multiply two nonnegative real numbers the result is a nonnegative real number.

### 13.12.2 Roots of Positive Self Adjoint Operators

With this preparation, it is time to give the theorem about roots.
Theorem 13.12.3 Let $T \in \mathscr{L}(H, H)$ be a positive self adjoint linear operator. Then for $m \in \mathbb{N}$, there exists a unique $m^{\text {th }}$ root $A$ with the following properties. $A^{m}=T, A$ is positive and self adjoint, A commutes with every operator which commutes with $T$.

Proof: Define the following sequence of operators:

$$
A_{0} \equiv 0, A_{n+1} \equiv A_{n}+\frac{1}{m}\left(T-A_{n}^{m}\right)
$$

Say $T \leq I$.
Claim 1: $A_{n} \leq I$.
Proof of Claim 1: True if $n=0$. Assume true for $n$. Then

$$
\begin{aligned}
I-A_{n+1} & =I-A_{n}+\frac{1}{m}\left(A_{n}^{m}-T\right) \geq I-A_{n}+\frac{1}{m}\left(A_{n}^{m}-I\right) \\
& =I-A_{n}-\frac{1}{m}\left(I-A_{n}^{m}\right) \\
& =\left(I-A_{m}\right)-\frac{1}{m}\left(I-A_{m}\right)\left(I+\cdots+A_{n}^{m-1}\right)
\end{aligned}
$$

Now, since $A_{n} \leq I, I+\cdots+A_{n}^{m-1} \leq m I$, it follows that

$$
=\left(I-A_{m}\right)\left(I-\frac{1}{m}\left(I+\cdots+A_{n}^{m-1}\right)\right) \geq\left(I-A_{m}\right)(I-I)=0
$$

so by induction, $A_{n} \leq I$.
Claim 2: $A_{n} \leq A_{n+1}$.
Proof of Claim 2: From the definition of $A_{n}$, this is true if $n=0$ because

$$
A_{1}=T \geq 0=A_{0}
$$

Suppose true for $n$. Then from Claim 1,

$$
\begin{aligned}
A_{n+2}-A_{n+1} & =A_{n+1}+\frac{1}{m}\left(T-A_{n+1}^{m}\right)-\left[A_{n}+\frac{1}{m}\left(T-A_{n}^{m}\right)\right] \\
& =A_{n+1}-A_{n}+\frac{1}{m}\left(A_{n}^{m}-A_{n+1}^{m}\right)
\end{aligned}
$$

$$
\begin{gathered}
=\left(A_{n+1}-A_{n}\right)-\left(A_{n+1}-A_{n}\right) \frac{1}{m}\left(A_{n+1}^{m-1}+A_{n+1}^{m-2} A_{n}+\cdots+A_{n}^{m-1}\right) \\
\geq\left(A_{n+1}-A_{n}\right)-\left(A_{n+1}-A_{n}\right) I=0
\end{gathered}
$$

since each $A_{n}, A_{n+1} \leq I$, so this proves the claim.
Claim 3: $A_{n} \geq 0$
Proof of Claim 3: This is true if $n=0$. Suppose it is true for $n$.

$$
\begin{aligned}
\left(A_{n+1} x, x\right) & =\left(A_{n} x, x\right)+\frac{1}{m}(T x, x)-\frac{1}{m}\left(A_{n}^{m} x, x\right) \\
& \geq\left(A_{n} x, x\right)+\frac{1}{m}(T x, x)-\frac{1}{m}\left(A_{n} x, x\right) \geq 0
\end{aligned}
$$

because by Proposition 13.12.2, $A_{n}-A_{n}^{m}=A_{n}\left(I-A_{n}^{m-1}\right) \geq 0$ because $A_{n} \leq I$.
Thus $\left(A_{n} x, x\right)$ is increasing and bounded above so it converges. Now let $n>k$. Using Proposition 13.12.2 $A_{n} A_{k} \geq A_{k}^{2}$ and also

$$
\left(A_{n}-A_{k}\right)\left(A_{n}+A_{k}\right) \leq 2\left(A_{n}-A_{k}\right) .
$$

Thus the following holds.

$$
\begin{gathered}
\left\|A_{n} x-A_{k} x\right\|^{2}=\left(\left(A_{n}-A_{k}\right)^{2} x, x\right)=\left(A_{n}^{2} x, x\right)-2\left(A_{n} A_{k} x, x\right)+\left(A_{k}^{2} x, x\right) \\
\leq\left(A_{n}^{2} x, x\right)-2\left(A_{k}^{2} x, x\right)+\left(A_{k}^{2} x, x\right)=\left(\left(A_{n}-A_{k}\right)\left(A_{n}+A_{k}\right) x, x\right) \\
\leq 2\left[\left(A_{n} x, x\right)-\left(A_{k} x, x\right)\right]
\end{gathered}
$$

which converges to 0 as $k, n \rightarrow \infty$. Therefore, $\lim _{n \rightarrow \infty} A_{n} x$ exists since $\left\{A_{n} x\right\}$ is a Cauchy sequence. Let this limit be $A x$. Then clearly $A$ is linear. Also, since each $A_{n} \geq 0$ and self adjoint, the Cauchy Schwarz inequality implies

$$
|(A x, y)|=\lim _{n \rightarrow \infty}\left|\left(A_{n} x, y\right)\right| \leq \lim \sup _{n \rightarrow \infty}\left|\left(A_{n} x, x\right)^{1 / 2}\left(A_{n} y, y\right)^{1 / 2}\right| \leq\|x\|\|y\|
$$

so $A$ is also continuous. Now $(A x, x)=\lim _{n \rightarrow \infty}\left(A_{n} x, x\right) \geq 0$ so $A$ is positive and it is clearly also self adjoint since each $A_{n}$ is. From passing to the limit in the definition of $A_{n}$,

$$
A x=A x+\frac{1}{m}\left(T x-A^{m} x\right)
$$

and so $T x=A^{m} x$. This proves the theorem in the case that $T \leq I$. Then if $T>I$, consider $T /\|T\| . T /\|T\| \leq I$ and so there is $B$ such that $B^{m}=T /\|T\|$. Let $A=\|T\|^{1 / m} B$. This proves the existence of the $m^{\text {th }}$ root. It is clear that $A$ commutes with every continuous linear operator that commutes with $T$ because this is true of each of the iterates. In fact, each of these is just a polynomial in $T$. It remains to verify uniqueness.

Next suppose both $A$ and $B$ are $m^{t h}$ roots of $T$ having all the properties stated in the theorem. Then $A B=B A$ because both $A$ and $B$ commute with every operator which commutes with $T$. Then from Proposition 13.12.2,

$$
\begin{equation*}
\left(\left(A^{m-1}+A^{m-2} B+\ldots+B^{m-1}\right)(A-B) x,(A-B) x\right) \geq 0 \tag{13.23}
\end{equation*}
$$

Therefore, $\left(\left(A^{m}-B^{m}\right) x,(A-B) x\right)=(0,(A-B) x)=0$.

Now this means $\left(A^{k} B^{l}(A-B) x,(A-B) x\right)=0$ for all $k+l=m-1$ since the sum of such terms is 0 and each of them is nonnegative. Now this implies

$$
\left(\sqrt{A^{k} B^{l}}(A-B) x, \sqrt{A^{k} B^{l}}(A-B) x\right)=0
$$

and so $\sqrt{A^{k} B^{l}}(A-B) x=0 \Rightarrow A^{k} B^{l}(A-B) x=0, k+l=m-1$. Then, using the binomial theorem,

$$
0=\sum_{j=0}^{m-1}\binom{m-1}{j} A^{m-1-j} B^{j}(-1)^{j}(A-B) x=(A-B)^{m} x
$$

This clearly implies $A=B$. To see this, consider $m=7$.
If $m=7,(A-B)^{7} x=0$ so $(A-B)^{8} x=0$ so $\left((A-B)^{4} x,(A-B)^{4} x\right)=0$ so it follows that $(A-B)^{4} x=0$ so

$$
\left((A-B)^{2} x,(A-B)^{2} x\right)=0
$$

so $\left((A-B)^{2} x, x\right)=0$ so $((A-B) x,(A-B) x)=0$ so $(A-B)=0$.

### 13.13 Spectral Theory of Self Adjoint Operators

First is some notation which may be useful since it will be used in the following presentation.

Definition 13.13.1 Let $X, Y$ be inner product space and let $u \in Y, v \in X$. Then define $u \otimes v \in$ $\mathscr{L}(X, Y)$ as follows.

$$
u \otimes v(w) \equiv(w, v) u
$$

where $(w, v)$ is the inner product in $X$. Then this is clearly linear. That it is continuous follows right away from

$$
|(w, v) u| \leq|u|_{Y}|w|_{X}|v|_{X}
$$

and so

$$
\sup _{|w|_{X} \leq 1}|u \otimes v(w)|_{Y} \leq|u|_{Y}|v|_{X}
$$

Sometimes this is called the tensor product, although much more can be said about the tensor product.

Note how this is similar to the rank one transformations used to consider the dimension of the space $\mathscr{L}(V, W)$ in Theorem 5.1.4. This is also a rank one transformation but here there is no restriction on the dimension of the vector spaces although, as usual, the interest is in finite dimensional spaces. In case you have $\left\{v_{1}, \cdots, v_{n}\right\}$ an orthonormal basis for $V$ and $\left\{u_{1}, \cdots, u_{m}\right\}$ an orthonormal basis for $Y$, (or even just a basis.) the linear transformations $u_{i} \otimes v_{j}$ are the same as those rank one transformations used before in the above theorem and are a basis for $\mathscr{L}(V, W)$. Thus for $A=\sum_{i, j} a_{i j} u_{i} \otimes v_{j}$, the matrix of $A$ with respect to the two bases has its $i j^{t h}$ entry equal to $a_{i j}$. This is stated as the following proposition.

Proposition 13.13.2 Suppose $\left\{v_{1}, \cdots, v_{n}\right\}$ is an orthonormal basis for $V$ and $\left\{u_{1}, \cdots, u_{m}\right\}$ is a basis for $W$. Then if $A \in \mathscr{L}(V, W)$ is given by $A=\sum_{i, j} a_{i j} u_{i} \otimes v_{j}$, then the matrix of $A$ with respect to these two bases is an $m \times n$ matrix whose $i j^{\text {th }}$ entry is $a_{i j}$.

In case $A$ is a Hermitian matrix, and you have an orthonormal basis of eigenvectors and $U$ is the unitary matrix having these eigenvectors as columns, recall that the matrix of $A$ with respect to this basis is diagonal. Recall why this is.

$$
\left(\begin{array}{ccc}
A \boldsymbol{u}_{1} & \cdots & A \boldsymbol{u}_{n}
\end{array}\right)=\left(\begin{array}{lll}
\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{n}
\end{array}\right) D
$$

where $D$ is the diagonal matrix having the eigenvalues down the diagonal. Thus $D=U^{*} A U$ and $A \boldsymbol{u}_{i}=\lambda_{i} \boldsymbol{u}_{i}$. It follows that as a linear transformation,

$$
A=\sum_{i} \lambda_{i} \boldsymbol{u}_{i} \otimes \boldsymbol{u}_{i}
$$

because both give the same answer when acting on elements of the orthonormal basis. This also says that the matrix of $A$ with respect to the given orthonormal basis is just the diagonal matrix having the eigenvalues down the main diagonal.

The following theorem is about the eigenvectors and eigenvalues of a self adjoint operator. Such operators may also be called Hermitian as in the case of matrices. The proof given generalizes to the situation of a compact self adjoint operator on a Hilbert space and leads to many very useful results. It is also a very elementary proof because it does not use the fundamental theorem of algebra and it contains a way, very important in applications, of finding the eigenvalues. This proof depends more directly on the methods of analysis than the preceding material. Recall the following notation.

Definition 13.13.3 Let $X$ be an inner product space and let $S \subseteq X$. Then

$$
S^{\perp} \equiv\{x \in X:(x, s)=0 \text { for all } s \in S\}
$$

Note that even if $S$ is not a subspace, $S^{\perp}$ is.
Theorem 13.13.4 Let $A \in \mathscr{L}(X, X)$ be self adjoint (Hermitian) where $X$ is a finite dimensional inner product space of dimension $n$. Thus $A=A^{*}$. Then there exists an orthonormal basis of eigenvectors, $\left\{v_{j}\right\}_{j=1}^{n}$.

Proof: Consider $(A x, x)$. This quantity is always a real number because

$$
\overline{(A x, x)}=(x, A x)=\left(x, A^{*} x\right)=(A x, x)
$$

thanks to the assumption that $A$ is self adjoint. Now define

$$
\lambda_{1} \equiv \inf \left\{(A x, x):|x|=1, x \in X_{1} \equiv X\right\}
$$

Claim: $\lambda_{1}$ is finite and there exists $v_{1} \in X$ with $\left|v_{1}\right|=1$ such that $\left(A v_{1}, v_{1}\right)=\lambda_{1}$.
Proof of claim: The set of vectors $\{x:|x|=1\}$ is a closed and bounded subset of the finite dimensional space $X$. Therefore, it is compact and so the vector $v_{1}$ exists by Theorem 10.6.3.

I claim that $\lambda_{1}$ is an eigenvalue and $v_{1}$ is an eigenvector. Letting $w \in X_{1} \equiv X$, the function of the real variable, $t$, given by

$$
f(t) \equiv \frac{\left(A\left(v_{1}+t w\right), v_{1}+t w\right)}{\left|v_{1}+t w\right|^{2}}=\frac{\left(A v_{1}, v_{1}\right)+2 t \operatorname{Re}\left(A v_{1}, w\right)+t^{2}(A w, w)}{\left|v_{1}\right|^{2}+2 t \operatorname{Re}\left(v_{1}, w\right)+t^{2}|w|^{2}}
$$

achieves its minimum when $t=0$. Therefore, the derivative of this function evaluated at $t=0$ must equal zero. Using the quotient rule, this implies, since $\left|v_{1}\right|=1$ that

$$
2 \operatorname{Re}\left(A v_{1}, w\right)\left|v_{1}\right|^{2}-2 \operatorname{Re}\left(v_{1}, w\right)\left(A v_{1}, v_{1}\right)=2\left(\operatorname{Re}\left(A v_{1}, w\right)-\operatorname{Re}\left(v_{1}, w\right) \lambda_{1}\right)=0
$$

Thus $\operatorname{Re}\left(A v_{1}-\lambda_{1} v_{1}, w\right)=0$ for all $w \in X$. This implies $A v_{1}=\lambda_{1} v_{1}$ by Proposition 12.1.5.
Continuing with the proof of the theorem, let $X_{2} \equiv\left\{v_{1}\right\}^{\perp}$. This is a closed subspace of $X$ and $A: X_{2} \rightarrow X_{2}$ because for $x \in X_{2}$,

$$
\left(A x, v_{1}\right)=\left(x, A v_{1}\right)=\lambda_{1}\left(x, v_{1}\right)=0
$$

Let

$$
\lambda_{2} \equiv \inf \left\{(A x, x):|x|=1, x \in X_{2}\right\}
$$

As before, there exists $v_{2} \in X_{2}$ such that $A v_{2}=\lambda_{2} v_{2}, \lambda_{1} \leq \lambda_{2}$. Now let $X_{3} \equiv\left\{v_{1}, v_{2}\right\}^{\perp}$ and continue in this way. As long as $k<n$, it will be the case that $\left\{v_{1}, \cdots, v_{k}\right\}^{\perp} \neq\{0\}$. This is because for $k<n$ these vectors cannot be a spanning set and so there exists some $w \notin \operatorname{span}\left(v_{1}, \cdots, v_{k}\right)$. Then letting $z$ be the closest point to $w$ from $\operatorname{span}\left(v_{1}, \cdots, v_{k}\right)$, it follows that $w-z \in\left\{v_{1}, \cdots, v_{k}\right\}^{\perp}$. Thus there is an decreasing sequence of eigenvalues $\left\{\lambda_{k}\right\}_{k=1}^{n}$ and a corresponding sequence of eigenvectors, $\left\{v_{1}, \cdots, v_{n}\right\}$ with this being an orthonormal set.

Contained in the proof of this theorem is the following important corollary.
Corollary 13.13.5 Let $A \in \mathscr{L}(X, X)$ be self adjoint where $X$ is a finite dimensional inner product space. Then all the eigenvalues are real and for $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$ the eigenvalues of $A$, there exists an orthonormal set of vectors $\left\{u_{1}, \cdots, u_{n}\right\}$ for which

$$
A u_{k}=\lambda_{k} u_{k}
$$

Furthermore,

$$
\lambda_{k} \equiv \inf \left\{(A x, x):|x|=1, x \in X_{k}\right\}
$$

where

$$
X_{k} \equiv\left\{u_{1}, \cdots, u_{k-1}\right\}^{\perp}, X_{1} \equiv X
$$

Corollary 13.13.6 Let $A \in \mathscr{L}(X, X)$ be self adjoint (Hermitian) where $X$ is a finite dimensional inner product space. Then the largest eigenvalue of $A$ is given by

$$
\begin{equation*}
\max \{(A \boldsymbol{x}, \boldsymbol{x}):|\boldsymbol{x}|=1\} \tag{13.24}
\end{equation*}
$$

and the minimum eigenvalue of $A$ is given by

$$
\begin{equation*}
\min \{(A x, x):|x|=1\} . \tag{13.25}
\end{equation*}
$$

Proof: The proof of this is just like the proof of Theorem 13.13.4. Simply replace inf with sup and obtain a decreasing list of eigenvalues. This establishes 13.24. The claim 13.25 follows from Theorem 13.13.4.

Another important observation is found in the following corollary.
Corollary 13.13.7 Let $A \in \mathscr{L}(X, X)$ where $A$ is self adjoint. Then $A=\sum_{i} \lambda_{i} v_{i} \otimes v_{i}$ where $A v_{i}=\lambda_{i} v_{i}$ and $\left\{v_{i}\right\}_{i=1}^{n}$ is an orthonormal basis.

Proof : If $v_{k}$ is one of the orthonormal basis vectors, $A v_{k}=\lambda_{k} v_{k}$. Also,

$$
\sum_{i} \lambda_{i} v_{i} \otimes v_{i}\left(v_{k}\right)=\sum_{i} \lambda_{i} v_{i}\left(v_{k}, v_{i}\right)=\sum_{i} \lambda_{i} \delta_{i k} v_{i}=\lambda_{k} v_{k} .
$$

Since the two linear transformations agree on a basis, it follows they must coincide.
By Proposition 13.13 .2 this says the matrix of $A$ with respect to this basis $\left\{v_{i}\right\}_{i=1}^{n}$ is the diagonal matrix having the eigenvalues $\lambda_{1}, \cdots, \lambda_{n}$ down the main diagonal.

The result of Courant and Fischer which follows resembles Corollary 13.13.5 but is more useful because it does not depend on a knowledge of the eigenvectors.

Theorem 13.13.8 Let $A \in \mathscr{L}(X, X)$ be self adjoint where $X$ is a finite dimensional inner product space. Then for $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$ the eigenvalues of $A$, there exist orthonormal vectors $\left\{u_{1}, \cdots, u_{n}\right\}$ for which

$$
A u_{k}=\lambda_{k} u_{k}
$$

## Furthermore,

$$
\begin{equation*}
\lambda_{k} \equiv \max _{w_{1}, \cdots, w_{k-1}}\left\{\min \left\{(A x, x):|x|=1, x \in\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp}\right\}\right\} \tag{13.26}
\end{equation*}
$$

where if $k=1,\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp} \equiv X$.
Proof: From Theorem 13.13.4, there exist eigenvalues and eigenvectors $\left\{u_{1}, \cdots, u_{n}\right\}$ which are orthonormal and $\lambda_{i} \leq \lambda_{i+1}$.

$$
(A x, x)=\sum_{j=1}^{n}\left(A x, u_{j}\right) \overline{\left(x, u_{j}\right)}=\sum_{j=1}^{n} \lambda_{j}\left(x, u_{j}\right)\left(u_{j}, x\right)=\sum_{j=1}^{n} \lambda_{j}\left|\left(x, u_{j}\right)\right|^{2}
$$

Recall that $(z, w)=\sum_{j}\left(z, u_{j}\right) \overline{\left(w, u_{i}\right)}$. Then let $Y=\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp}$

$$
\begin{align*}
& \inf \{(A x, x):|x|=1, x \in Y\}=\inf \left\{\sum_{j=1}^{n} \lambda_{j}\left|\left(x, u_{j}\right)\right|^{2}:|x|=1, x \in Y\right\} \\
\leq & \inf \left\{\sum_{j=1}^{k} \lambda_{j}\left|\left(x, u_{j}\right)\right|^{2}:|x|=1,\left(x, u_{j}\right)=0 \text { for } j>k, \text { and } x \in Y\right\} . \tag{13.27}
\end{align*}
$$

The reason this is so is that the infimum is taken over a smaller set. Therefore, the infimum gets larger. Now 13.27 is no larger than

$$
\inf \left\{\lambda_{k} \sum_{j=1}^{n}\left|\left(x, u_{j}\right)\right|^{2}:|x|=1,\left(x, u_{j}\right)=0 \text { for } j>k, \text { and } x \in Y\right\} \leq \lambda_{k}
$$

because since $\left\{u_{1}, \cdots, u_{n}\right\}$ is an orthonormal basis, $|x|^{2}=\sum_{j=1}^{n}\left|\left(x, u_{j}\right)\right|^{2}$. It follows, since

$$
\left\{w_{1}, \cdots, w_{k-1}\right\}
$$

is arbitrary,

$$
\begin{equation*}
\sup _{w_{1}, \cdots, w_{k-1}}\left\{\inf \left\{(A x, x):|x|=1, x \in\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp}\right\}\right\} \leq \lambda_{k} \tag{13.28}
\end{equation*}
$$

Then from Corollary 13.13.5,

$$
\begin{gathered}
\lambda_{k}=\inf \left\{(A x, x):|x|=1, x \in\left\{u_{1}, \cdots, u_{k-1}\right\}^{\perp}\right\} \leq \\
\sup _{w_{1}, \cdots, w_{k-1}}\left\{\inf \left\{(A x, x):|x|=1, x \in\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp}\right\}\right\} \leq \lambda_{k}
\end{gathered}
$$

Hence these are all equal and this proves the theorem.
The following corollary is immediate.
Corollary 13.13.9 Let $A \in \mathscr{L}(X, X)$ be self adjoint where $X$ is a finite dimensional inner product space. Then for $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$ the eigenvalues of $A$, there exist orthonormal vectors $\left\{u_{1}, \cdots, u_{n}\right\}$ for which

$$
A u_{k}=\lambda_{k} u_{k}
$$

Furthermore,

$$
\begin{equation*}
\lambda_{k} \equiv \max _{w_{1}, \cdots, w_{k-1}}\left\{\min \left\{\frac{(A x, x)}{|x|^{2}}: x \neq 0, x \in\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp}\right\}\right\} \tag{13.29}
\end{equation*}
$$

where if $k=1,\left\{w_{1}, \cdots, w_{k-1}\right\}^{\perp} \equiv X$.
Here is a version of this for which the roles of max and min are reversed.
Corollary 13.13.10 Let $A \in \mathscr{L}(X, X)$ be self adjoint where $X$ is a finite dimensional inner product space. Then for $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$ the eigenvalues of $A$, there exist orthonormal vectors $\left\{u_{1}, \cdots, u_{n}\right\}$ for which

$$
A u_{k}=\lambda_{k} u_{k}
$$

## Furthermore,

$$
\begin{equation*}
\lambda_{k} \equiv \min _{w_{1}, \cdots, w_{n-k}}\left\{\max \left\{\frac{(A x, x)}{|x|^{2}}: x \neq 0, x \in\left\{w_{1}, \cdots, w_{n-k}\right\}^{\perp}\right\}\right\} \tag{13.30}
\end{equation*}
$$

where if $k=n,\left\{w_{1}, \cdots, w_{n-k}\right\}^{\perp} \equiv X$.

### 13.14 Positive and Negative Linear Transformations

The notion of a positive definite or negative definite linear transformation is very important in many applications. In particular it is used in versions of the second derivative test for functions of many variables. Here the main interest is the case of a linear transformation which is an $n \times n$ matrix but the theorem is stated and proved using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

Definition 13.14.1 $A$ self adjoint $A \in \mathscr{L}(X, X)$, is positive definite if whenever $\boldsymbol{x} \neq \mathbf{0}$, $(A \boldsymbol{x}, \boldsymbol{x})>0$ and $A$ is negative definite if for all $\boldsymbol{x} \neq \mathbf{0},(A \boldsymbol{x}, \boldsymbol{x})<0$. $A$ is positive semidefinite or just nonnegative for short if for all $\boldsymbol{x},(A \boldsymbol{x}, \boldsymbol{x}) \geq 0$. $A$ is negative semidefinite or nonpositive for short if for all $\boldsymbol{x},(A \boldsymbol{x}, \boldsymbol{x}) \leq 0$.

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

Lemma 13.14.2 Let $X$ be a finite dimensional inner product space. A self adjoint $A \in$ $\mathscr{L}(X, X)$ is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.

Proof: Suppose first that $A$ is positive definite and let $\lambda$ be an eigenvalue. Then for $\boldsymbol{x}$ an eigenvector corresponding to $\lambda, \lambda(\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{A x}, \boldsymbol{x})>0$. Therefore, $\lambda>0$ as claimed.

Now suppose all the eigenvalues of $A$ are positive. From Theorem 13.13.4 and Corollary 13.13.7, $A=\sum_{i=1}^{n} \lambda_{i} \boldsymbol{u}_{i} \otimes \boldsymbol{u}_{i}$ where the $\lambda_{i}$ are the positive eigenvalues and $\left\{\boldsymbol{u}_{i}\right\}$ are an orthonormal set of eigenvectors. Therefore, letting $\boldsymbol{x} \neq \mathbf{0}$,

$$
\begin{aligned}
(A \boldsymbol{x}, \boldsymbol{x}) & =\left(\left(\sum_{i=1}^{n} \lambda_{i} \boldsymbol{u}_{i} \otimes \boldsymbol{u}_{i}\right) \boldsymbol{x}, \boldsymbol{x}\right)=\left(\sum_{i=1}^{n} \lambda_{i} \boldsymbol{u}_{i}\left(\boldsymbol{x}, \boldsymbol{u}_{i}\right), \boldsymbol{x}\right) \\
& =\left(\sum_{i=1}^{n} \lambda_{i}\left(\boldsymbol{x}, \boldsymbol{u}_{i}\right)\left(\boldsymbol{u}_{i}, \boldsymbol{x}\right)\right)=\sum_{i=1}^{n} \lambda_{i}\left|\left(\boldsymbol{u}_{i}, \boldsymbol{x}\right)\right|^{2}>0
\end{aligned}
$$

because, since $\left\{\boldsymbol{u}_{i}\right\}$ is an orthonormal basis, $|\boldsymbol{x}|^{2}=\sum_{i=1}^{n}\left|\left(\boldsymbol{u}_{i}, \boldsymbol{x}\right)\right|^{2}$.
To establish the claim about negative definite, it suffices to note that $A$ is negative definite if and only if $-A$ is positive definite and the eigenvalues of $A$ are $(-1)$ times the eigenvalues of $-A$. The claims about positive semidefinite and negative semidefinite are obtained similarly.

The next theorem is about a way to recognize whether a self adjoint $n \times n$ complex matrix $A$ is positive or negative definite without having to find the eigenvalues. In order to state this theorem, here is some notation.

Definition 13.14.3 Let $A$ be an $n \times n$ matrix. Denote by $A_{k}$ the $k \times k$ matrix obtained by deleting the $k+1, \cdots, n$ columns and the $k+1, \cdots, n$ rows from $A$. Thus $A_{n}=A$ and $A_{k}$ is the $k \times k$ submatrix of $A$ which occupies the upper left corner of $A$. The determinants of these submatrices are called the principle minors.

The following theorem is proved in [10]. For the sake of simplicity, we state this for real matrices since this is also where the main interest lies.

Theorem 13.14.4 Let $A$ be a self adjoint $n \times n$ matrix. Then $A$ is positive definite if and only if $\operatorname{det}\left(A_{k}\right)>0$ for every $k=1, \cdots, n$.

Proof: This theorem is proved by induction on $n$. It is clearly true if $n=1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\operatorname{det}(A)>0$, it follows that all the eigenvalues are nonzero. Are they all positive? Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\operatorname{det}(A)$. Pick two, $\lambda_{1}$ and $\lambda_{2}$ and let $A \boldsymbol{u}_{i}=\lambda_{i} \boldsymbol{u}_{i}$ where $\boldsymbol{u}_{i} \neq \mathbf{0}$ for $i=1,2$ and $\left(\boldsymbol{u}_{1}, \boldsymbol{u}_{2}\right)=0$. Now if $\boldsymbol{y} \equiv \alpha_{1} \boldsymbol{u}_{1}+\alpha_{2} \boldsymbol{u}_{2}$ is an element of span $\left(\boldsymbol{u}_{1}, \boldsymbol{u}_{2}\right)$, then since these are eigenvalues and $\left(\boldsymbol{u}_{1}, \boldsymbol{u}_{2}\right)_{\mathbb{R}^{n}}=0$, a short computation shows

$$
\left(A\left(\alpha_{1} \boldsymbol{u}_{1}+\alpha_{2} \boldsymbol{u}_{2}\right), \alpha_{1} \boldsymbol{u}_{1}+\alpha_{2} \boldsymbol{u}_{2}\right)=\left|\alpha_{1}\right|^{2} \lambda_{1}\left|\boldsymbol{u}_{1}\right|^{2}+\left|\alpha_{2}\right|^{2} \lambda_{2}\left|\boldsymbol{u}_{2}\right|^{2}<0
$$

Now letting $\boldsymbol{x} \in \mathbb{R}^{n-1}, \boldsymbol{x} \neq \mathbf{0}$, the induction hypothesis implies

$$
\left(\boldsymbol{x}^{T}, 0\right) A\binom{\boldsymbol{x}}{0}=\boldsymbol{x}^{T} A_{n-1} \boldsymbol{x}=\left(A_{n-1} \boldsymbol{x}, \boldsymbol{x}\right)>0
$$

The dimension of $\left\{\boldsymbol{z} \in \mathbb{R}^{n}: z_{n}=0\right\}$ is $n-1$ and the dimension of $\operatorname{span}\left(\boldsymbol{u}_{1}, \boldsymbol{u}_{2}\right)=2$ and so there must be some nonzero $\boldsymbol{x} \in \mathbb{R}^{n}$ which is in both of these subspaces of $\mathbb{R}^{n}$. However, the first computation would require that $(A \boldsymbol{x}, \boldsymbol{x})<0$ while the second would require that $(A x, x)>0$. This contradiction shows that all the eigenvalues must be positive. This proves the if part of the theorem.

To show the converse, note that, as above, $(A \boldsymbol{x}, \boldsymbol{x})=\boldsymbol{x}^{T} A \boldsymbol{x}$. Suppose that $A$ is positive definite. Then this is equivalent to having

$$
\boldsymbol{x}^{T} A \boldsymbol{x} \geq \delta\|\boldsymbol{x}\|^{2}
$$

Note that for $\boldsymbol{x} \in \mathbb{R}^{k}$,

$$
\left(\begin{array}{ll}
\boldsymbol{x}^{T} & \mathbf{0}
\end{array}\right) A\binom{\boldsymbol{x}}{\mathbf{0}}=\boldsymbol{x}^{T} A_{k} \boldsymbol{x} \geq \boldsymbol{\delta}\|\boldsymbol{x}\|^{2}
$$

From Lemma 13.14.2, this implies that all the eigenvalues of $A_{k}$ are positive. Hence from Lemma 13.14.2, it follows that $\operatorname{det}\left(A_{k}\right)>0$, being the product of its eigenvalues.

Corollary 13.14.5 Let $A$ be a self adjoint $n \times n$ matrix. Then $A$ is negative definite if and only if $\operatorname{det}\left(A_{k}\right)(-1)^{k}>0$ for every $k=1, \cdots, n$.

Proof: This is immediate from the above theorem by noting that, as in the proof of Lemma 13.14.2, $A$ is negative definite if and only if $-A$ is positive definite. Therefore, $\operatorname{det}\left(-A_{k}\right)>0$ for all $k=1, \cdots, n$, is equivalent to having $A$ negative definite. However, $\operatorname{det}\left(-A_{k}\right)=(-1)^{k} \operatorname{det}\left(A_{k}\right)$.

### 13.15 The Singular Value Decomposition

In this section, $A$ will be an $m \times n$ matrix. To begin with, here is a simple lemma observed earlier.

Lemma 13.15.1 Let $A$ be an $m \times n$ matrix. Then $A^{*} A$ is self adjoint and all its eigenvalues are nonnegative.

Proof: It is obvious that $A^{*} A$ is self adjoint. Suppose $A^{*} A \boldsymbol{x}=\lambda \boldsymbol{x}$. Then $\lambda|\boldsymbol{x}|^{2}=$ $(\lambda \boldsymbol{x}, \boldsymbol{x})=\left(A^{*} A x, x\right)=(A x, A x) \geq 0$.

Definition 13.15.2 Let $A$ be an $m \times n$ matrix. The singular values of $A$ are the square roots of the positive eigenvalues of $A^{*} A$.

With this definition and lemma here is the main theorem on the singular value decomposition. In all that follows, I will write the following partitioned matrix

$$
\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

where $\sigma$ denotes an $r \times r$ diagonal matrix of the form

$$
\left(\begin{array}{ccc}
\sigma_{1} & & 0 \\
& \ddots & \\
0 & & \sigma_{k}
\end{array}\right)
$$

and the bottom row of zero matrices in the partitioned matrix, as well as the right columns of zero matrices are each of the right size so that the resulting matrix is $m \times n$. Either could vanish completely. However, I will write it in the above form. It is easy to make the necessary adjustments in the other two cases.

Theorem 13.15.3 Let $A$ be an $m \times n$ matrix. Then there exist unitary matrices, $U$ and $V$ of the appropriate size such that

$$
U^{*} A V=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

where $\sigma$ is of the form

$$
\sigma=\left(\begin{array}{ccc}
\sigma_{1} & & 0 \\
& \ddots & \\
0 & & \sigma_{k}
\end{array}\right)
$$

for the $\sigma_{i}$ the singular values of $A$, arranged in order of decreasing size.
Proof: By the above lemma and Theorem 13.13.4 there exists an orthonormal basis, $\left\{\boldsymbol{v}_{i}\right\}_{i=1}^{n}$ for $\mathbb{F}^{n}$ such that $A^{*} A \boldsymbol{v}_{i}=\sigma_{i}^{2} \boldsymbol{v}_{i}$ where $\sigma_{i}^{2}>0$ for $i=1, \cdots, k, \sigma_{i}>0$, and equals zero if $i>k$. Let the eigenvalues $\sigma_{i}^{2}$ be arranged in decreasing order. It is desired to have

$$
A V=U\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

and so if $U=\left(\begin{array}{lll}\boldsymbol{u}_{1} & \cdots & \boldsymbol{u}_{m}\end{array}\right)$, one needs to have for $j \leq k, \sigma_{j} \boldsymbol{u}_{j}=A \boldsymbol{v}_{j}$. Thus let

$$
\boldsymbol{u}_{j} \equiv \sigma_{j}^{-1} A \boldsymbol{v}_{j}, j \leq k
$$

Then for $i, j \leq k$,

$$
\begin{aligned}
\left(\boldsymbol{u}_{i}, \boldsymbol{u}_{j}\right) & =\sigma_{j}^{-1} \sigma_{i}^{-1}\left(A \boldsymbol{v}_{i}, A \boldsymbol{v}_{j}\right)=\sigma_{j}^{-1} \sigma_{i}^{-1}\left(A^{*} A \boldsymbol{v}_{i}, \boldsymbol{v}_{j}\right) \\
& =\sigma_{j}^{-1} \sigma_{i}^{-1} \sigma_{i}^{2}\left(\boldsymbol{v}_{i}, \boldsymbol{v}_{j}\right)=\delta_{i j}
\end{aligned}
$$

Now extend to an orthonormal basis of $\mathbb{F}^{m},\left\{\boldsymbol{u}_{1}, \cdots, \boldsymbol{u}_{k}, \boldsymbol{u}_{k+1}, \cdots, \boldsymbol{u}_{m}\right\}$. If $i>k$,

$$
\left(A \boldsymbol{v}_{i}, A \boldsymbol{v}_{i}\right)=\left(A^{*} A \boldsymbol{v}_{i}, \boldsymbol{v}_{i}\right)=0\left(\boldsymbol{v}_{i}, \boldsymbol{v}_{i}\right)=0
$$

so $A \boldsymbol{v}_{i}=\mathbf{0}$. Then for $\sigma$ given as above in the statement of the theorem, it follows that

$$
A V=U\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right), U^{*} A V=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 13.15.4 Let $A$ be an $m \times n$ matrix. Then the rank of both $A$ and $A^{*}$ equals the number of singular values.

Proof: Since $V$ and $U$ are unitary, they are each one to one and onto and so it follows that

$$
\operatorname{rank}(A)=\operatorname{rank}\left(U^{*} A V\right)=\operatorname{rank}\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)=\text { number of singular values. }
$$

Also since $U, V$ are unitary,

$$
\begin{aligned}
& \operatorname{rank}\left(A^{*}\right)=\operatorname{rank}\left(V^{*} A^{*} U\right)=\operatorname{rank}\left(\left(U^{*} A V\right)^{*}\right) \\
= & \operatorname{rank}\left(\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)^{*}\right)=\text { number of singular values. }
\end{aligned}
$$

### 13.16 Approximation In The Frobenius Norm

The Frobenius norm is one of many norms for a matrix. It is arguably the most obvious of all norms. Here is its definition.

Definition 13.16.1 Let $A$ be a complex $m \times n$ matrix. Then

$$
\|A\|_{F} \equiv\left(\operatorname{trace}\left(A A^{*}\right)\right)^{1 / 2}
$$

Also this norm comes from the inner product

$$
(A, B)_{F} \equiv \operatorname{trace}\left(A B^{*}\right)
$$

Thus $\|A\|_{F}^{2}$ is easily seen to equal $\sum_{i j}\left|a_{i j}\right|^{2}$ so essentially, it treats the matrix as a vector in $\mathbb{F}^{m \times n}$.

Lemma 13.16.2 Let $A$ be an $m \times n$ complex matrix with singular matrix

$$
\Sigma=\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

with $\sigma$ as defined above, $U^{*} A V=\Sigma$. Then

$$
\begin{equation*}
\|\Sigma\|_{F}^{2}=\|A\|_{F}^{2} \tag{13.31}
\end{equation*}
$$

and the following hold for the Frobenius norm. If $U, V$ are unitary and of the right size,

$$
\begin{equation*}
\|U A\|_{F}=\|A\|_{F},\|U A V\|_{F}=\|A\|_{F} . \tag{13.32}
\end{equation*}
$$

Proof: From the definition and letting $U, V$ be unitary and of the right size,

$$
\|U A\|_{F}^{2} \equiv \operatorname{trace}\left(U A A^{*} U^{*}\right)=\operatorname{trace}\left(U^{*} U A A^{*}\right)=\operatorname{trace}\left(A A^{*}\right)=\|A\|_{F}^{2}
$$

Also,

$$
\|A V\|_{F}^{2} \equiv \operatorname{trace}\left(A V V^{*} A^{*}\right)=\operatorname{trace}\left(A A^{*}\right)=\|A\|_{F}^{2}
$$

It follows

$$
\|\Sigma\|_{F}^{2}=\left\|U^{*} A V\right\|_{F}^{2}=\|A V\|_{F}^{2}=\|A\|_{F}^{2}
$$

Of course, this shows that

$$
\|A\|_{F}^{2}=\sum_{i} \sigma_{i}^{2}
$$

the sum of the squares of the singular values of $A$.
Why is the singular value decomposition important? It implies

$$
A=U\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right) V^{*}
$$

where $\sigma$ is the diagonal matrix having the singular values down the diagonal. Now sometimes $A$ is a huge matrix, $1000 \times 2000$ or something like that. This happens in applications to situations where the entries of $A$ describe a picture. What also happens is that most of the singular values are very small. What if you deleted those which were very small, say for all $i \geq l$ and got a new matrix

$$
A^{\prime} \equiv U\left(\begin{array}{cc}
\sigma^{\prime} & 0 \\
0 & 0
\end{array}\right) V^{*} ?
$$

Then the entries of $A^{\prime}$ would end up being close to the entries of $A$ but there is much less information to keep track of. This turns out to be very useful. More precisely, letting

$$
\begin{gathered}
\sigma=\left(\begin{array}{ccc}
\sigma_{1} & & 0 \\
& \ddots & \\
0 & & \sigma_{r}
\end{array}\right), U^{*} A V=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right), \\
\left\|A-A^{\prime}\right\|_{F}^{2}=\left\|U\left(\begin{array}{cc}
\sigma-\sigma^{\prime} & 0 \\
0 & 0
\end{array}\right) V^{*}\right\|_{F}^{2}=\sum_{k=l+1}^{r} \sigma_{k}^{2}
\end{gathered}
$$

Thus $A$ is approximated by $A^{\prime}$ where $A^{\prime}$ has rank $l<r$. In fact, it is also true that out of all matrices of rank $l$, this $A^{\prime}$ is the one which is closest to $A$ in the Frobenius norm. Here is roughly why this is so. Suppose $\tilde{B}$ approximates $A=\left(\begin{array}{cc}\sigma_{r \times r} & 0 \\ 0 & 0\end{array}\right)$ as well as possible out of all matrices $\tilde{B}$ having rank no more than $l<r$ the size of the matrix $\sigma_{r \times r}$.

Suppose the rank of $\tilde{B}$ is $l$. Then obviously no column $x_{j}$ of $\tilde{B}$ in a basis for the column space can have $j>r$ since if so, the approximation of $A$ could be improved by simply making this column into a zero column. Therefore there are $\binom{r}{l}$ choices for columns for a basis for the column space of $\tilde{B}$.

Let $\boldsymbol{x}$ be a column in the basis for the column space of $\tilde{B}$ and let it be column $j$ in the matrix $\tilde{B}$. Denote the diagonal entry by $x_{j}=\sigma_{j}+h$. Then the error incurred due to approximating with this column is

$$
h^{2}+\sum_{i \neq j} x_{i}^{2}
$$

One obviously minimizes this error by letting $h=0=x_{i}$ for all $i \neq j$. That is, the column should have all zeroes with $\sigma_{j}$ in the diagonal position. As to any columns of $\tilde{B}$ which are not pivot columns, such a column is a linear combination of these basis columns which have exactly one entry, in the diagonal position. These non pivot columns must have a 0 in the diagonal position since if not, the rank of the matrix would be more than $l$. Then the off diagonal entries should equal zero to make the approximation as good as possible. Thus the non basis columns are columns consisting of zeros and $\tilde{B}$ is a diagonal matrix with $l$ nonzero diagonal entries selected from the first $r$ columns of $A$. It only remains to observe that, since the singular values decrease in size from upper left to lower right in $A$, to minimize the error, one should pick the first $l$ columns for the basis for $\tilde{B}$ in order to use the sum of the squares of the smallest possible singular values in the error. That is, you would replace $\sigma_{r \times r}$ with the upper left $l \times l$ corner of $\sigma_{r \times r}$.

$$
A=\left(\begin{array}{cc}
\sigma_{r \times r} & 0 \\
0 & 0
\end{array}\right), \Rightarrow \tilde{B}=\left(\begin{array}{cc}
\sigma_{l \times l} & 0 \\
0 & 0
\end{array}\right)
$$

For example, consider

$$
\left(\begin{array}{llll}
3 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right)
$$

The best rank 2 approximation is

$$
\left(\begin{array}{llll}
3 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}\right)
$$

Now suppose $A$ is an $m \times n$ matrix. Let $U, V$ be unitary and of the right size such that

$$
U^{*} A V=\left(\begin{array}{cc}
\sigma_{r \times r} & 0 \\
0 & 0
\end{array}\right)
$$

Then suppose $B$ approximates $A$ as well as possible in the Frobenius norm, $B$ having rank $l<r$. Then you would want

$$
\|A-B\|=\left\|U^{*} A V-U^{*} B V\right\|=\left\|\left(\begin{array}{cc}
\sigma_{r \times r} & 0 \\
0 & 0
\end{array}\right)-U^{*} B V\right\|
$$

to be as small as possible. Therefore, from the above discussion, you should have

$$
\tilde{B} \equiv U^{*} B V=\left(\begin{array}{cc}
\sigma_{l \times l} & 0 \\
0 & 0
\end{array}\right), B=U\left(\begin{array}{cc}
\sigma_{l \times l} & 0 \\
0 & 0
\end{array}\right) V^{*}
$$

whereas

$$
A=U\left(\begin{array}{cc}
\sigma_{r \times r} & 0 \\
0 & 0
\end{array}\right) V^{*}
$$

### 13.17 Least Squares And Singular Value Decomposition

The singular value decomposition also has a very interesting connection to the problem of least squares solutions. Recall that it was desired to find $\boldsymbol{x}$ such that $|\boldsymbol{A x}-\boldsymbol{y}|$ is as small as possible. Lemma 12.4.1 shows that there is a solution to this problem which can be found by solving the system $A^{*} A \boldsymbol{x}=A^{*} \boldsymbol{y}$. Each $\boldsymbol{x}$ which solves this system solves the minimization problem as was shown in the lemma just mentioned. Now consider this equation for the solutions of the minimization problem in terms of the singular value decomposition.

$$
\overbrace{V\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right) U^{*} U\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) V^{*} \boldsymbol{x}}^{A}=\overbrace{V\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y}}^{A^{*}} .
$$

Therefore, this yields the following upon using block multiplication and multiplying on the left by $V^{*}$.

$$
\left(\begin{array}{cc}
\sigma^{2} & 0  \tag{13.33}\\
0 & 0
\end{array}\right) V^{*} \boldsymbol{x}=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y}
$$

One solution to this equation which is very easy to spot is

$$
\boldsymbol{x}=V\left(\begin{array}{cc}
\sigma^{-1} & 0  \tag{13.34}\\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y}
$$

### 13.18 The Moore Penrose Inverse

The particular solution of the least squares problem given in 13.34 is important enough that it motivates the following definition.

Definition 13.18.1 Let A be an $m \times n$ matrix. Then the Moore Penrose inverse of $A$, denoted by $A^{+}$is defined as

$$
A^{+} \equiv V\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*}
$$

Here

$$
U^{*} A V=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

as above.
Thus $A^{+} \boldsymbol{y}$ is a solution to find $\boldsymbol{x}$ which minimizes $|A \boldsymbol{x}-\boldsymbol{y}|$. In fact, one can say more about this. In the following picture $M_{\boldsymbol{y}}$ denotes the set of least squares solutions $\boldsymbol{x}$ such that $A^{*} A \boldsymbol{x}=A^{*} \boldsymbol{y}$.


Then $A^{+}(\boldsymbol{y})$ is as given in the picture.
Proposition 13.18.2 $A^{+} \boldsymbol{y}$ is the solution to the problem of minimizing $|A \boldsymbol{x}-\boldsymbol{y}|$ for all $\boldsymbol{x}$ which has smallest norm. Thus

$$
\left|A A^{+} \boldsymbol{y}-\boldsymbol{y}\right| \leq|A \boldsymbol{x}-\boldsymbol{y}| \text { for all } \boldsymbol{x}
$$

and if $\boldsymbol{x}_{1}$ satisfies $\left|A \boldsymbol{x}_{1}-\boldsymbol{y}\right| \leq|A \boldsymbol{x}-\boldsymbol{y}|$ for all $\boldsymbol{x}$, then $\left|A^{+} \boldsymbol{y}\right| \leq\left|\boldsymbol{x}_{1}\right|$.
Proof: Consider $\boldsymbol{x}$ satisfying 13.33, equivalently $A^{*} A \boldsymbol{x}=A^{*} \boldsymbol{y}$,

$$
\left(\begin{array}{cc}
\sigma^{2} & 0 \\
0 & 0
\end{array}\right) V^{*} \boldsymbol{x}=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y}
$$

which has smallest norm. This is equivalent to making $\left|V^{*} \boldsymbol{x}\right|$ as small as possible because $V^{*}$ is unitary and so it preserves norms. For $\boldsymbol{z}$ a vector, denote by $(\boldsymbol{z})_{k}$ the vector in $\mathbb{F}^{k}$ which consists of the first $k$ entries of $\boldsymbol{z}$. Then if $\boldsymbol{x}$ is a solution to 13.33

$$
\binom{\sigma^{2}\left(V^{*} \boldsymbol{x}\right)_{k}}{\mathbf{0}}=\binom{\boldsymbol{\sigma}\left(U^{*} \boldsymbol{y}\right)_{k}}{\mathbf{0}}
$$

and so $\left(V^{*} \boldsymbol{x}\right)_{k}=\sigma^{-1}\left(U^{*} \boldsymbol{y}\right)_{k}$. Thus the first $k$ entries of $V^{*} \boldsymbol{x}$ are determined. In order to make $\left|V^{*} \boldsymbol{x}\right|$ as small as possible, the remaining $n-k$ entries should equal zero. Therefore,

$$
V^{*} \boldsymbol{x}=\binom{\left(V^{*} \boldsymbol{x}\right)_{k}}{0}=\binom{\sigma^{-1}\left(U^{*} \boldsymbol{y}\right)_{k}}{0}=\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y}
$$

and so

$$
\boldsymbol{x}=V\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*} \boldsymbol{y} \equiv A^{+} \boldsymbol{y}
$$

Lemma 13.18.3 The matrix $A^{+}$satisfies the following conditions.

$$
\begin{equation*}
A A^{+} A=A, A^{+} A A^{+}=A^{+}, A^{+} A \text { and } A A^{+} \text {are Hermitian. } \tag{13.35}
\end{equation*}
$$

Proof: This is routine. Recall

$$
A=U\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right) V^{*}
$$

and

$$
A^{+}=V\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*}
$$

so you just plug in and verify it works.
A much more interesting observation is that $A^{+}$is characterized as being the unique matrix which satisfies 13.35 . This is the content of the following Theorem. The conditions are sometimes called the Penrose conditions.

Theorem 13.18.4 Let $A$ be an $m \times n$ matrix. Then a matrix $A_{0}$, is the Moore Penrose inverse of $A$ if and only if $A_{0}$ satisfies

$$
\begin{equation*}
A A_{0} A=A, A_{0} A A_{0}=A_{0}, A_{0} A \text { and } A A_{0} \text { are Hermitian } \tag{13.36}
\end{equation*}
$$

Proof: From the above lemma, the Moore Penrose inverse satisfies 13.36. Suppose then that $A_{0}$ satisfies 13.36. It is necessary to verify that $A_{0}=A^{+}$. Recall that from the singular value decomposition, there exist unitary matrices, $U$ and $V$ such that

$$
U^{*} A V=\Sigma \equiv\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right), A=U \Sigma V^{*}
$$

Recall that

$$
A^{+}=V\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*}
$$

Let

$$
A_{0}=V\left(\begin{array}{ll}
P & Q  \tag{13.37}\\
R & S
\end{array}\right) U^{*}
$$

where $P$ is $r \times r$, the same size as the diagonal matrix composed of the singular values on the main diagonal.

Next use the first equation of 13.36 to write

$$
\overbrace{U \Sigma V^{*} V}^{A} \overbrace{\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}}^{\overbrace{U \Sigma V^{*}}} \overbrace{U \Delta V^{*}}^{A} .
$$

Then multiplying both sides on the left by $U^{*}$ and on the right by $V$,

$$
\left(\begin{array}{ll}
\sigma & 0  \tag{13.38}\\
0 & 0
\end{array}\right)\left(\begin{array}{ll}
P & Q \\
R & S
\end{array}\right)\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)=\left(\begin{array}{cc}
\sigma P \sigma & 0 \\
0 & 0
\end{array}\right)=\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)
$$

Therefore, $P=\sigma^{-1}$. From the requirement that $A A_{0}$ is Hermitian,

$$
\overbrace{U \Sigma V^{*} V}^{A} \overbrace{\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}}^{A_{0}}=U\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right)\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}
$$

must be Hermitian. Therefore, it is necessary that

$$
\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)\left(\begin{array}{ll}
P & Q \\
R & S
\end{array}\right)=\left(\begin{array}{cc}
\sigma P & \sigma Q \\
0 & 0
\end{array}\right)=\left(\begin{array}{cc}
I & \sigma Q \\
0 & 0
\end{array}\right)
$$

is Hermitian. Then

$$
\left(\begin{array}{cc}
I & \sigma Q \\
0 & 0
\end{array}\right)=\left(\begin{array}{cc}
I & 0 \\
Q^{*} \sigma & 0
\end{array}\right)
$$

and so $Q=0$.
Next,

$$
\overbrace{V\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*} \overbrace{U \Sigma V^{*}}^{A}}^{A_{0}}=V\left(\begin{array}{cc}
P \sigma & 0 \\
R \sigma & 0
\end{array}\right) V^{*}=V\left(\begin{array}{cc}
I & 0 \\
R \sigma & 0
\end{array}\right) V^{*}
$$

is Hermitian. Therefore, also

$$
\left(\begin{array}{cc}
I & 0 \\
R \sigma & 0
\end{array}\right)
$$

is Hermitian. Thus $R=0$ because

$$
\left(\begin{array}{cc}
I & 0 \\
R \sigma & 0
\end{array}\right)^{*}=\left(\begin{array}{cc}
I & \sigma^{*} R^{*} \\
0 & 0
\end{array}\right)
$$

which requires $R \sigma=0$. Now multiply on right by $\sigma^{-1}$ to find that $R=0$.
Use 13.37 and the second equation of 13.36 to write

$$
\overbrace{V\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}}^{A_{U \Sigma V^{*} V}} \overbrace{\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}}^{A}=\overbrace{V\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right) U^{*}}^{A_{0}} .
$$

which implies

$$
\left(\begin{array}{ll}
P & Q \\
R & S
\end{array}\right)\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)\left(\begin{array}{ll}
P & Q \\
R & S
\end{array}\right)=\left(\begin{array}{ll}
P & Q \\
R & S
\end{array}\right) .
$$

This yields from the above in which is was shown that $R, Q$ are both 0

$$
\begin{align*}
\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & S
\end{array}\right)\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right)\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & S
\end{array}\right) & =\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right)  \tag{13.39}\\
& =\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & S
\end{array}\right) \tag{13.40}
\end{align*}
$$

Therefore, $S=0$ also and so

$$
V^{*} A_{0} U \equiv\left(\begin{array}{cc}
P & Q \\
R & S
\end{array}\right)=\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right)
$$

which says

$$
A_{0}=V\left(\begin{array}{cc}
\sigma^{-1} & 0 \\
0 & 0
\end{array}\right) U^{*} \equiv A^{+}
$$

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in 13.36. It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 3.

### 13.19 The Spectral Norm And The Operator Norm

Another way of describing a norm for an $n \times n$ matrix is as follows.
Definition 13.19.1 Let $A$ be an $m \times n$ matrix. Define the spectral norm of $A$, written as $\|A\|_{2}$ to be

$$
\max \left\{\lambda^{1 / 2}: \lambda \text { is an eigenvalue of } A^{*} A\right\}
$$

That is, the largest singular value of $A$. (Note the eigenvalues of $A^{*} A$ are all positive because if $A^{*} A \boldsymbol{x}=\lambda \boldsymbol{x}$, then

$$
\left.\lambda|x|^{2}=\lambda(x, x)=\left(A^{*} A x, x\right)=(A x, A x) \geq 0 .\right)
$$

Actually, this is nothing new. It turns out that $\|\cdot\|_{2}$ is nothing more than the operator norm for $A$ taken with respect to the usual Euclidean norm,

$$
|x|=\left(\sum_{k=1}^{n}\left|x_{k}\right|^{2}\right)^{1 / 2}
$$

Proposition 13.19.2 The following holds.

$$
\|A\|_{2}=\sup \{|A \boldsymbol{x}|:|\boldsymbol{x}|=1\} \equiv\|A\| .
$$

Proof: Note that $A^{*} A$ is Hermitian and so by Corollary 13.13.6,

$$
\begin{aligned}
\|A\|_{2} & =\max \left\{\left(A^{*} A \boldsymbol{x}, \boldsymbol{x}\right)^{1 / 2}:|\boldsymbol{x}|=1\right\}=\max \left\{(A \boldsymbol{x}, A \boldsymbol{x})^{1 / 2}:|\boldsymbol{x}|=1\right\} \\
& =\max \{|A \boldsymbol{x}|:|\boldsymbol{x}|=1\}=\|A\| .
\end{aligned}
$$

Here is another proof of this proposition. Recall there are unitary matrices of the right size $U, V$ such that $A=U\left(\begin{array}{cc}\sigma & 0 \\ 0 & 0\end{array}\right) V^{*}$ where the matrix on the inside is as described in the section on the singular value decomposition. Then since unitary matrices preserve norms,

$$
\begin{aligned}
\|A\| & =\sup _{|\boldsymbol{x}| \leq 1}\left|U\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) V^{*} \boldsymbol{x}\right|=\sup _{\left|V^{*} \boldsymbol{x}\right| \leq 1}\left|U\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) V^{*} \boldsymbol{x}\right| \\
& =\sup _{|\boldsymbol{y}| \leq 1}\left|U\left(\begin{array}{ll}
\sigma & 0 \\
0 & 0
\end{array}\right) \boldsymbol{y}\right|=\sup _{|\boldsymbol{y}| \leq 1}\left|\left(\begin{array}{cc}
\sigma & 0 \\
0 & 0
\end{array}\right) \boldsymbol{y}\right|=\sigma_{1} \equiv\|A\|_{2}
\end{aligned}
$$

This completes the alternate proof.
From now on, $\|A\|_{2}$ will mean either the operator norm of $A$ taken with respect to the usual Euclidean norm or the largest singular value of $A$, whichever is most convenient.

### 13.20 The Positive Part Of A Hermitian Matrix

Actually, some of the most interesting functions of matrices do not come as a power series expanded about 0 which was presented earlier. One example of this situation has already been encountered in the proof of the right polar decomposition with the square root of an Hermitian transformation which had all nonnegative eigenvalues. Another example is that of taking the positive part of an Hermitian matrix. This is important in some physical models where something may depend on the positive part of the strain which is a symmetric real matrix. Obviously there is no way to consider this as a power series expanded about 0 because the function $f(r)=r^{+} \equiv \frac{|r|+r}{2}$ is not even differentiable at 0 . Therefore, a totally different approach must be considered. Actually, the only use of this I know of involves real symmetric matrices but the general case is considered here. First the notion of a positive part is defined.

Definition 13.20.1 Let $A$ be an Hermitian matrix. Thus it suffices to consider $A$ as an element of $\mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{n}\right)$ according to the usual notion of matrix multiplication. Then there is a unitary matrix $U$ such that

$$
A=U D U^{*}
$$

where $D$ is a diagonal matrix. Then

$$
A_{+} \equiv U D_{+} U^{*}
$$

where $D_{+}$is obtained from $D$ by replacing each diagonal entry with its positive part.
This gives us a nice definition of what is meant but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix $A$. The following addresses this question. Then

$$
A \boldsymbol{x}=\sum_{i} \lambda_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*} \boldsymbol{x}
$$

You can see this is the case by checking on the $\boldsymbol{u}_{j}$. $A$ agrees with $\sum_{i} \lambda_{i} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}$ on a basis and so they give the same result for all vectors. Thus similarly

$$
A_{+}=\sum_{i} \lambda_{i}^{+} \boldsymbol{u}_{i} \boldsymbol{u}_{i}^{*}
$$

Theorem 13.20.2 If $A, B$ be Hermitian matrices, then for $|\cdot|$ the Frobenius norm,

$$
\left|A_{+}-B_{+}\right| \leq|A-B|
$$

Proof: Let $A=\sum_{i} \boldsymbol{\lambda}_{i} \boldsymbol{v}_{i} \boldsymbol{v}_{i}^{*}$ and let $B=\sum_{j} \mu_{j} \boldsymbol{w}_{j} \boldsymbol{w}_{j}^{*}$ where $\left\{\boldsymbol{v}_{i}\right\}$ and $\left\{\boldsymbol{w}_{j}\right\}$ are orthonormal bases of eigenvectors. Now $A_{+}, B_{+}$are Hermitian and so their difference is also. It follows that

$$
\begin{gathered}
\left|A_{+}-B_{+}\right|^{2}=\operatorname{trace}\left(\sum_{i} \lambda_{i}^{+} \boldsymbol{v}_{i} \boldsymbol{v}_{i}^{*}-\sum_{j} \mu_{j}^{+} \boldsymbol{w}_{j} \boldsymbol{w}_{j}^{*}\right)^{2}= \\
\operatorname{trace}\left[\sum_{i}\left(\lambda_{i}^{+}\right)^{2} \boldsymbol{v}_{i} \boldsymbol{v}_{i}^{*}+\sum_{j}\left(\mu_{j}^{+}\right)^{2} \boldsymbol{w}_{j} \boldsymbol{w}_{j}^{*}\right.
\end{gathered}
$$

$$
\left.-\sum_{i, j} \lambda_{i}^{+} \mu_{j}^{+}\left(\boldsymbol{w}_{j}, \boldsymbol{v}_{i}\right) \boldsymbol{v}_{i} \boldsymbol{w}_{j}^{*}-\sum_{i, j} \lambda_{i}^{+} \mu_{j}^{+}\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right) \boldsymbol{w}_{j} \boldsymbol{v}_{i}^{*}\right]
$$

The trace satisfies trace $(A B)=\operatorname{trace}(B A)$ when both products make sense. Therefore,

$$
\operatorname{trace}\left(\boldsymbol{v}_{i} \boldsymbol{w}_{j}^{*}\right)=\operatorname{trace}\left(\boldsymbol{w}_{j}^{*} \boldsymbol{v}_{i}\right)=\boldsymbol{w}_{j}^{*} \boldsymbol{v}_{i} \equiv\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)
$$

a similar formula for $\boldsymbol{w}_{j} \boldsymbol{v}_{i}^{*}$. Therefore, this equals

$$
\begin{equation*}
=\sum_{i}\left(\lambda_{i}^{+}\right)^{2}+\sum_{j}\left(\mu_{j}^{+}\right)^{2}-2 \sum_{i, j} \lambda_{i}^{+} \mu_{j}^{+}\left|\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)\right|^{2} . \tag{13.41}
\end{equation*}
$$

Since these are orthonormal bases,

$$
\sum_{i}\left|\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)\right|^{2}=1=\sum_{j}\left|\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)\right|^{2}
$$

and so 13.41 equals

$$
=\sum_{i} \sum_{j}\left(\left(\lambda_{i}^{+}\right)^{2}+\left(\mu_{j}^{+}\right)^{2}-2 \lambda_{i}^{+} \mu_{j}^{+}\right)\left|\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)\right|^{2}
$$

Similarly,

$$
|A-B|^{2}=\sum_{i} \sum_{j}\left(\left(\lambda_{i}\right)^{2}+\left(\mu_{j}\right)^{2}-2 \lambda_{i} \mu_{j}\right)\left|\left(\boldsymbol{v}_{i}, \boldsymbol{w}_{j}\right)\right|^{2} .
$$

Now it is easy to check that $\left(\lambda_{i}\right)^{2}+\left(\mu_{j}\right)^{2}-2 \lambda_{i} \mu_{j} \geq\left(\lambda_{i}^{+}\right)^{2}+\left(\mu_{j}^{+}\right)^{2}-2 \lambda_{i}^{+} \mu_{j}^{+}$.

### 13.21 Exercises

1. Show $\left(A^{*}\right)^{*}=A$ and $(A B)^{*}=B^{*} A^{*}$.
2. Prove Corollary 13.13.10.
3. Show that if $A$ is an $n \times n$ matrix which has an inverse then $A^{+}=A^{-1}$.
4. Using the singular value decomposition, show that for any square matrix $A$, it follows that $A^{*} A$ is unitarily similar to $A A^{*}$.
5. Let $A, B$ be a $m \times n$ matrices. Define an inner product on the set of $m \times n$ matrices by

$$
(A, B)_{F} \equiv \operatorname{trace}\left(A B^{*}\right)
$$

Show this is an inner product satisfying all the inner product axioms. Recall for $M$ an $n \times n$ matrix, trace $(M) \equiv \sum_{i=1}^{n} M_{i i}$. The resulting norm, $\|\cdot\|_{F}$ is called the Frobenius norm and it can be used to measure the distance between two matrices.
6. It was shown that a matrix $A$ is normal if and only if it is unitarily similar to a diagonal matrix. It was also shown that if a matrix is Hermitian, then it is unitarily similar to a real diagonal matrix. Show the converse of this last statement is also true. If a matrix is unitarily similar to a real diagonal matrix, then it is Hermitian.
7. Let $A$ be an $m \times n$ matrix. Show $\|A\|_{F}^{2} \equiv(A, A)_{F}=\sum_{j} \sigma_{j}^{2}$ where the $\sigma_{j}$ are the singular values of $A$.
8. If $A$ is a general $n \times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\left\{A_{k}\right\}$ of $n \times n$ matrices having distinct eigenvalues which has the property that the $i j^{t h}$ entry of $A_{k}$ converges to the $i j^{t h}$ entry of $A$ for all $i j$. Hint: Use Schur's theorem.
9. Prove the Cayley Hamilton theorem as follows. First suppose $A$ has a basis of eigenvectors $\left\{\boldsymbol{v}_{k}\right\}_{k=1}^{n}, A \boldsymbol{v}_{k}=\lambda_{k} \boldsymbol{v}_{k}$. Let $p(\boldsymbol{\lambda})$ be the characteristic polynomial. Show $p(A) \boldsymbol{v}_{k}=p\left(\lambda_{k}\right) \boldsymbol{v}_{k}=\mathbf{0}$. Then since $\left\{\boldsymbol{v}_{k}\right\}$ is a basis, it follows $p(A) \boldsymbol{x}=\mathbf{0}$ for all $\boldsymbol{x}$ and so $p(A)=0$. Next in the general case, use Problem 8 to obtain a sequence $\left\{A_{k}\right\}$ of matrices whose entries converge to the entries of $A$ such that $A_{k}$ has $n$ distinct eigenvalues and therefore by Theorem 6.6.1 on Page $147 A_{k}$ has a basis of eigenvectors. Therefore, from the first part and for $p_{k}(\lambda)$ the characteristic polynomial for $A_{k}$, it follows $p_{k}\left(A_{k}\right)=0$. Now explain why and the sense in which $\lim _{k \rightarrow \infty} p_{k}\left(A_{k}\right)=p(A)$.
10. Show directly that if $A$ is an $n \times n$ matrix and $A=A^{*}$ ( $A$ is Hermitian) then all the eigenvalues are real and eigenvectors can be assumed to be real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).
11. Let $\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}$ be an orthonormal basis for $\mathbb{F}^{n}$. Let $Q$ be a matrix whose $i^{t h}$ column is $\boldsymbol{v}_{i}$. Show

$$
Q^{*} Q=Q Q^{*}=I
$$

12. Show that an $n \times n$ matrix $Q$ is unitary if and only if it preserves distances. This means $|Q \boldsymbol{v}|=|\boldsymbol{v}|$. This was done in the text but you should try to do it for yourself.
13. Suppose $\left\{\boldsymbol{v}_{1}, \cdots, \boldsymbol{v}_{n}\right\}$ and $\left\{\boldsymbol{w}_{1}, \cdots, \boldsymbol{w}_{n}\right\}$ are two orthonormal bases for $\mathbb{F}^{n}$ and suppose $Q$ is an $n \times n$ matrix satisfying $Q \boldsymbol{v}_{i}=\boldsymbol{w}_{i}$. Then show $Q$ is unitary. If $|\boldsymbol{v}|=1$, show there is a unitary transformation which maps $v$ to $e_{1}$. This is done in the text but do it yourself with all details.
14. Let $A$ be a Hermitian matrix so $A=A^{*}$ and suppose all eigenvalues of $A$ are larger than $\delta^{2}$. Show

$$
(A v, v) \geq \delta^{2}|v|^{2}
$$

Where here, the inner product is $(\boldsymbol{v}, \boldsymbol{u}) \equiv \sum_{j=1}^{n} v_{j} \overline{u_{j}}$.
15. The discrete Fourier transform maps $\mathbb{C}^{n} \rightarrow \mathbb{C}^{n}$ as follows.

$$
F(\boldsymbol{x})=\boldsymbol{z} \text { where } z_{k}=\frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{-i \frac{2 \pi}{n} j k} x_{j}
$$

Show that $F^{-1}$ exists and is given by the formula

$$
F^{-1}(\boldsymbol{z})=\boldsymbol{x} \text { where } x_{j}=\frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{i \frac{2 \pi}{n} j k_{k}}
$$

Here is one way to approach this problem. Note $\boldsymbol{z}=U \boldsymbol{x}$ where $U=$

$$
\frac{1}{\sqrt{n}}\left(\begin{array}{ccccc}
e^{-i \frac{2 \pi}{n} 0 \cdot 0} & e^{-i \frac{2 \pi}{n} 1 \cdot 0} & e^{-i \frac{2 \pi}{n} 2 \cdot 0} & \cdots & e^{-i \frac{2 \pi}{n}(n-1) \cdot 0} \\
e^{-i \frac{2 \pi}{n} 0 \cdot 1} & e^{-i \frac{2 \pi}{n} 1 \cdot 1} & e^{-i \frac{2 \pi}{n} 2 \cdot 1} & \cdots & e^{-i \frac{2 \pi}{n}(n-1) \cdot 1} \\
e^{-i \frac{2 \pi}{n} 0 \cdot 2} & e^{-i \frac{2 \pi}{n} 1 \cdot 2} & e^{-i \frac{2 \pi}{n} 2 \cdot 2} & \cdots & e^{-i \frac{2 \pi}{n}(n-1) \cdot 2} \\
\vdots & \vdots & \vdots & & \vdots \\
e^{-i \frac{2 \pi}{n} 0 \cdot(n-1)} & e^{-i \frac{2 \pi}{n} 1 \cdot(n-1)} & e^{-i \frac{2 \pi}{n} 2 \cdot(n-1)} & \cdots & e^{-i \frac{2 \pi}{n}(n-1) \cdot(n-1)}
\end{array}\right)
$$

Now argue $U$ is unitary and use this to establish the result. To show this verify each row has length 1 and the inner product of two different rows gives 0 . Now $U_{k j}=e^{-i \frac{2 \pi}{n} j k}$ and so $\left(U^{*}\right)_{k j}=e^{i \frac{2 \pi}{n} j k}$.
16. Let $f$ be a periodic function having period $2 \pi$. The Fourier series of $f$ is an expression of the form

$$
\sum_{k=-\infty}^{\infty} c_{k} e^{i k x} \equiv \lim _{n \rightarrow \infty} \sum_{k=-n}^{n} c_{k} e^{i k x}
$$

and the idea is to find $c_{k}$ such that the above sequence converges in some way to $f$. If

$$
f(x)=\sum_{k=-\infty}^{\infty} c_{k} e^{i k x}
$$

and you formally multiply both sides by $e^{-i m x}$ and then integrate from 0 to $2 \pi$, interchanging the integral with the sum without any concern for whether this makes sense, show it is reasonable from this to expect

$$
c_{m}=\frac{1}{2 \pi} \int_{0}^{2 \pi} f(x) e^{-i m x} d x
$$

Now suppose you only know $f(x)$ at equally spaced points $2 \pi j / n$ for $j=0,1, \cdots, n$. Consider the Riemann sum for this integral obtained from using the left endpoint of the subintervals determined from the partition $\left\{\frac{2 \pi}{n} j\right\}_{j=0}^{n}$. How does this compare with the discrete Fourier transform? What happens as $n \rightarrow \infty$ to this approximation?
17. Suppose $A$ is a real $3 \times 3$ orthogonal matrix (Recall this means $A A^{T}=A^{T} A=I$.) having determinant 1 . Show it must have an eigenvalue equal to 1 . Note this shows there exists a vector $\boldsymbol{x} \neq \mathbf{0}$ such that $A \boldsymbol{x}=\boldsymbol{x}$. Hint: Show first or recall that any orthogonal matrix must preserve lengths. That is, $|\boldsymbol{A x}|=|\boldsymbol{x}|$.
18. Let $A$ be a complex $m \times n$ matrix. Using the description of the Moore Penrose inverse in terms of the singular value decomposition, show that

$$
\lim _{\delta \rightarrow 0+}\left(A^{*} A+\delta I\right)^{-1} A^{*}=A^{+}
$$

where the convergence happens in the Frobenius norm. Also verify, using the singular value decomposition, that the inverse exists in the above formula. Observe that this shows that the Moore Penrose inverse is unique.
19. Show that $A^{+}=\left(A^{*} A\right)^{+} A^{*}$. Hint: You might use the description of $A^{+}$in terms of the singular value decomposition.
20. In Theorem 13.11.1. Show that every matrix which commutes with $A$ also commutes with $A^{1 / k}$ the unique nonnegative self adjoint $k^{\text {th }}$ root.
21. Let $X$ be a finite dimensional inner product space and let $\beta=\left\{u_{1}, \cdots, u_{n}\right\}$ be an orthonormal basis for $X$. Let $A \in \mathscr{L}(X, X)$ be self adjoint and nonnegative and let $M$ be its matrix with respect to the given orthonormal basis. Show that $M$ is nonnegative, self adjoint also. Use this to show that $A$ has a unique nonnegative self adjoint $k^{\text {th }}$ root.
22. Let $A$ be a complex $m \times n$ matrix having singular value decomposition $U^{*} A V=$ $\left(\begin{array}{cc}\sigma & 0 \\ 0 & 0\end{array}\right)$ as explained above, where $\sigma$ is $k \times k$. Show that

$$
\operatorname{ker}(A)=\operatorname{span}\left(V e_{k+1}, \cdots, V e_{n}\right)
$$

the last $n-k$ columns of $V$.
23. The principal submatrices of an $n \times n$ matrix $A$ are $A_{k}$ where $A_{k}$ consists those entries which are in the first $k$ rows and first $k$ columns of $A$. Suppose $A$ is a real symmetric matrix and that $\boldsymbol{x} \rightarrow\langle\boldsymbol{A x}, \boldsymbol{x}\rangle$ is positive definite. This means that if $\boldsymbol{x} \neq \boldsymbol{0}$, then $\langle A x, x\rangle>0$. Show that each of the principal submatrices are positive definite. Hint: Consider $\left(\begin{array}{ll}\boldsymbol{x}^{T} & \mathbf{0}\end{array}\right) A\binom{\boldsymbol{x}}{\mathbf{0}}$ where $\boldsymbol{x}$ consists of $k$ entries.
24. $\uparrow$ A matrix $A$ has an $L U$ factorization if it there exists a lower triangular matrix $L$ having all ones on the diagonal and an upper triangular matrix $U$ such that $A=L U$. Show that if $A$ is a symmetric positive definite $n \times n$ real matrix, then $A$ has an $L U$ factorization with the property that each entry on the main diagonal in $U$ is positive. Hint: This is pretty clear if $A$ is $1 \times 1$. Assume true for $(n-1) \times(n-1)$. Then

$$
A=\left(\begin{array}{cc}
\hat{A} & \boldsymbol{a} \\
\boldsymbol{a}^{T} & a_{n n}
\end{array}\right)
$$

Then as above, $\hat{A}$ is positive definite. Thus it has an $L U$ factorization with all positive entries on the diagonal of $U$. Notice that, using block multiplication,

$$
A=\left(\begin{array}{cc}
L U & \boldsymbol{a} \\
\boldsymbol{a}^{T} & a_{n n}
\end{array}\right)=\left(\begin{array}{cc}
L & 0 \\
0 & 1
\end{array}\right)\left(\begin{array}{cc}
U & L^{-1} \boldsymbol{a} \\
\boldsymbol{a}^{T} & a_{n n}
\end{array}\right)
$$

Now consider that matrix on the right. Argue that it is of the form $\tilde{L} \tilde{U}$ where $\tilde{U}$ has all positive diagonal entries except possibly for the one in the $n^{\text {th }}$ row and $n^{\text {th }}$ column. Now explain why $\operatorname{det}(A)>0$ and argue that in fact all diagonal entries of $\tilde{U}$ are positive.
25. $\uparrow$ Let $A$ be a real symmetric $n \times n$ matrix and $A=L U$ where $L$ has all ones down the diagonal and $U$ has all positive entries down the main diagonal. Show that $A=L D H$ where $L$ is lower triangular and $H$ is upper triangular, each having all ones down the diagonal and $D$ a diagonal matrix having all positive entries down the main diagonal. In fact, these are the diagonal entries of $U$.
26. $\uparrow$ Show that if $L, L_{1}$ are lower triangular with ones down the main diagonal and $H, H_{1}$ are upper triangular with all ones down the main diagonal and $D, D_{1}$ are diagonal matrices having all positive diagonal entries, and if $L D H=L_{1} D_{1} H_{1}$, then $L=L_{1}, H=H_{1}, D=D_{1}$. Hint: Explain why $D_{1}^{-1} L_{1}^{-1} L D=H_{1} H^{-1}$. Then explain why the right side is upper triangular and the left side is lower triangular. Conclude these are both diagonal matrices. However, there are all ones down the diagonal in the expression on the right. Hence $H=H_{1}$. Do something similar to conclude that $L=L_{1}$ and then that $D=D_{1}$.
27. $\uparrow$ Show that if $A$ is a symmetric real matrix such that $\boldsymbol{x} \rightarrow\langle A \boldsymbol{x}, \boldsymbol{x}\rangle$ is positive definite, then there exists a lower triangular matrix $L$ having all positive entries down the diagonal such that $A=L L^{T}$. Hint: From the above, $A=L D H$ where $L, H$ are respectively lower and upper triangular having all ones down the diagonal and $D$ is a diagonal matrix having all positive entries. Then argue from the above problem and symmetry of $A$ that $H=L^{T}$. Now modify $L$ by making it equal to $L D^{1 / 2}$. This is called the Cholesky factorization.
28. Given $F \in \mathscr{L}(X, Y)$ where $X, Y$ are inner product spaces and $\operatorname{dim}(X)=n \leq m=$ $\operatorname{dim}(Y)$, there exists $R, U$ such that $U$ is nonnegative and Hermitian ( $U=U^{*}$ ) and $R^{*} R=I$ such that $F=R U$. Show that $U$ is actually unique and that $R$ is determined on $U(X)$. This was done in the book, but try to remember why this is so.
29. If $A$ is a complex Hermitian $n \times n$ matrix which has all eigenvalues nonnegative, show that there exists a complex Hermitian matrix $B$ such that $B B=A$.
30. $\uparrow$ Suppose $A, B$ are $n \times n$ real Hermitian matrices and they both have all nonnegative eigenvalues. Show that $\operatorname{det}(A+B) \geq \operatorname{det}(A)+\operatorname{det}(B)$. Hint: Use the above problem and the Cauchy Binet theorem. Let $P^{2}=A, Q^{2}=B$ where $P, Q$ are Hermitian and nonnegative. Then

$$
A+B=\left(\begin{array}{ll}
P & Q
\end{array}\right)\binom{P}{Q}
$$

31. Suppose $B=\left(\begin{array}{cc}\alpha & c^{*} \\ b & A\end{array}\right)$ is an $(n+1) \times(n+1)$ Hermitian nonnegative matrix where $\alpha$ is a scalar and $A$ is $n \times n$. Show that $\alpha$ must be real, $c=b$, and $A=A^{*}, A$ is nonnegative, and that if $\alpha=0$, then $\boldsymbol{b}=\mathbf{0}$. Otherwise, $\alpha>0$.
32. $\uparrow$ If $A$ is an $n \times n$ complex Hermitian and nonnegative matrix, show that there exists an upper triangular matrix $B$ such that $B^{*} B=A$. Hint: Prove this by induction. It is obviously true if $n=1$. Now if you have an $(n+1) \times(n+1)$ Hermitian nonnegative matrix, then from the above problem, it is of the form $\left(\begin{array}{cc}\alpha^{2} & \alpha b^{*} \\ \alpha b & A\end{array}\right), \alpha$ real.
33. $\uparrow$ Suppose $A$ is a nonnegative Hermitian matrix (all eigenvalues are nonnegative) which is partitioned as

$$
A=\left(\begin{array}{ll}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{array}\right)
$$

where $A_{11}, A_{22}$ are square matrices. Show that $\operatorname{det}(A) \leq \operatorname{det}\left(A_{11}\right) \operatorname{det}\left(A_{22}\right)$. Hint: Use the above problem to factor $A$ getting

$$
A=\left(\begin{array}{cc}
B_{11}^{*} & 0^{*} \\
B_{12}^{*} & B_{22}^{*}
\end{array}\right)\left(\begin{array}{cc}
B_{11} & B_{12} \\
0 & B_{22}
\end{array}\right)
$$

Next argue that $A_{11}=B_{11}^{*} B_{11}, A_{22}=B_{12}^{*} B_{12}+B_{22}^{*} B_{22}$. Use the Cauchy Binet theorem to argue that $\operatorname{det}\left(A_{22}\right)=\operatorname{det}\left(B_{12}^{*} B_{12}+B_{22}^{*} B_{22}\right) \geq \operatorname{det}\left(B_{22}^{*} B_{22}\right)$. Then explain why

$$
\begin{aligned}
\operatorname{det}(A) & =\operatorname{det}\left(B_{11}^{*}\right) \operatorname{det}\left(B_{22}^{*}\right) \operatorname{det}\left(B_{11}\right) \operatorname{det}\left(B_{22}\right) \\
& =\operatorname{det}\left(B_{11}^{*} B_{11}\right) \operatorname{det}\left(B_{22}^{*} B_{22}\right)
\end{aligned}
$$

34. $\uparrow$ Prove the inequality of Hadamard. If $A$ is a Hermitian matrix which is nonnegative (all eigenvalues are nonnegative), then $\operatorname{det}(A) \leq \prod_{i} A_{i i}$.

## Chapter 14

## Analysis Of Linear Transformations

### 14.1 The Condition Number

Let $A \in \mathscr{L}(X, X)$ be a linear transformation where $X$ is a finite dimensional vector space and consider the problem $A x=b$ where it is assumed there is a unique solution to this problem. How does the solution change if $A$ is changed a little bit and if $b$ is changed a little bit? This is clearly an interesting question because you often do not know $A$ and $b$ exactly. If a small change in these quantities results in a large change in the solution, $x$, then it seems clear this would be undesirable. In what follows $\|\cdot\|$ when applied to a linear transformation will always refer to the operator norm. Recall the following property of the operator norm in Theorem 10.7.3.

Lemma 14.1.1 Let $A, B \in \mathscr{L}(X, X)$ where $X$ is a normed vector space as above. Then for $\|\cdot\|$ denoting the operator norm, $\|A B\| \leq\|A\|\|B\|$.

Lemma 14.1.2 Let $A, B \in \mathscr{L}(X, X), A^{-1} \in \mathscr{L}(X, X)$, and suppose

$$
\|B\|<1 /\left\|A^{-1}\right\|
$$

Then $(A+B)^{-1},\left(I+A^{-1} B\right)^{-1}$ exists and

$$
\begin{gather*}
\left\|\left(I+A^{-1} B\right)^{-1}\right\| \leq\left(1-\left\|A^{-1} B\right\|\right)^{-1}  \tag{14.1}\\
\left\|(A+B)^{-1}\right\| \leq\left\|A^{-1}\right\|\left|\frac{1}{1-\left\|A^{-1} B\right\|}\right| \tag{14.2}
\end{gather*}
$$

The above formula makes sense because $\left\|A^{-1} B\right\|<1$.
Proof: By Lemma 10.7.3,

$$
\begin{equation*}
\left\|A^{-1} B\right\| \leq\left\|A^{-1}\right\|\|B\|<\left\|A^{-1}\right\| \frac{1}{\left\|A^{-1}\right\|}=1 \tag{14.3}
\end{equation*}
$$

Then from the triangle inequality,

$$
\begin{aligned}
\left\|\left(I+A^{-1} B\right) x\right\| & \geq\|x\|-\left\|A^{-1} B x\right\| \\
& \geq\|x\|-\left\|A^{-1} B\right\|\|x\|=\left(1-\left\|A^{-1} B\right\|\right)\|x\|
\end{aligned}
$$

It follows that $I+A^{-1} B$ is one to one because from $14.3,1-\left\|A^{-1} B\right\|>0$. Thus if $\left(I+A^{-1} B\right) x=0$, then $x=0$. Thus $I+A^{-1} B$ is also onto, taking a basis to a basis. Then a generic $y \in X$ is of the form $y=\left(I+A^{-1} B\right) x$ and the above shows that

$$
\left\|\left(I+A^{-1} B\right)^{-1} y\right\| \leq\left(1-\left\|A^{-1} B\right\|\right)^{-1}\|y\|
$$

which verifies 14.1. Thus $(A+B)=A\left(I+A^{-1} B\right)$ is one to one and this with Lemma 10.7.3 implies 14.2.

Proposition 14.1.3 Suppose $A$ is invertible, $b \neq 0, A x=b$, and $(A+B) x_{1}=b_{1}$ where $\|B\|<1 /\left\|A^{-1}\right\|$. Then

$$
\frac{\left\|x_{1}-x\right\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\|\|A\|}{1-\left\|A^{-1} B\right\|}\left(\frac{\left\|b_{1}-b\right\|}{\|b\|}+\frac{\|B\|}{\|A\|}\right)
$$

Proof: This follows from the above lemma.

$$
\begin{aligned}
\frac{\left\|x_{1}-x\right\|}{\|x\|} & =\frac{\left\|\left(I+A^{-1} B\right)^{-1} A^{-1} b_{1}-A^{-1} b\right\|}{\left\|A^{-1} b\right\|} \\
& \leq \frac{1}{1-\left\|A^{-1} B\right\|} \frac{\left\|A^{-1} b_{1}-\left(I+A^{-1} B\right) A^{-1} b\right\|}{\left\|A^{-1} b\right\|} \\
& \leq \frac{1}{1-\left\|A^{-1} B\right\|} \frac{\left\|A^{-1}\left(b_{1}-b\right)\right\|+\left\|A^{-1} B A^{-1} b\right\|}{\left\|A^{-1} b\right\|} \\
& \leq \frac{\left\|A^{-1}\right\|}{1-\left\|A^{-1} B\right\|}\left(\frac{\left\|b_{1}-b\right\|}{\left\|A^{-1} b\right\|}+\|B\|\right)
\end{aligned}
$$

because $A^{-1} b /\left\|A^{-1} b\right\|$ is a unit vector. Now multiply and divide by $\|A\|$. Then

$$
\begin{aligned}
& \leq \frac{\left\|A^{-1}\right\|\|A\|}{1-\left\|A^{-1} B\right\|}\left(\frac{\left\|b_{1}-b\right\|}{\|A\|\left\|A^{-1} b\right\|}+\frac{\|B\|}{\|A\|}\right) \\
& \leq \frac{\left\|A^{-1}\right\|\|A\|}{1-\left\|A^{-1} B\right\|}\left(\frac{\left\|b_{1}-b\right\|}{\|b\|}+\frac{\|B\|}{\|A\|}\right) .
\end{aligned}
$$

This shows that the number, $\left\|A^{-1}\right\|\|A\|$, controls how sensitive the relative change in the solution of $A x=b$ is to small changes in $A$ and $b$. This number is called the condition number. It is bad when this number is large because a small relative change in $b$, for example could yield a large relative change in $x$.

Recall that for $A$ an $n \times n$ matrix, $\|A\|_{2}=\sigma_{1}$ where $\sigma_{1}$ is the largest singular value. The largest singular value of $A^{-1}$ is therefore, $1 / \sigma_{n}$ where $\sigma_{n}$ is the smallest singular value of $A$. Therefore, the condition number is controlled by $\sigma_{1} / \sigma_{n}$, the ratio of the largest to the smallest singular value of $A$ provided the norm is the usual Euclidean norm.

### 14.2 The Spectral Radius

Even though it is in general impractical to compute the Jordan form, its existence is all that is needed in order to prove an important theorem about something which is relatively easy to compute. This is the spectral radius of a matrix.

Definition 14.2.1 Define $\sigma(A)$ to be the eigenvalues of A. Also,

$$
\rho(A) \equiv \max (|\lambda|: \lambda \in \sigma(A))
$$

The number, $\rho(A)$ is known as the spectral radius of $A$.

Recall the following symbols and their meaning. $\lim \sup _{n \rightarrow \infty} a_{n}, \liminf _{n \rightarrow \infty} a_{n}$. They are respectively the largest and smallest limit points of the sequence $\left\{a_{n}\right\}$ where $\pm \infty$ is allowed in the case where the sequence is unbounded. They are also defined as

$$
\begin{aligned}
\lim \sup _{n \rightarrow \infty} a_{n} & \equiv \lim _{n \rightarrow \infty}\left(\sup \left\{a_{k}: k \geq n\right\}\right) \\
\lim \inf _{n \rightarrow \infty} a_{n} & \equiv \lim _{n \rightarrow \infty}\left(\inf \left\{a_{k}: k \geq n\right\}\right)
\end{aligned}
$$

Thus, the limit of the sequence exists if and only if these are both equal to the same real number.

Lemma 14.2.2 Let J be a $p \times p$ Jordan block

$$
J=\left(\begin{array}{llll}
\lambda & 1 & & 0 \\
& \lambda & \ddots & \\
& & \ddots & 1 \\
0 & & & \lambda
\end{array}\right)
$$

Then $\lim _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n}=|\lambda|$
Proof: The norm on matrices can be any norm. It could be operator norm for example. If $\lambda=0$, there is nothing to show because $J^{p}=0$ and so the limit is obviously 0 . Therefore, assume $\lambda \neq 0$.

$$
J^{n}=\sum_{i=0}^{p}\binom{n}{i} N^{i} \lambda^{n-i}
$$

Then

$$
\begin{align*}
\left\|J^{n}\right\| & \leq \sum_{i=0}^{p}\binom{n}{i}\left\|N^{i}\right\||\lambda|^{n-i}=|\lambda|^{n}+C|\lambda|^{n} \sum_{i=1}^{p}\binom{n}{i}|\lambda|^{-i}  \tag{14.4}\\
& \leq|\lambda|^{n}\left(1+C n^{p}\right) \leq|\lambda|^{n} \tilde{C} n^{p} \tag{14.5}
\end{align*}
$$

where the $C$ depends on $\sum_{i=1}^{p}|\lambda|^{-i}$ and the $\left\|N^{i}\right\|$. Therefore,

$$
\lim \sup _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n} \leq|\lambda| \lim \sup _{n \rightarrow \infty}\left(\tilde{C} n^{p}\right)^{1 / n}=|\lambda|
$$

Next let $\boldsymbol{x}$ be an eigenvector for $\lambda$ such that $\|\boldsymbol{x}\|=1$, the norm being whatever norm is desired. Then $J \boldsymbol{x}=\lambda \boldsymbol{x}$. It follows that $J^{n} \boldsymbol{x}=\lambda^{n} \boldsymbol{x}$. Thus $\left\|J^{n}\right\| \geq\left\|J^{n} \boldsymbol{x}\right\|=|\lambda|^{n}\|\boldsymbol{x}\|=|\lambda|^{n}$. It follows that $\liminf _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n} \geq|\lambda|$. Therefore,

$$
|\lambda| \leq \lim \inf _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n} \leq \lim \sup _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n} \leq|\lambda|
$$

which shows that $\lim _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n}=|\lambda|$. The same conclusion holds for any other norm. Indeed, if $|||\cdot|||$ were another norm, there are constants $\delta, \Delta$ such that

$$
\delta^{1 / n}\left\|J^{n}\right\|^{1 / n} \leq\| \| J^{n}\left\|^{1 / n} \leq \Delta^{1 / n}\right\| J^{n} \|^{1 / n}
$$

Then since $\lim _{n \rightarrow \infty} \delta^{1 / n}=\lim _{n \rightarrow \infty} \Delta^{1 / n}=1$, the squeezing theorem from calculus implies that $\lim _{n \rightarrow \infty}\left|\left\|J^{n}\right\|\right|^{1 / n}=\rho$.

Corollary 14.2.3 Let J be in Jordan canonical form

$$
J=\left(\begin{array}{lll}
J_{1} & & \\
& \ddots & \\
& & J_{s}
\end{array}\right)
$$

where each $J_{k}$ is a block diagonal having $\lambda_{k}$ on the main diagonal and strings of ones on the super diagonal, as described earlier. Also let $\rho \equiv \max \left\{\left|\lambda_{i}\right|: \lambda_{i} \in \sigma(J)\right\}$. Then for any norm $\|\cdot\|$

$$
\lim _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n}=\rho
$$

Proof: For convenience, take the norm to be given as $\|A\| \equiv \max \left\{\left|A_{i j}\right|, i, j\right\}$. Then with this norm,

$$
\left\|J^{n}\right\|^{1 / n}=\max \left\{\left\|J_{k}^{n}\right\|^{1 / n}, k=1, \cdots, s\right\}
$$

From Lemma 14.2.2,

$$
\lim _{n \rightarrow \infty}\left\|J_{k}^{n}\right\|^{1 / n}=\left|\lambda_{k}\right|
$$

Therefore,

$$
\begin{aligned}
\lim _{n \rightarrow \infty}\left\|J^{n}\right\|^{1 / n} & =\lim _{n \rightarrow \infty}\left(\max \left\{\left\|J_{k}^{n}\right\|^{1 / n}, k=1, \cdots, s\right\}\right) \\
& =\max _{k}\left(\lim _{n \rightarrow \infty}\left\|J_{k}^{n}\right\|^{1 / n}\right)=\max _{k}\left|\lambda_{k}\right|=\rho
\end{aligned}
$$

Now let the norm on the matrices be any other norm say $\|\|\cdot\|\|$. By equivalence of norms, there are $\delta, \Delta$ such that

$$
\delta\|A\| \leq\| \| A\| \| \leq \Delta\|A\|
$$

for all matrices $A$. Therefore,

$$
\delta^{1 / n}\left\|J^{n}\right\|^{1 / n} \leq\| \| J^{n}\left\|^{1 / n} \leq \Delta^{1 / n}\right\| J^{n} \|^{1 / n}
$$

and so, passing to a limit, it follows that, since $\lim _{n \rightarrow \infty} \delta^{1 / n}=\lim _{n \rightarrow \infty} \Delta^{1 / n}=1$,

$$
\rho=\lim _{n \rightarrow \infty}\| \| J^{n}\| \|^{1 / n}
$$

Theorem 14.2.4 (Gelfand) Let A be a complex $p \times p$ matrix. Then if $\rho$ is the absolute value of its largest eigenvalue,

$$
\lim _{n \rightarrow \infty}\left\|A^{n}\right\|^{1 / n}=\rho
$$

Here $\|\cdot\|$ is any norm on $\mathscr{L}\left(\mathbb{C}^{n}, \mathbb{C}^{n}\right)$.
Proof: Let $\|\cdot\|$ be the operator norm on $\mathscr{L}\left(\mathbb{C}^{n}, \mathbb{C}^{n}\right)$. Then letting $J$ denote the Jordan form of $A, S^{-1} A S=J$ and these two $J, A$ have the same eigenvalues. Thus it follows from Corollary 14.2.3

$$
\begin{aligned}
\lim \sup _{n \rightarrow \infty}\left\|A^{n}\right\|^{1 / n} & =\limsup _{n \rightarrow \infty}\left\|S J^{n} S^{-1}\right\|^{1 / n} \leq \lim \sup _{n \rightarrow \infty}\left(\|S\|\left\|S^{-1}\right\|\left\|J^{n}\right\|\right)^{1 / n} \\
& =\lim _{n \rightarrow \infty}\left(\|S\|\left\|S^{-1}\right\|\right)^{1 / n}\left\|J^{n}\right\|^{1 / n}=\rho
\end{aligned}
$$

Letting $\lambda$ be the largest eigenvalue of $A,|\lambda|=\rho$, and $A \boldsymbol{x}=\lambda \boldsymbol{x}$ where $\|\boldsymbol{x}\|=1$,

$$
\left\|A^{n}\right\| \geq\left\|A^{n} \boldsymbol{x}\right\|=\rho^{n}
$$

and so

$$
\lim \inf _{n \rightarrow \infty}\left\|A^{n}\right\|^{1 / n} \geq \rho \geq \lim \sup _{n \rightarrow \infty}\left\|A^{n}\right\|^{1 / n}
$$

If follows that $\liminf _{n \rightarrow \infty}\left\|A^{n}\right\|^{1 / n}=\limsup p_{n \rightarrow \infty}\left\|\left.A^{n}\right|^{1 / n}=\lim _{n \rightarrow \infty}\right\| A^{n} \|^{1 / n}=\rho$. As in Corollary 14.2.3, there is no difference if any other norm is used because they are all equivalent.

I would argue that a better way to prove this theorem is to use the theory of complex analysis and tie it in with a Laurent series. However, there is a non trivial issue related to the set of convergence of the Laurent series which involves the theory of functions of a complex variable and this knowledge is not being assumed here. Thus the above gives an algebraic proof which does not involve so much hard analysis.

Example 14.2.5 Consider $\left(\begin{array}{ccc}9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8\end{array}\right)$. Estimate the absolute value of the largest eigenvalue.

A laborious computation reveals the eigenvalues are 5 and 10. Therefore, the right answer in this case is 10 . Consider $\left\|A^{9}\right\|^{1 / 9}$ where the norm is obtained by taking the maximum of all the absolute values of the entries. Thus

$$
\left(\begin{array}{ccc}
9 & -1 & 2 \\
-2 & 8 & 4 \\
1 & 1 & 8
\end{array}\right)^{9}=\left(\begin{array}{ccc}
800390625 & -199609375 & 399218750 \\
-399218750 & 600781250 & 798437500 \\
199609375 & 199609375 & 600781250
\end{array}\right)
$$

and taking the seventh root of the largest entry gives

$$
\rho(A) \approx 800390625^{1 / 9}=9.7556
$$

Of course the interest lies primarily in matrices for which the exact roots to the characteristic equation are not known and in the theoretical significance.

### 14.3 Series And Sequences Of Linear Operators

Before beginning this discussion, it is necessary to define what is meant by convergence in $\mathscr{L}(X, Y)$.

Definition 14.3.1 Let $\left\{A_{k}\right\}_{k=1}^{\infty}$ be a sequence in $\mathscr{L}(X, Y)$ where $X, Y$ are finite dimensional normed linear spaces. Then $\lim _{n \rightarrow \infty} A_{k}=A$ if for every $\varepsilon>0$ there exists $N$ such that if $n>N$, then

$$
\left\|A-A_{n}\right\|<\varepsilon
$$

Here the norm refers to any of the norms defined on $\mathscr{L}(X, Y)$. By Corollary 10.6.4 and Theorem 5.1.4 it doesn't matter which one is used. Define the symbol for an infinite sum in the usual way. Thus

$$
\sum_{k=1}^{\infty} A_{k} \equiv \lim _{n \rightarrow \infty} \sum_{k=1}^{n} A_{k}
$$

Lemma 14.3.2 Suppose $\left\{A_{k}\right\}_{k=1}^{\infty}$ is a sequence in $\mathscr{L}(X, Y)$ where $X, Y$ are finite dimensional normed linear spaces. Then if $\sum_{k=1}^{\infty}\left\|A_{k}\right\|<\infty$, It follows that

$$
\begin{equation*}
\sum_{k=1}^{\infty} A_{k} \tag{14.6}
\end{equation*}
$$

exists (converges). In words, absolute convergence implies convergence. Also,

$$
\left\|\sum_{k=1}^{\infty} A_{k}\right\| \leq \sum_{k=1}^{\infty}\left\|A_{k}\right\|
$$

Proof: For $p \leq m \leq n$,

$$
\left\|\sum_{k=1}^{n} A_{k}-\sum_{k=1}^{m} A_{k}\right\| \leq \sum_{k=p}^{\infty}\left\|A_{k}\right\|
$$

and so for $p$ large enough, this term on the right in the above inequality is less than $\varepsilon$. Since $\varepsilon$ is arbitrary, this shows the partial sums of 14.6 are a Cauchy sequence. Therefore by Corollary 10.6 .4 it follows that these partial sums converge. As to the last claim,

$$
\left\|\sum_{k=1}^{n} A_{k}\right\| \leq \sum_{k=1}^{n}\left\|A_{k}\right\| \leq \sum_{k=1}^{\infty}\left\|A_{k}\right\|
$$

Therefore, passing to the limit,

$$
\left\|\sum_{k=1}^{\infty} A_{k}\right\| \leq \sum_{k=1}^{\infty}\left\|A_{k}\right\| .
$$

Why is this last step justified? (Recall the triangle inequality $|\|A\|-\|B\|| \leq\|A-B\|$.)
Now here is a useful result for differential equations.
Theorem 14.3.3 Let $X$ be a finite dimensional inner product space and let $A \in \mathscr{L}(X, X)$. Define

$$
\Phi(t) \equiv \sum_{k=0}^{\infty} \frac{t^{k} A^{k}}{k!}
$$

Then the series converges for each $t \in \mathbb{R}$. Also

$$
\Phi^{\prime}(t) \equiv \lim _{h \rightarrow 0} \frac{\Phi(t+h)-\Phi(t)}{h}=\sum_{k=1}^{\infty} \frac{t^{k-1} A^{k}}{(k-1)!}=A \sum_{k=0}^{\infty} \frac{t^{k} A^{k}}{k!}=A \Phi(t)
$$

Also $A \Phi(t)=\Phi(t) A$ and for all $t, \Phi(t) \Phi(-t)=I$ so $\Phi(t)^{-1}=\Phi(-t), \Phi(0)=I$. (It is understood that $A^{0}=I$ in the above formula.)

Proof: First consider the claim about convergence.

$$
\sum_{k=0}^{\infty}\left\|\frac{t^{k} A^{k}}{k!}\right\| \leq \sum_{k=0}^{\infty} \frac{|t|^{k}\|A\|^{k}}{k!}=e^{|t|\|A\|}<\infty
$$

so it converges by Lemma 14.3.2.

$$
\begin{aligned}
\frac{\Phi(t+h)-\Phi(t)}{h} & =\frac{1}{h} \sum_{k=0}^{\infty} \frac{\left((t+h)^{k}-t^{k}\right) A^{k}}{k!} \\
& =\frac{1}{h} \sum_{k=0}^{\infty} \frac{\left(k\left(t+\theta_{k} h\right)^{k-1} h\right) A^{k}}{k!}=\sum_{k=1}^{\infty} \frac{\left(t+\theta_{k} h\right)^{k-1} A^{k}}{(k-1)!}
\end{aligned}
$$

this by the mean value theorem. Note that the series converges thanks to Lemma 14.3.2. Here $\theta_{k} \in(0,1)$. Thus

$$
\begin{gathered}
\left\|\frac{\Phi(t+h)-\Phi(t)}{h}-\sum_{k=1}^{\infty} \frac{t^{k-1} A^{k}}{(k-1)!}\right\|=\left\|\sum_{k=1}^{\infty} \frac{\left(\left(t+\theta_{k} h\right)^{k-1}-t^{k-1}\right) A^{k}}{(k-1)!}\right\| \\
=\left\|\sum_{k=1}^{\infty} \frac{\left((k-1)\left(t+\tau_{k} \theta_{k} h\right)^{k-2} \theta_{k} h\right) A^{k}}{(k-1)!}\right\|=|h|\left\|\sum_{k=2}^{\infty} \frac{\left(\left(t+\tau_{k} \theta_{k} h\right)^{k-2} \theta_{k}\right) A^{k}}{(k-2)!}\right\| \\
\leq|h| \sum_{k=2}^{\infty} \frac{(|t|+|h|)^{k-2}\|A\|^{k-2}}{(k-2)!}\|A\|^{2}=|h| e^{(|t|+|h|)\|A\|}\|A\|^{2}
\end{gathered}
$$

so letting $|h|<1$, this is no larger than $|h| e^{(|t|+1)\|A\|}\|A\|^{2}$. Hence the desired limit is valid. It is obvious that $A \Phi(t)=\Phi(t) A$. Also the formula shows that

$$
\Phi^{\prime}(t)=A \Phi(t)=\Phi(t) A, \Phi(0)=I .
$$

Now consider the claim about $\Phi(-t)$. The above computation shows that

$$
\Phi^{\prime}(-t)=A \Phi(-t)
$$

and so $\frac{d}{d t}(\Phi(-t))=-\Phi^{\prime}(-t)=-A \Phi(-t)$. Now let $x, y$ be two vectors in $X$. Consider

$$
(\Phi(-t) \Phi(t) x, y)_{X}
$$

Then this equals $(x, y)$ when $t=0$. Take its derivative.

$$
\begin{aligned}
& \left(\left(-\Phi^{\prime}(-t) \Phi(t)+\Phi(-t) \Phi^{\prime}(t)\right) x, y\right)_{X} \\
= & ((-A \Phi(-t) \Phi(t)+\Phi(-t) A \Phi(t)) x, y)_{X} \\
= & (0, y)_{X}=0
\end{aligned}
$$

Hence this scalar valued function equals a constant and so the constant must be $(x, y)_{X}$. Hence for all $x, y,(\Phi(-t) \Phi(t) x-x, y)_{X}=0$ for all $x, y$ and this is so in particular for $y=\Phi(-t) \Phi(t) x-x$ which shows that $\Phi(-t) \Phi(t)=I$.

In fact, one can prove a group identity of the form $\Phi(t+s)=\Phi(t) \Phi(t)$ for all $t, s \in \mathbb{R}$.
Corollary 14.3.4 Let $\Phi(t)$ be given as above. Then $\Phi(t+s)=\Phi(t) \Phi(s)$ for any $s, t \in \mathbb{R}$.

Proof: Let $\boldsymbol{y}(t) \equiv(\Phi(t) \Phi(s)-\Phi(t+s)) \boldsymbol{x}$. Thus $\boldsymbol{y}(0)=\mathbf{0}$. Now pick $\boldsymbol{z} \in X$.

$$
\boldsymbol{y}^{\prime}(t)=A \boldsymbol{y}(t), \quad\left(\boldsymbol{y}^{\prime}(t), \boldsymbol{z}\right)=(A \boldsymbol{y}(t), \boldsymbol{z})
$$

Now, as above, $\frac{d}{d t}(\Phi(-t))=-A \Phi(-t)$ and so

$$
\begin{gathered}
\frac{d}{d t}(\Phi(-t) \boldsymbol{y}(t), \boldsymbol{z})=\left(-A \Phi(-t) \boldsymbol{y}(t)+\Phi(-t) \boldsymbol{y}^{\prime}(t), \boldsymbol{z}\right) \\
=\left(-A \Phi(-t) \boldsymbol{y}(t)+\Phi(-t) \boldsymbol{y}^{\prime}(t), \boldsymbol{z}\right)=(-A \Phi(-t) \boldsymbol{y}(t)+\Phi(-t) A \boldsymbol{y}(t), \boldsymbol{z}) \\
=\quad(-A \Phi(-t) \boldsymbol{y}(t)+A \Phi(-t) \boldsymbol{y}(t), \boldsymbol{z})=0
\end{gathered}
$$

It follows from beginning calculus that

$$
(\Phi(-t) \boldsymbol{y}(t), \boldsymbol{z})=C
$$

However, $\boldsymbol{y}(0)=\mathbf{0}$ and so $C=0$. Since $\boldsymbol{z}$ is arbitrary, it follows that $\Phi(-t) \boldsymbol{y}(t)=\mathbf{0}$. By Theorem 14.3.3, $\boldsymbol{y}(t)=\mathbf{0}$. Now $\boldsymbol{x}$ was arbitrary so also $\Phi(t) \Phi(s)-\Phi(t+s)=0$ which proves the corollary.

Note how this also shows that $\Phi(t)$ commutes with $\Phi(s)$ for any $t, s$. Also note that all of this works with no change if $A \in \mathscr{L}(X, X)$ where $X$ is a Hilbert space, possibly not finite dimensional. In fact you don't even need a Hilbert space. It would work fine with a Banach space, and you would replace the inner product with the pairing with the dual space but this requires more functional analysis than what is considered here.

As a special case, suppose $\lambda \in \mathbb{C}$ and consider $\sum_{k=0}^{\infty} \frac{t^{k} \lambda^{k}}{k!}$ where $t \in \mathbb{R}$. In this case, $A_{k}=\frac{t^{k} \lambda^{k}}{k!}$ and you can think of it as being in $\mathscr{L}(\mathbb{C}, \mathbb{C})$. Then the following corollary is of great interest.

Corollary 14.3.5 Let

$$
f(t) \equiv \sum_{k=0}^{\infty} \frac{t^{k} \lambda^{k}}{k!} \equiv 1+\sum_{k=1}^{\infty} \frac{t^{k} \lambda^{k}}{k!}
$$

Then this function is a well defined complex valued function and furthermore, it satisfies the initial value problem,

$$
y^{\prime}=\lambda y, y(0)=1
$$

Furthermore, if $\lambda=a+i b$,

$$
|f|(t)=e^{a t}
$$

Proof: The first part is a special case of the above theorem. Then

$$
\bar{y}^{\prime}=\bar{\lambda} \bar{y}, \bar{y}(0)=1
$$

It follows

$$
\begin{aligned}
\frac{d}{d y}|y(t)|^{2} & =y^{\prime}(t) \bar{y}(t)+y(t) \bar{y}^{\prime}(t) \\
& =(a+i b)|y(t)|^{2}+(a-i b)|y(t)|^{2} \\
& =2 a|y(t)|^{2},|y(0)|^{2}=1
\end{aligned}
$$

It follows $|y(t)|^{2}=e^{2 a t},|y(t)|=e^{a t}$ as claimed.
This follows because in general, if

$$
z^{\prime}=c z, z(0)=1
$$

with $c$ real it follows $z(t)=e^{c t}$. To see this, $z^{\prime}-c z=0$ and so, multiplying both sides by $e^{-c t}$ you get

$$
\frac{d}{d t}\left(z e^{-c t}\right)=0
$$

and so $z e^{-c t}$ equals a constant which must be 1 because of the initial condition $z(0)=1$.
Definition 14.3.6 The function in Corollary 14.3 .5 given by that power series is denoted as

$$
\exp (\lambda t) \text { or } e^{\lambda t}
$$

The next lemma is normally discussed in advanced calculus courses but is proved here for the convenience of the reader. It is known as the root test.

Definition 14.3.7 For $\left\{a_{n}\right\}$ any sequence of real numbers

$$
\lim \sup _{n \rightarrow \infty} a_{n} \equiv \lim _{n \rightarrow \infty}\left(\sup \left\{a_{k}: k \geq n\right\}\right)
$$

Similarly

$$
\lim _{n \rightarrow \infty} \inf _{n} a_{n} \equiv \lim _{n \rightarrow \infty}\left(\inf \left\{a_{k}: k \geq n\right\}\right)
$$

In case $A_{n}$ is an increasing (decreasing) sequence which is unbounded above (below) then it is understood that $\lim _{n \rightarrow \infty} A_{n}=\infty(-\infty)$ respectively. Thus either of limsup or liminf can equal $+\infty$ or $-\infty$. However, the important thing about these is that unlike the limit, these always exist.

It is convenient to think of these as the largest point which is the limit of some subsequence of $\left\{a_{n}\right\}$ and the smallest point which is the limit of some subsequence of $\left\{a_{n}\right\}$ respectively. Thus $\lim _{n \rightarrow \infty} a_{n}$ exists and equals some point of $[-\infty, \infty]$ if and only if the two are equal.

Lemma 14.3.8 Let $\left\{a_{p}\right\}$ be a sequence of nonnegative terms and let

$$
r=\limsup _{p \rightarrow \infty} a_{p}^{1 / p}
$$

Then if $r<1$, it follows the series, $\sum_{k=1}^{\infty} a_{k}$ converges and if $r>1$, then $a_{p}$ fails to converge to 0 so the series diverges. If $A$ is an $n \times n$ matrix and

$$
\begin{equation*}
r=\lim \sup _{p \rightarrow \infty}\left\|A^{p}\right\|^{1 / p}, \tag{14.7}
\end{equation*}
$$

then if $r>1$, then $\sum_{k=0}^{\infty} A^{k}$ fails to converge and if $r<1$ then the series converges. Note that the series converges when the spectral radius is less than one and diverges if the spectral radius is larger than one. In fact, $\limsup _{p \rightarrow \infty}\left\|A^{p}\right\|^{1 / p}=\lim _{p \rightarrow \infty}\left\|A^{p}\right\|^{1 / p}$ from Theorem 14.2.4.

Proof: Suppose $r<1$. Then there exists $N$ such that if $p>N$,

$$
a_{p}^{1 / p}<R
$$

where $r<R<1$. Therefore, for all such $p, a_{p}<R^{p}$ and so by comparison with the geometric series, $\sum R^{p}$, it follows $\sum_{p=1}^{\infty} a_{p}$ converges.

Next suppose $r>1$. Then letting $1<R<r$, it follows there are infinitely many values of $p$ at which

$$
R<a_{p}^{1 / p}
$$

which implies $R^{p}<a_{p}$, showing that $a_{p}$ cannot converge to 0 and so the series cannot converge either.

To see the last claim, if $r>1$, then $\left\|A^{p}\right\|$ fails to converge to 0 so $\left\{\sum_{k=0}^{m} A^{k}\right\}_{m=0}^{\infty}$ is not a Cauchy sequence. Hence $\sum_{k=0}^{\infty} A^{k} \equiv \lim _{m \rightarrow \infty} \sum_{k=0}^{m} A^{k}$ cannot exist. If $r<1$, then for all $n$ large enough, $\left\|A^{n}\right\|^{1 / n} \leq r<1$ for some $r$ so $\left\|A^{n}\right\| \leq r^{n}$. Hence $\sum_{n}\left\|A^{n}\right\|$ converges and so by Lemma 14.3.2, it follows that $\sum_{k=1}^{\infty} A^{k}$ also converges.

Now denote by $\sigma(A)^{p}$ the collection of all numbers of the form $\lambda^{p}$ where $\lambda \in \sigma(A)$.
Lemma 14.3.9 $\sigma\left(A^{p}\right)=\sigma(A)^{p} \equiv\left\{\lambda^{p}: \lambda \in \sigma(A)\right\}$.
Proof: In dealing with $\sigma\left(A^{p}\right)$, it suffices to deal with $\sigma\left(J^{p}\right)$ where $J$ is the Jordan form of $A$ because $J^{p}$ and $A^{p}$ are similar. Thus if $\lambda \in \sigma\left(A^{p}\right)$, then $\lambda \in \sigma\left(J^{p}\right)$ and so $\lambda=\alpha$ where $\alpha$ is one of the entries on the main diagonal of $J^{p}$. These entries are of the form $\lambda^{p}$ where $\lambda \in \sigma(A)$. Thus $\lambda \in \sigma(A)^{p}$ and this shows $\sigma\left(A^{p}\right) \subseteq \sigma(A)^{p}$.

Now take $\alpha \in \sigma(A)$ and consider $\alpha^{p}$.

$$
\alpha^{p} I-A^{p}=\left(\alpha^{p-1} I+\cdots+\alpha A^{p-2}+A^{p-1}\right)(\alpha I-A)
$$

and so $\alpha^{p} I-A^{p}$ fails to be one to one which shows that $\alpha^{p} \in \sigma\left(A^{p}\right)$ which shows that $\sigma(A)^{p} \subseteq \sigma\left(A^{p}\right)$.

### 14.4 Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$
\begin{equation*}
A x=b \tag{14.8}
\end{equation*}
$$

where $A$ is an $n \times n$ matrix. In many applications, the matrix $A$ is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. The method is to write

$$
A=B-C
$$

where $B^{-1}$ exists. Then the system is of the form

$$
B x=C x+b
$$

and so the solution is solves

$$
\boldsymbol{x}=B^{-1} C \boldsymbol{x}+B^{-1} \boldsymbol{b} \equiv T \boldsymbol{x}
$$

In other words, you look for a fixed point of $T$. There are standard methods for finding such fixed points which hold in general Banach spaces which is the term for a complete normed linear space.

Definition 14.4.1 A normed vector space, $E$ with norm $\|\cdot\|$ is called a Banach space if it is also complete. This means that every Cauchy sequence converges. Recall that a sequence $\left\{x_{n}\right\}_{n=1}^{\infty}$ is a Cauchy sequence iffor every $\varepsilon>0$ there exists $N$ such that whenever $m, n>N$,

$$
\left\|x_{n}-x_{m}\right\|<\varepsilon .
$$

Thus whenever $\left\{x_{n}\right\}$ is a Cauchy sequence, there exists $x$ such that

$$
\lim _{n \rightarrow \infty}\left\|x-x_{n}\right\|=0
$$

The following is an example of an infinite dimensional Banach space. We have already observed that finite dimensional normed linear spaces are Banach spaces.

Example 14.4.2 Let $E$ be a Banach space and let $\Omega$ be a nonempty subset of a normed linear space $F$. Let $B(\Omega ; E)$ denote those functions $f$ for which

$$
\|f\| \equiv \sup \left\{\|f(x)\|_{E}: x \in \Omega\right\}<\infty
$$

Denote by $B C(\Omega ; E)$ the set of functions in $B(\Omega ; E)$ which are also continuous.
Lemma 14.4.3 The above $\|\cdot\|$ is a norm on $B(\Omega ; E)$. The subspace $B C(\Omega ; E)$ with the given norm is a Banach space.

Proof: It is obvious $\|\cdot\|$ is a norm. It only remains to verify $B C(\Omega ; E)$ is complete. Let $\left\{f_{n}\right\}$ be a Cauchy sequence. Since $\left\|f_{n}-f_{m}\right\| \rightarrow 0$ as $m, n \rightarrow \infty$, it follows that $\left\{f_{n}(x)\right\}$ is a Cauchy sequence in $E$ for each $x$. Let $f(x) \equiv \lim _{n \rightarrow \infty} f_{n}(x)$. Then for any $x \in \Omega$.

$$
\left\|f_{n}(x)-f_{m}(x)\right\|_{E} \leq\left\|f_{n}-f_{m}\right\|<\varepsilon
$$

whenever $m, n$ are large enough, say as large as $N$. For $n \geq N$, let $m \rightarrow \infty$. Then passing to the limit, it follows that for all $x$,

$$
\left\|f_{n}(x)-f(x)\right\|_{E} \leq \varepsilon
$$

and so for all $x$,

$$
\|f(x)\|_{E} \leq \varepsilon+\left\|f_{n}(x)\right\|_{E} \leq \varepsilon+\left\|f_{n}\right\|
$$

It follows that $\|f\| \leq\left\|f_{n}\right\|+\varepsilon$ and $\left\|f-f_{n}\right\| \leq \varepsilon$.
It remains to verify that $f$ is continuous.

$$
\begin{aligned}
\|f(x)-f(y)\|_{E} & \leq\left\|f(x)-f_{n}(x)\right\|_{E}+\left\|f_{n}(x)-f_{n}(y)\right\|_{E}+\left\|f_{n}(y)-f(y)\right\|_{E} \\
& \leq 2\left\|f-f_{n}\right\|+\left\|f_{n}(x)-f_{n}(y)\right\|_{E}<\frac{2 \varepsilon}{3}+\left\|f_{n}(x)-f_{n}(y)\right\|_{E}
\end{aligned}
$$

for all $n$ large enough. Now pick such an $n$. By continuity, the last term is less than $\frac{\varepsilon}{3}$ if $\|x-y\|$ is small enough. Hence $f$ is continuous as well.

The most familiar example of a Banach space is $\mathbb{F}^{n}$. The following lemma is of great importance so it is stated in general.

Lemma 14.4.4 Suppose $T: E \rightarrow E$ where $E$ is a Banach space with norm $|\cdot|$. Also suppose

$$
\begin{equation*}
|T \boldsymbol{x}-T \boldsymbol{y}| \leq r|\boldsymbol{x}-\boldsymbol{y}| \tag{14.9}
\end{equation*}
$$

for some $r \in(0,1)$. Then there exists a unique fixed point, $\boldsymbol{x} \in E$ such that

$$
\begin{equation*}
T x=x \tag{14.10}
\end{equation*}
$$

Letting $\boldsymbol{x}^{1} \in E$, this fixed point $\boldsymbol{x}$, is the limit of the sequence of iterates,

$$
\begin{equation*}
\boldsymbol{x}^{1}, \boldsymbol{T} \boldsymbol{x}^{1}, \boldsymbol{T}^{2} \boldsymbol{x}^{1}, \cdots \tag{14.11}
\end{equation*}
$$

In addition to this, there is a nice estimate which tells how close $\boldsymbol{x}^{1}$ is to $\boldsymbol{x}$ in terms of things which can be computed.

$$
\begin{equation*}
\left|\boldsymbol{x}^{1}-\boldsymbol{x}\right| \leq \frac{1}{1-r}\left|\boldsymbol{x}^{1}-T \boldsymbol{x}^{1}\right| \tag{14.12}
\end{equation*}
$$

Proof: This follows easily when it is shown that the above sequence, $\left\{T^{k} \boldsymbol{x}^{1}\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Note that

$$
\left|T^{2} \boldsymbol{x}^{1}-T \boldsymbol{x}^{1}\right| \leq r\left|T \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right|
$$

Suppose

$$
\begin{equation*}
\left|T^{k} \boldsymbol{x}^{1}-T^{k-1} \boldsymbol{x}^{1}\right| \leq r^{k-1}\left|\boldsymbol{T} \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right| \tag{14.13}
\end{equation*}
$$

Then

$$
\begin{aligned}
\left|T^{k+1} \boldsymbol{x}^{1}-T^{k} \boldsymbol{x}^{1}\right| & \leq r\left|T^{k} \boldsymbol{x}^{1}-T^{k-1} \boldsymbol{x}^{1}\right| \\
& \leq r r^{k-1}\left|\boldsymbol{T} \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right|=r^{k}\left|\boldsymbol{T} \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right|
\end{aligned}
$$

By induction, this shows that for all $k \geq 2,14.13$ is valid. Now let $k>l \geq N$.

$$
\begin{aligned}
\left|T^{k} \boldsymbol{x}^{1}-T^{l} \boldsymbol{x}^{1}\right| & =\left|\sum_{j=l}^{k-1}\left(T^{j+1} \boldsymbol{x}^{1}-T^{j} \boldsymbol{x}^{1}\right)\right| \leq \sum_{j=l}^{k-1}\left|T^{j+1} \boldsymbol{x}^{1}-T^{j} \boldsymbol{x}^{1}\right| \\
& \leq \sum_{j=N}^{k-1} r^{j}\left|T \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right| \leq\left|T \boldsymbol{x}^{1}-\boldsymbol{x}^{1}\right| \frac{r^{N}}{1-r}
\end{aligned}
$$

which converges to 0 as $N \rightarrow \infty$. Therefore, this is a Cauchy sequence so it must converge to $\boldsymbol{x} \in E$. Then

$$
\boldsymbol{x}=\lim _{k \rightarrow \infty} T^{k} \boldsymbol{x}^{1}=\lim _{k \rightarrow \infty} T^{k+1} \boldsymbol{x}^{1}=T \lim _{k \rightarrow \infty} T^{k} \boldsymbol{x}^{1}=T \boldsymbol{x}
$$

This shows the existence of the fixed point. To show it is unique, suppose there were another one, $\boldsymbol{y}$. Then

$$
|\boldsymbol{x}-\boldsymbol{y}|=|T \boldsymbol{x}-\boldsymbol{T} \boldsymbol{y}| \leq r|\boldsymbol{x}-\boldsymbol{y}|
$$

and so $\boldsymbol{x}=\boldsymbol{y}$.
It remains to verify the estimate.

$$
\begin{aligned}
\left|\boldsymbol{x}^{1}-\boldsymbol{x}\right| & \leq\left|\boldsymbol{x}^{1}-\boldsymbol{T} \boldsymbol{x}^{1}\right|+\left|\boldsymbol{T} \boldsymbol{x}^{1}-\boldsymbol{x}\right|=\left|\boldsymbol{x}^{1}-\boldsymbol{T} \boldsymbol{x}^{1}\right|+\left|\boldsymbol{T} \boldsymbol{x}^{1}-\boldsymbol{T} \boldsymbol{x}\right| \\
& \leq\left|\boldsymbol{x}^{1}-\boldsymbol{T} \boldsymbol{x}^{1}\right|+r\left|\boldsymbol{x}^{1}-\boldsymbol{x}\right|
\end{aligned}
$$

and solving the inequality for $\left|\boldsymbol{x}^{1}-\boldsymbol{x}\right|$ gives the estimate desired.
The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

Corollary 14.4.5 Suppose $T: E \rightarrow E$, for some constant $C$

$$
|T \boldsymbol{x}-T \boldsymbol{y}| \leq C|\boldsymbol{x}-\boldsymbol{y}|
$$

for all $\boldsymbol{x}, \boldsymbol{y} \in E$, and for some $N \in \mathbb{N}$,

$$
\left|T^{N} \boldsymbol{x}-T^{N} \boldsymbol{y}\right| \leq r|\boldsymbol{x}-\boldsymbol{y}|
$$

for all $\boldsymbol{x}, \boldsymbol{y} \in E$ where $r \in(0,1)$. Then there exists a unique fixed point for $T$ and it is still the limit of the sequence, $\left\{T^{k} x^{1}\right\}$ for any choice of $x^{1}$.

Proof: From Lemma 14.4.4 there exists a unique fixed point for $T^{N}$ denoted here as $\boldsymbol{x}$. Therefore, $T^{N} \boldsymbol{x}=\boldsymbol{x}$. Now doing $T$ to both sides,

$$
T^{N} T \boldsymbol{x}=T \boldsymbol{x}
$$

By uniqueness, $\boldsymbol{T} \boldsymbol{x}=\boldsymbol{x}$ because the above equation shows $\boldsymbol{T} \boldsymbol{x}$ is a fixed point of $T^{N}$ and there is only one fixed point of $T^{N}$. In fact, there is only one fixed point of $T$ because a fixed point of $T$ is automatically a fixed point of $T^{N}$.

It remains to show $T^{k} \boldsymbol{x}^{1} \rightarrow \boldsymbol{x}$, the unique fixed point of $T^{N}$. If this does not happen, there exists $\varepsilon>0$ and a subsequence, still denoted by $T^{k}$ such that

$$
\left|T^{k} x^{1}-x\right| \geq \varepsilon
$$

Now $k=j_{k} N+r_{k}$ where $r_{k} \in\{0, \cdots, N-1\}$ and $j_{k}$ is a positive integer with $\lim _{k \rightarrow \infty} j_{k}=\infty$. Then there exists a single $r \in\{0, \cdots, N-1\}$ such that for infinitely many $k, r_{k}=r$. Taking a further subsequence, still denoted by $T^{k}$ it follows

$$
\begin{equation*}
\left|T^{j_{k} N+r} \boldsymbol{x}^{1}-\boldsymbol{x}\right| \geq \varepsilon \tag{14.14}
\end{equation*}
$$

However,

$$
T^{j_{k} N+r} \boldsymbol{x}^{1}=T^{r} T^{j_{k} N} \boldsymbol{x}^{1} \rightarrow T^{r} \boldsymbol{x}=\boldsymbol{x}
$$

and this contradicts 14.14.
Now return to our system $A \boldsymbol{x}=\boldsymbol{b}$. Recall it was a fixed point of $T$ where

$$
\boldsymbol{x}=B^{-1} C \boldsymbol{x}+B^{-1} \boldsymbol{b} \equiv T \boldsymbol{x}
$$

Then the fundamental theorem on convergence is the following. First note the following.

$$
\begin{aligned}
T^{2} \boldsymbol{x} & =B^{-1} C\left(B^{-1} C \boldsymbol{x}+\boldsymbol{b}\right)+B^{-1} \boldsymbol{b} \\
& =\left(B^{-1} C\right)^{2}+e_{2}(\boldsymbol{b})
\end{aligned}
$$

where $e_{2}(\boldsymbol{b})$ does not depend on $\boldsymbol{x}$. Similarly,

$$
\begin{equation*}
T^{n} \boldsymbol{x}=\left(B^{-1} C\right)^{n}+e_{n}(\boldsymbol{b}) \tag{14.15}
\end{equation*}
$$

where $e_{n}(\boldsymbol{b})$ does not depend on $\boldsymbol{x}$. Thus

$$
\left|T^{n} \boldsymbol{x}-T^{n} \boldsymbol{y}\right| \leq\left\|\left(B^{-1} C\right)^{n}\right\||\boldsymbol{x}-\boldsymbol{y}|
$$

Theorem 14.4.6 Suppose $\rho\left(B^{-1} C\right)<1$. Then the iterates described above converge to the unique solution of $A \boldsymbol{x}=\boldsymbol{b}$.

Proof: Consider the above iterates. Let $\boldsymbol{T} \boldsymbol{x}=\boldsymbol{B}^{-1} \boldsymbol{C} \boldsymbol{x}+B^{-1} \boldsymbol{b}$. Then

$$
\left|T^{k} \boldsymbol{x}-T^{k} \boldsymbol{y}\right|=\left|\left(B^{-1} C\right)^{k} \boldsymbol{x}-\left(B^{-1} C\right)^{k} \boldsymbol{y}\right| \leq\left\|\left(B^{-1} C\right)^{k}\right\||\boldsymbol{x}-\boldsymbol{y}|
$$

Here $\|\cdot\|$ refers to any of the operator norms. It doesn't matter which one you pick because they are all equivalent. I am writing the proof to indicate the operator norm taken with respect to the usual norm on $E$. Since $\rho\left(B^{-1} C\right)<1$, it follows from Gelfand's theorem, Theorem 14.2.4 on Page 392, there exists $N$ such that if $k \geq N$, then $\left\|\left(B^{-1} C\right)^{k}\right\| \leq r<1$. Consequently,

$$
\left|T^{N} \boldsymbol{x}-T^{N} \boldsymbol{y}\right| \leq r|\boldsymbol{x}-\boldsymbol{y}|
$$

Also $|T \boldsymbol{x}-T \boldsymbol{y}| \leq\left|\left|B^{-1} C\right|\right||\boldsymbol{x}-\boldsymbol{y}|$ and so Corollary 14.4.5 applies and gives the conclusion of this theorem.

In the Jacobi method, you have

$$
A=\left(\begin{array}{lll}
* & & * \\
& \ddots & \\
* & & *
\end{array}\right)
$$

and you let $B$ be the diagonal matrix whose diagonal entries are those of $A$ and you let $C$ be $(-1)$ times the matrix obtained from $A$ by making the diagonal entries 0 and retaining all the other entries of $A$. Thus

$$
B=\left(\begin{array}{ccc}
* & & 0 \\
& \ddots & \\
0 & & *
\end{array}\right), C=-\left(\begin{array}{ccc}
0 & & * \\
& \ddots & \\
* & & 0
\end{array}\right)
$$

In the Gauss Seidel method, you let

$$
B=\left(\begin{array}{ccc}
* & & 0 \\
& \ddots & \\
* & & *
\end{array}\right), C=-\left(\begin{array}{ccc}
0 & & * \\
& \ddots & \\
0 & & 0
\end{array}\right)
$$

Thus you keep the entries of $A$ which are on or below the main diagonal in order to get $B$. To get $C$ you take -1 times the matrix obtained from $A$ by replacing all entries below and on the main diagonal with zeros.

Observation 14.4.7 Note that if A is diagonally dominant, meaning

$$
\left|a_{i i}\right|>\sum_{j \neq i}\left|a_{i j}\right|
$$

then in both cases above, $\rho\left(B^{-1} C\right)<1$ so the two iterative procedures will converge.
To see this, suppose $B^{-1} C \boldsymbol{x}=\lambda \boldsymbol{x},|\lambda| \geq 1$. Then you get $(\lambda B-C) \boldsymbol{x}=\mathbf{0}$ However, in either the case of Jacobi iteration or Gauss Seidel iteration, the matrix $\lambda B-C$ will be diagonally dominant and so by Gerschgorin's theorem will have no zero eigenvalues which requires that this matrix be one to one. Thus there are no eigenvectors for such $\lambda$ and hence $\rho\left(B^{-1} C\right)<1$.

### 14.5 Exercises

1. Solve the system

$$
\left(\begin{array}{lll}
4 & 1 & 1 \\
1 & 5 & 2 \\
0 & 2 & 6
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
3
\end{array}\right)
$$

using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.
2. Solve the system

$$
\left(\begin{array}{lll}
4 & 1 & 1 \\
1 & 7 & 2 \\
0 & 2 & 4
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
3
\end{array}\right)
$$

using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.
3. Solve the system

$$
\left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 7 & 2 \\
0 & 2 & 4
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
2 \\
3
\end{array}\right)
$$

using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.
4. If you are considering a system of the form $\boldsymbol{A} \boldsymbol{x}=\boldsymbol{b}$ and $A^{-1}$ does not exist, will either the Gauss Seidel or Jacobi methods work? Explain. What does this indicate about finding eigenvectors for a given eigenvalue?
5. For $\|x\|_{\infty} \equiv \max \left\{\left|x_{j}\right|: j=1,2, \cdots, n\right\}$, the parallelogram identity does not hold. Explain.
6. A norm $\|\cdot\|$ is said to be strictly convex if whenever $\|x\|=\|y\|, x \neq y$, it follows

$$
\left\|\frac{x+y}{2}\right\|<\|x\|=\|y\| .
$$

Show the norm $|\cdot|$ which comes from an inner product is strictly convex.
7. A norm $\|\cdot\|$ is said to be uniformly convex if whenever $\left\|x_{n}\right\|,\left\|y_{n}\right\|$ are equal to 1 for all $n \in \mathbb{N}$ and $\lim _{n \rightarrow \infty}\left\|x_{n}+y_{n}\right\|=2$, it follows $\lim _{n \rightarrow \infty}\left\|x_{n}-y_{n}\right\|=0$. Show the norm $|\cdot|$ coming from an inner product is always uniformly convex. Also show that uniform convexity implies strict convexity which is defined in Problem 6.
8. Suppose $A: \mathbb{C}^{n} \rightarrow \mathbb{C}^{n}$ is a one to one and onto matrix. Define

$$
\|x\| \equiv|A x|
$$

Show this is a norm.
9. If $X$ is a finite dimensional normed vector space and $A, B \in \mathscr{L}(X, X)$ such that $\|B\|<$ $\|A\|$ and $A^{-1}$ exists, can it be concluded that $\left\|A^{-1} B\right\|<1$ ? Either give a counter example or a proof.
10. Let $X$ be a vector space with a norm $\|\cdot\|$ and let $V=\operatorname{span}\left(v_{1}, \cdots, v_{m}\right)$ be a finite dimensional subspace of $X$ such that $\left\{v_{1}, \cdots, v_{m}\right\}$ is a basis for $V$. Show $V$ is a closed subspace of $X$. This means that if $w_{n} \rightarrow w$ and each $w_{n} \in V$, then so is $w$. Next show that if $w \notin V$,

$$
\operatorname{dist}(w, V) \equiv \inf \{\|w-v\|: v \in V\}>0
$$

is a continuous function of $w$ and

$$
\left|\operatorname{dist}(w, V)-\operatorname{dist}\left(w_{1}, V\right)\right| \leq\left\|w_{1}-w\right\|
$$

Next show that if $w \notin V$, there exists $z$ such that $\|z\|=1$ and dist $(z, V)>1 / 2$. For those who know some advanced calculus, show that if $X$ is an infinite dimensional vector space having norm $\|\cdot\|$, then the closed unit ball in $X$ cannot be compact. Thus closed and bounded is never compact in an infinite dimensional normed vector space.
11. Suppose $\rho(A)<1$ for $A \in \mathscr{L}(V, V)$ where $V$ is a $p$ dimensional vector space having a norm $\|\cdot\|$. You can use $\mathbb{R}^{p}$ or $\mathbb{C}^{p}$ if you like. Show there exists a new norm $\|\|\cdot\|\|$ such that with respect to this new norm, $\|\|A\|\|<1$ where $|\|A \mid\|$ denotes the operator norm of $A$ taken with respect to this new norm on $V$,

$$
|\|A\|| \equiv \sup \{\mid\|A \boldsymbol{x}\|\|:\| \boldsymbol{x}\| \| \leq 1\}
$$

Hint: You know from Gelfand's theorem that

$$
\left\|A^{n}\right\|^{1 / n}<r<1
$$

provided $n$ is large enough, this operator norm taken with respect to $\|\cdot\|$. Show there exists $0<\lambda<1$ such that

$$
\rho\left(\frac{A}{\lambda}\right)<1 \text {. }
$$

You can do this by arguing the eigenvalues of $A / \lambda$ are the scalars $\mu / \lambda$ where $\mu \in$ $\sigma(A)$. Now let $\mathbb{Z}_{+}$denote the nonnegative integers.

$$
\||\boldsymbol{x}|\| \equiv \sup _{n \in \mathbb{Z}_{+}}\left\|\frac{A^{n}}{\lambda^{n}} \boldsymbol{x}\right\|
$$

First show this is actually a norm. Next explain why

$$
\|\mid A \boldsymbol{x}\|\left\|\equiv \lambda \sup _{n \in \mathbb{Z}_{+}}\right\| \frac{A^{n+1}}{\lambda^{n+1}} \boldsymbol{x}\|\leq \lambda\|\|\boldsymbol{x}\| \| .
$$

12. Establish a similar result to Problem 11 without using Gelfand's theorem. Use an argument which depends directly on the Jordan form or a modification of it.
13. Using Problem 11 give an easier proof of Theorem 14.4.6 without having to use Corollary 14.4 .5 . It would suffice to use a different norm of this problem and the contraction mapping principle of Lemma 14.4.4.
14. A matrix $A$ is diagonally dominant if $\left|a_{i i}\right|>\sum_{j \neq i}\left|a_{i j}\right|$. Show that the Gauss Seidel method converges if $A$ is diagonally dominant.
15. Suppose $f(\lambda)=\sum_{n=0}^{\infty} a_{n} \lambda^{n}$ converges if $|\lambda|<R$. Show that if $\rho(A)<R$ where $A$ is an $n \times n$ matrix, then

$$
f(A) \equiv \sum_{n=0}^{\infty} a_{n} A^{n}
$$

converges in $\mathscr{L}\left(\mathbb{F}^{n}, \mathbb{F}^{n}\right)$. Hint: Use Gelfand's theorem and the root test.
16. Referring to Corollary 14.3 .5 , for $\lambda=a+i b$ show

$$
\exp (\lambda t)=e^{a t}(\cos (b t)+i \sin (b t))
$$

Hint: Let $y(t)=\exp (\lambda t)$ and let $z(t)=e^{-a t} y(t)$. Show

$$
z^{\prime \prime}+b^{2} z=0, z(0)=1, z^{\prime}(0)=i b
$$

Now letting $z=u+i v$ where $u, v$ are real valued, show

$$
\begin{aligned}
u^{\prime \prime}+b^{2} u & =0, u(0)=1, u^{\prime}(0)=0 \\
v^{\prime \prime}+b^{2} v & =0, v(0)=0, v^{\prime}(0)=b
\end{aligned}
$$

Next show $u(t)=\cos (b t)$ and $v(t)=\sin (b t)$ work in the above and that there is at most one solution to

$$
w^{\prime \prime}+b^{2} w=0 w(0)=\alpha, w^{\prime}(0)=\beta
$$

Thus $z(t)=\cos (b t)+i \sin (b t)$ and so $y(t)=e^{a t}(\cos (b t)+i \sin (b t))$. To show there is at most one solution to the above problem, suppose you have two, $w_{1}, w_{2}$. Subtract them. Let $f=w_{1}-w_{2}$. Thus

$$
f^{\prime \prime}+b^{2} f=0
$$

and $f$ is real valued. Multiply both sides by $f^{\prime}$ and conclude

$$
\frac{d}{d t}\left(\frac{\left(f^{\prime}\right)^{2}}{2}+b^{2} \frac{f^{2}}{2}\right)=0
$$

Thus the expression in parenthesis is constant. Explain why this constant must equal 0.
17. Let $A \in \mathscr{L}\left(\mathbb{R}^{n}, \mathbb{R}^{n}\right)$. Show the following power series converges in $\mathscr{L}\left(\mathbb{R}^{n}, \mathbb{R}^{n}\right)$.

$$
\Psi(t) \equiv \sum_{k=0}^{\infty} \frac{t^{k} A^{k}}{k!}
$$

This was done in the chapter. Go over it and be sure you understand it. This is how you can define $\exp (t A)$. Next show that $\Psi^{\prime}(t)=A \Psi(t), \Psi(0)=I$. Next let $\Phi(t)=\sum_{k=0}^{\infty} \frac{t^{k}(-A)^{k}}{k!}$. Show each $\Phi(t), \Psi(t)$ each commute with $A$. Next show that $\Phi(t) \Psi(t)=I$ for all $t$. Finally, solve the initial value problem

$$
x^{\prime}=A x+f, x(0)=x_{0}
$$

in terms of $\Phi$ and $\Psi$. This yields most of the substance of a typical differential equations course.
18. In Problem $17 \Psi(t)$ is defined by the given series. Denote by $\exp (t \sigma(A))$ the numbers $\exp (t \lambda)$ where $\lambda \in \sigma(A)$. Show $\exp (t \sigma(A))=\sigma(\Psi(t))$. This is like Lemma 14.3.9. Letting $J$ be the Jordan canonical form for $A$, explain why

$$
\Psi(t) \equiv \sum_{k=0}^{\infty} \frac{t^{k} A^{k}}{k!}=S \sum_{k=0}^{\infty} \frac{t^{k} J^{k}}{k!} S^{-1}
$$

and you note that in $J^{k}$, the diagonal entries are of the form $\lambda^{k}$ for $\lambda$ an eigenvalue of $A$. Also $J=D+N$ where $N$ is nilpotent and commutes with $D$. Argue then that

$$
\sum_{k=0}^{\infty} \frac{t^{k} J^{k}}{k!}
$$

is an upper triangular matrix which has on the diagonal the expressions $e^{\lambda t}$ where $\lambda \in \sigma(A)$. Thus conclude

$$
\sigma(\Psi(t)) \subseteq \exp (t \sigma(A))
$$

Next take $e^{t \lambda} \in \exp (t \sigma(A))$ and argue it must be in $\sigma(\Psi(t))$. You can do this as follows:

$$
\begin{aligned}
\Psi(t)-e^{t \lambda} I & =\sum_{k=0}^{\infty} \frac{t^{k} A^{k}}{k!}-\sum_{k=0}^{\infty} \frac{t^{k} \lambda^{k}}{k!} I=\sum_{k=0}^{\infty} \frac{t^{k}}{k!}\left(A^{k}-\lambda^{k} I\right) \\
& =\left(\sum_{k=0}^{\infty} \frac{t^{k}}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^{j}\right)(A-\lambda I)
\end{aligned}
$$

Now you need to argue

$$
\sum_{k=0}^{\infty} \frac{t^{k}}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^{j}
$$

converges to something in $\mathscr{L}\left(\mathbb{R}^{n}, \mathbb{R}^{n}\right)$. To do this, use the ratio test and Lemma 14.3.2 after first using the triangle inequality. Since $\lambda \in \sigma(A), \Psi(t)-e^{t \lambda} I$ is not one to one and so this establishes the other inclusion. You fill in the details. This theorem is a special case of theorems which go by the name "spectral mapping theorem" which was discussed in the text. However, go through it yourself.
19. Suppose $\Psi(t) \in \mathscr{L}(V, W)$ where $V, W$ are finite dimensional inner product spaces and $t \rightarrow \Psi(t)$ is continuous for $t \in[a, b]$ : For every $\varepsilon>0$ there there exists $\delta>0$ such that if $|s-t|<\delta$ then $\|\Psi(t)-\Psi(s)\|<\varepsilon$. Show $t \rightarrow(\Psi(t) v, w)$ is continuous. Here it is the inner product in $W$. Also define what it means for $t \rightarrow \Psi(t) v$ to be continuous and show this is continuous. Do it all for differentiable in place of continuous. Next show $t \rightarrow\|\Psi(t)\|$ is continuous.
20. If $z(t) \in W$, a finite dimensional inner product space, what does it mean for $t \rightarrow z(t)$ to be continuous or differentiable? If $z$ is continuous, define

$$
\int_{a}^{b} z(t) d t \in W
$$

as follows.

$$
\left(w, \int_{a}^{b} z(t) d t\right) \equiv \int_{a}^{b}(w, z(t)) d t .
$$

Show that this definition is well defined and furthermore the triangle inequality,

$$
\left|\int_{a}^{b} z(t) d t\right| \leq \int_{a}^{b}|z(t)| d t
$$

and fundamental theorem of calculus,

$$
\frac{d}{d t}\left(\int_{a}^{t} z(s) d s\right)=z(t)
$$

hold along with any other interesting properties of integrals which are true.
21. For $V, W$ two inner product spaces, define

$$
\int_{a}^{b} \Psi(t) d t \in \mathscr{L}(V, W)
$$

as follows.

$$
\left(w, \int_{a}^{b} \Psi(t) d t(v)\right) \equiv \int_{a}^{b}(w, \Psi(t) v) d t
$$

Show this is well defined and does indeed give $\int_{a}^{b} \Psi(t) d t \in \mathscr{L}(V, W)$. Also show the triangle inequality

$$
\left\|\int_{a}^{b} \Psi(t) d t\right\| \leq \int_{a}^{b}\|\Psi(t)\| d t
$$

where $\|\cdot\|$ is the operator norm and verify the fundamental theorem of calculus holds.

$$
\left(\int_{a}^{t} \Psi(s) d s\right)^{\prime}=\Psi(t)
$$

Also verify the usual properties of integrals continue to hold such as the fact the integral is linear and

$$
\int_{a}^{b} \Psi(t) d t+\int_{b}^{c} \Psi(t) d t=\int_{a}^{c} \Psi(t) d t
$$

and similar things. Hint: On showing the triangle inequality, it will help if you use the fact that

$$
|w|_{W}=\sup _{|v| \leq 1}|(w, v)| .
$$

You should show this also.
22. Prove Gronwall's inequality. Suppose $u(t) \geq 0$ and for all $t \in[0, T]$,

$$
u(t) \leq u_{0}+\int_{0}^{t} K u(s) d s
$$

where $K$ is some nonnegative constant. Then

$$
u(t) \leq u_{0} e^{K t}
$$

Hint: $w(t)=\int_{0}^{t} u(s) d s$. Then using the fundamental theorem of calculus, $w(t)$ satisfies the following.

$$
u(t)-K w(t)=w^{\prime}(t)-K w(t) \leq u_{0}, w(0)=0
$$

Now use the usual techniques you saw in an introductory differential equations class. Multiply both sides of the above inequality by $e^{-K t}$ and note the resulting left side is now a total derivative. Integrate both sides from 0 to $t$ and see what you have got.
23. With Gronwall's inequality and the integral defined in Problem 21 with its properties listed there, prove there is at most one solution to the initial value problem

$$
\boldsymbol{y}^{\prime}=A \boldsymbol{y}, \boldsymbol{y}(0)=\boldsymbol{y}_{0} .
$$

Hint: If there are two solutions, subtract them and call the result $\boldsymbol{z}$. Then

$$
z^{\prime}=A \boldsymbol{z}, \boldsymbol{z}(0)=\mathbf{0}
$$

It follows

$$
\boldsymbol{z}(t)=0+\int_{0}^{t} A \boldsymbol{z}(s) d s
$$

and so

$$
\|\boldsymbol{z}(t)\| \leq \int_{0}^{t}\|A\|\|\boldsymbol{z}(s)\| d s
$$

Now consider Gronwall's inequality of Problem 22.
24. Suppose $A$ is a matrix which has the property that whenever $\mu \in \sigma(A), \operatorname{Re} \mu<0$. Consider the initial value problem

$$
\boldsymbol{y}^{\prime}=A \boldsymbol{y}, \boldsymbol{y}(0)=\boldsymbol{y}_{0}
$$

The existence and uniqueness of a solution to this equation has been established above in preceding problems, Problem 17 to 23 . Show that in this case where the real parts of the eigenvalues are all negative, the solution to the initial value problem satisfies

$$
\lim _{t \rightarrow \infty} \boldsymbol{y}(t)=\mathbf{0}
$$

Hint: A nice way to approach this problem is to show you can reduce it to the consideration of the initial value problem

$$
z^{\prime}=J_{\varepsilon} z, z(0)=z_{0}
$$

where $J_{\varepsilon}$ is the modified Jordan canonical form where instead of ones down the main diagonal, there are $\varepsilon$ down the main diagonal (Problem 14). Then

$$
z^{\prime}=D z+N_{\varepsilon} \boldsymbol{z}
$$

where $D$ is the diagonal matrix obtained from the eigenvalues of $A$ and $N_{\varepsilon}$ is a nilpotent matrix commuting with $D$ which is very small provided $\varepsilon$ is chosen very small. Now let $\Psi(t)$ be the solution of

$$
\Psi^{\prime}=-D \Psi, \Psi(0)=I
$$

described earlier as

$$
\sum_{k=0}^{\infty} \frac{(-1)^{k} t^{k} D^{k}}{k!}
$$

Thus $\Psi(t)$ commutes with $D$ and $N_{\mathcal{\varepsilon}}$. Tell why. Next argue

$$
(\Psi(t) \boldsymbol{z})^{\prime}=\Psi(t) N_{\varepsilon} \boldsymbol{z}(t)
$$

and integrate from 0 to $t$. Then

$$
\Psi(t) \boldsymbol{z}(t)-\boldsymbol{z}_{0}=\int_{0}^{t} \Psi(s) N_{\mathcal{\varepsilon}} \boldsymbol{z}(s) d s
$$

It follows

$$
\|\Psi(t) \boldsymbol{z}(t)\| \leq\left\|z_{0}\right\|+\int_{0}^{t}\left\|N_{\mathcal{E}}\right\|\|\Psi(s) \boldsymbol{z}(s)\| d s
$$

It follows from Gronwall's inequality

$$
\|\Psi(t) \boldsymbol{z}(t)\| \leq\left\|z_{0}\right\| e^{\left\|N_{\varepsilon}\right\| t}
$$

Now look closely at the form of $\Psi(t)$ to get an estimate which is interesting. Explain why

$$
\Psi(t)=\left(\begin{array}{ccc}
e^{\mu_{1} t} & & 0 \\
& \ddots & \\
0 & & e^{\mu_{n} t}
\end{array}\right)
$$

and now observe that if $\varepsilon$ is chosen small enough, $\left\|N_{\mathcal{\varepsilon}}\right\|$ is so small that each component of $\boldsymbol{z}(t)$ converges to 0 .
25. Using Problem 24 show that if $A$ is a matrix having the real parts of all eigenvalues less than 0 then if

$$
\Psi^{\prime}(t)=A \Psi(t), \Psi(0)=I
$$

it follows

$$
\lim _{t \rightarrow \infty} \Psi(t)=0
$$

Hint: Consider the columns of $\Psi(t)$ ?
26. Let $\Psi(t)$ be a fundamental matrix satisfying

$$
\Psi^{\prime}(t)=A \Psi(t), \Psi(0)=I
$$

Show $\Psi(t)^{n}=\Psi(n t)$. Hint: Subtract and show the difference satisfies $\Phi^{\prime}=A \Phi$, and $\Phi(0)=0$. Use uniqueness.
27. If the real parts of the eigenvalues of $A$ are all negative, show that for every positive $t$,

$$
\lim _{n \rightarrow \infty} \Psi(n t)=0
$$

Hint: Pick $\operatorname{Re}(\sigma(A))<-\lambda<0$ and use Problem 18 about the spectrum of $\Psi(t)$ and Gelfand's theorem for the spectral radius along with Problem 26 to argue that $\left\|\Psi(n t) / e^{-\lambda n t}\right\|<1$ for all $n$ large enough.
28. Let $H$ be a Hermitian matrix. $\left(H=H^{*}\right)$. Show that $e^{i H} \equiv \sum_{n=0}^{\infty} \frac{(i H)^{n}}{n!}$ is unitary.
29. Show the converse of the above exercise. If $V$ is unitary, then $V=e^{i H}$ for some $H$ Hermitian.

Hint: First verify that $V$ is normal. Thus $U^{*} V U=D$. Now verify that $D^{*} D=I$. What does this mean for the diagonal entries of $D$ ? If you have a complex number which has magnitude 1 , what form does it take?
30. If $U$ is unitary and does not have -1 as an eigenvalue so that $(I+U)^{-1}$ exists, show that

$$
H=i(I-U)(I+U)^{-1}
$$

is Hermitian. Then, verify that

$$
U=(I+i H)(I-i H)^{-1}
$$

31. Suppose that $A \in \mathscr{L}(V, V)$ where $V$ is a normed linear space. Also suppose that $\|A\|<1$ where this refers to the operator norm on $A$. Verify that

$$
(I-A)^{-1}=\sum_{i=0}^{\infty} A^{i}
$$

This is called the Neumann series. Suppose now that you only know the algebraic condition $\rho(A)<1$. Is it still the case that the Neumann series converges to $(I-A)^{-1}$ ?

## Chapter 15

## Numerical Methods, Eigenvalues

### 15.1 The Power Method For Eigenvalues

This chapter discusses numerical methods for finding eigenvalues. However, to do this correctly, you must include numerical analysis considerations which are distinct from linear algebra. The purpose of this chapter is to give an introduction to some numerical methods without leaving the context of linear algebra. In addition, some examples are given which make use of computer algebra systems. For a more thorough discussion, you should see books on numerical methods in linear algebra like some listed in the references.

Let $A$ be a complex $p \times p$ matrix and suppose that it has distinct eigenvalues

$$
\left\{\lambda_{1}, \cdots, \lambda_{m}\right\}
$$

and that $\left|\lambda_{1}\right|>\left|\lambda_{k}\right|$ for all $k$. Also let the Jordan form of $A$ be

$$
J=\left(\begin{array}{lll}
J_{1} & & \\
& \ddots & \\
& & J_{m}
\end{array}\right)
$$

with $J_{1}$ an $m_{1} \times m_{1}$ matrix. $J_{k}=\lambda_{k} I_{k}+N_{k}$ where $N_{k}^{r_{k}} \neq 0$ but $N_{k}^{r_{k}+1}=0$. Also let $P^{-1} A P=$ $J, A=P J P^{-1}$. Now fix $\boldsymbol{x} \in \mathbb{F}^{p}$. Take $A \boldsymbol{x}$ and let $s_{1}$ be the entry of the vector $A \boldsymbol{x}$ which has largest absolute value. Thus $A \boldsymbol{x} / s_{1}$ is a vector $\boldsymbol{y}_{1}$ which has a component of 1 and every other entry of this vector has magnitude no larger than 1 . If the scalars $\left\{s_{1}, \cdots, s_{n-1}\right\}$ and vectors $\left\{\boldsymbol{y}_{1}, \cdots, \boldsymbol{y}_{n-1}\right\}$ have been obtained, let $\boldsymbol{y}_{n} \equiv A \boldsymbol{y}_{n-1} / s_{n}$ where $s_{n}$ is the entry of $A \boldsymbol{y}_{n-1}$ which has largest absolute value. Thus

$$
\begin{gather*}
\boldsymbol{y}_{n}=\frac{A A \boldsymbol{y}_{n-2}}{s_{n} s_{n-1}} \cdots=\frac{A^{n} \boldsymbol{x}}{s_{n} s_{n-1} \cdots s_{1}}=  \tag{15.1}\\
\frac{1}{s_{n} s_{n-1} \cdots s_{1}} P\left(\begin{array}{ccc}
J_{1}^{n} & & \\
& \ddots & \\
& & J_{m}^{n}
\end{array}\right) P^{-1} \boldsymbol{x}= \\
\frac{\lambda_{1}^{n}}{s_{n} s_{n-1} \cdots s_{1}} P\left(\begin{array}{lll}
\lambda_{1}^{-n} J_{1}^{n} & & \\
& \ddots & \\
& & \lambda_{1}^{-n} J_{m}^{n}
\end{array}\right) P^{-1} \boldsymbol{x} \tag{15.2}
\end{gather*}
$$

Consider one of the blocks in the Jordan form. First consider the $k^{t h}$ of these blocks, $k>1$. It equals

$$
\lambda_{1}^{-n} J_{k}^{n}=\sum_{i=0}^{r_{k}}\binom{n}{i} \lambda_{1}^{-n} \lambda_{k}^{n-i} N_{k}^{i}
$$

which clearly converges to 0 as $n \rightarrow \infty$ since $\left|\lambda_{1}\right|>\left|\lambda_{k}\right|$. An application of the ratio test or root test for each term in the sum will show this. When $k=1$, this block is

$$
\lambda_{1}^{-n} J_{1}^{n}=\lambda_{1}^{-n} J_{k}^{n}=\sum_{i=0}^{r_{1}}\binom{n}{i} \lambda_{1}^{-n} \lambda_{1}^{n-i} N_{1}^{i}=\binom{n}{r_{1}}\left[\lambda_{1}^{-r_{1}} N_{1}^{r_{1}}+e_{n}\right]
$$

where $\lim _{n \rightarrow \infty} e_{n}=0$ because it is a sum of bounded matrices which are multiplied by $\binom{n}{i} /\binom{n}{r_{1}}$. This quotient converges to 0 as $n \rightarrow \infty$ because $i<r_{1}$. It follows that 15.2 is of the form

$$
\boldsymbol{y}_{n}=\frac{\lambda_{1}^{n}}{s_{n} s_{n-1} \cdots s_{1}}\binom{n}{r_{1}} P\left(\begin{array}{cc}
\lambda_{1}^{-r_{1}} N_{1}^{r_{1}}+e_{n} & 0 \\
0 & E_{n}
\end{array}\right) P^{-1} \boldsymbol{x} \equiv \frac{\lambda_{1}^{n}}{s_{n} s_{n-1} \cdots s_{1}}\binom{n}{r_{1}} \boldsymbol{w}_{n}
$$

where $E_{n} \rightarrow 0, e_{n} \rightarrow 0$. Let $\left(P^{-1} \boldsymbol{x}\right)_{m_{1}}$ denote the first $m_{1}$ entries of the vector $P^{-1} \boldsymbol{x}$. Unless a very unlucky choice for $\boldsymbol{x}$ was picked, it will follow that $\left(P^{-1} \boldsymbol{x}\right)_{m_{1}} \notin \operatorname{ker}\left(N_{1}^{r_{1}}\right)$. Then for large $n, \boldsymbol{y}_{n}$ is close to the vector

$$
\frac{\lambda_{1}^{n}}{s_{n} s_{n-1} \cdots s_{1}}\binom{n}{r_{1}} P\left(\begin{array}{cc}
\lambda_{1}^{-r_{1}} N_{1}^{r_{1}} & 0 \\
0 & 0
\end{array}\right) P^{-1} \boldsymbol{x} \equiv \frac{\lambda_{1}^{n}}{s_{n} s_{n-1} \cdots s_{1}}\binom{n}{r_{1}} \boldsymbol{w} \equiv \boldsymbol{z} \neq \mathbf{0}
$$

However, this is an eigenvector because

$$
\begin{aligned}
& \left(A-\lambda_{1} I\right) \boldsymbol{w}=\overbrace{P\left(J-\lambda_{1} I\right) P^{-1}}^{A-\lambda_{1} I} P\left(\begin{array}{ccc}
\lambda_{1}^{-r_{1}} N_{1}^{r_{1}} & 0 \\
0 & 0
\end{array}\right) P^{-1} \boldsymbol{x}= \\
& P\left(\begin{array}{ccc}
N_{1} & & \\
& \ddots & \\
& & J_{m}-\lambda_{1} I
\end{array}\right) P^{-1} P\left(\begin{array}{cc}
\lambda_{1}^{-r_{1}} N_{1}^{r_{1}} & \\
& \ddots \\
& \\
& \\
& \\
& \\
& \\
& \\
N_{1} \lambda_{1}^{-r_{1}} N_{1}^{r_{1}} & 0 \\
0 & 0
\end{array}\right) P^{-1} \boldsymbol{x}=\mathbf{0}
\end{aligned}
$$

Recall $N_{1}^{r_{1}+1}=0$. Now you could recover an approximation to the eigenvalue as follows.

$$
\frac{\left(A \boldsymbol{y}_{n}, \boldsymbol{y}_{n}\right)}{\left(\boldsymbol{y}_{n}, \boldsymbol{y}_{n}\right)} \approx \frac{(A \boldsymbol{z}, \boldsymbol{z})}{(\boldsymbol{z}, \boldsymbol{z})}=\lambda_{1}
$$

Here $\approx$ means "approximately equal". However, there is a more convenient way to identify the eigenvalue in terms of the scaling factors $s_{k}$.

$$
\left\|\frac{\lambda_{1}^{n}}{s_{n} \cdots s_{1}}\binom{n}{r_{1}}\left(\boldsymbol{w}_{n}-\boldsymbol{w}\right)\right\|_{\infty} \approx 0
$$

Pick the largest nonzero entry of $\boldsymbol{w}, w_{l}$. Then for large $n$, it is also likely the case that the largest entry of $\boldsymbol{w}_{n}$ will be in the $l^{\text {th }}$ position because $\boldsymbol{w}_{m}$ is close to $\boldsymbol{w}$. From the construction,

$$
\frac{\lambda_{1}^{n}}{s_{n} \cdots s_{1}}\binom{n}{r_{1}} w_{n l}=1 \approx \frac{\lambda_{1}^{n}}{s_{n} \cdots s_{1}}\binom{n}{r_{1}} w_{l}
$$

In other words, for large $n, \frac{\lambda_{1}^{n}}{s_{n} \cdots s_{1}}\binom{n}{r_{1}} \approx 1 / w_{l}$. Therefore, for large $n$,

$$
\frac{\lambda_{1}^{n}}{s_{n} \cdots s_{1}}\binom{n}{r_{1}} \approx \frac{\lambda_{1}^{n+1}}{s_{n+1} s_{n} \cdots s_{1}}\binom{n+1}{r_{1}}
$$

and so $\binom{n}{r_{1}} /\binom{n+1}{r_{1}} \approx \frac{\lambda_{1}}{s_{n+1}}$. But $\lim _{n \rightarrow \infty}\binom{n}{r_{1}} /\binom{n+1}{r_{1}}=1$ and so, for large $n$ it must be the case that $\lambda_{1} \approx s_{n+1}$.

This has proved the following theorem which justifies the power method.
Theorem 15.1.1 Let A be a complex $p \times p$ matrix such that the eigenvalues are

$$
\left\{\lambda_{1}, \lambda_{2}, \cdots, \lambda_{r}\right\}
$$

with $\left|\lambda_{1}\right|>\left|\lambda_{j}\right|$ for all $j \neq 1$. Then for $\boldsymbol{x}$ a given vector, let

$$
\boldsymbol{y}_{1}=\frac{A \boldsymbol{x}}{s_{1}}
$$

where $s_{1}$ is an entry of $A \boldsymbol{x}$ which has the largest absolute value. If the scalars $\left\{s_{1}, \cdots, s_{n-1}\right\}$ and vectors $\left\{\boldsymbol{y}_{1}, \cdots, \boldsymbol{y}_{n-1}\right\}$ have been obtained, let

$$
\boldsymbol{y}_{n} \equiv \frac{A \boldsymbol{y}_{n-1}}{s_{n}}
$$

where $s_{n}$ is the entry of $A \boldsymbol{y}_{n-1}$ which has largest absolute value. Then it is probably the case that $\left\{s_{n}\right\}$ will converge to $\lambda_{1}$ and $\left\{\boldsymbol{y}_{n}\right\}$ will converge to an eigenvector associated with $\lambda_{1}$. If it doesn't, you picked an incredibly inauspicious initial vector $\boldsymbol{x}$.

In summary, here is the procedure.

## Finding the largest eigenvalue with its eigenvector.

1. Start with a vector, $\boldsymbol{u}_{1}$ which you hope is not unlucky.
2. If $\boldsymbol{u}_{k}$ is known, $\boldsymbol{u}_{k+1}=\frac{A \boldsymbol{u}_{k}}{s_{k+1}}$ where $s_{k+1}$ is the entry of $A \boldsymbol{u}_{k}$ which has largest absolute value.
3. When the scaling factors $s_{k}$ are not changing much, $s_{k+1}$ will be close to the eigenvalue and $\boldsymbol{u}_{k+1}$ will be close to an eigenvector.
4. Check your answer to see if it worked well. If things don't work well, try another $\boldsymbol{u}_{1}$. You were miraculously unlucky in your choice.

Example 15.1.2 Find the largest eigenvalue of $A=\left(\begin{array}{ccc}5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3\end{array}\right)$.
You can begin with $\boldsymbol{u}_{1}=(1, \cdots, 1)^{T}$ and apply the above procedure. However, you can accelerate the process if you begin with $A^{n} \boldsymbol{u}_{1}$ and then divide by the largest entry to get the first approximate eigenvector. Thus

$$
\left(\begin{array}{ccc}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{array}\right)^{20}\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)=\left(\begin{array}{c}
2.5558 \times 10^{21} \\
-1.2779 \times 10^{21} \\
-3.6562 \times 10^{15}
\end{array}\right)
$$

Divide by the largest entry to obtain a good aproximation.

$$
\left(\begin{array}{c}
2.5558 \times 10^{21} \\
-1.2779 \times 10^{21} \\
-3.6562 \times 10^{15}
\end{array}\right) \frac{1}{2.5558 \times 10^{21}}=\left(\begin{array}{c}
1.0 \\
-0.5 \\
-1.4306 \times 10^{-6}
\end{array}\right)
$$

Now begin with this one.

$$
\left(\begin{array}{ccc}
5 & -14 & 11 \\
-4 & 4 & -4 \\
3 & 6 & -3
\end{array}\right)\left(\begin{array}{c}
1.0 \\
-0.5 \\
-1.4306 \times 10^{-6}
\end{array}\right)=\left(\begin{array}{c}
12.000 \\
-6.0000 \\
4.2918 \times 10^{-6}
\end{array}\right)
$$

Divide by 12 to get the next iterate.

$$
\left(\begin{array}{c}
12.000 \\
-6.0000 \\
4.2918 \times 10^{-6}
\end{array}\right) \frac{1}{12}=\left(\begin{array}{c}
1.0 \\
-0.5 \\
3.5765 \times 10^{-7}
\end{array}\right)
$$

Another iteration will reveal that the scaling factor is still 12 . Thus this is an approximate eigenvalue. In fact, it is the largest eigenvalue and the corresponding eigenvector is $\left(\begin{array}{lll}1.0 & -0.5 & 0\end{array}\right)$. The process has worked very well.

### 15.1.1 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. One can find complex eigenvalues using this method. The situation is this: You have a number $\alpha$ which is close to $\lambda$, some eigenvalue of an $n \times n$ matrix $A$. You don't know $\lambda$ but you know that $\alpha$ is closer to $\lambda$ than to any other eigenvalue. Your problem is to find both $\lambda$ and an eigenvector which goes with $\lambda$. Another way to look at this is to start with $\alpha$ and seek the eigenvalue $\lambda$, which is closest to $\alpha$ along with an eigenvector associated with $\lambda$. If $\alpha$ is an eigenvalue of $A$, then you have what you want. Therefore, I will always assume $\alpha$ is not an eigenvalue of $A$ and so $(A-\alpha I)^{-1}$ exists. The method is based on the following lemma.
Lemma 15.1.3 Let $\left\{\lambda_{k}\right\}_{k=1}^{n}$ be the eigenvalues of $A$. If $\boldsymbol{x}_{k}$ is an eigenvector of $A$ for the eigenvalue $\lambda_{k}$, then $\boldsymbol{x}_{k}$ is an eigenvector for $(A-\alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_{k}-\alpha}$. Conversely, if

$$
\begin{equation*}
(A-\alpha I)^{-1} \boldsymbol{y}=\frac{1}{\lambda-\alpha} \boldsymbol{y} \tag{15.3}
\end{equation*}
$$

and $\boldsymbol{y} \neq \mathbf{0}$, then $A \boldsymbol{y}=\lambda \boldsymbol{y}$.
Proof: Let $\lambda_{k}$ and $\boldsymbol{x}_{k}$ be as described in the statement of the lemma. Then

$$
(A-\alpha I) x_{k}=\left(\lambda_{k}-\alpha\right) x_{k}
$$

and so $\frac{1}{\lambda_{k}-\alpha} \boldsymbol{x}_{k}=(A-\alpha I)^{-1} \boldsymbol{x}_{k}$.Suppose 15.3. Then $\boldsymbol{y}=\frac{1}{\lambda-\alpha}[A \boldsymbol{y}-\alpha \boldsymbol{y}]$. Solving for $A \boldsymbol{y}$ leads to $A \boldsymbol{y}=\lambda \boldsymbol{y}$.

Now assume $\alpha$ is closer to $\lambda$ than to any other eigenvalue. Then the magnitude of $\frac{1}{\lambda-\alpha}$ is greater than the magnitude of all the other eigenvalues of $(A-\alpha I)^{-1}$. Therefore, the power method applied to $(A-\alpha I)^{-1}$ will yield $\frac{1}{\lambda-\alpha}$. You end up with $s_{n+1} \approx \frac{1}{\lambda-\alpha}$ and solve for $\lambda$.

### 15.1.2 The Explicit Description Of The Method

Here is how you use this method to find the eigenvalue closest to $\alpha$ and the corresponding eigenvector.

1. Find $(A-\alpha I)^{-1}$.
2. Pick $\boldsymbol{u}_{1}$. If you are not phenomenally unlucky, the iterations will converge.
3. If $\boldsymbol{u}_{k}$ has been obtained, $\boldsymbol{u}_{k+1}=\frac{(A-\alpha I)^{-1} \boldsymbol{u}_{k}}{s_{k+1}}$ where $s_{k+1}$ is the entry of $(A-\alpha I)^{-1} \boldsymbol{u}_{k}$ which has largest absolute value.
4. When the scaling factors, $s_{k}$ are not changing much and the $\boldsymbol{u}_{k}$ are not changing much, find the approximation to the eigenvalue by solving $s_{k+1}=\frac{1}{\lambda-\alpha}$ for $\lambda$. The eigenvector is approximated by $\boldsymbol{u}_{k+1}$.
5. Check your work by multiplying by the original matrix to see how well what you have found works.

Thus this amounts to the power method for the matrix $(A-\alpha I)^{-1}$ but you are free to pick $\alpha$.

### 15.2 Automation With Matlab

You can do the above example and other examples using Matlab. Here are some commands which will do this. It is done here for a $3 \times 3$ matrix but you adapt for any size.

```
a=[5-8 6;1 0 0;0 0 1 0]; b=i; F=inv(a-b*eye(3));
S=1; u=[1;1;1]; d=1; k=1;
while d >.00001 & k<1000
w=F*u; [M,I]=max(abs(w));T=w(I); u=w/T;
d=abs(T-S); S=T; k=k+1;
end
u
b+1/T
k
a*u-(b+1/T)*u
```

eye(3) signifies the $3 \times 3$ identity. It is less trouble to write this.
Note how the "while loop" is limited to 1000 iterations. That way it won't go on forever if there is something wrong. This asks for the eigenvalue closest to $b=i$. When Matlab stalls, to get it to quit, you type control c. The last line checks the answer and the line with $k$ tells the number of iterations used. Also, the funny notation $[\mathrm{M}, \mathrm{I}]=\max (\mathrm{abs}(\mathrm{w}))$; $\mathrm{T}=\mathrm{w}(\mathrm{I})$; gets it to pick out the entry which has largest absolute value $\mathrm{w}(\mathrm{I})$ and keep that entry unchanged. The above iteration finds the eigenvalue closest to $i$ along with the corresponding eigenvector. When the procedure does not work well for $b$ real, you might imagine that there are complex eigenvalues and so, since the above procedure is going to give you real approximations, it can't find the complex eigenvalues. Thus you should take $b$ to be complex as done above.

If you have Matlab work the above iteration, you get the following for the eigenvector eigenvalue and number of iterations, and error .

$$
\left(\begin{array}{c}
1 \\
.5-.5 i \\
-.5 i
\end{array}\right), 1+i, k=18,10^{-5}\left(\begin{array}{c}
0 \\
-0.1321+0.1862 i \\
-0.1325+0.1863 i
\end{array}\right)
$$

In fact, this eigenvector is exactly right as is the eigenvalue $1+i$.
Thus this method will find eigenvalues real or complex along with an eigenvector associated with the eigenvalue. Note that the characteristic polynomial of the above matrix is $\lambda^{3}-5 \lambda^{2}+8 \lambda-6$ and the above finds a complex root to this polynomial. More generally, if you have a polynomial $\lambda^{n}+a_{n-1} \lambda^{n-1}+\cdots+a_{1} \lambda+a_{0}$, a matrix which has this as its characteristic polynomial is called a companion matrix and you can show a matrix which works for this polynomial is of the form

$$
\left(\begin{array}{cccc}
-a_{n-1} & -a_{n-2} & \cdots & a_{0} \\
1 & 0 & & \\
& \ddots & \ddots & \\
& & 1 & 0
\end{array}\right)
$$

You could use this or the earlier companion matrix described in the material on the rational canonical form. Thus this method is capable of finding roots to a polynomial equation which are close to a given complex number. Of course there is a problem with determining which number you should pick. A way to determine this will be discussed later. It involves something called the QR algorithm.

Example 15.2.1 Find the eigenvalue of $A=\left(\begin{array}{ccc}5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3\end{array}\right)$ which is closest to -7 . Also find an eigenvector which goes with this eigenvalue.

We use the algorithm described above.

```
a=[5 -14 11;-4 4 -4;3 6 -3]; b=-7; F=inv(a-b*eye(3));
S=1; u=[1;1;1]; d=1;k=1;
while d >.0001 & k<1000
w=F*u; [M,I]=max(abs(w)); T=w(I); u=w/T;
d=abs(T-S); S=T; k=k+1;
end
u
k
b+1/T
a*u-(b+1/T)*u
```

This yields the following after 8 iterations.

$$
\left(\begin{array}{c}
1 \\
0 \\
-1
\end{array}\right),-6
$$

for the eigenvector and eigenvalue. In fact, this is exactly correct.
Example 15.2.2 Consider the symmetric matrix $A=\left(\begin{array}{lll}1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2\end{array}\right)$. Find the middle eigenvalue and an eigenvector which goes with it.

Since $A$ is symmetric, it follows it has three real eigenvalues which are solutions to

$$
\begin{aligned}
p(\lambda) & =\operatorname{det}\left(\lambda\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)-\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{array}\right)\right) \\
& =\lambda^{3}-4 \lambda^{2}-24 \lambda-17=0
\end{aligned}
$$

If you use your graphing calculator to graph this polynomial, you find there is an eigenvalue somewhere between -.9 and -.8 and that this is the middle eigenvalue. Using -.8 as the number close to the eigenvalue desired, after 7 iterations, you get

$$
\boldsymbol{u}=\left(\begin{array}{c}
1 \\
-.5878 \\
-.2271
\end{array}\right), \lambda=-.8569
$$

Note that

$$
\begin{aligned}
& \left(\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 1 & 4 \\
3 & 4 & 2
\end{array}\right)-(-.8569)\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)\right)\left(\begin{array}{c}
1 \\
-.5878 \\
-.2271
\end{array}\right) \\
= & \left(\begin{array}{c}
2.5244 \times 10^{-29} \\
1.1418 \times 10^{-4} \\
-1.99 \times 10^{-6}
\end{array}\right)
\end{aligned}
$$

There is an easy to use trick which will eliminate some of the fuss and bother in using the shifted inverse power method. If you have

$$
(A-\alpha I)^{-1} \boldsymbol{x}=\mu \boldsymbol{x}
$$

then multiplying through by $(A-\alpha I)$, one finds that $\boldsymbol{x}$ will be an eigenvector for $A$ with eigenvalue $\alpha+\mu^{-1}$. Hence you could simply take $(A-\alpha I)^{-1}$ to a high power and multiply by a vector to get a vector which points in the direction of an eigenvalue of $A$. Then divide by the largest entry and identify the eigenvalue directly by multiplying the eigenvector by $A$. This is illustrated in the next example.

Example 15.2.3 Find the eigenvalue near -1.2 along with an eigenvector.

$$
A=\left(\begin{array}{lll}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{array}\right)
$$

This is only a $3 \times 3$ matrix and so it is not hard to estimate the eigenvalues. Just get the characteristic equation, graph it using a calculator and zoom in to find the eigenvalues. If you do this, you find there is an eigenvalue near -1.2 , one near -.4 , and one near 5.5. (The characteristic equation is $2+8 \lambda+4 \lambda^{2}-\lambda^{3}=0$.) Of course we have no idea what the eigenvectors are.

Lets first try to find the eigenvector and an approximation for the eigenvalue near -1.2. In this case, let $\alpha=-1.2$. Then

$$
(A-\alpha I)^{-1}=\left(\begin{array}{ccc}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{array}\right)
$$

Then

$$
\begin{aligned}
& \left(\begin{array}{ccc}
-25.357143 & -33.928571 & 50.0 \\
12.5 & 17.5 & -25.0 \\
23.214286 & 30.357143 & -45.0
\end{array}\right)^{17}\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right) \\
= & \left(\begin{array}{c}
-4.9432 \times 10^{28} \\
2.4312 \times 10^{28} \\
4.4928 \times 10^{28}
\end{array}\right)
\end{aligned}
$$

The initial approximation for an eigenvector will then be the above divided by its largest entry.

$$
\left(\begin{array}{c}
-4.9432 \times 10^{28} \\
2.4312 \times 10^{28} \\
4.4928 \times 10^{28}
\end{array}\right) \frac{1}{-4.9432 \times 10^{28}}=\left(\begin{array}{c}
1.0 \\
-0.49183 \\
-0.90888
\end{array}\right)
$$

How close is this to being an eigenvector?

$$
\begin{gathered}
\left(\begin{array}{lll}
2 & 1 & 3 \\
2 & 1 & 1 \\
3 & 2 & 1
\end{array}\right)\left(\begin{array}{c}
1.0 \\
-0.49183 \\
-0.90888
\end{array}\right)=\left(\begin{array}{c}
-1.2185 \\
0.59929 \\
1.1075
\end{array}\right) \\
-1.2185\left(\begin{array}{c}
1.0 \\
-0.49183 \\
-0.90888
\end{array}\right)=\left(\begin{array}{c}
-1.2185 \\
0.59929 \\
1.1075
\end{array}\right)
\end{gathered}
$$

For all practical purposes, this has found the eigenvector and eigenvalue of -1.2185 .

### 15.2.1 Complex Eigenvalues

What about complex eigenvalues? If your matrix is real, you won't see these by graphing the characteristic equation on your calculator. Will the shifted inverse power method find these eigenvalues and their associated eigenvectors? The answer is yes. However, for a real matrix, you must pick $\alpha$ to be complex. This is because the eigenvalues occur in conjugate pairs so if you don't pick it complex, it will be the same distance between any conjugate pair of complex numbers and so nothing in the above argument for convergence implies you will get convergence to a complex number. Also, the process of iteration will yield only real vectors and scalars.

Example 15.2.4 Find the complex eigenvalues and corresponding eigenvectors for the matrix

$$
\left(\begin{array}{ccc}
5 & -8 & 6 \\
1 & 0 & 0 \\
0 & 1 & 0
\end{array}\right)
$$

Here the characteristic equation is $\lambda^{3}-5 \lambda^{2}+8 \lambda-6=0$. One solution is $\lambda=3$. The other two are $1+i$ and $1-i$. I will apply the process to $\alpha=i$ to find the eigenvalue closest to $i$. The above algorithm yields the following after 15 iterations.

$$
u=\left(\begin{array}{c}
1 \\
.5-.5 i \\
-.5 i
\end{array}\right), \lambda=1+i
$$

This illustrates an interesting topic which leads to many related topics. If you have a polynomial, $x^{4}+a x^{3}+b x^{2}+c x+d$, you can consider it as the characteristic polynomial of a certain matrix, called a companion matrix. In this case,

$$
\left(\begin{array}{cccc}
-a & -b & -c & -d \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right)
$$

The above example was just a companion matrix for $\lambda^{3}-5 \lambda^{2}+8 \lambda-6$. You can see the pattern which will enable you to obtain a companion matrix for any polynomial of the form $\lambda^{n}+a_{1} \lambda^{n-1}+\cdots+a_{n-1} \lambda+a_{n}$. This illustrates that one way to find the complex zeros of a polynomial is to use the shifted inverse power method on a companion matrix for the polynomial. Doubtless there are better ways but this does illustrate how impressive this procedure is. Do you have a better way?

Note that the shifted inverse power method is a way you can begin with something close but not equal to an eigenvalue and end up with something close to an eigenvector.

### 15.2.2 Rayleigh Quotients and Estimates for Eigenvalues

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. Recall a matrix $A$ is Hermitian if $A=A^{*}$ where $A^{*}$ means to take the transpose of the conjugate of $A$. In the case of a real matrix, Hermitian reduces to symmetric. Recall also that for $\boldsymbol{x} \in \mathbb{F}^{n}$,

$$
|\boldsymbol{x}|^{2}=\boldsymbol{x}^{*} \boldsymbol{x}=\sum_{j=1}^{n}\left|x_{j}\right|^{2}
$$

Recall the following corollary found on Page 331 which is stated here for convenience.
Corollary 15.2.5 If $A$ is Hermitian, then all the eigenvalues of $A$ are real and there exists an orthonormal basis of eigenvectors.

Thus for $\left\{\boldsymbol{x}_{k}\right\}_{k=1}^{n}$ this orthonormal basis,

$$
\boldsymbol{x}_{i}^{*} \boldsymbol{x}_{j}=\boldsymbol{\delta}_{i j} \equiv\left\{\begin{array}{l}
1 \text { if } i=j \\
0 \text { if } i \neq j
\end{array}\right.
$$

For $\boldsymbol{x} \in \mathbb{F}^{n}, \boldsymbol{x} \neq \mathbf{0}$, the Rayleigh quotient is defined by $\frac{\boldsymbol{x}^{*} A \boldsymbol{x}}{|\boldsymbol{x}|^{2}}$. Now let the eigenvalues of $A$ be $\lambda_{1} \leq \lambda_{2} \leq \cdots \leq \lambda_{n}$ and $A \boldsymbol{x}_{k}=\lambda_{k} \boldsymbol{x}_{k}$ where $\left\{\boldsymbol{x}_{k}\right\}_{k=1}^{n}$ is the above orthonormal basis of eigenvectors mentioned in the corollary. Then if $\boldsymbol{x}$ is an arbitrary vector, there exist constants, $a_{i}$ such that $\boldsymbol{x}=\sum_{i=1}^{n} a_{i} \boldsymbol{x}_{i}$. Also,

$$
|\boldsymbol{x}|^{2}=\sum_{i=1}^{n} \bar{a}_{i} \boldsymbol{x}_{i}^{*} \sum_{j=1}^{n} a_{j} \boldsymbol{x}_{j}=\sum_{i j} \bar{a}_{i} a_{j} \boldsymbol{x}_{i}^{*} \boldsymbol{x}_{j}=\sum_{i j} \bar{a}_{i} a_{j} \delta_{i j}=\sum_{i=1}^{n}\left|a_{i}\right|^{2} .
$$

Therefore,

$$
\begin{aligned}
\frac{\boldsymbol{x}^{*} A \boldsymbol{x}}{|\boldsymbol{x}|^{2}} & =\frac{\left(\sum_{i=1}^{n} \bar{a}_{i} x_{i}^{*}\right)\left(\sum_{j=1}^{n} a_{j} \lambda_{j} \boldsymbol{x}_{j}\right)}{\sum_{i=1}^{n}\left|a_{i}\right|^{2}}=\frac{\sum_{i j} \bar{a}_{i} a_{j} \lambda_{j} \boldsymbol{x}_{i}^{*} \boldsymbol{x}_{j}}{\sum_{i=1}^{n}\left|a_{i}\right|^{2}} \\
& =\frac{\sum_{i j} \bar{a}_{i} a_{j} \lambda_{j} \delta_{i j}}{\sum_{i=1}^{n}\left|a_{i}\right|^{2}}=\frac{\sum_{i=1}^{n}\left|a_{i}\right|^{2} \lambda_{i}}{\sum_{i=1}^{n}\left|a_{i}\right|^{2}} \in\left[\lambda_{1}, \lambda_{n}\right] .
\end{aligned}
$$

In other words, the Rayleigh quotient is always between the largest and the smallest eigenvalues of $A$. When $\boldsymbol{x}=\boldsymbol{x}_{n}$, the Rayleigh quotient equals the largest eigenvalue and when $\boldsymbol{x}=\boldsymbol{x}_{1}$ the Rayleigh quotient equals the smallest eigenvalue. Suppose you calculate a Rayleigh quotient. How close is it to some eigenvalue?

Theorem 15.2.6 Let $\boldsymbol{x} \neq \mathbf{0}$ and form the Rayleigh quotient,

$$
\frac{x^{*} A \boldsymbol{x}}{|\boldsymbol{x}|^{2}} \equiv q
$$

Then there exists an eigenvalue of $A$, denoted here by $\lambda_{q}$ such that

$$
\begin{equation*}
\left|\lambda_{q}-q\right| \leq \frac{|A \boldsymbol{x}-q \boldsymbol{x}|}{|\boldsymbol{x}|} \tag{15.4}
\end{equation*}
$$

Proof: Let $\boldsymbol{x}=\sum_{k=1}^{n} a_{k} \boldsymbol{x}_{k}$ where $\left\{\boldsymbol{x}_{k}\right\}_{k=1}^{n}$ is the orthonormal basis of eigenvectors.

$$
\begin{gathered}
|A \boldsymbol{x}-q \boldsymbol{x}|^{2}=(A \boldsymbol{x}-q \boldsymbol{x})^{*}(A \boldsymbol{x}-q \boldsymbol{x}) \\
=\left(\sum_{k=1}^{n} a_{k} \lambda_{k} \boldsymbol{x}_{k}-q a_{k} \boldsymbol{x}_{k}\right)^{*}\left(\sum_{k=1}^{n} a_{k} \lambda_{k} \boldsymbol{x}_{k}-q a_{k} \boldsymbol{x}_{k}\right) \\
=\left(\sum_{j=1}^{n}\left(\boldsymbol{\lambda}_{j}-q\right) \bar{a}_{j} \boldsymbol{x}_{j}^{*}\right)\left(\sum_{k=1}^{n}\left(\lambda_{k}-q\right) a_{k} \boldsymbol{x}_{k}\right) \\
=\sum_{j, k}\left(\lambda_{j}-q\right) \bar{a}_{j}\left(\lambda_{k}-q\right) a_{k} \boldsymbol{x}_{j}^{*} \boldsymbol{x}_{k}=\sum_{k=1}^{n}\left|a_{k}\right|^{2}\left(\boldsymbol{\lambda}_{k}-q\right)^{2}
\end{gathered}
$$

Now pick the eigenvalue $\lambda_{q}$ which is closest to $q$. Then

$$
|A \boldsymbol{x}-q \boldsymbol{x}|^{2}=\sum_{k=1}^{n}\left|a_{k}\right|^{2}\left(\lambda_{k}-q\right)^{2} \geq\left(\lambda_{q}-q\right)^{2} \sum_{k=1}^{n}\left|a_{k}\right|^{2}=\left(\lambda_{q}-q\right)^{2}|\boldsymbol{x}|^{2}
$$

which implies 15.4.

Example 15.2.7 Consider the symmetric matrix

$$
A=\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{array}\right)
$$

Let $\boldsymbol{x}=(1,1,1)^{T}$. How close is the Rayleigh quotient to some eigenvalue of $A$ ? Find the eigenvector and eigenvalue to several decimal places.

Everything is real and so there is no need to worry about taking conjugates. Therefore, the Rayleigh quotient is

$$
\frac{\left(\begin{array}{lll}
1 & 1 & 1
\end{array}\right)\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{array}\right)\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)}{3}=\frac{19}{3}
$$

According to the above theorem, there is some eigenvalue of this matrix $\lambda_{q}$ such that

$$
\begin{aligned}
\left|\lambda_{q}-\frac{19}{3}\right| & \leq \frac{\left|\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{array}\right)\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)-\frac{19}{3}\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)\right|}{\sqrt{3}}=\frac{1}{\sqrt{3}}\left(\begin{array}{c}
-\frac{1}{3} \\
-\frac{4}{3} \\
\frac{5}{3}
\end{array}\right) \\
& =\frac{\sqrt{\frac{1}{9}+\left(\frac{4}{3}\right)^{2}+\left(\frac{5}{3}\right)^{2}}}{\sqrt{3}}=1.2472
\end{aligned}
$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the shifted inverse power method is all about.

Solve

$$
\left(\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{array}\right)-\frac{19}{3}\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)
$$

In other words solve

$$
\left(\begin{array}{ccc}
-\frac{16}{3} & 2 & 3 \\
2 & -\frac{13}{3} & 1 \\
3 & 1 & -\frac{7}{3}
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{l}
1 \\
1 \\
1
\end{array}\right)
$$

and divide by the entry which is largest, 3.8707, to get

$$
u_{2}=\left(\begin{array}{c}
.69925 \\
.49389 \\
1.0
\end{array}\right)
$$

Now solve

$$
\left(\begin{array}{ccc}
-\frac{16}{3} & 2 & 3 \\
2 & -\frac{13}{3} & 1 \\
3 & 1 & -\frac{7}{3}
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{c}
.69925 \\
.49389 \\
1.0
\end{array}\right)
$$

and divide by the largest entry, 2.9979 to get

$$
\boldsymbol{u}_{3}=\left(\begin{array}{c}
.71473 \\
.52263 \\
1.0
\end{array}\right)
$$

Now solve

$$
\left(\begin{array}{ccc}
-\frac{16}{3} & 2 & 3 \\
2 & -\frac{13}{3} & 1 \\
3 & 1 & -\frac{7}{3}
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{c}
.71473 \\
.52263 \\
1.0
\end{array}\right)
$$

and divide by the largest entry, 3.0454 , to get

$$
\boldsymbol{u}_{4}=\left(\begin{array}{c}
.7137 \\
.52056 \\
1.0
\end{array}\right)
$$

Solve

$$
\left(\begin{array}{ccc}
-\frac{16}{3} & 2 & 3 \\
2 & -\frac{13}{3} & 1 \\
3 & 1 & -\frac{7}{3}
\end{array}\right)\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\left(\begin{array}{c}
.7137 \\
.52056 \\
1.0
\end{array}\right)
$$

and divide by the largest entry, 3.0421 to get

$$
\boldsymbol{u}_{5}=\left(\begin{array}{c}
.71378 \\
.52073 \\
1.0
\end{array}\right)
$$

You can see these scaling factors are not changing much. The predicted eigenvalue is then about

$$
\frac{1}{3.0421}+\frac{19}{3}=6.6621 .
$$

How close is this?

$$
\left(\begin{array}{lll}
1 & 2 & 3 \\
2 & 2 & 1 \\
3 & 1 & 4
\end{array}\right)\left(\begin{array}{c}
.71378 \\
.52073 \\
1.0
\end{array}\right)=\left(\begin{array}{c}
4.7552 \\
3.469 \\
6.6621
\end{array}\right)
$$

while

$$
6.6621\left(\begin{array}{c}
.71378 \\
.52073 \\
1.0
\end{array}\right)=\left(\begin{array}{l}
4.7553 \\
3.4692 \\
6.6621
\end{array}\right)
$$

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

### 15.3 The $Q R$ Algorithm

### 15.3.1 Basic Properties And Definition

Recall the theorem about the $Q R$ factorization in Theorem 12.3.9. It says that given an $n \times n$ real matrix $A$, there exists a real orthogonal matrix $Q$ and an upper triangular matrix $R$ such that $A=Q R$ and that this factorization can be accomplished by a systematic procedure. One such procedure was given in proving this theorem.

Theorem 15.3.1 Let $A$ be an $n \times n$ complex matrix. Then there exists a unitary $Q$ and upper triangular $R$ such that $A=Q R$.

Proof: This is obvious if $n=1$. Suppose true for $n$ and let

$$
A=\left(\begin{array}{llll}
a_{1} & \cdots & a_{n} & a_{n+1}
\end{array}\right)
$$

Let $Q_{1}$ be a unitary matrix such that $Q_{1} \boldsymbol{a}_{1}=\left|\boldsymbol{a}_{1}\right| \boldsymbol{e}_{1}$ in case $\boldsymbol{a}_{1} \neq \mathbf{0}$. If $\boldsymbol{a}_{1}=\mathbf{0}$, let $Q_{1}=I$. Thus

$$
Q_{1} A=\left(\begin{array}{cc}
a & b \\
\mathbf{0} & A_{1}
\end{array}\right)
$$

where $A_{1}$ is $(n-1) \times(n-1)$. By induction, there exists $Q_{2}^{\prime}$ an $(n-1) \times(n-1)$ unitary matrix such that $Q_{2}^{\prime} A_{1}=R^{\prime}$, an upper triangular matrix. Then

$$
\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & Q_{2}^{\prime}
\end{array}\right) Q_{1} A=\left(\begin{array}{cc}
a & \boldsymbol{b} \\
\mathbf{0} & R^{\prime}
\end{array}\right)=R
$$

Since the product of unitary matrices is unitary, there exists $Q$ unitary such that $Q^{*} A=R$ and so $A=Q R$.

The $Q R$ algorithm is described in the following definition.
Definition 15.3.2 The $Q R$ algorithm is the following. In the description of this algorithm, $Q$ is unitary and $R$ is upper triangular having nonnegative entries on the main diagonal. Starting with A an $n \times n$ matrix, form

$$
\begin{equation*}
A_{0} \equiv A=Q_{1} R_{1} \tag{15.5}
\end{equation*}
$$

Then

$$
\begin{equation*}
A_{1} \equiv R_{1} Q_{1} . \tag{15.6}
\end{equation*}
$$

In general given

$$
\begin{equation*}
A_{k}=R_{k} Q_{k} \tag{15.7}
\end{equation*}
$$

$\operatorname{obtain} A_{k+1}$ by

$$
\begin{equation*}
A_{k}=Q_{k+1} R_{k+1}, A_{k+1}=R_{k+1} Q_{k+1} \tag{15.8}
\end{equation*}
$$

This algorithm was proposed by Francis in 1961. The sequence $\left\{A_{k}\right\}$ is the desired sequence of iterates. Now with the above definition of the algorithm, here are its properties. The next lemma shows each of the $A_{k}$ is unitarily similar to $A$ and the amazing thing about this algorithm is that often it becomes increasingly easy to find the eigenvalues of the $A_{k}$.

Lemma 15.3.3 Let $A$ be an $n \times n$ matrix and let the $Q_{k}$ and $R_{k}$ be as described in the algorithm. Then each $A_{k}$ is unitarily similar to $A$ and denoting by $Q^{(k)}$ the product $Q_{1} Q_{2} \cdots Q_{k}$ and $R^{(k)}$ the product $R_{k} R_{k-1} \cdots R_{1}$, it follows that $A^{k}=Q^{(k)} R^{(k)}$ (The matrix on the left is $A$ raised to the $k^{\text {th }}$ power.)

$$
A=Q^{(k)} A_{k} Q^{(k) *}, A_{k}=Q^{(k) *} A Q^{(k)}
$$

Proof: From the algorithm, $R_{k+1}=A_{k+1} Q_{k+1}^{*}$ and so

$$
A_{k}=Q_{k+1} R_{k+1}=Q_{k+1} A_{k+1} Q_{k+1}^{*}
$$

Now iterating this, it follows

$$
\begin{gathered}
A_{k-1}=Q_{k} A_{k} Q_{k}^{*}=Q_{k} Q_{k+1} A_{k+1} Q_{k+1}^{*} Q_{k}^{*} \\
A_{k-2}=Q_{k-1} A_{k-1} Q_{k-1}^{*}=Q_{k-1} Q_{k} Q_{k+1} A_{k+1} Q_{k+1}^{*} Q_{k}^{*} Q_{k-1}^{*}
\end{gathered}
$$

etc. Thus, after $k-2$ more iterations,

$$
A=Q^{(k+1)} A_{k+1} Q^{(k+1) *}
$$

The product of unitary matrices is unitary and so this proves the first claim of the lemma.
Now consider the part about $A^{k}$. From the algorithm, this is clearly true for $k=1$. $\left(A^{1}=Q R\right)$ Suppose then that

$$
A^{k}=Q_{1} Q_{2} \cdots Q_{k} R_{k} R_{k-1} \cdots R_{1}
$$

What was just shown indicated

$$
A=Q_{1} Q_{2} \cdots Q_{k+1} A_{k+1} Q_{k+1}^{*} Q_{k}^{*} \cdots Q_{1}^{*}
$$

and now from the algorithm, $A_{k+1}=R_{k+1} Q_{k+1}$ and so

$$
A=Q_{1} Q_{2} \cdots Q_{k+1} R_{k+1} Q_{k+1} Q_{k+1}^{*} Q_{k}^{*} \cdots Q_{1}^{*}
$$

Then

$$
\begin{gathered}
A^{k+1}=A A^{k}= \\
\overbrace{Q_{1} Q_{2} \cdots Q_{k+1} R_{k+1} Q_{k+1} Q_{k+1}^{*} Q_{k}^{*} \cdots Q_{1}^{*} Q_{1} \cdots Q_{k} R_{k} R_{k-1} \cdots R_{1}}^{A}=Q_{1} Q_{2} \cdots Q_{k+1} R_{k+1} R_{k} R_{k-1} \cdots R_{1} \equiv Q^{(k+1)} R^{(k+1)}
\end{gathered}
$$

Here is another very interesting lemma.
Lemma 15.3.4 Suppose $Q^{(k)}, Q$ are unitary and $R_{k}$ is upper triangular such that the diagonal entries on $R_{k}$ are all positive and

$$
Q=\lim _{k \rightarrow \infty} Q^{(k)} R_{k}
$$

Then

$$
\lim _{k \rightarrow \infty} Q^{(k)}=Q, \lim _{k \rightarrow \infty} R_{k}=I
$$

Also the $Q R$ factorization of $A$ is unique.

Proof: Let

$$
Q=\left(\boldsymbol{q}_{1}, \cdots, \boldsymbol{q}_{n}\right), Q^{(k)}=\left(\boldsymbol{q}_{1}^{k}, \cdots, \boldsymbol{q}_{n}^{k}\right)
$$

where the $\boldsymbol{q}$ are the columns. Also denote by $r_{i j}^{k}$ the $i j^{t h}$ entry of $R_{k}$. Thus

$$
Q^{(k)} R_{k}=\left(\boldsymbol{q}_{1}^{k}, \cdots, \boldsymbol{q}_{n}^{k}\right)\left(\begin{array}{ccc}
r_{11}^{k} & & * \\
& \ddots & \\
0 & & r_{n n}^{k}
\end{array}\right)
$$

It follows $r_{11}^{k} \boldsymbol{q}_{1}^{k} \rightarrow \boldsymbol{q}_{1}$ and so $r_{11}^{k}=\left|r_{11}^{k} \boldsymbol{q}_{1}^{k}\right| \rightarrow 1$. Therefore, $\boldsymbol{q}_{1}^{k} \rightarrow \boldsymbol{q}_{1}$. Next consider the second column.

$$
r_{12}^{k} \boldsymbol{q}_{1}^{k}+r_{22}^{k} \boldsymbol{q}_{2}^{k} \rightarrow \boldsymbol{q}_{2}
$$

Taking the inner product of both sides with $\boldsymbol{q}_{1}^{k}$ it follows

$$
\lim _{k \rightarrow \infty} r_{12}^{k}=\lim _{k \rightarrow \infty}\left(\boldsymbol{q}_{2} \cdot \boldsymbol{q}_{1}^{k}\right)=\left(\boldsymbol{q}_{2} \cdot \boldsymbol{q}_{1}\right)=0 .
$$

Therefore, $\lim _{k \rightarrow \infty} r_{22}^{k} \boldsymbol{q}_{2}^{k}=\boldsymbol{q}_{2}$ and since $r_{22}^{k}>0$, it follows as in the first part that $r_{22}^{k} \rightarrow 1$. Hence $\lim _{k \rightarrow \infty} \boldsymbol{q}_{2}^{k}=\boldsymbol{q}_{2}$. Continuing this way, it follows $\lim _{k \rightarrow \infty} r_{i j}^{k}=0$ for all $i \neq j$ and

$$
\lim _{k \rightarrow \infty} r_{j j}^{k}=1, \lim _{k \rightarrow \infty} \boldsymbol{q}_{j}^{k}=\boldsymbol{q}_{j}
$$

Thus $R_{k} \rightarrow I$ and $Q^{(k)} \rightarrow Q$. This proves the first part of the lemma.
The second part follows immediately. If $Q R=Q^{\prime} R^{\prime}=A$ where $A^{-1}$ exists, then $Q^{*} Q^{\prime}=$ $R\left(R^{\prime}\right)^{-1}$ and I need to show both sides of the above are equal to $I$. The left side of the above is unitary and the right side is upper triangular having positive entries on the diagonal. This is because the inverse of such an upper triangular matrix having positive entries on the main diagonal is still upper triangular having positive entries on the main diagonal and the product of two such upper triangular matrices gives another of the same form having positive entries on the main diagonal. Suppose then that $Q=R$ where $Q$ is unitary and $R$ is upper triangular having positive entries on the main diagonal. Let $Q_{k}=Q$ and $R_{k}=R$. It follows $I R_{k} \rightarrow R=Q$ and so from the first part, $R_{k} \rightarrow I$ but $R_{k}=R$ and so $R=I$. Thus applying this to $Q^{*} Q^{\prime}=R\left(R^{\prime}\right)^{-1}$ yields both sides equal $I$.

A case of all this is of great interest. Suppose $A$ has a largest eigenvalue $\lambda$ which is real. Then $A^{n}$ is of the form $\left(A^{n-1} \boldsymbol{a}_{1}, \cdots, A^{n-1} \boldsymbol{a}_{n}\right)$ and so likely each of these columns will be pointing roughly in the direction of an eigenvector of $A$ which corresponds to this eigenvalue. Then when you do the $Q R$ factorization of this, it follows from the fact that $R$ is upper triangular, that the first column of $Q$ will be a multiple of $A^{n-1} \boldsymbol{a}_{1}$ and so will end up being roughly parallel to the eigenvector desired. Also this will require the entries below the top in the first column of $A_{n}=Q^{T} A Q$ will all be small because they will be of the form $\boldsymbol{q}_{i}^{T} A \boldsymbol{q}_{1} \approx \lambda \boldsymbol{q}_{i}^{T} \boldsymbol{q}_{1}=0$. Therefore, $A_{n}$ will be of the form

$$
\left(\begin{array}{cc}
\lambda^{\prime} & a \\
e & B
\end{array}\right)
$$

where $\boldsymbol{e}$ is small. It follows that $\lambda^{\prime}$ will be close to $\lambda$ and $\boldsymbol{q}_{1}$ will be close to an eigenvector for $\lambda$. Then if you like, you could do the same thing with the matrix $B$ to obtain approximations for the other eigenvalues. Finally, you could use the shifted inverse power method to get more exact solutions.

### 15.3.2 The Case Of Real Eigenvalues

With these lemmas, it is possible to prove that for the $Q R$ algorithm and certain conditions, the sequence $A_{k}$ converges pointwise to an upper triangular matrix having the eigenvalues of $A$ down the diagonal. I will assume all the matrices are real here.

This convergence won't always happen. Consider for example the matrix

$$
\left(\begin{array}{ll}
0 & 1 \\
1 & 0
\end{array}\right)
$$

You can verify quickly that the algorithm will return this matrix for each $k$. The problem here is that, although the matrix has the two eigenvalues $-1,1$, they have the same absolute value. The $Q R$ algorithm works in somewhat the same way as the power method, exploiting differences in the size of the eigenvalues.

If $A$ has all real eigenvalues and you are interested in finding these eigenvalues along with the corresponding eigenvectors, you could always consider $A+\lambda I$ instead where $\lambda$ is sufficiently large and positive that $A+\lambda I$ has all positive eigenvalues. (Recall Gerschgorin's theorem.) Then if $\mu$ is an eigenvalue of $A+\lambda I$ with

$$
(A+\lambda I) x=\mu x
$$

then $A \boldsymbol{x}=(\mu-\lambda) \boldsymbol{x}$ so to find the eigenvalues of $A$ you just subtract $\lambda$ from the eigenvalues of $A+\lambda I$. Thus there is no loss of generality in assuming at the outset that the eigenvalues of $A$ are all positive. Here is the theorem. It involves a technical condition which will often hold. The proof presented here follows [42] and is a special case of that presented in this reference.

Before giving the proof, note that the product of upper triangular matrices is upper triangular. If they both have positive entries on the main diagonal so will the product. Furthermore, the inverse of an upper triangular matrix is upper triangular. I will use these simple facts without much comment whenever convenient.

Theorem 15.3.5 Let A be a real matrix having eigenvalues

$$
\lambda_{1}>\lambda_{2}>\cdots>\lambda_{n}>0
$$

and let $A=S D S^{-1}$ where

$$
D=\left(\begin{array}{ccc}
\lambda_{1} & & 0 \\
& \ddots & \\
0 & & \lambda_{n}
\end{array}\right)
$$

and suppose $S^{-1}$ has an $L U$ factorization. Then the matrices $A_{k}$ in the $Q R$ algorithm described above converge to an upper triangular matrix $T^{\prime}$ having the eigenvalues of $A$, $\lambda_{1}, \cdots, \lambda_{n}$ descending on the main diagonal. The matrices $Q^{(k)}$ converge to $Q^{\prime}$, an orthogonal matrix which equals $Q$ except for possibly having some columns multiplied by -1 for $Q$ the unitary part of the $Q R$ factorization of $S, S=Q R$,and

$$
\lim _{k \rightarrow \infty} A_{k}=T^{\prime}=Q^{\prime T} A Q^{\prime}
$$

Proof: From Lemma 15.3.3

$$
\begin{equation*}
A^{k}=Q^{(k)} R^{(k)}=S D^{k} S^{-1} \tag{15.9}
\end{equation*}
$$

Let $S=Q R$ where this is just a $Q R$ factorization which is known to exist and let $S^{-1}=L U$ which is assumed to exist. Thus

$$
\begin{equation*}
Q^{(k)} R^{(k)}=Q R D^{k} L U \tag{15.10}
\end{equation*}
$$

and so $Q^{(k)} R^{(k)}=Q R D^{k} L U=Q R D^{k} L D^{-k} D^{k} U$. That matrix in the middle, $D^{k} L D^{-k}$ satisfies

$$
\left(D^{k} L D^{-k}\right)_{i j}=\lambda_{i}^{k} L_{i j} \lambda_{j}^{-k} \text { for } j \leq i, 0 \text { if } j>i
$$

Thus for $j<i$ the expression converges to 0 because $\lambda_{j}>\lambda_{i}$ when this happens. When $i=j$ it reduces to 1 . Thus the matrix in the middle is of the form $I+E_{k}$ where $E_{k} \rightarrow 0$. Then it follows

$$
\begin{gathered}
A^{k}=Q^{(k)} R^{(k)}=Q R\left(I+E_{k}\right) D^{k} U \\
=Q\left(I+R E_{k} R^{-1}\right) R D^{k} U \equiv Q\left(I+F_{k}\right) R D^{k} U
\end{gathered}
$$

where $F_{k} \rightarrow 0$. Then let $I+F_{k}=Q_{k} R_{k}$ where this is another $Q R$ factorization. Then it reduces to

$$
Q^{(k)} R^{(k)}=Q Q_{k} R_{k} R D^{k} U
$$

This looks really interesting because by Lemma 15.3.4 $Q_{k} \rightarrow I$ and $R_{k} \rightarrow I$ because $Q_{k} R_{k}=\left(I+F_{k}\right) \rightarrow I$. So it follows $Q Q_{k}$ is an orthogonal matrix converging to $Q$ while $R_{k} R D^{k} U\left(R^{(k)}\right)^{-1}$ is upper triangular, being the product of upper triangular matrices. Unfortunately, it is not known that the diagonal entries of this matrix are nonnegative because of the $U$. Let $\Lambda$ be just like the identity matrix but having some of the ones replaced with -1 in such a way that $\Lambda U$ is an upper triangular matrix having positive diagonal entries. Note $\Lambda^{2}=I$ and also $\Lambda$ commutes with a diagonal matrix. Thus

$$
Q^{(k)} R^{(k)}=Q Q_{k} R_{k} R D^{k} \Lambda^{2} U=Q Q_{k} R_{k} R \Lambda D^{k}(\Lambda U)
$$

At this point, one does some inspired massaging to write the above in the form

$$
\begin{aligned}
& Q Q_{k}\left(\Lambda D^{k}\right)\left[\left(\Lambda D^{k}\right)^{-1} R_{k} R \Lambda D^{k}\right](\Lambda U) \\
= & Q\left(Q_{k} \Lambda\right) D^{k}\left[\left(\Lambda D^{k}\right)^{-1} R_{k} R \Lambda D^{k}\right](\Lambda U) \\
= & Q\left(Q_{k} \Lambda\right) \overbrace{D^{k}\left[\left(\Lambda D^{k}\right)^{-1} R_{k} R \Lambda D^{k}\right](\Lambda U)}^{\equiv G_{k}}
\end{aligned}
$$

Now I claim the middle matrix in [.] is upper triangular and has all positive entries on the diagonal. This is because it is an upper triangular matrix which is similar to the upper triangular matrix $R_{k} R$ and so it has the same eigenvalues (diagonal entries) as $R_{k} R$. Thus the matrix $G_{k} \equiv D^{k}\left[\left(\Lambda D^{k}\right)^{-1} R_{k} R \Lambda D^{k}\right](\Lambda U)$ is upper triangular and has all positive entries on the diagonal. Multiply on the right by $G_{k}^{-1}$ to get

$$
Q^{(k)} R^{(k)} G_{k}^{-1}=Q Q_{k} \Lambda \rightarrow Q^{\prime}
$$

where $Q^{\prime}$ is essentially equal to $Q$ but might have some of the columns multiplied by -1 . This is because $Q_{k} \rightarrow I$ and so $Q_{k} \Lambda \rightarrow \Lambda$. Now by Lemma 15.3.4, it follows

$$
Q^{(k)} \rightarrow Q^{\prime}, R^{(k)} G_{k}^{-1} \rightarrow I
$$

It remains to verify $A_{k}$ converges to an upper triangular matrix. Recall that from 15.9 and the definition below this $(S=Q R)$

$$
A=S D S^{-1}=(Q R) D(Q R)^{-1}=Q R D R^{-1} Q^{T}=Q T Q^{T}
$$

Where $T$ is an upper triangular matrix. This is because it is the product of upper triangular matrices $R, D, R^{-1}$. Thus $Q^{T} A Q=T$. If you replace $Q$ with $Q^{\prime}$ in the above, it still results in an upper triangular matrix $T^{\prime}$ having the same diagonal entries as $T$. This is because

$$
T=Q^{T} A Q=\left(Q^{\prime} \Lambda\right)^{T} A\left(Q^{\prime} \Lambda\right)=\Lambda Q^{\prime T} A Q^{\prime} \Lambda
$$

and considering the $i i^{t h}$ entry yields

$$
\left(Q^{T} A Q\right)_{i i} \equiv \sum_{j, k} \Lambda_{i j}\left(Q^{\prime T} A Q^{\prime}\right)_{j k} \Lambda_{k i}=\Lambda_{i i} \Lambda_{i i}\left(Q^{\prime T} A Q^{\prime}\right)_{i i}=\left(Q^{\prime T} A Q^{\prime}\right)_{i i}
$$

Recall from Lemma 15.3.3, $A_{k}=Q^{(k) T} A Q^{(k)}$. Thus taking a limit and using the first part,

$$
A_{k}=Q^{(k) T} A Q^{(k)} \rightarrow Q^{T} A Q^{\prime}=T^{\prime}
$$

An easy case is for $A$ symmetric. Recall Corollary 13.1.6. By this corollary, there exists an orthogonal (real unitary) matrix $Q$ such that

$$
Q^{T} A Q=D
$$

where $D$ is diagonal having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right.

Corollary 15.3.6 Let A be a real symmetric $n \times n$ matrix having eigenvalues

$$
\lambda_{1}>\lambda_{2}>\cdots>\lambda_{n}>0
$$

and let $Q$ be defined by

$$
\begin{equation*}
Q D Q^{T}=A, D=Q^{T} A Q \tag{15.11}
\end{equation*}
$$

where $Q$ is orthogonal and $D$ is a diagonal matrix having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right. Let $Q^{T}$ have an $L U$ factorization. Then in the $Q R$ algorithm, the matrices $Q^{(k)}$ converge to $Q^{\prime}$ where $Q^{\prime}$ is the same as $Q$ except having some columns multiplied by $(-1)$. Thus the columns of $Q^{\prime}$ are eigenvectors of $A$. The matrices $A_{k}$ converge to $D$.

Proof: This follows from Theorem 15.3.5. Here $S=Q, S^{-1}=Q^{T}$. Thus

$$
Q=S=Q R
$$

and $R=I$. By Theorem 15.3.5 and Lemma 15.3.3,

$$
A_{k}=Q^{(k) T} A Q^{(k)} \rightarrow Q^{\prime T} A Q^{\prime}=Q^{T} A Q=D .
$$

because formula 15.11 is unaffected by replacing $Q$ with $Q^{\prime}$.
When using the $Q R$ algorithm, it is not necessary to check technical condition about $S^{-1}$ having an $L U$ factorization. The algorithm delivers a sequence of matrices which are similar to the original one. If that sequence converges to an upper triangular matrix, then the algorithm worked. Furthermore, the technical condition is sufficient but not necessary. The algorithm will work even without the technical condition.

Example 15.3.7 Find the eigenvalues and eigenvectors of the matrix

$$
A=\left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{array}\right)
$$

It is a symmetric matrix but other than that, I just pulled it out of the air. By Lemma 15.3.3 it follows $A_{k}=Q^{(k) T} A Q^{(k)}$. And so to get to the answer quickly I could have the computer raise $A$ to a power and then take the $Q R$ factorization of what results to get the $k^{t h}$ iteration using the above formula. Lets pick $k=10$.

$$
\left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{array}\right)^{10}=\left(\begin{array}{lll}
4.2273 \times 10^{7} & 2.5959 \times 10^{7} & 1.8611 \times 10^{7} \\
2.5959 \times 10^{7} & 1.6072 \times 10^{7} & 1.1506 \times 10^{7} \\
1.8611 \times 10^{7} & 1.1506 \times 10^{7} & 8.2396 \times 10^{6}
\end{array}\right)
$$

Now take $Q R$ factorization of this. The computer will do that also.
This yields

$$
\begin{aligned}
& \left(\begin{array}{cccc}
.79785 & -.59912 & -6.6943 \times 10^{-2} \\
.48995 & .70912 & -.50706 \\
.35126 & .37176 & .85931
\end{array}\right) \\
& \left(\begin{array}{ccc}
5.2983 \times 10^{7} & 3.2627 \times 10^{7} & 2.338 \times 10^{7} \\
0 & 1.2172 \times 10^{5} & 71946 \\
0 & 0 & 277.03
\end{array}\right)
\end{aligned}
$$

Next it follows

$$
\begin{aligned}
A_{10}= & \left(\begin{array}{ccc}
.79785 & -.59912 & -6.6943 \times 10^{-2} \\
.48995 & .70912 & -.50706 \\
.35126 & .37176 & .85931
\end{array}\right)^{T} \\
& \left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{array}\right)\left(\begin{array}{ccc}
.79785 & -.59912 & -6.6943 \times 10^{-2} \\
.48995 & .70912 & -.50706 \\
.35126 & .37176 & .85931
\end{array}\right)
\end{aligned}
$$

and this equals

$$
\left(\begin{array}{ccc}
6.0571 & 3.698 \times 10^{-3} & 3.4346 \times 10^{-5} \\
3.698 \times 10^{-3} & 3.2008 & -4.0643 \times 10^{-4} \\
3.4346 \times 10^{-5} & -4.0643 \times 10^{-4} & -.2579
\end{array}\right)
$$

By Gerschgorin's theorem, the eigenvalues are pretty close to the diagonal entries of the above matrix. Note I didn't use the theorem, just Lemma 15.3.3 and Gerschgorin's theorem to verify the eigenvalues are close to the above numbers. The eigenvectors are close to

$$
\left(\begin{array}{l}
.79785 \\
.48995 \\
.35126
\end{array}\right),\left(\begin{array}{c}
-.59912 \\
.70912 \\
.37176
\end{array}\right),\left(\begin{array}{c}
-6.6943 \times 10^{-2} \\
-.50706 \\
.85931
\end{array}\right)
$$

Lets check one of these.

$$
\begin{aligned}
& \left(\left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{array}\right)-6.0571\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)\right)\left(\begin{array}{l}
.79785 \\
.48995 \\
.35126
\end{array}\right) \\
= & \left(\begin{array}{c}
-2.1972 \times 10^{-3} \\
2.5439 \times 10^{-3} \\
1.3931 \times 10^{-3}
\end{array}\right) \approx\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
\end{aligned}
$$

Now lets see how well the smallest approximate eigenvalue and eigenvector works.

$$
\begin{gathered}
\left(\left(\begin{array}{lll}
5 & 1 & 1 \\
1 & 3 & 2 \\
1 & 2 & 1
\end{array}\right)-(-.2579)\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)\right)\left(\begin{array}{c}
-6.6943 \times 10^{-2} \\
-.50706 \\
.85931
\end{array}\right) \\
=\left(\begin{array}{c}
2.704 \times 10^{-4} \\
-2.7377 \times 10^{-4} \\
-1.3695 \times 10^{-4}
\end{array}\right) \approx\left(\begin{array}{l}
0 \\
0 \\
0
\end{array}\right)
\end{gathered}
$$

For practical purposes, this has found the eigenvalues and eigenvectors.

### 15.3.3 The $Q R$ Algorithm In The General Case

In the case where $A$ has distinct positive eigenvalues it was shown above that under reasonable conditions related to a certain matrix having an $L U$ factorization the $Q R$ algorithm produces a sequence of matrices $\left\{A_{k}\right\}$ which converges to an upper triangular matrix. What if $A$ is just an $n \times n$ matrix having possibly complex eigenvalues but $A$ is nondefective? What happens with the $Q R$ algorithm in this case? The short answer to this question is that the $A_{k}$ of the algorithm typically cannot converge. However, this does not mean the algorithm is not useful in finding eigenvalues. It turns out the sequence of matrices $\left\{A_{k}\right\}$ have the appearance of a block upper triangular matrix for large $k$ in the sense that the entries below the blocks on the main diagonal are small. Then looking at these blocks gives a way to approximate the eigenvalues.

First it is important to note a simple fact about unitary diagonal matrices. In what follows $\Lambda$ will denote a unitary matrix which is also a diagonal matrix. These matrices are just the identity matrix with some of the ones replaced with a number of the form $e^{i \theta}$ for some $\theta$. The important property of multiplication of any matrix by $\Lambda$ on either side is that it leaves all the zero entries the same and also preserves the absolute values of the
other entries. Thus a block triangular matrix multiplied by $\Lambda$ on either side is still block triangular. If the matrix is close to being block triangular this property of being close to a block triangular matrix is also preserved by multiplying on either side by $\Lambda$. Other patterns depending only on the size of the absolute value occurring in the matrix are also preserved by multiplying on either side by $\Lambda$. In other words, in looking for a pattern in a matrix, multiplication by $\Lambda$ is irrelevant.

Now let $A$ be an $n \times n$ matrix having real or complex entries. By Lemma 15.3.3 and the assumption that $A$ is nondefective, there exists an invertible $S$,

$$
\begin{equation*}
A^{k}=Q^{(k)} R^{(k)}=S D^{k} S^{-1} \tag{15.12}
\end{equation*}
$$

where

$$
D=\left(\begin{array}{ccc}
\lambda_{1} & & 0 \\
& \ddots & \\
0 & & \lambda_{n}
\end{array}\right)
$$

and by rearranging the columns of $S, D$ can be made such that $\left|\lambda_{1}\right| \geq\left|\lambda_{2}\right| \geq \cdots \geq\left|\lambda_{n}\right|$. Assume $S^{-1}$ has an $L U$ factorization. Then

$$
A^{k}=S D^{k} L U=S D^{k} L D^{-k} D^{k} U
$$

Consider the matrix in the middle, $D^{k} L D^{-k}$. The $i j^{t h}$ entry is of the form

$$
\left(D^{k} L D^{-k}\right)_{i j}=\left\{\begin{array}{l}
\lambda_{i}^{k} L_{i j} \lambda_{j}^{-k} \text { if } j<i \\
1 \text { if } i=j \\
0 \text { if } j>i
\end{array}\right.
$$

and these all converge to 0 whenever $\left|\lambda_{i}\right|<\left|\lambda_{j}\right|$. Thus $D^{k} L D^{-k}=\left(L_{k}+E_{k}\right)$ where $L_{k}$ is a lower triangular matrix which has all ones down the diagonal and some subdiagonal terms of the form

$$
\begin{equation*}
\lambda_{i}^{k} L_{i j} \lambda_{j}^{-k} \tag{15.13}
\end{equation*}
$$

for which $\left|\lambda_{i}\right|=\left|\lambda_{j}\right|$ while $E_{k} \rightarrow 0$. (Note the entries of $L_{k}$ are all bounded independent of $k$ but some may fail to converge.) Then

$$
Q^{(k)} R^{(k)}=S\left(L_{k}+E_{k}\right) D^{k} U
$$

Let

$$
\begin{equation*}
S L_{k}=Q_{k} R_{k} \tag{15.14}
\end{equation*}
$$

where this is the $Q R$ factorization of $S L_{k}$. Then

$$
\begin{aligned}
Q^{(k)} R^{(k)} & =\left(Q_{k} R_{k}+S E_{k}\right) D^{k} U \\
& =Q_{k}\left(I+Q_{k}^{*} S E_{k} R_{k}^{-1}\right) R_{k} D^{k} U \\
& =Q_{k}\left(I+F_{k}\right) R_{k} D^{k} U
\end{aligned}
$$

where $F_{k} \rightarrow 0$. Let $I+F_{k}=Q_{k}^{\prime} R_{k}^{\prime}$. Then $Q^{(k)} R^{(k)}=Q_{k} Q_{k}^{\prime} R_{k}^{\prime} R_{k} D^{k} U$. By Lemma 15.3.4

$$
\begin{equation*}
Q_{k}^{\prime} \rightarrow I \text { and } R_{k}^{\prime} \rightarrow I \tag{15.15}
\end{equation*}
$$

Now let $\Lambda_{k}$ be a diagonal unitary matrix which has the property that $\Lambda_{k}^{*} D^{k} U$ is an upper triangular matrix which has all the diagonal entries positive. Then

$$
Q^{(k)} R^{(k)}=Q_{k} Q_{k}^{\prime} \Lambda_{k}\left(\Lambda_{k}^{*} R_{k}^{\prime} R_{k} \Lambda_{k}\right) \Lambda_{k}^{*} D^{k} U
$$

That matrix in the middle has all positive diagonal entries because it is itself an upper triangular matrix, being the product of such, and is similar to the matrix $R_{k}^{\prime} R_{k}$ which is upper triangular with positive diagonal entries. By Lemma 15.3.4 again, this time using the uniqueness assertion,

$$
Q^{(k)}=Q_{k} Q_{k}^{\prime} \Lambda_{k}, R^{(k)}=\left(\Lambda_{k}^{*} R_{k}^{\prime} R_{k} \Lambda_{k}\right) \Lambda_{k}^{*} D^{k} U
$$

Note the term $Q_{k} Q_{k}^{\prime} \Lambda_{k}$ must be real because the algorithm gives all $Q^{(k)}$ as real matrices. By 15.15 it follows that for $k$ large enough $Q^{(k)} \approx Q_{k} \Lambda_{k}$ where $\approx$ means the two matrices are close. Recall $A_{k}=Q^{(k) T} A Q^{(k)}$ and so for large $k$,

$$
A_{k} \approx\left(Q_{k} \Lambda_{k}\right)^{*} A\left(Q_{k} \Lambda_{k}\right)=\Lambda_{k}^{*} Q_{k}^{*} A Q_{k} \Lambda_{k}
$$

As noted above, the form of $\Lambda_{k}^{*} Q_{k}^{*} A Q_{k} \Lambda_{k}$ in terms of which entries are large and small is not affected by the presence of $\Lambda_{k}$ and $\Lambda_{k}^{*}$. Thus, in considering what form this is in, it suffices to consider $Q_{k}^{*} A Q_{k}$.

This could get pretty complicated but I will consider the case where

$$
\begin{equation*}
\text { if }\left|\lambda_{i}\right|=\left|\lambda_{i+1}\right| \text {, then }\left|\lambda_{i+2}\right|<\left|\lambda_{i+1}\right| \tag{15.16}
\end{equation*}
$$

This is typical of the situation where the eigenvalues are all distinct and the matrix $A$ is real so the eigenvalues occur as conjugate pairs. Then in this case, $L_{k}$ above is lower triangular with some nonzero terms on the diagonal right below the main diagonal but zeros everywhere else. Thus maybe $\left(L_{k}\right)_{s+1, s} \neq 0$ Recall 15.14 which implies

$$
\begin{equation*}
Q_{k}=S L_{k} R_{k}^{-1} \tag{15.17}
\end{equation*}
$$

where $R_{k}^{-1}$ is upper triangular. Also recall from the definition of $S$ in 15.12 , it follows that $S^{-1} A S=D$. Thus the columns of $S$ are eigenvectors of $A$, the $i^{\text {th }}$ being an eigenvector for $\lambda_{i}$. Now from the form of $L_{k}$, it follows $L_{k} R_{k}^{-1}$ is a block upper triangular matrix denoted by $T_{B}$ and so $Q_{k}=S T_{B}$. It follows from the above construction in 15.13 and the given assumption on the sizes of the eigenvalues, there are finitely many $2 \times 2$ blocks centered on the main diagonal along with possibly some diagonal entries. Therefore, for large $k$ the matrix $A_{k}=Q^{(k) T} A Q^{(k)}$ is approximately of the same form as that of

$$
Q_{k}^{*} A Q_{k}=T_{B}^{-1} S^{-1} A S T_{B}=T_{B}^{-1} D T_{B}
$$

which is a block upper triangular matrix. As explained above, multiplication by the various diagonal unitary matrices does not affect this form. Therefore, for large $k, A_{k}$ is approximately a block upper triangular matrix.

How would this change if the above assumption on the size of the eigenvalues were relaxed but the matrix was still nondefective with appropriate matrices having an $L U$ factorization as above? It would mean the blocks on the diagonal would be larger. This immediately makes the problem more cumbersome to deal with. However, in the case that the eigenvalues of $A$ are distinct, the above situation really is typical of what occurs and in any case can be quickly reduced to this case.

To see this, suppose condition 15.16 is violated and $\lambda_{j}, \cdots, \lambda_{j+p}$ are complex eigenvalues having nonzero imaginary parts such that each has the same absolute value but they are all distinct. Then let $\mu>0$ and consider the matrix $A+\mu I$. Thus the corresponding eigenvalues of $A+\mu I$ are $\lambda_{j}+\mu, \cdots, \lambda_{j+p}+\mu$. A short computation shows $\left|\lambda_{j}+\mu\right|, \cdots,\left|\lambda_{j+p}+\mu\right|$ are all distinct and so the above situation of 15.16 is obtained. Of course, if there are repeated eigenvalues, it may not be possible to reduce to the case above and you would end up with large blocks on the main diagonal which could be difficult to deal with.

So how do you identify the eigenvalues? You know $A_{k}$ and behold that it is close to a block upper triangular matrix $T_{B}^{\prime}$. You know $A_{k}$ is also similar to $A$. Therefore, $T_{B}^{\prime}$ has eigenvalues which are close to the eigenvalues of $A_{k}$ and hence those of $A$ provided $k$ is sufficiently large. See Theorem 13.4.2 which depends on complex analysis or the exercise on Page 347 which gives another way to see this. Thus you find the eigenvalues of this block triangular matrix $T_{B}^{\prime}$ and assert that these are good approximations of the eigenvalues of $A_{k}$ and hence to those of $A$. How do you find the eigenvalues of a block triangular matrix? This is easy from Lemma 13.1.4. Say

$$
T_{B}^{\prime}=\left(\begin{array}{ccc}
B_{1} & \cdots & * \\
& \ddots & \vdots \\
0 & & B_{m}
\end{array}\right)
$$

Then forming $\lambda I-T_{B}^{\prime}$ and taking the determinant, it follows from Lemma 13.1.4 this equals

$$
\prod_{j=1}^{m} \operatorname{det}\left(\lambda I_{j}-B_{j}\right)
$$

and so all you have to do is take the union of the eigenvalues for each $B_{j}$. In the case emphasized here this is very easy because these blocks are just $2 \times 2$ matrices.

How do you identify approximate eigenvectors from this? First try to find the approximate eigenvectors for $A_{k}$. Pick an approximate eigenvalue $\lambda$, an exact eigenvalue for $T_{B}^{\prime}$. Then find $\boldsymbol{v}$ solving $T_{B}^{\prime} \boldsymbol{v}=\lambda \boldsymbol{v}$. It follows since $T_{B}^{\prime}$ is close to $A_{k}$ that $A_{k} \boldsymbol{v} \cong \lambda \boldsymbol{v}$ and so

$$
Q^{(k)} A Q^{(k) T} \boldsymbol{v}=A_{k} \boldsymbol{v} \approx \lambda \boldsymbol{v}
$$

Hence

$$
A Q^{(k) T} \boldsymbol{v} \approx \lambda Q^{(k) T} \boldsymbol{v}
$$

and so $Q^{(k) T} v$ is an approximation to the eigenvector which goes with the eigenvalue of $A$ which is close to $\lambda$.

Example 15.3.8 Here is a matrix.

$$
\left(\begin{array}{ccc}
3 & 2 & 1 \\
-2 & 0 & -1 \\
-2 & -2 & 0
\end{array}\right)
$$

It happens that the eigenvalues of this matrix are $1,1+i, 1-i$. Lets apply the $Q R$ algorithm as if the eigenvalues were not known.

Applying the $Q R$ algorithm to this matrix yields the following sequence of matrices.

$$
\begin{gathered}
A_{1}=\left(\begin{array}{ccc}
1.2353 & 1.9412 & 4.3657 \\
-.39215 & 1.5425 & 5.3886 \times 10^{-2} \\
-.16169 & -.18864 & .22222
\end{array}\right) \\
\vdots \\
A_{12}=\left(\begin{array}{ccc}
9.1772 \times 10^{-2} & .63089 & -2.0398 \\
-2.8556 & 1.9082 & -3.1043 \\
1.0786 \times 10^{-2} & 3.4614 \times 10^{-4} & 1.0
\end{array}\right)
\end{gathered}
$$

At this point the bottom two terms on the left part of the bottom row are both very small so it appears the real eigenvalue is near 1.0. The complex eigenvalues are obtained from solving

$$
\operatorname{det}\left(\lambda\left(\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right)-\left(\begin{array}{cc}
9.1772 \times 10^{-2} & .63089 \\
-2.8556 & 1.9082
\end{array}\right)\right)=0
$$

This yields

$$
\lambda=1.0-.98828 i, 1.0+.98828 i
$$

Example 15.3.9 The equation $x^{4}+x^{3}+4 x^{2}+x-2=0$ has exactly two real solutions. You can see this by graphing it. However, the rational root theorem from algebra shows neither of these solutions are rational. Also, graphing it does not yield any information about the complex solutions. Lets use the QR algorithm to approximate all the solutions, real and complex.

A matrix whose characteristic polynomial is the given polynomial is

$$
\left(\begin{array}{cccc}
-1 & -4 & -1 & 2 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right)
$$

Using the $Q R$ algorithm yields the following sequence of iterates for $A_{k}$

$$
\begin{gathered}
A_{1}=\left(\begin{array}{cccc}
.99999 & -2.5927 & -1.7588 & -1.2978 \\
2.1213 & -1.7778 & -1.6042 & -.99415 \\
0 & .34246 & -.32749 & -.91799 \\
0 & 0 & -.44659 & .10526
\end{array}\right) \\
\vdots \\
A_{9}=\left(\begin{array}{cccc}
-.83412 & -4.1682 & -1.939 & -.7783 \\
1.05 & .14514 & .2171 & 2.5474 \times 10^{-2} \\
0 & 4.0264 \times 10^{-4} & -.85029 & -.61608 \\
0 & 0 & -1.8263 \times 10^{-2} & .53939
\end{array}\right)
\end{gathered}
$$

Now this is similar to $A$ and the eigenvalues are close to the eigenvalues obtained from the two blocks on the diagonal,

$$
\left(\begin{array}{cc}
-.83412 & -4.1682 \\
1.05 & .14514
\end{array}\right),\left(\begin{array}{cc}
-.85029 & -.61608 \\
-1.8263 \times 10^{-2} & .53939
\end{array}\right)
$$

since $4.0264 \times 10^{-4}$ is small. After routine computations involving the quadratic formula, these are seen to be

$$
-.85834, .54744,-.34449-2.0339 i,-.34449+2.0339 i
$$

When these are plugged in to the polynomial equation, you see that each is close to being a solution of the equation.

### 15.3.4 Upper Hessenberg Matrices

It seems like most of the attention to the $Q R$ algorithm has to do with finding ways to get it to "converge" faster. Great and marvelous are the clever tricks which have been proposed to do this but my intent is to present the basic ideas, not to go in to the numerous refinements of this algorithm. However, there is one thing which should be done. It involves reducing to the case of an upper Hessenberg matrix which is one which is zero below the main sub diagonal. The following shows that any square matrix is unitarily similar to such an upper Hessenberg matrix.

Let $A$ be an invertible $n \times n$ matrix. Let $Q_{1}^{\prime}$ be a unitary matrix

$$
Q_{1}^{\prime}\left(\begin{array}{c}
a_{21} \\
\vdots \\
a_{n 1}
\end{array}\right)=\left(\begin{array}{c}
\sqrt{\sum_{j=2}^{n}\left|a_{j 1}\right|^{2}} \\
0 \\
\vdots \\
0
\end{array}\right) \equiv\left(\begin{array}{c}
a \\
0 \\
\vdots \\
0
\end{array}\right)
$$

The vector $Q_{1}^{\prime}$ is multiplying is just the bottom $n-1$ entries of the first column of $A$. Then let $Q_{1}$ be

$$
\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & Q_{1}^{\prime}
\end{array}\right)
$$

It follows

$$
\begin{aligned}
& Q_{1} A Q_{1}^{*}=\left(\begin{array}{cc}
1 & \mathbf{0} \\
\mathbf{0} & Q_{1}^{\prime}
\end{array}\right) A Q_{1}^{*}=\left(\begin{array}{cccc}
a_{11} & a_{12} & \cdots & a_{1 n} \\
a & & & \\
\vdots & & A_{1}^{\prime} & \\
0 & &
\end{array}\right)\left(\begin{array}{ccc}
1 & \mathbf{0} \\
\mathbf{0} & Q_{1}^{\prime *}
\end{array}\right) \\
&=\left(\begin{array}{cccc}
* & * & \cdots & * \\
a & & \\
\vdots & & A_{1} \\
0 & &
\end{array}\right)
\end{aligned}
$$

Now let $Q_{2}^{\prime}$ be the $n-2 \times n-2$ matrix which does to the first column of $A_{1}$ the same sort of thing that the $n-1 \times n-1$ matrix $Q_{1}^{\prime}$ did to the first column of $A$. Let

$$
Q_{2} \equiv\left(\begin{array}{cc}
I & 0 \\
0 & Q_{2}^{\prime}
\end{array}\right)
$$

where $I$ is the $2 \times 2$ identity. Then applying block multiplication,

$$
Q_{2} Q_{1} A Q_{1}^{*} Q_{2}^{*}=\left(\begin{array}{ccccc}
* & * & \cdots & * & * \\
* & * & \cdots & * & * \\
0 & * & & & \\
\vdots & \vdots & & A_{2} & \\
0 & 0 & & &
\end{array}\right)
$$

where $A_{2}$ is now an $n-2 \times n-2$ matrix. Continuing this way you eventually get a unitary matrix $Q$ which is a product of those discussed above such that

$$
Q A Q^{T}=\left(\begin{array}{ccccc}
* & * & \cdots & * & * \\
* & * & \cdots & * & * \\
0 & * & * & & \vdots \\
\vdots & \vdots & \ddots & \ddots & * \\
0 & 0 & & * & *
\end{array}\right)
$$

This matrix equals zero below the subdiagonal. It is called an upper Hessenberg matrix.
It happens that in the $Q R$ algorithm, if $A_{k}$ is upper Hessenberg, so is $A_{k+1}$. To see this, note that the matrix is upper Hessenberg means that $A_{i j}=0$ whenever $i-j \geq 2$.

$$
A_{k+1}=R_{k} Q_{k}
$$

where $A_{k}=Q_{k} R_{k}$. Therefore as shown before,

$$
A_{k+1}=R_{k} A_{k} R_{k}^{-1}
$$

Let the $i j^{t h}$ entry of $A_{k}$ be $a_{i j}^{k}$. Then if $i-j \geq 2$

$$
a_{i j}^{k+1}=\sum_{p=i}^{n} \sum_{q=1}^{j} r_{i p} a_{p q}^{k} r_{q j}^{-1}
$$

It is given that $a_{p q}^{k}=0$ whenever $p-q \geq 2$. However, from the above sum,

$$
p-q \geq i-j \geq 2
$$

and so the sum equals 0 .
Since upper Hessenberg matrices stay that way in the algorithm and it is closer to being upper triangular, it is reasonable to suppose the $Q R$ algorithm will yield good results more quickly for this upper Hessenberg matrix than for the original matrix. This would be especially true if the matrix is good sized. The other important thing to observe is that, starting with an upper Hessenberg matrix, the algorithm will restrict the size of the blocks which occur to being $2 \times 2$ blocks which are easy to deal with. These blocks allow you to identify the eigenvalues.

Example 15.3.10 Let $A=\left(\begin{array}{cccc}1 & 2 & 3 & 4 \\ 2 & -2 & -3 & 3 \\ 3 & -3 & 5 & 1 \\ 4 & 3 & 1 & -3\end{array}\right)$ a symmetric matrix. Thus it has real eigenvalues and can be diagonalized. Find its eigenvalues.

As explained above, there is an upper Hessenberg matrix. Matlab can find it using the techniques given above pretty quickly. The syntax is as follows.

$$
\begin{aligned}
& \mathrm{A}=\left[\begin{array}{lll}
2 & 1 & 3 ;-5,3,-2 ; 1,2,3] \\
{[\mathrm{P}, \mathrm{H}]=\operatorname{hess}(\mathrm{A})}
\end{array}\right.
\end{aligned}
$$

Then the Hessenberg matrix similar to $A$ is

$$
H=\left(\begin{array}{cccc}
-1.4476 & -4.9048 & 0 & 0 \\
-4.9048 & 3.2553 & -2.0479 & 0 \\
0 & -2.0479 & 2.1923 & -5.0990 \\
0 & 0 & -5.0990 & -3
\end{array}\right)
$$

Note how it is symmetric also. This will always happen when you begin with a symmetric matrix. Now use the $Q R$ algorithm on this matrix. The syntax is as follows in Matlab.
$\mathrm{H}=[$ enter H here]
hold on
for $\mathrm{k}=1: 100$
$[\mathrm{Q}, \mathrm{R}]=\mathrm{qr}(\mathrm{H}) ;$
$\mathrm{H}=\mathrm{R} * \mathrm{Q} ;$
end
Q
R
H

You already have $H$ and matlab knows about it so you don't need to enter $H$ again. This yields the following matrix similar to the original one.

$$
\left(\begin{array}{cccc}
7.4618 & 0 & 0 & 0 \\
0 & -6.3804 & 0 & 0 \\
0 & 0 & -4.419 & -.3679 \\
0 & 0 & -.3679 & 4.3376
\end{array}\right)
$$

The eigenvalues of this matrix are

$$
7.4618,-6.3804,4.353,-4.4344
$$

You might want to check that the product of these equals the determinant of the matrix and that the sum equals the trace of the matrix. In fact, this works out very well. To find eigenvectors, you could use the shifted inverse power method. They will be different for the Hessenberg matrix than for the original matrix $A$.

### 15.4 Exercises

In these exercises which call for a computation, don't waste time on them unless you use a computer or calculator which can raise matrices to powers and take $Q R$ factorizations.

1. In Example 15.2.7 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.
2. Find the eigenvalues and eigenvectors of the matrix $A=\left(\begin{array}{ccc}3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 2\end{array}\right)$ numerically. In this case the exact eigenvalues are $\pm \sqrt{3}, 6$. Compare with the exact answers.
3. Find the eigenvalues and eigenvectors of the matrix $A=\left(\begin{array}{ccc}3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2\end{array}\right)$ numerically. The exact eigenvalues are $2,4+\sqrt{15}, 4-\sqrt{15}$. Compare your numerical results with the exact values. Is it much fun to compute the exact eigenvectors?
4. Find the eigenvalues and eigenvectors of the matrix $A=\left(\begin{array}{ccc}0 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2\end{array}\right)$ numerically. I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.
5. Find the eigenvalues and eigenvectors of the matrix $A=\left(\begin{array}{ccc}0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2\end{array}\right)$ numerically. I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.
6. Consider the matrix $A=\left(\begin{array}{lll}3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0\end{array}\right)$ and the vector $(1,1,1)^{T}$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of $A$.
7. Consider the matrix $A=\left(\begin{array}{lll}1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5\end{array}\right)$ and the vector $(1,1,1)^{T}$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of $A$.
8. Consider the matrix $A=\left(\begin{array}{ccc}3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3\end{array}\right)$ and the vector $(1,1,1)^{T}$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of $A$.
9. Using Gerschgorin's theorem, find upper and lower bounds for the eigenvalues of $A=\left(\begin{array}{ccc}3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3\end{array}\right)$.
10. Tell how to find a matrix whose characteristic polynomial is a given monic polynomial. This is called a companion matrix. Find the roots of the polynomial $x^{3}+7 x^{2}+$ $3 x+7$.
11. Find the roots to $x^{4}+3 x^{3}+4 x^{2}+x+1$. It has two complex roots.
12. Suppose $A$ is a real symmetric matrix and the technique of reducing to an upper Hessenberg matrix is followed. Show the resulting upper Hessenberg matrix is actually equal to 0 on the top as well as the bottom.

## Chapter 16

## Approximation of Functions and the Integral

These topics are not about linear algebra, but linear algebra is used in a very essential manner so these topics are applications of Linear algebra to analysis. Many more examples could be included but this book is long enough.

This chapter is just what the title indicates. It will involve approximating functions and a simple definition of the integral. This definition is sufficient to consider all piecewise continuous functions and it does not depend on Riemann sums. Thus it is closer to what was done in the 1700 's than in the 1800 's. However, it is based on the Weierstrass approximation theorem so it is definitely dependent on material which originated in the 1800's. After this, is a very interesting application of ideas from linear algebra to prove Müntz's theorems.

The notation $C([0, b] ; X)$ will denote the functions which are continuous with values in $[0, b]$ with values in $X$ which will always be a normed vector space. It could be $\mathbb{C}$ or $\mathbb{R}$ for example.

### 16.1 Weierstrass Approximation Theorem

An arbitrary continuous function defined on an interval can be approximated uniformly by a polynomial, there exists a similar theorem which is just a generalization of this which will hold for continuous functions defined on a box or more generally a closed and bounded set. However, we will settle for the case of a box first. The proof is based on the following lemma.

Lemma 16.1. 1 The following estimate holds for $x \in[0,1]$ and $m \geq 2$.

$$
\sum_{k=0}^{m}\binom{m}{k}(k-m x)^{2} x^{k}(1-x)^{m-k} \leq \frac{1}{4} m
$$

Proof: First of all, from the binomial theorem

$$
\sum_{k=0}^{m}\binom{m}{k}(t x)^{k}(1-x)^{m-k}=(1-x+t x)^{m}
$$

Take a derivative and then let $t=1$.

$$
\begin{gathered}
\sum_{k=0}^{m}\binom{m}{k} k(t x)^{k-1} x(1-x)^{m-k}=m x(t x-x+1)^{m-1} \\
\sum_{k=0}^{m}\binom{m}{k} k(x)^{k}(1-x)^{m-k}=m x
\end{gathered}
$$

Then also,

$$
\sum_{k=0}^{m}\binom{m}{k} k(t x)^{k}(1-x)^{m-k}=m x t(t x-x+1)^{m-1}
$$

Take another time derivative of both sides.

$$
\begin{aligned}
& \sum_{k=0}^{m}\binom{m}{k} k^{2}(t x)^{k-1} x(1-x)^{m-k} \\
= & m x\left((t x-x+1)^{m-1}-t x(t x-x+1)^{m-2}+m t x(t x-x+1)^{m-2}\right)
\end{aligned}
$$

Plug in $t=1$.

$$
\sum_{k=0}^{m}\binom{m}{k} k^{2} x^{k}(1-x)^{m-k}=m x(m x-x+1)
$$

Then it follows

$$
\begin{gathered}
\sum_{k=0}^{m}\binom{m}{k}(k-m x)^{2} x^{k}(1-x)^{m-k} \\
=\sum_{k=0}^{m}\binom{m}{k}\left(k^{2}-2 k m x+x^{2} m^{2}\right) x^{k}(1-x)^{m-k}
\end{gathered}
$$

and from what was just shown, this equals

$$
x^{2} m^{2}-x^{2} m+m x-2 m x(m x)+x^{2} m^{2}=-x^{2} m+m x=\frac{m}{4}-m\left(x-\frac{1}{2}\right)^{2}
$$

Thus the expression is maximized when $x=1 / 2$ and yields $m / 4$ in this case. This proves the lemma.

With this preparation, here is the first version of the Weierstrass approximation theorem. I will allow $f$ to have values in a complete, real or complex normed linear space. Thus, $f \in C([0,1] ; X)$ where $X$ is a Banach space, Definition 14.4.1. Thus this is a function which is continuous with values in $X$ as discussed earlier with metric spaces.

Theorem 16.1.2 Let $f \in C([0,1] ; X)$ and let the norm on $X$ be denoted by $\|\cdot\|$.

$$
p_{m}(x) \equiv \sum_{k=0}^{m}\binom{m}{k} x^{k}(1-x)^{m-k} f\left(\frac{k}{m}\right) .
$$

Then these polynomials having coefficients in $X$ converge uniformly to $f$ on $[0,1]$.
Proof: Let $\|f\|_{\infty}$ denote the largest value of $\|f(x)\|$. By uniform continuity of $f$, there exists a $\delta>0$ such that if $\left|x-x^{\prime}\right|<\delta$, then $\left\|f(x)-f\left(x^{\prime}\right)\right\|<\varepsilon / 2$. By the binomial theorem,

$$
\begin{aligned}
& \left\|p_{m}(x)-f(x)\right\| \leq \sum_{k=0}^{m}\binom{m}{k} x^{k}(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\| \\
& \leq \sum_{\left|\frac{k}{m}-x\right|<\delta}\binom{m}{k} x^{k}(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\|+ \\
& 2\|f\|_{\infty} \sum_{\left|\frac{k}{m}-x\right| \geq \delta}\binom{m}{k} x^{k}(1-x)^{m-k}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \leq \sum_{k=0}^{m}\binom{m}{k} x^{k}(1-x)^{m-k} \frac{\varepsilon}{2}+2\|f\|_{\infty} \sum_{(k-m x)^{2} \geq m^{2} \delta^{2}}\binom{m}{k} x^{k}(1-x)^{m-k} \\
& \quad \leq \frac{\varepsilon}{2}+2\|f\|_{\infty} \frac{1}{m^{2} \delta^{2}} \sum_{k=0}^{m}\binom{m}{k}(k-m x)^{2} x^{k}(1-x)^{m-k} \\
& \quad \leq \frac{\varepsilon}{2}+2\|f\|_{\infty} \frac{1}{4} m \frac{1}{\delta^{2} m^{2}}<\varepsilon
\end{aligned}
$$

provided $m$ is large enough. Thus $\left\|p_{m}-f\right\|_{\infty}<\varepsilon$ when $m$ is large enough.
Note that we do not need to have $X$ be complete in order for this to hold. It would have sufficed to have simply let $X$ be a normed linear space.

Corollary 16.1.3 If $f \in C([a, b] ; X)$ where $X$ is a normed linear space, then there exists $a$ sequence of polynomials which converge uniformly to $f$ on $[a, b]$.

Proof: Let $l:[0,1] \rightarrow[a, b]$ be one to one, linear and onto. Then $f \circ l$ is continuous on $[0,1]$ and so if $\varepsilon>0$ is given, there exists a polynomial $p$ such that for all $x \in[0,1],\|p(x)-f \circ l(x)\|<\varepsilon$. Therefore, letting $y=l(x)$, it follows that for all $y \in[a, b]$,

$$
\left\|p\left(l^{-1}(y)\right)-f(y)\right\|<\varepsilon
$$

The exact form of the polynomial is as follows.

$$
\begin{gather*}
p(x)=\sum_{k=0}^{m}\binom{m}{k} x^{k}(1-x)^{m-k} f\left(l\left(\frac{k}{m}\right)\right) \\
p\left(l^{-1}(y)\right)=\sum_{k=0}^{m}\binom{m}{k}\left(l^{-1}(y)\right)^{k}\left(1-l^{-1}(y)\right)^{m-k} f\left(l\left(\frac{k}{m}\right)\right) \tag{16.1}
\end{gather*}
$$

Here is a corollary.
Corollary 16.1.4 Let $f$ be a continuous function defined on $[-M, M]$ with $f(0)=0$. Then there exists a sequence of polynomials $\left\{p_{m}\right\}, p_{m}(0)=0$ and

$$
\lim _{m \rightarrow \infty}\left\|p_{m}-f\right\|_{\infty}=0
$$

Proof: From Corollary 16.1.3 there exists a sequence of polynomials $\left\{\widehat{p_{m}}\right\}$ such that $\left\|\widehat{p_{m}}-f\right\|_{\infty} \rightarrow 0$. Simply consider $p_{m}=\widehat{p_{m}}-\widehat{p_{m}}(0)$.

### 16.2 Functions of Many Variables

First note that if $h: K \times H \rightarrow \mathbb{R}$ is a real valued continuous function where $K, H$ are compact sets in metric spaces,

$$
\max _{x \in K} h(x, y) \geq h(x, y), \text { so } \max _{y \in H} \max _{x \in K} h(x, y) \geq h(x, y)
$$

which implies $\max _{y \in H} \max _{x \in K} h(x, y) \geq \max _{(x, y) \in K \times H} h(x, y)$. The other inequality is also obtained.

Let $\boldsymbol{f} \in C\left(R_{p} ; X\right)$ where $R_{p}=[0,1]^{p}$. Then let $\hat{\boldsymbol{x}}_{p} \equiv\left(x_{1}, \ldots, x_{p-1}\right)$. By Theorem 16.1.2, if $n$ is large enough,

$$
\max _{x_{p} \in[0,1]}\left\|\sum_{k=0}^{n} f\left(\cdot, \frac{k}{n}\right)\binom{n}{k} x_{p}^{k}\left(1-x_{p}\right)^{n-k}-\boldsymbol{f}\left(\cdot, x_{p}\right)\right\|_{C\left([0,1]^{p-1} ; X\right)}<\frac{\varepsilon}{2}
$$

Now $\boldsymbol{f}\left(\cdot, \frac{k}{n}\right) \in C\left(R_{p-1} ; X\right)$ and so by induction, there is a polynomial $\boldsymbol{p}_{k}\left(\hat{\boldsymbol{x}}_{p}\right)$ such that

$$
\max _{\hat{\boldsymbol{x}}_{p} \in R_{p-1}}\left\|\boldsymbol{p}_{k}\left(\hat{\boldsymbol{x}}_{p}\right)-\binom{n}{k} \boldsymbol{f}\left(\hat{\boldsymbol{x}}_{p}, \frac{k}{n}\right)\right\|_{X}<\frac{\varepsilon}{(n+1) 2}
$$

Thus, letting $\boldsymbol{p}(\boldsymbol{x}) \equiv \sum_{k=0}^{n} \boldsymbol{p}_{k}\left(\hat{\boldsymbol{x}}_{p}\right) x_{p}^{k}\left(1-x_{p}\right)^{n-k}$,

$$
\|\boldsymbol{p}-\boldsymbol{f}\|_{C\left(R_{p} ; X\right)} \leq \max _{x_{p} \in[0,1]} \max _{\hat{\boldsymbol{x}}_{p} \in R_{p-1}}\left\|\boldsymbol{p}\left(\hat{\boldsymbol{x}}_{p}, x_{p}\right)-\boldsymbol{f}\left(\hat{\boldsymbol{x}}_{p}, x_{p}\right)\right\|_{X}<\boldsymbol{\varepsilon}
$$

where $\boldsymbol{p}$ is a polynomial with coefficients in $X$.
In general, if $R_{p} \equiv \prod_{k=1}^{p}\left[a_{k}, b_{k}\right]$, note that there is a linear function $l_{k}:[0,1] \rightarrow\left[a_{k}, b_{k}\right]$ which is one to one and onto. Thus $\boldsymbol{l}(\boldsymbol{x}) \equiv\left(l_{1}\left(x_{1}\right), \ldots, l_{p}\left(x_{p}\right)\right)$ is a one to one and onto map from $[0,1]^{p}$ to $R_{p}$ and the above result can be applied to $\boldsymbol{f} \circ \boldsymbol{l}$ to obtain a polynomial $\boldsymbol{p}$ with $\|\boldsymbol{p}-\boldsymbol{f} \circ \boldsymbol{l}\|_{C\left([0,1]^{p} ; X\right)}<\boldsymbol{\varepsilon}$. Thus $\left\|\boldsymbol{p} \circ \boldsymbol{l}^{-1}-\boldsymbol{f}\right\|_{C\left(R_{p} ; X\right)}<\boldsymbol{\varepsilon}$ and $\boldsymbol{p} \circ \boldsymbol{l}^{-1}$ is a polynomial. This proves the following theorem.

Theorem 16.2.1 Let $\boldsymbol{f}$ be a function in $C(R ; X)$ for $X$ a normed linear space where $R \equiv$ $\prod_{k=1}^{p}\left[a_{k}, b_{k}\right]$. Then for any $\varepsilon>0$ there exists a polynomial $\boldsymbol{p}$ having coefficients in $X$ such that $\|\boldsymbol{p}-\boldsymbol{f}\|_{C(R ; X)}<\varepsilon$.

These Bernstein polynomials are very remarkable approximations. It turns out that if $f$ is $C^{1}([0,1] ; X)$, then $\lim _{n \rightarrow \infty} p_{n}^{\prime}(x) \rightarrow f^{\prime}(x)$ uniformly on $[0,1]$. This all works for functions of many variables as well, but here I will only show it for functions of one variable. I assume the reader knows about the derivative of a function of one variable.

Lemma 16.2.2 Let $f \in C^{1}([0,1])$ and let

$$
p_{m}(x) \equiv \sum_{k=0}^{m}\binom{m}{k} x^{k}(1-x)^{m-k} f\left(\frac{k}{m}\right)
$$

be the $m^{\text {th }}$ Bernstein polynomial. Then in addition to $\left\|p_{m}-f\right\|_{[0,1]} \rightarrow 0$, it also follows that

$$
\left\|p_{m}^{\prime}-f^{\prime}\right\|_{[0,1]} \rightarrow 0
$$

Proof: From simple computations,

$$
\begin{aligned}
p_{m}^{\prime}(x)= & \sum_{k=1}^{m}\binom{m}{k} k x^{k-1}(1-x)^{m-k} f\left(\frac{k}{m}\right) \\
& -\sum_{k=0}^{m-1}\binom{m}{k} x^{k}(m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \frac{m(m-1)!}{(m-k)!(k-1)!} x^{k-1}(1-x)^{m-k} f\left(\frac{k}{m}\right) \\
&-\sum_{k=0}^{m-1}\binom{m}{k} x^{k}(m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right) \\
&= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^{k}(1-x)^{m-1-k} f\left(\frac{k+1}{m}\right) \\
& \quad-\sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^{k}(1-x)^{m-1-k} f\left(\frac{k}{m}\right) \\
&=\sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^{k}(1-x)^{m-1-k}\left(f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)\right) \\
&= \sum_{k=0}^{m-1}\binom{m-1}{k} x^{k}(1-x)^{m-1-k}\left(\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1 / m}\right)
\end{aligned}
$$

By the mean value theorem,

$$
\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1 / m}=f^{\prime}\left(x_{k, m}\right), x_{k, m} \in\left(\frac{k}{m}, \frac{k+1}{m}\right)
$$

Now the desired result follows as before from the uniform continuity of $f^{\prime}$ on $[0,1]$. Let $\delta>0$ be such that if

$$
|x-y|<\delta, \text { then }\left|f^{\prime}(x)-f^{\prime}(y)\right|<\varepsilon
$$

and let $m$ be so large that $1 / m<\delta / 2$. Then if $\left|x-\frac{k}{m}\right|<\delta / 2$, it follows that $\left|x-x_{k, m}\right|<\delta$ and so

$$
\left|f^{\prime}(x)-f^{\prime}\left(x_{k, m}\right)\right|=\left|f^{\prime}(x)-\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1 / m}\right|<\varepsilon
$$

Now as before, letting $M \geq\left|f^{\prime}(x)\right|$ for all $x$,

$$
\begin{aligned}
& \left|p_{m}^{\prime}(x)-f^{\prime}(x)\right| \leq \sum_{k=0}^{m-1}\binom{m-1}{k} x^{k}(1-x)^{m-1-k}\left|f^{\prime}\left(x_{k, m}\right)-f^{\prime}(x)\right| \\
& \leq \quad \sum_{\left\{x:\left|x-\frac{k}{m}\right|<\frac{\delta}{2}\right\}}\binom{m-1}{k} x^{k}(1-x)^{m-1-k} \varepsilon \\
& +M \sum_{k=0}^{m-1}\binom{m-1}{k} \frac{4(k-m x)^{2}}{m^{2} \delta^{2}} x^{k}(1-x)^{m-1-k} \\
& \quad \leq \varepsilon+4 M \frac{1}{4} m \frac{1}{m^{2} \delta^{2}}=\varepsilon+M \frac{1}{m \delta^{2}}<2 \varepsilon
\end{aligned}
$$

whenever $m$ is large enough. Thus this proves uniform convergence.

### 16.3 A Generalization with Tietze Extension Theorem

The following is the Tietze extension theorem, Theorem 10.5.5 presented earlier.
Theorem 16.3.1 Let $M$ be a closed nonempty subset of $X$ and let $f: M \rightarrow[a, b]$ be continuous at every point of $M$. Then there exists a function, $g$ continuous on all of $X$ which coincides with $f$ on $M$ such that $g(X) \subseteq[a, b]$.

With the Tietze extension theorem, here is a better version of the Weierstrass approximation theorem.

Theorem 16.3.2 Let $K$ be a closed and bounded subset of $\mathbb{R}^{p}$ and let $f: K \rightarrow \mathbb{R}$ be continuous. Then there exists a sequence of polynomials $\left\{p_{m}\right\}$ such that

$$
\lim _{m \rightarrow \infty}\left(\sup \left\{\left|f(\boldsymbol{x})-p_{m}(\boldsymbol{x})\right|: \boldsymbol{x} \in K\right\}\right)=0
$$

In other words, the sequence of polynomials converges uniformly to $f$ on $K$.
Proof: By the Tietze extension theorem, there exists an extension of $f$ to a continuous function $g$ defined on all $\mathbb{R}^{p}$ such that $g=f$ on $K$. Now since $K$ is bounded, there exist intervals, $\left[a_{k}, b_{k}\right]$ such that $K \subseteq \prod_{k=1}^{p}\left[a_{k}, b_{k}\right]=R$. Then by the Weierstrass approximation theorem, Theorem 16.2.1 there exists a sequence of polynomials $\left\{p_{m}\right\}$ converging uniformly to $g$ on $R$. Therefore, this sequence of polynomials converges uniformly to $g=f$ on $K$ as well. This proves the theorem.

By considering the real and imaginary parts of a function which has values in $\mathbb{C}$ one can generalize the above theorem.

Corollary 16.3.3 Let $K$ be a closed and bounded subset of $\mathbb{R}^{p}$ and let $f: K \rightarrow \mathbb{F}$ be continuous. Then there exists a sequence of polynomials $\left\{p_{m}\right\}$ such that

$$
\lim _{m \rightarrow \infty}\left(\sup \left\{\left|f(\boldsymbol{x})-p_{m}(\boldsymbol{x})\right|: \boldsymbol{x} \in K\right\}\right)=0
$$

In other words, the sequence of polynomials converges uniformly to $f$ on $K$.
More generally, the function $f$ could have values in $\mathbb{R}^{p}$. There is no change in the proof. You just use norm symbols rather than absolute values and nothing at all changes in the theorem where the function is defined on a rectangle. Then you apply the Tietze extension theorem to each component in the case the function has values in $\mathbb{R}^{p}$. Using a better extension theorem than what is presented in this book, one could generalize this to a function having values in a Banach space.

### 16.4 An Approach to the Integral

First is a short review of the derivative of a function of one variable.
Definition 16.4.1 Let $f:[a, b] \rightarrow \mathbb{R}$. Then $f^{\prime}(x) \equiv \lim _{x \rightarrow 0} \frac{f(x+h)-f(x)}{h}$ where $h$ is always such that $x, x+h$ are both in the interval $[a, b]$ so we include derivatives at the right and left end points in this definition.

The most important theorem about derivatives of functions of one variable is the mean value theorem.

Theorem 16.4.2 Let $f:[a, b] \rightarrow \mathbb{R}$ be continuous. Then if the maximum value of $f$ occurs at a point $x \in(a, b)$, it follows that if $f^{\prime}(x)=0$. If $f$ achieves a minimum at $x \in(a, b)$ where $f^{\prime}(x)$ exists, it also follows that $f^{\prime}(x)=0$.

Proof: By Theorem 10.1.39, $f$ achieves a maximum at some point $x$. If $f^{\prime}(x)$ exists, then

$$
f^{\prime}(x)=\lim _{h \rightarrow 0+} \frac{f(x+h)-f(x)}{h}=\lim _{h \rightarrow 0-} \frac{f(x+h)-f(x)}{h}
$$

However, the first limit is non-positive while the second is non-negative and so $f^{\prime}(x)=0$. The situation is similar if the minimum occurs at $x \in(a, b)$.

The Cauchy mean value theorem follows. The usual one is obtained by letting $g(x)=x$.
Theorem 16.4.3 Let $f, g$ be continuous on $[a, b]$ and differentiable on $(a, b)$. Then there exists $x \in(a, b)$ such that $f^{\prime}(x)(g(b)-g(a))=g^{\prime}(x)(f(b)-f(a))$. If $g(x)=x$, this yields $f(b)-f(a)=f^{\prime}(x)(b-a)$, also $f(a)-f(b)=f^{\prime}(x)(a-b)$.

Proof: Let $h(x) \equiv f(x)(g(b)-g(a))-g(x)(f(b)-f(a))$. Then

$$
h(a)=h(b)=f(a) g(b)-g(a) f(b) .
$$

If $h$ is constant, then pick any $x \in(a, b)$ and $h^{\prime}(x)=0$. If $h$ is not constant, then it has either a maximum or a minimum on $(a, b)$ and so if $x$ is the point where either occurs, then $h^{\prime}(x)=0$ which proves the theorem.

Recall that an antiderivative of a function $f$ is just a function $F$ such that $F^{\prime}=f$. You know how to find an antiderivative for a polynomial. $\left(\frac{x^{n+1}}{n+1}\right)^{\prime}=x^{n}$ so $\int \sum_{k=1}^{n} a_{k} x^{k}=$ $\sum_{k=1}^{n} a_{k} \frac{x^{k+1}}{k+1}+C$. With this information and the Weierstrass theorem, it is easy to define integrals of continuous functions with all the properties presented in elementary calculus courses. It is an approach which does not depend on Riemann sums yet still gives the fundamental theorem of calculus. Note that if $F^{\prime}(x)=0$ for $x$ in an interval, then for $x, y$ in that interval, $F(y)-F(x)=0(y-x)$ so $F$ is a constant. Thus, if $F^{\prime}=G^{\prime}$ on an open interval, $F, G$ continuous on the closed interval, it follows that $F-G$ is a constant and so $F(b)-F(a)=G(b)-G(a)$.

Definition 16.4.4 For $p(x)$ a polynomial on $[a, b]$, let $P^{\prime}(x)=p(x)$. Thus, by the mean value theorem if $P^{\prime}, \hat{P}^{\prime}$ both equal $p$, it follows that $P(b)-P(a)=\hat{P}(b)-\hat{P}(a)$. Then define $\int_{a}^{b} p(x) d x \equiv P(b)-P(a)$. If $f \in C([a, b])$, define $\int_{a}^{b} f(x) d x \equiv \lim _{n \rightarrow \infty} \int_{a}^{b} p_{n}(x) d x$ where $\lim _{n \rightarrow \infty}\left\|p_{n}-f\right\| \equiv \lim _{n \rightarrow \infty} \max _{x \in[a, b]}\left|f(x)-p_{n}(x)\right|=0$.

Proposition 16.4.5 The above integral is well defined and satisfies the following properties.

1. $\int_{a}^{b} f d x=f(\hat{x})(b-a)$ for some $\hat{x}$ between $a$ and $b$. Thus $\left|\int_{a}^{b} f d x\right| \leq\|f\||b-a|$.
2. If $f$ is continuous on an interval which contains all necessary intervals,

$$
\int_{a}^{c} f d x+\int_{c}^{b} f d x=\int_{a}^{b} f d x, \text { so } \int_{a}^{b} f d x+\int_{b}^{a} f d x=\int_{b}^{b} f d x=0
$$

3. If $F(x) \equiv \int_{a}^{x} f d t$, Then $F^{\prime}(x)=f(x)$ so any continuous function has an antiderivative, and for any $a \neq b, \int_{a}^{b} f d x=G(b)-G(a)$ whenever $G^{\prime}=f$ on the open interval determined by $a, b$ and $G$ continuous on the closed interval determined by $a, b$. Also,

$$
\int_{a}^{b}(\alpha f(x)+\beta g(x)) d x=\alpha \int_{a}^{b} f(x) d x+\beta \int_{a} \beta g(x) d x
$$

If $a<b$, and $f(x) \geq 0$, then $\int_{a}^{b} f d x \geq 0$. Also $\left|\int_{a}^{b} f d x\right| \leq\left|\int_{a}^{b}\right| f|d x|$.
4. $\int_{a}^{b} 1 d x=b-a$.

Proof: First, why is the integral well defined? With notation as in the above definition, the mean value theorem implies

$$
\begin{equation*}
\int_{a}^{b} p(x) d x \equiv P(b)-P(a)=p(\hat{x})(b-a) \tag{16.2}
\end{equation*}
$$

where $\hat{x}$ is between $a$ and $b$ and so $\left|\int_{a}^{b} p(x) d x\right| \leq\|p\||b-a|$. If $\left\|p_{n}-f\right\| \rightarrow 0$, then

$$
\lim _{m, n \rightarrow \infty}\left\|p_{n}-p_{m}\right\|=0
$$

and so

$$
\begin{aligned}
\left|\int_{a}^{b} p_{n}(x) d x-\int_{a}^{b} p_{m}(x) d x\right| & =\left|\left(P_{n}(b)-P_{n}(a)\right)-\left(P_{m}(b)-P_{m}(a)\right)\right| \\
& =\left|\left(P_{n}(b)-P_{m}(b)\right)-\left(P_{n}(a)-P_{m}(a)\right)\right| \\
& =\left|\int_{a}^{b}\left(p_{n}-p_{m}\right) d x\right| \leq\left\|p_{n}-p_{m}\right\||b-a|
\end{aligned}
$$

Thus the limit exists because $\left\{\int_{a}^{b} p_{n} d x\right\}_{n}$ is a Cauchy sequence and $\mathbb{R}$ is complete.
From $16.2,1$. holds for a polynomial $p(x)$. Let $\left\|p_{n}-f\right\| \rightarrow 0$. Then by definition,

$$
\begin{equation*}
\int_{a}^{b} f d x \equiv \lim _{n \rightarrow \infty} \int_{a}^{b} p_{n} d x=p_{n}\left(x_{n}\right)(b-a) \tag{16.3}
\end{equation*}
$$

for some $x_{n}$ in the open interval determined by $(a, b)$. By compactness, there is a further subsequence, still denoted with $n$ such that $x_{n} \rightarrow x \in[a, b]$. Then fixing $m$ such that $\left\|f-p_{n}\right\|<\varepsilon$ whenever $n \geq m$, assume $n>m$. Then $\left\|p_{m}-p_{n}\right\| \leq\left\|p_{m}-f\right\|+\left\|f-p_{n}\right\|<$ $2 \varepsilon$ and so

$$
\begin{gathered}
\left|f(x)-p_{n}\left(x_{n}\right)\right| \leq\left|f(x)-f\left(x_{n}\right)\right|+\left|f\left(x_{n}\right)-p_{m}\left(x_{n}\right)\right|+\left|p_{m}\left(x_{n}\right)-p_{n}\left(x_{n}\right)\right| \\
\leq\left|f(x)-f\left(x_{n}\right)\right|+\left\|f-p_{m}\right\|+\left\|p_{m}-p_{n}\right\|<\left|f(x)-f\left(x_{n}\right)\right|+3 \varepsilon
\end{gathered}
$$

Now if $n$ is still larger, continuity of $f$ shows that $\left|f(x)-p_{n}\left(x_{n}\right)\right|<4 \varepsilon$. Since $\varepsilon$ is arbitrary, $p_{n}\left(x_{n}\right) \rightarrow f(x)$ and so, passing to the limit with this subsequence in 16.3 yields 1 .

Now consider 2. It holds for polynomials $p(x)$ obviously. So let $\left\|p_{n}-f\right\| \rightarrow 0$. Then

$$
\int_{a}^{c} p_{n} d x+\int_{c}^{b} p_{n} d x=\int_{a}^{b} p_{n} d x
$$

Pass to a limit as $n \rightarrow \infty$ and use the definition to get 2 . Also note that $\int_{b}^{b} f(x) d x=0$ follows from the definition.

Next consider 3. Let $h \neq 0$ and let $x$ be in the open interval determined by $a$ and $b$. Then for small $h$,

$$
\frac{F(x+h)-F(x)}{h}=\frac{1}{h} \int_{x}^{x+h} f(t) d t=f\left(x_{h}\right)
$$

where $x_{h}$ is between $x$ and $x+h$. Let $h \rightarrow 0$. By continuity of $f$, it follows that the limit of the right side exists and so

$$
\lim _{h \rightarrow 0} \frac{F(x+h)-F(x)}{h}=\lim _{h \rightarrow 0} f\left(x_{h}\right)=f(x)
$$

If $x$ is either end point, the argument is the same except you have to pay attention to the sign of $h$ so that both $x$ and $x+h$ are in $[a, b]$. Thus $F$ is continuous on $[a, b]$ and $F^{\prime}$ exists on $(a, b)$ so if $G$ is an antiderivative,

$$
\int_{a}^{b} f(t) d t \equiv F(b)=F(b)-F(a)=G(b)-G(a)
$$

The claim that the integral is linear is obvious from this. Indeed, if $F^{\prime}=f, G^{\prime}=g$,

$$
\begin{aligned}
\int_{a}^{b}(\alpha f(t)+\beta g(t)) d t & =\alpha F(b)+\beta G(b)-(\alpha F(a)+\beta G(a)) \\
& =\alpha(F(b)-F(a))+\beta(G(b)-G(a)) \\
& =\alpha \int_{a}^{b} f(t) d t+\beta \int_{a}^{b} g(t) d t
\end{aligned}
$$

If $f \geq 0$, then the mean value theorem implies that for some

$$
t \in(a, b), F(b)-F(a)=\int_{a}^{b} f d x=f(t)(b-a) \geq 0
$$

Thus $\int_{a}^{b}(|f|-f) d x \geq 0, \int_{a}^{b}(|f|+f) d x \geq 0$ and so

$$
\int_{a}^{b}|f| d x \geq \int_{a}^{b} f d x, \int_{a}^{b}|f| d x \geq-\int_{a}^{b} f d x
$$

so this proves $\left|\int_{a}^{b} f d x\right| \leq \int_{a}^{b}|f| d x$. This, along with part 2 implies the other claim that $\left|\int_{a}^{b} f d x\right| \leq\left|\int_{a}^{b}\right| f|d x|$.

The last claim is obvious because an antiderivative of 1 is $F(x)=x$.
Note also that the usual change of variables theorem is available because if $F^{\prime}=f$, then $f(g(x)) g^{\prime}(x)=\frac{d}{d x} F(g(x))$ so that, from the above proposition,

$$
F(g(b))-F(g(a))=\int_{g(a)}^{g(b)} f(y) d y=\int_{a}^{b} f(g(x)) g^{\prime}(x) d x
$$

We usually let $y=g(x)$ and $d y=g^{\prime}(x) d x$ and then change the limits as indicated above, equivalently we massage the expression to look like the above. Integration by parts also follows from differentiation rules.

Consider the iterated integral $\int_{a_{1}}^{b_{1}} \cdots \int_{a_{p}}^{b_{p}} \alpha x_{1}^{\alpha_{1}} \cdots x_{p}^{\alpha_{p}} d x_{p} \cdots d x_{1}$. It means just what it meant in calculus. You do the integral with respect to $x_{p}$ first, keeping the other variables constant, obtaining a polynomial function of the other variables. Then you do this one with respect to $x_{p-1}$ and so forth. Thus, doing the computation, it reduces to

$$
\alpha \prod_{k=1}^{p}\left(\int_{a_{k}}^{b_{k}} x_{k}^{\alpha_{k}} d x_{k}\right)=\alpha \prod_{k=1}^{p}\left(\frac{b^{\alpha_{k}+1}}{\alpha_{k}+1}-\frac{a^{\alpha_{k}+1}}{\alpha_{k}+1}\right)
$$

and the same thing would be obtained for any other order of the iterated integrals. Since each of these integrals is linear, it follows that if $\left(i_{1}, \cdots, i_{p}\right)$ is any permutation of $(1, \cdots, p)$, then for any polynomial $q$,

$$
\int_{a_{1}}^{b_{1}} \cdots \int_{a_{p}}^{b_{p}} q\left(x_{1}, \ldots, x_{p}\right) d x_{p} \cdots d x_{1}=\int_{a_{i_{p}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} q\left(x_{1}, \ldots, x_{p}\right) d x_{i_{p}} \cdots d x_{i_{1}}
$$

Now let $f: \prod_{k=1}^{p}\left[a_{k}, b_{k}\right] \rightarrow \mathbb{R}$ be continuous. Then each iterated integral results in a continuous function of the remaining variables and so the iterated integral makes sense. For example, by Proposition 16.4.5, $\left|\int_{c}^{d} f(x, y) d y-\int_{c}^{d} f(\hat{x}, y) d y\right|=$

$$
\left|\int_{c}^{d}(f(x, y)-f(\hat{x}, y)) d y\right| \leq \max _{y \in[c, d]}|f(x, y)-f(\hat{x}, y)|<\varepsilon
$$

if $|x-\hat{x}|$ is sufficiently small, thanks to uniform continuity of $f$ on the compact set $[a, b] \times$ $[c, d]$. Thus it makes perfect sense to consider the iterated integral $\int_{a}^{b} \int_{c}^{d} f(x, y) d y d x$. Then using Proposition 16.4.5 on the iterated integrals along with Theorem 16.2.1, there exists a sequence of polynomials which converges to $f$ uniformly $\left\{p_{n}\right\}$. Then applying Proposition 16.4.5 repeatedly,

$$
\begin{gather*}
\left|\int_{a_{i_{p}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} f(\boldsymbol{x}) d x_{p} \cdots d x_{1}-\int_{a_{i_{p}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} p_{n}(\boldsymbol{x}) d x_{p} \cdots d x_{1}\right| \\
\leq\left\|f-p_{n}\right\| \prod_{k=1}^{p}\left|b_{k}-a_{k}\right| \tag{16.4}
\end{gather*}
$$

With this, it is easy to prove a rudimentary Fubini theorem valid for continuous functions.
Theorem 16.4.6 $f: \prod_{k=1}^{p}\left[a_{k}, b_{k}\right] \rightarrow \mathbb{R}$ be continuous. Then for $\left(i_{1}, \cdots, i_{p}\right)$ any permutation of $(1, \cdots, p)$,

$$
\int_{a_{i_{p}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} f(\boldsymbol{x}) d x_{i_{p}} \cdots d x_{i_{1}}=\int_{a_{1}}^{b_{1}} \cdots \int_{a_{p}}^{b_{p}} f(\boldsymbol{x}) d x_{p} \cdots d x_{1}
$$

If $f \geq 0$, then the iterated integrals are nonnegative if each $a_{k} \leq b_{k}$.
Proof: Let $\left\|p_{n}-f\right\| \rightarrow 0$ where $p_{n}$ is a polynomial. Then from 16.4,

$$
\int_{a_{i_{1}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} f(\boldsymbol{x}) d x_{i_{p}} \cdots d x_{i_{1}}=\lim _{n \rightarrow \infty} \int_{a_{i_{p}}}^{b_{i_{1}}} \cdots \int_{a_{i_{p}}}^{b_{i_{p}}} p_{n}(\boldsymbol{x}) d x_{i_{p}} \cdots d x_{i_{1}}
$$

$$
=\lim _{n \rightarrow \infty} \int_{a_{1}}^{b_{1}} \cdots \int_{a_{p}}^{b_{p}} p_{n}(\boldsymbol{x}) d x_{p} \cdots d x_{1}=\int_{a_{1}}^{b_{1}} \cdots \int_{a_{p}}^{b_{p}} f(\boldsymbol{x}) d x_{p} \cdots d x_{1}
$$

You could replace $f$ with $f \mathscr{X}_{G}$ where $\mathscr{X}_{G}(\boldsymbol{x})=1$ if $\boldsymbol{x} \in G$ and 0 otherwise provided each section of $G$ consisting of holding all variables constant but 1 , consists of finitely many intervals. Thus you can integrate over all the usual sets encountered in beginning calculus.

Definition 16.4.7 A function $f:[a, b] \rightarrow \mathbb{R}$ is piecewise continuous if there are $z_{i}$ with $a=z_{0}<z_{1}<\cdots<z_{n}=b$, called a partition of $[a, b]$, and functions $f_{i}$ continuous on $\left[z_{i-1}, z_{i}\right]$ such that $f=f_{i}$ on $\left(z_{i-1}, z_{i}\right)$. For $f$ piecewise continuous, define

$$
\int_{a}^{b} f(t) d t \equiv \sum_{i=1}^{n} \int_{z_{i-1}}^{z_{i}} f_{i}(s) d s
$$

Of course this gives what appears to be a new definition because if $f$ is continuous on $[a, b]$, then it is piecewise continuous for any such partition. However, it gives the same answer because, from this new definition,

$$
\int_{a}^{b} f(t) d t=\sum_{i=1}^{n}\left(F\left(z_{i}\right)-F\left(z_{i-1}\right)\right)=F(b)-F(a)
$$

Does this give the main properties of the integral? In particular, is the integral still linear? Suppose $f, g$ are piecewise continuous. Then let $\left\{z_{i}\right\}_{i=1}^{n}$ include all the partition points of both of these functions. Then, since it was just shown that no harm is done by including more partition points,

$$
\begin{gathered}
\int_{a}^{b} \alpha f(t)+\beta g(t) d t \equiv \sum_{i=1}^{n} \int_{z_{i-1}}^{z_{i}}\left(\alpha f_{i}(s)+\beta g_{i}(s)\right) d s \\
=\sum_{i=1}^{n} \alpha \int_{z_{i-1}}^{z_{i}} f_{i}(s) d s+\sum_{i=1}^{n} \beta \int_{z_{i-1}}^{z_{i}} g_{i}(s) d s \\
=\alpha \sum_{i=1}^{n} \int_{z_{i-1}}^{z_{i}} f_{i}(s) d s+\beta \sum_{i=1}^{n} \int_{z_{i-1}}^{z_{i}} g_{i}(s) d s \\
=\alpha \int_{a}^{b} f(t) d t+\beta \int_{a}^{b} g(t) d t
\end{gathered}
$$

Also, the claim that $\int_{a}^{b} f d t=\int_{a}^{c} f d t+\int_{c}^{b} f d t$ is obtained exactly as before by considering all partition points on each integral preserving the order of the limits in the small intervals determined by the partition points. That is, if $a>c$, you would have $z_{i-1}>z_{i}$. Notice how this automatically takes care of orientation.

Is this as general as a complete treatment of Riemann integration? No it is not. In particular, it does not include the well known example where $f(x)=\sin \left(\frac{1}{x}\right)$ for $x \in(0,1]$ and $f(0) \equiv 0$. However, it is sufficiently general to include all cases which are typically of interest. It would be enough to build a theory of ordinary differential equations. It would also be enough to provide the theory of convergence of Fourier series to the midpoint of the jump and so forth. Also, the Riemann integral is woefully inadequate when it comes to a need to handle limits. You need the Lebesgue integral and to obtain this, it is enough to consider knowledge of integrals of continuous functions.

### 16.5 The Müntz Theorems

All this about to be presented would work on any interval, but it would involve fussy considerations involved with extra constants. Therefore, I will only present what happens on $[0,1]$. These theorems have to do with considering linear combinations of the functions $f_{p}(x) \equiv x^{p}$ for $p=p_{1}, p_{2}, \ldots$ and whether one can approximate an arbitrary continuous function with such a linear combination. Linear algebra techniques are what make this possible, at least in this book. I am following Cheney [13]. In what follows $m$ will be a nonnegative integer. I will consider the real inner product space $X$ consisting of functions in $C([0,1])$ with the inner product $\int_{0}^{1} f g d x=(f, g)$. Thus, as shown earlier, the Cauchy Schwarz inequality holds

$$
\int_{0}^{1}|f||g| d x \leq\left(\int_{0}^{1}|f|^{2} d x\right)^{1 / 2}\left(\int_{0}^{1}|g|^{2} d x\right)^{1 / 2}
$$

I will write $|f| \equiv\left(\int_{0}^{1}|f|^{2} d x\right)^{1 / 2}$. The above treatment of the integral of continuous functions is sufficient for the needs here. Also let $V_{n} \equiv \operatorname{span}\left(f_{p_{1}}, \ldots, f_{p_{n}}\right)$.

The main idea is to estimate the distance between $f_{m}$ and $V_{m}$ in $X$. The Grammian matrix of $\left\{f_{p_{1}}, \ldots, f_{p_{n}}\right\}$ is easily seen to be

$$
G\left(f_{p_{1}}, \ldots, f_{p_{n}}\right)=\left(\begin{array}{ccc}
\frac{1}{p_{1}+p_{1}+1} & \cdots & \frac{1}{p_{1}+p_{n}+1} \\
\vdots & & \vdots \\
\frac{1}{p_{1}+p_{n}+1} & \cdots & \frac{1}{p_{n}+p_{n}+1}
\end{array}\right)
$$

I will assume $p_{j}>-\frac{1}{2}$ to avoid any possibility of terms which make no sense in the Grammian matrix given above. I will also assume none of these $p_{j}$ are integers so that $V_{n}$ never contains $f_{m}, f_{m}(x)=x^{m}, m$ a positive integer. If such is in your list, it simply makes the approximation easier to obtain. By Theorem 8.6.5, the Cauchy identity for determinants,

$$
\operatorname{det} G\left(f_{p_{1}}, \ldots, f_{p_{n}}\right)=\frac{\prod_{j<i \leq n}\left(p_{i}-p_{j}\right)\left(p_{i}-p_{j}\right)}{\prod_{i, j \leq n}\left(p_{i}+p_{j}+1\right)}
$$

You let $a_{i}=p_{i}, b_{i}=p_{i}+1$. Thus from Proposition 12.2.2 $\left\{f_{p_{1}}, \ldots, f_{p_{n}}\right\}$ is linearly independent if and only if the exponents $p_{j}$ are distinct. Assume this happens. Then from Theorem 12.2.3 about the distance to a subspace, if $d_{n}$ is this distance between $f_{m}$ and $V_{n}$,

$$
d_{n}^{2}=\frac{\operatorname{det}\left(G\left(f_{p_{1}}, \ldots, f_{p_{n}}, f_{m}\right)\right)}{\operatorname{det}\left(G\left(f_{p_{1}}, \ldots, f_{p_{n}}\right)\right)}
$$

By the Cauchy identity Theorem 8.6 .5 again, letting $p_{n+1} \equiv m$,

$$
\begin{aligned}
d_{n}^{2} & =\frac{\left(\frac{\Pi_{j<i \leq n+1}\left(p_{i}-p_{j}\right)\left(p_{i}-p_{j}\right)}{\Pi_{i, j \leq n+1}\left(p_{i}+p_{j}+1\right)}\right)}{\left(\frac{\Pi_{j<i \leq n}\left(p_{i}-p_{j}\right)\left(p_{i}-p_{j}\right)}{\Pi_{i, j \leq n}\left(p_{i}+p_{j}+1\right)}\right)} \\
& =\left(\frac{\prod_{j<n+1}\left(p_{n+1}-p_{j}\right)\left(p_{n+1}-p_{j}\right)}{\prod_{i<n+1}\left(p_{i}+p_{n+1}+1\right) \Pi_{j<n+1}\left(p_{n+1}+p_{j}+1\right)\left(2 p_{n+1}+1\right)}\right)
\end{aligned}
$$

It follows that, since $p_{n+1}=m$

$$
d_{n}=\frac{\prod_{j<n+1}\left|m-p_{j}\right|}{\prod_{j<n+1}\left(m+p_{j}+1\right) \sqrt{(2 m+1)}}=\frac{1}{\sqrt{(2 m+1)}} \prod_{j<n+1} \frac{\left|m-p_{j}\right|}{\left(m+p_{j}+1\right)}
$$

The idea is to let $n \rightarrow \infty$ and see whether the distance between the best approximation and $x \rightarrow x^{m}$ converges to 0 . That is, to determine whether $d_{n} \rightarrow 0$. I want $m$ an integer to be arbitrary and for the sake of convenience, I want $\left|m-p_{j}\right|=p_{j}-m$ for all $j$ large enough. Therefore, one needs to have $\lim _{k \rightarrow \infty} p_{k}=\infty$. From now on, this is assumed. It is desired to have

$$
-\infty=\lim _{n \rightarrow \infty} \ln \left(d_{n} \sqrt{2 m+1}\right)=\lim _{n \rightarrow \infty} \sum_{j<n+1} \ln \left(1-\left(1-\frac{\left|m-p_{j}\right|}{\left(m+p_{j}+1\right)}\right)\right)
$$

Note that $0<\frac{\left|m-p_{j}\right|}{\left(m+p_{j}+1\right)}=r_{j}<1$ and so it is easily shown that

$$
\ln \left(1-\left(1-r_{j}\right)\right) \in\left(-2\left(1-r_{j}\right),-\left(1-r_{j}\right)\right)
$$

It follows that $d_{n} \rightarrow 0$ if and only if

$$
\sum_{j} 1-\frac{p_{j}-m}{m+p_{j}+1}=\sum_{j} \frac{2 m+1}{m+p_{j}+1}
$$

diverges. But by the limit comparison test from calculus, this happens if and only if $\sum_{j} \frac{1}{p_{j}}=$ $\infty$. This proves most of the following theorem.

Theorem 16.5.1 Let $g \in C([0,1])$. Then there exists a sequence $h_{k}$ consisting of a linear combination of functions $f_{p_{j}}$ such that $\left|g-h_{n}\right| \rightarrow 0$. Here we define $V_{n} \equiv \operatorname{span}\left(f_{p_{1}}, \ldots, f_{p_{n}}\right)$ where no $p_{k}$ is an integer, all are larger than $-1 / 2$, and $\lim _{n \rightarrow \infty} p_{n}=\infty$ and $\sum_{k} \frac{1}{p_{k}}=\infty$.

Proof: Let $\varepsilon>0$ be given. By the Weierstrass approximation theorem, there is a polynomial $p(x)$ such that $|g-p| \leq\left\|g-p_{n}\right\|<\varepsilon$ where the second norm is the supremum norm of the Weierstrass theorem. Thus $p$ is a linear combination of functions $f_{m}$ for $m$ an integer. $p=\sum_{k=1}^{L} c_{k} f_{k}$. Let $h_{n_{k}} \in V_{n_{k}}$ such that $\left|h_{n_{k}}-f_{k}\right|\left|c_{k}\right|<\frac{\varepsilon}{L}$. Let $n>$ $\max \left\{n_{k}: k \leq L\right\}$. Consider $h \in V_{n}$ defined by $h \equiv \sum_{k=1}^{L} c_{k} h_{n_{k}}$. Then $|h-g| \leq|h-p|+$ $|p-g| \leq \sum_{k=1}^{\bar{L}}\left|c_{k} h_{n_{k}}-c_{k} f_{k}\right|+\varepsilon<L_{\bar{L}}^{\varepsilon}+\varepsilon=2 \varepsilon$. It follows that letting $\varepsilon_{k} \rightarrow 0$, there exists $h_{k}$ consisting of a finite linear combination of functions of the form $f_{p_{j}}$ such that $\left|h_{k}-g\right| \rightarrow 0$.

This is the first Müntz theorem. The second one involves approximation in the usual norm for continuous function $\|f\| \equiv \max \{|f(x)|: x \in[0,1]\}$. It also depends on linear algebra techniques.

Theorem 16.5.2 Let $\frac{1}{2}<p_{k}$, none of the $p_{k}$ are zero, and $\lim _{k \rightarrow \infty} p_{k}=\infty$ and $\sum_{k} \frac{1}{p_{k}}=\infty$. Let $V_{n} \equiv \operatorname{span}\left(1, f_{p_{1}}, \ldots, f_{p_{n}}\right)$. Then if $g \in C([0,1])$ and $\varepsilon>0$, there exists $h \in V_{n}$ for some $n$ such that $\|g-h\| \leq \varepsilon$.

Proof: This follows in the same way as above if I can show that for all $m$ a nonnegative integer, there is a function $h \in V_{n}$ with $\left\|h-f_{m}\right\|<\varepsilon$. Since 1 is included, there is nothing to
show if $m=0$. Thus, assume $m>0$. From the above theorem, consider $n$ large enough that $\left|m f_{m-1}-h\right|<\varepsilon$ for some $h \in \operatorname{span}\left(f_{p_{1}-1}, \ldots, f_{p_{n}-1}\right), h(x)=\sum_{k=1}^{n} c_{k} f_{p_{k}-1}(x)$. Then note that $x^{m}=\int_{0}^{x} m f_{m-1}(t) d t$ and also $H(x) \equiv \int_{0}^{x} h(t) d t \in V_{n}$. Therefore, from the Cauchy Schwarz inequality,

$$
\begin{aligned}
\left|x^{m}-H(x)\right| & =\left|\int_{0}^{x}\left(m f_{m-1}(t)-h(t)\right) d t\right| \leq \int_{0}^{1} 1\left|m f_{m-1}(t)-h(t)\right| d t \\
& \leq 1\left(\int_{0}^{1}\left|m f_{m-1}(t)-h(t)\right|^{2} d t\right)^{1 / 2}=\left|m f_{m-1}-h\right|<\varepsilon
\end{aligned}
$$

Since this is true for each $x$, it follows that $\left\|f_{m}-H\right\| \leq \varepsilon$. Then the same argument used above, depending on the triangle inequality proves the theorem.

Note that, as before, this shows that if $g$ is continuous, there is a sequence of $h_{n}$ consisting of linear combinations of the $f_{p_{k}}$ which converges uniformly to $g$.

Example 16.5.3 Let $p_{k} \equiv \ln (1+k)$. Then if $g$ is continuous on $[0,1]$, there is a function of the form $c_{0}+\sum_{k=1}^{L} c_{k} x^{\ln (1+k)} \equiv h(x)$ such that $\|h-g\|<\varepsilon$. You could replace $\ln (1+k)$ with $k \ln (1+k)$ or $5 k$ and draw the same conclusion.

### 16.6 Exercises

1. Show the above argument used to obtain Theorems 16.5.1, 16.5.2 works to give both of these theorems if your interest is in $[-1,1]$ rather than $[0,1]$ provided you replace $x$ with $|x|$ whenever $x^{p}$ occurs for $p$ not an integer. You must do something like this because if $x<0$, maybe $x^{p}$ is not well defined in the context of real analysis.
2. Generalize Theorems $16.5 .1,16.5 .2$ to any interval $[a, b]$.
3. Show that the same arguments will work for proving Theorems 16.5.1, 16.5.2 in approximating functions in $C([0,1] ; X)$ where $X$ is an inner product space if you define a new inner product $(f, g) \equiv \int_{0}^{1}(f, g) d x$. Show how to use this to generalize Theorems 16.5.1, 16.5.2 to the case of many variables as was done for the Weierstrass theorem. Then, using the ideas of the above problem, show how to consider approximation of functions $C(R ; X)$ where $R$ is some $n$ dimensional box of the form $\prod_{i=1}^{n}\left[a_{i}, b_{i}\right]$. How would you generalize to the case of $C(R, \mathbb{R})$ where $R$ is just some closed and bounded set?

## Appendix A

## Homological Methods*

There is a lot more on homology in [40], [39]. That worthwhile topological theorems can be obtained through using free Abelian groups and abstract algebra arguments is very counter-intuitive to me, but this is in fact the case, enough that I concluded to give an introduction. This is certainly only an introduction to this subject which amounts to my attempts to understand it myself. These methods make possible an approach to the hard topology theorems like the Brouwer fixed point theorem or Jordan curve theorem. They are algebraic in nature and so I think this fits in with this book. All that I present will be basic homology and the homology of spheres with applications. The approach is to reduce questions about homeomorphisms to algebraic questions about homology groups which are a type of quotient group. To be specific, let the topological spaces be metric spaces. I have not discussed general topological spaces in this book but many of the theorems apply to this more general situation. A homeomorphism is a one to one and onto mapping $f$ between two metric spaces such that $f$ and $f^{-1}$ are both continuous. When two metric spaces are homeomorphic, it means all considerations relative to open sets, such as convergence of sequences will coincide. That is, a sequence converges in one space if and only if its image in the other space also converges. More generally, in the case of topological space, the functions $f, f^{-1}$ map the topology of one space to the topology of another.

To begin with here is a definition of a free Abelian group.
Definition A.0.1 Let $G$ be a nonempty set. A free Abelian group featuring $G$ consists of all formal sums of this form $\sum_{\phi} n_{\phi} \phi$ where the $n_{\phi}$ are integers and only finitely many are non-zero and the $\phi$ are elements of $G$. Then we add two of these in the obvious way

$$
\sum_{\phi} n_{\phi} \phi+\sum_{\phi} m_{\phi} \phi=\sum_{\phi}\left(n_{\phi}+m_{\phi}\right) \phi
$$

The inverse of $\sum_{\phi} n_{\phi} \phi$ is $\sum_{\phi}\left(-n_{\phi}\right) \phi$ and the zero is the one where all the $n_{\phi}$ equal 0 . Each $\sum_{\phi} n_{\phi} \phi$ is determined uniquely by the integers $n_{\phi}$. If we know $f(\phi)$ for each $\phi \in G$, we can define a homomorphism on the whole Abelian group as follows.

$$
f_{\#}\left(\sum_{\phi} n_{\phi} \phi\right) \equiv \sum_{\phi} n_{\phi} f_{\#}(\phi)
$$

If the entire group is obtained as the form kg for $k \in \mathbb{Z}$ and $g$ is an element of the group, then $g$ is said to generate the group.

Lemma A.0.2 Let $A, B$ be two Abelian groups generated by $a, b$ respectively. Also let $\gamma: A \rightarrow B$ be an isomorphism meaning that $\gamma$ is a homomorphism, $\gamma(x+y)=\gamma(x)+\gamma(y)$, and $\gamma$ is one to one and onto. Then $\gamma(a)= \pm b$.

Proof: It follows easily that $\gamma^{-1}$ is also a homomorphism. $b=\gamma(x)$ for a unique $x \in A$. Then there is $l \in \mathbb{Z}$ such that $x=l a$ and so $b=\gamma(l a)$ and so $\gamma^{-1}(b)=l a$. Similarly $\gamma(a)=k b$. Therefore, $a=k \gamma^{-1}(b)=k l a$ showing that $k l=1$. Therefore, both $k, l=1$ or both are -1 showing what was claimed.

In particular, if $a$ is a generator of one of these Abelian free groups and so is $b$, then, letting $\gamma=\mathrm{id}, a= \pm b$.

This free Abelian group is a kind of module

## A. 1 Singular Simplices and Boundaries

Definition A.1.1 Let $X$ be a topological space. Whenever convenient let it be a metric space. Let

$$
\sigma_{p} \equiv\left\{t \in \mathbb{R}^{p+1}: \sum_{i=0}^{p+1} t_{i}=1, t_{i} \geq 0 \text { for each } i\right\}
$$

Then $S_{p}(X)$ will denote expressions of the form $\sum_{\phi} n_{\phi} \phi$ where $\phi: \sigma_{p} \rightarrow X$ is a continuous mapping, called a singular simplex, and only finitely many of the integers $n_{\phi}$ are nonzero. It is the free Abelian group featuring the singular simplices with values in $X$ which have domain $\sigma_{p}$ just described. It follows that if you have a function defined on each $\phi$, a continuous mapping from $\sigma_{p}$ to $X$, then this extends uniquely to a homomorphism on $S_{p}(X)$. These distribute across + signs. The first of these to consider is the boundary operator. Define a homomorphism $\partial_{i}: S_{p}(X) \rightarrow S_{p-1}(X)$ as

$$
\partial_{i} \phi\left(t_{0}, \ldots, t_{p-1}\right) \equiv \phi\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{p-1}\right)
$$

where 0 replaces $t_{i}$ and then the $t_{i}$ is in the following slot followed by $t_{j}$ in order for $j \leq p-$ 1. Note that the subscripts end with $p-1$ rather than $p$. Then $\partial$ will be a homomorphism given by the following on a singular simplex $\phi$

$$
\partial \phi\left(t_{0}, \ldots, t_{p-1}\right) \equiv \sum_{i=0}^{p}(-1)^{i} \partial_{i} \phi\left(t_{0}, \ldots, t_{p-1}\right)
$$

In case $p=1$, you would only have the end points left and so for $\phi \in S_{1}(X)$,

$$
\partial \phi\left(t_{0}\right)=\phi(1,0)-\phi(0,1)
$$

As to $S_{0}(X)$, these are of the form $\phi(1)$ and we define $\partial \phi \equiv 0$ in $S_{0}(X)$. This means the zero homomorphism. In this case $S_{0}(X)=\left\{\sum_{\phi} n_{\phi} \phi\right\}$. We make $\partial$ a homomorphism on $S_{p}(X)$ by defining $\partial\left(\sum_{\phi} n_{\phi} \phi\right) \equiv \sum_{\phi} n_{\phi} \partial \phi$. I will sometimes refer to something in $S_{p}(X)$ as a chain.

Note how this means that the 0 simplices are essentially points of $X$. The order of the $t_{i}$ is very important so that orientation will be preserved.

In all of this $\sigma_{p}$ will be the simplex having vertices

$$
(1,0, \ldots),(0,1, \ldots), \ldots(0,0, \ldots, 1,0, \ldots)
$$

the 1 in the $p^{\text {th }}$ position. I may write these points as $(1,0, \ldots 0),(0,1, \ldots 0), \ldots(0,0, \ldots, 1)$.
Can we tie these formal sums to something more familiar? They are functions defined on a single simplex $\sigma_{p}$, but I don't know that one can even do addition in $X$ and so this free group is pretty formal at this point. As to $-\phi$ for $\phi$ a simplex, it doesn't have any intrinsic meaning either unless maybe $\phi$ has values in a vector space. Thus, we understand this to mean $(-1) \phi$ which will amount to the same thing in the vector space setting. What is the 0 in this Abelian group? It would be $0 \phi$ or there is simply nothing there to add to another element of the Abelian group.

Here is a picture of $\sigma_{2}$ in $\mathbb{R}^{3}$.


The big news about $\partial$ is in the following lemma.
Lemma A.1.2 For $p \geq 2, \partial^{2}: S_{p}(X) \rightarrow S_{p-2}(X)$ is 0 .
Proof: Consider some cases. Now consider $p=2$ so $\phi\left(t_{0}, t_{1}, t_{2}\right)$

$$
\partial \phi\left(t_{0}, t_{1}\right) \equiv \phi\left(0, t_{0}, t_{1}\right)-\phi\left(t_{0}, 0, t_{1}\right)+\phi\left(t_{0}, t_{1}, 0\right)
$$

and

$$
\begin{gathered}
\partial^{2} \phi\left(t_{0}\right) \equiv\left(\phi\left(0,0, t_{0}\right)-\phi\left(0, t_{0}, 0\right)\right)-\left(\phi\left(0,0, t_{0}\right)-\phi\left(t_{0}, 0,0\right)\right) \\
+\left(\phi\left(0, t_{0}, 0\right)-\phi\left(t_{0}, 0,0\right)\right)=0
\end{gathered}
$$

Thus it produces the zero map from $S_{1}(X)$ to $X$. Next let $p=3$. We have $\phi\left(t_{0}, t_{1}, t_{2}, t_{3}\right)$

$$
\partial \phi\left(t_{0}, t_{1}, t_{2}\right) \equiv \phi\left(0, t_{0}, t_{1}, t_{2}\right)-\phi\left(t_{0}, 0, t_{1}, t_{2}\right)+\phi\left(t_{0}, t_{1}, 0, t_{2}\right)-\phi\left(t_{0}, t_{1}, t_{2}, 0\right)
$$

Then $\partial^{2} \phi\left(t_{0}, t_{1}\right)=$

$$
\begin{aligned}
& \phi\left(0,0, t_{0}, t_{1}\right)-\phi\left(0, t_{0}, 0, t_{1}\right)+\phi\left(0, t_{0}, t_{1}, 0\right) \\
& -\left(\phi\left(0,0, t_{0}, t_{1}\right)-\phi\left(t_{0}, 0,0, t_{1}\right)+\phi\left(t_{0}, 0, t_{1}, 0\right)\right) \\
& +\left(\phi\left(0, t_{0}, 0, t_{1}\right)-\phi\left(t_{0}, 0,0, t_{1}\right)+\phi\left(t_{0}, t_{1}, 0,0\right)\right) \\
& -\left(\phi\left(0, t_{0}, t_{1}, 0\right)-\phi\left(t_{0}, 0, t_{1}, 0\right)+\phi\left(t_{0}, t_{1}, 0,0\right)\right)
\end{aligned}
$$

which equals 0 .
Now suppose $\phi \in S_{p+1}(X), p+1>2$. I will indicate the position of various entries in the following by an index written above it.

$$
\begin{align*}
& \partial \phi\left(t_{0}, \ldots, t_{p}\right) \equiv \sum_{i=0}^{p+1}(-1)^{i} \phi\left(\begin{array}{c}
0 \\
t_{0}
\end{array}, \ldots, \stackrel{i}{0}, \stackrel{i+1}{t_{i}}, \ldots, \stackrel{p+1}{t_{p}}\right) \\
& =\phi\left(0, t_{0}, \ldots, t_{p}\right)+(-1)^{1} \phi\left(t_{0}, 0, t_{1}, \ldots, t_{p}\right)+\sum_{i=2}^{p-1}(-1)^{i} \phi\binom{0}{t_{0}, \ldots, \stackrel{i}{0}, \stackrel{i+1}{t_{i}}, \ldots, \stackrel{p+1}{t_{p}}} \\
& +(-1)^{p+1} \phi\left(t_{0}, \ldots, t_{p}, 0\right)+(-1)^{p} \phi\left(t_{0}, \ldots, 0, t_{p}\right) \tag{1.1}
\end{align*}
$$

Then doing another $\partial$, the first two terms and the last two will cancel. These give

$$
(-1)^{0} \phi\left(0,0, t_{0}, \ldots, t_{p-1}\right)+(-1)^{1} \phi\left(0,0, t_{0}, \ldots, t_{p-1}\right)
$$

and

$$
(-1)^{p+1}(-1)^{p} \phi\left(t_{0}, \ldots, 0,0\right)+(-1)^{p}(-1)^{p} \phi\left(t_{0}, \ldots, 0,0\right)
$$

while the middle term in 1.1 will consist of sums of the form $\phi\left(t_{0}, \ldots, 0, \ldots, 0, \ldots t_{p-1}\right)$ multiplied by -1 raised to a power where the first 0 is in position $r$ and the second is in position $s$. There will be a pair of these. One is of the form $\phi\left(t_{0}, \ldots, 0, \ldots, 0, \ldots t_{p-1}\right)(-1)^{r+s}$ and the other is of the form $\phi\left(t_{0}, \ldots, 0, \ldots, 0, \ldots t_{p-1}\right)(-1)^{r+s-1}$ so these cancel and the result is that there is nothing left. The functions just cancelled out. This is what we mean by 0 .

## A. 2 The Homology Groups

Definition A.2.1 A cycle in $S_{n}(X)$ will be $c$ such that $\partial c=0$. The set of all cycles will be denoted by $Z_{n}(X)$. Also $c$ will be a boundary if there exists $\psi \in S_{n+1}(X), \psi=\sum_{\eta} n_{\eta} \eta$ such that $c=\partial \psi$. Boundaries are denoted as $B_{n}(X)$. It follows from the above lemma that $B_{n}(X) \subseteq Z_{n}(X)$. Note that $S_{0}(X)=Z_{0}(X)$ because for any $\phi \in S_{0}(X), \partial \phi=0$.

So what is this referring to? Can we consider a picture which would depict or give some meaning to the ideas of these definitions? Let $X=\mathbb{R}^{2}$. The following would be a picture of the image of a cycle. In this picture the $\phi_{i}$ are such that as $\left(t_{0}, t_{1}\right)$ goes from $(1,0)$ to $(0,1)$ the motion of $\phi_{i}\left(t_{0}, t_{1}\right)$ is counter clockwise around the solid curve. It depicts some arcs strung together to go around something, ending where it started.


In this picture $\left(t_{0}, t_{1}\right) \rightarrow \phi_{i}\left(t_{0}, t_{1}\right)$ gives the curved lines shown and this is a cycle because

$$
\begin{aligned}
& \partial\left(\phi_{1}+\phi_{2}+\phi_{3}+\phi_{4}\right) \equiv \phi_{1}(1,0)-\phi_{1}(1,0) \\
&=\phi_{1}(0,1)+\phi_{2}(1,0)-\phi_{2}(1,0,1) \\
&=\phi_{2}(0,1) \\
&+\phi_{3}(1,0)-\phi_{3}(0,1)+\phi_{4}(1,0)-\phi_{4}(0,1)=0
\end{aligned}
$$

Notice how this being a cycle in this special case includes orientation in this observation. Assuming motion around the two curvy triangles having the dotted line as one side is counter clockwise, $\partial\left(\psi_{1}+\psi_{2}\right)$ would be a boundary which we could regard as essentially the same as the original cycle. Think of motion along the shared side canceling. More generally, if you have two cycles whose difference is a boundary, we will regard these as equivalent. This leads to the definition of homology groups.
Definition A.2.2 $H_{n}(X)=Z_{n}(X) / B_{n}(X)$. Thus $H_{n}(X)$ will be the equivalence classes where two things are equivalent means their difference is a boundary. $H_{n}(X)$ is called the homology group. I will denote as $[c]$ the element of $H_{n}(X)$ which is determined by the cycle c. As in linear algebra, we define addition by $[c]+[d] \equiv[c+d]$.

The addition is well defined because if $[c]=[\hat{c}],[d]=[\hat{d}]$ this means $c-\hat{c}, d-\hat{d}$ are both boundaries and so clearly $c+d-(\hat{c}+\hat{d})$ is also a boundary since $\partial$ is linear.

Lemma A.2.3 Suppose $f: X \rightarrow Y$ is continuous. Define

$$
f_{\#}(\phi)\left(t_{0}, \ldots, t_{n}\right) \equiv f \circ \phi\left(t_{0}, \ldots, t_{n}\right)
$$

Then $f_{\#}: S_{n}(X) \rightarrow S_{n}(Y)$ is the resulting homomorphism. Also, the following diagram commutes.

$$
\begin{array}{ccc}
S_{n}(X) & \xrightarrow{f_{\#}} & S_{n}(Y) \\
\downarrow \partial & & \downarrow \partial \\
S_{n-1}(X) & \xrightarrow{f_{\#}} & S_{n-1}(Y)
\end{array}
$$

We can define $f_{*}$ as a homomorphism of homology groups as follows: $f_{*}: H_{n}(X) \rightarrow H_{n}(Y)$. $f_{*}[c] \equiv\left[f_{\#}(c)\right]$.

Proof: The first claim is obvious that $f_{\#}(\phi) \in S_{n}(Y)$. Thus, defining $f_{\#}$ on the singular simplices, it extends uniquely to a homomorphism on $S_{n}(X)$.

$$
f_{\#}\left(\sum_{\phi} n_{\phi} \phi\right) \equiv \sum_{\phi} n_{\phi} f_{\#}(\phi)
$$

As to the diagram,

$$
\partial_{i} f_{\#}(\phi)\left(t_{0}, \ldots t_{n-1}\right) \equiv \partial_{i}(f \circ \phi)\left(t_{0}, \ldots, t_{n-1}\right) \equiv(f \circ \phi)\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{n-1}\right)
$$

and

$$
f_{\#}\left(\partial_{i} \phi\right)\left(t_{0}, \ldots, t_{n-1}\right) \equiv f_{\#} \phi\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{n-1}\right) \equiv(f \circ \phi)\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{n-1}\right)
$$

Consider the claim about $f_{*}$ as a homomorphism of homology groups. If $[c]=[\hat{c}]$, is it true that $\left[f_{\#} c\right]=\left[f_{\#} \hat{c}\right]$ ? Since $[c]=[\hat{c}]$, this means $c-\hat{c}$ is a boundary. Say $c-\hat{c}=\partial d$. Then

$$
f_{\#} c-f_{\#} \hat{c}=f_{\#}(c-\hat{c})=f_{\#}(\partial d)=\partial f_{\#}(d)
$$

and so it is indeed the case that $\left[f_{\#} c\right]=\left[f_{\#} \hat{c}\right]$ because the above difference is a boundary. Then with this being well defined, it is clear that $f_{*}$ is a homomorphism.

It appears that one could simply let $f_{\#}$ have the same meaning as $f_{*}$ but it looks to me like this new notation is preferred.

Then what if $f: X \rightarrow Y$ is a homeomorphism? In this case, $f$ is one to one and onto and it and its inverse are continuous. Thus we have the following corollary which says that homeomorphisms lead to isomorphisms of homology groups.This is fairly interesting because it connects a topological concept to one which is purely algebraic. Thus if you knew that two homology groups are not isomorphic, then you would also know that the two topological spaces are not homeomorphic.

Corollary A.2.4 Let $f: X \rightarrow Y$ be a homeomorphism. Then $f_{*}: H_{n}(X) \rightarrow H_{n}(Y)$ is an isomorphism, a homomorphism which is one to one and onto and its inverse is also a homomorphism which is one to one and onto.

Theorem A.2.5 Let $X \neq \emptyset$ be pathwise connected meaning that if $x, y$ are two points of $X$, there is a continuous map from $\sigma_{1}$ denoted as $\phi$ such that $\phi((1,0))=x, \phi((0,1))=y$. Then $H_{0}(X)$ is isomorphic to $\mathbb{Z}$ the additive group of the integers.

Proof: $S_{0}(X)$ consists of mappings $\phi$ which take $1=\sigma_{0}$ to points of $X$ so we could essentially identify $S_{0}(X)$ with the points of $X$. Also $Z_{0}(X)=S_{0}(X)$ by definition since $\partial$ maps all things in $S_{0}(X)$ to 0 . Pick $x \in X$. Let $\psi_{\phi}$ be the 1 simplex which "goes from" $\phi(1)$ to $x$ a point of $X$ so $\partial\left(\psi_{\phi}\right)=\phi(1)-x=\phi-x$. Then

$$
\sum_{\phi} n_{\phi} \phi=\sum_{\phi} n_{\phi}(\phi-x)+\left(\sum_{\phi} n_{\phi}\right) x=\partial\left(\sum_{\phi} n_{\phi} \psi_{\phi}\right)+\left(\sum_{\phi} n_{\phi}\right) x
$$

Let $\alpha: S_{0}(X) \rightarrow \mathbb{Z}$ be defined as $\alpha\left(\sum_{\phi} n_{\phi} \phi\right) \equiv \sum_{\phi} n_{\phi}$. Thus $\sum_{\phi} n_{\phi} \phi$ is a boundary if $\operatorname{ker}(\alpha)=0$ so $\operatorname{ker}(\alpha) \subseteq B_{0}(X)$. Next consider $\alpha$ acting on a boundary.

$$
\alpha\left(\partial\left(\sum_{\psi} n_{\psi} \psi\right)\right)=\alpha\left(\sum_{\psi} n_{\psi}\binom{\mathrm{a} 0 \text { simplex }}{\psi(1,0)-\psi(0,1)}\right)=0
$$

since $n_{\psi}-n_{\psi}=0$ so $B_{0}(X) \subseteq \operatorname{ker}(\alpha)$. Thus $\alpha$ is onto $\mathbb{Z}$ obviously, is a homomorphism, and $\operatorname{ker}(\alpha)=B_{0}(X)$ so $Z_{0}(X) / B_{0}(X) \approx \mathbb{Z}$. The isomorphism is $\hat{\alpha}([c]) \equiv \alpha(c)$.

Here and elsewhere, the notation $\approx$ means isomorphic. That is there is a one to one onto homomorphism with inverse also a homomorphism.

The tendency is to simply state that $Z_{0}(X) / B_{0}(X)=\mathbb{Z}$ because they are isomorphic and it seems to be all about algebraic considerations.

Definition A.2.6 Let $X$ be a topological space. We can say $x \sim y$ means there is a continuous curve in $X$ from $x$ to $y$. This is clearly an equivalence relation. The equivalence classes are called path components.

Definition A.2.7 When we have subgroups $G_{\alpha}, \alpha \in A$ of a free Abelian group $G, \sum_{\alpha \in A} G_{\alpha}$ signifies all finite sums $g_{\alpha_{1}}+g_{\alpha_{2}}+\cdots+g_{\alpha_{m}}$ where $g_{\alpha_{j}} \in G_{\alpha_{j}}$ and if two of these sums are equal, then the corresponding $g_{\alpha_{k}}$ are equal. $\sum_{\alpha \in A} G_{\alpha}$ is a new group. If you have

$$
g_{\alpha_{1}}+g_{\alpha_{2}}+\cdots+g_{\alpha_{m}}, \quad g_{\alpha_{1}}+g_{\beta_{2}}+\cdots+g_{\beta_{n}}
$$

and you wanted to add them, you would get

$$
2 g_{\alpha_{1}}+g_{\alpha_{2}}+\cdots+g_{\alpha_{m}}+g_{\beta_{2}}+\cdots+g_{\beta_{n}}
$$

One could define a group $\prod_{\alpha \in A} G_{\alpha}$ with the understanding that addition is componentwise, and then this direct sum $\sum_{\alpha \in A} G_{\alpha}$ is the subgroup of this product which involves only finitely many components being nonzero.

Proposition A.2.8 Let $X_{\alpha}, \alpha \in A$ be the path components of $X$. Then

$$
H_{n}(X)=\sum_{\alpha \in A} H_{n}\left(X_{\alpha}\right)
$$

Proof: Let $\phi$ be an $n$ simplex. Then since $\sigma_{n}$ is path connected, $\phi\left(\sigma_{n}\right)$ is contained in some $X_{\alpha}$ and has empty intersection with $X_{\beta}$ for any $\beta \neq \alpha$. Thus if you have any $c$ in $S_{n}(X)$ which is a cycle, so $c=\sum_{\phi} n_{\phi} \phi$, you can split this sum into finitely many pieces according to which $X_{\alpha}$ contains the image of $\phi$. Thus $c=\sum_{i=1}^{m} c_{i}$ where the simplices in $c_{i}$ have values in $X_{\alpha_{i}}$. Then each $c_{i}$ must also be a cycle. To see this note that $\partial_{i} \phi\left(\sigma_{n-1}\right)$ is contained in one of these $X_{\alpha}$, the same one which contains $\phi\left(\sigma_{n}\right)$. Also note that $\partial c=0$ amounts to having $|\partial c|=\emptyset$ where $|\partial \phi|$ denotes the points of $\phi\left(\sigma_{n-1}\right)$ a path connected set. When this happens, I will say that $c$ is supported in $X_{\alpha}$. How do we obtain that $c$ is a cycle? $\partial c$ must involve integer multiples of $\phi\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{n-1}\right)$ and these will cancel. However, all of $t_{0}, t_{i}, \ldots, t_{n-1} \rightarrow \phi\left(t_{0}, \ldots, 0, t_{i}, \ldots, t_{n-1}\right)$ describe path connected sets so each of these sets $\phi\left(\sigma_{n-1}\right)$ is contained in a single $X_{\alpha}$. Hence, if $\partial c=0$, then each $\partial c_{i}=0$ also.

Now suppose $d=\partial b$. Then $b=\sum_{\psi} n_{\psi} \psi$ where each $\psi\left(\sigma_{n}\right)$ is contained in a single $X_{\alpha}$ and in particular, $|\partial \psi| \subseteq X_{\alpha}$. Thus $\partial b=d=\sum_{\psi} n_{\psi} \partial \psi$ where $|\partial \psi|,|\psi|$ are contained in one of those $X_{\alpha}$. Thus the boundaries are of the same form $d=\sum_{\psi} n_{\psi} \partial \psi$ where $\psi$ has all
values in a single $X_{\alpha}$. If $c, \tilde{c}$ differ by a boundary, so $c-\tilde{c}=\partial b$, then $b=b_{1}+\cdots+b_{m}$ where each $\left|b_{j}\right|$ is contained in a single $X_{\alpha_{j}}, j=1,2, \ldots, m$. It follows that the $c_{i}, \tilde{c}_{i}$ just described, those supported in $X_{\alpha_{i}}$ satisfy $c_{j}-\tilde{c}_{j}=\partial b_{j}$ and so we can define $H_{n}(X)=\sum_{\alpha \in A} H_{n}\left(X_{\alpha}\right)$. In other words, if $c_{i}$ is a boundary supported on $X_{\alpha_{i}}$ then it is the boundary of some $b_{i}$ supported on $X_{\alpha_{i}}$ since the boundaries of the other $b_{j}$ will not intersect $X_{\alpha_{i}}$.

Next is something pretty interesting in the case of a convex topological space meaning that if $x, y \in X$, then it makes sense to form $x t+y(1-t) \in X$ for $t \in[0,1]$ and + is continuous from $X \times X$ to $X$. Here is a picture which suggests the construction used.


Theorem A.2.9 Let $X$ be convex. For all $n>0, H_{n}(X)=0$. Also, there exists a map $T: \phi \rightarrow \theta$ where $\phi$ is a singular $n$ simplex and $\theta$ is a singular $n+1$ simplex. This map can be obtained from a simple formula for $n \geq 0$. For $X$ convex cycles and boundaries are the same for $n>0$. Also for $\phi$ a simplex, $\phi=\partial T \phi+T \partial \phi$ where $T$ denotes the homomorphism from extending $T$ to all of $S_{n}(X)$.

Proof: For $n>0$ and $\phi \in S_{n}(X)$. Then define $\theta$ in $S_{n+1}(X)$ as follows.

$$
\theta\left(t_{0}, \ldots, t_{n}, t_{n+1}\right) \equiv\left\{\begin{array}{l}
\left(1-t_{0}\right) \phi\left(\frac{t_{1}}{\left(1-t_{0}\right)}, \frac{t_{2}}{\left(1-t_{0}\right)}, \ldots, \frac{t_{n+1}}{\left(1-t_{0}\right)}\right)+t_{0} x \text { if } t_{0}<1 \\
x \text { if } t_{0}=1
\end{array}\right.
$$

Note that we assume $\sum_{j=0}^{n+1} t_{j}=1$ each $t_{j} \geq 0$. Therefore, this makes perfect sense because $\sum_{j=1}^{n+1} t_{j}=1-t_{0}$ and so $\sum_{j=1}^{n+1} \frac{t_{j}}{\left(1-t_{0}\right)}=1$. The thing which might not be clear is that this $\theta$ is continuous. There is clearly no problem at any point of $\sigma_{n+1}$ where $t_{0}<1$ so let $\left(t_{0}^{n}, \ldots, t_{n}^{n}, t_{n+1}^{n}\right) \rightarrow\left(1, t_{1}, \ldots, t_{n+1}\right)$. Then, since the sum is always 1 , it follows that $t_{0}^{n} \rightarrow 1$ and $\sum_{j=1}^{n+1} t_{j}^{n} \rightarrow \sum_{j=1}^{n+1} t_{j}=0$. Also each $t_{j} \geq 0$ so they each converge to 0 . The set $\phi\left(\sigma_{n}\right)$ is bounded and so

$$
\left(1-t_{0}\right) \phi\left(\frac{t_{1}}{\left(1-t_{0}\right)}, \frac{t_{2}}{\left(1-t_{0}\right)}, \ldots, \frac{t_{n+1}}{\left(1-t_{0}\right)}\right) \rightarrow 0
$$

and $t_{0}^{n} x \rightarrow x$. Thus this is indeed continuous.
Let $T \phi \equiv \theta$ and $T$ a homomorphism. Thus $\partial_{0} T \phi=\phi$. Also $\partial_{1} \theta\left(t_{0}, \ldots, t_{n}\right)=$

$$
\theta\left(t_{0}, 0, t_{1}, \ldots, t_{n}\right) \equiv\left\{\begin{array}{l}
\left(1-t_{0}\right) \phi\left(0, \frac{t_{1}}{\left(1-t_{0}\right)}, \ldots, \frac{t_{n}}{\left(1-t_{0}\right)}\right)+t_{0} x, t_{0}<1 \\
x \text { for } t_{0}=1
\end{array}\right.
$$

and $\partial_{2} T \phi\left(t_{0}, \ldots, t_{n}\right) \equiv \theta\left(t_{0}, t_{1}, 0, t_{2}, \ldots, t_{n}\right)$

$$
=\left\{\begin{array}{l}
\left(1-t_{0}\right) \phi\left(\frac{t_{1}}{\left(1-t_{0}\right)}, 0, \frac{t_{2}}{\left(1-t_{0}\right)}, \ldots, \frac{t_{n}}{\left(1-t_{0}\right)}\right)+t_{0} x, t_{0}<1 \\
x \text { for } t_{0}=1
\end{array}\right.
$$

One sees the pattern from this. Now

$$
\begin{aligned}
\partial_{0} \phi\left(t_{0}, \ldots, t_{n-1}\right) & \equiv \phi\left(0, t_{0}, \ldots, t_{n-1}\right) \\
\partial_{1} \phi\left(t_{0}, \ldots, t_{n-1}\right) & \equiv \phi\left(t_{0}, 0, t_{1}, \ldots, t_{n-}\right)
\end{aligned}
$$

Thus

$$
\begin{gathered}
T \partial_{0} \phi\left(t_{0}, \ldots, t_{n}\right) \equiv\left\{\begin{array}{l}
\left(1-t_{0}\right) \phi\left(0, \frac{t_{1}}{1-t_{0}}, \ldots, \frac{t_{n}}{1-t_{0}}\right)+t_{0} x \text { if } t_{0}<1 \\
x \text { if } t_{0}=1
\end{array}\right. \\
T \partial_{1} \phi\left(t_{0}, \ldots, t_{n}\right)=\left\{\begin{array}{l}
\left(1-t_{0}\right) \phi\left(\frac{t_{1}}{1-t_{0}}, 0, \frac{t_{2}}{\left(1-t_{0}\right)}, \ldots, \frac{t_{n}}{1-t_{0}}\right)+t_{0} x \text { if } t_{0}<1 \\
x \text { if } t_{0}=1
\end{array}\right.
\end{gathered}
$$

etc. That is, $T \partial_{0} \phi=\partial_{1} T \phi$ and in general, $T \partial_{i-1} \phi=\partial_{i} T \phi$ until $T \partial_{n} \phi=\partial_{n+1} T \phi$.
Then $\partial T \phi+T \partial \phi=\phi+\sum_{i=1}^{n+1}(-1)^{i} \partial_{i} T \phi+\sum_{i=0}^{n}(-1)^{i} T \partial_{i} \phi$

$$
\begin{aligned}
& =\phi+\sum_{i=1}^{n+1}(-1)^{i} \partial_{i} T \phi+\sum_{i=1}^{n+1}(-1)^{i-1} T \partial_{i-1} \phi \\
& =\phi+\sum_{i=1}^{n+1}(-1)^{i} \partial_{i} T \phi+\sum_{i=1}^{n+1}(-1)^{i-1} \partial_{i} T \phi=\phi
\end{aligned}
$$

Thus $\partial T+T \partial$ is the identity. It follows that if $\phi \in Z_{n}(X)$, then $\phi=\partial T \phi+T \partial \phi=\partial T \phi$ and so $\phi$ is a boundary. Therefore, $Z_{n}(X) / B_{n}(X)=0$ because $[\phi]=0$ for all $\phi \in Z_{n}(X)$. Boundaries and cycles are the same thing for $n>0$ and $X$ convex.

Definition A.2.10 A set is star shaped if there is a special point $x$ called the star center such that segments from $x$ to other points are contained in the set.

An examination of the proof of the above shows the following corollary.
Corollary A.2.11 Let $X \subseteq \mathbb{R}^{p}$ be star shaped with star center $x$. Then if $n>0, H_{n}(X)=0$.


The above picture is of a star shaped set which is definitely not convex.

## A. 3 Homotopy

This is about homology groups of homotopic maps.
Lemma A.3.1 Let $X, Y$ be convex and let $g_{0}, g_{1}: X \rightarrow Y$ be given continuous functions. Then for each $n>0$ there exists $\hat{T}: S_{n}(X) \rightarrow S_{n+1}(Y)$ such that

$$
g_{0 \#}-g_{1 \#}=\hat{T} \partial+\partial \hat{T}
$$

Proof: From Theorem A.2.9, if $\phi \in S_{n}(Y)$ there exists a homomorphism on $S_{n}(Y)$ called $\tilde{T}$ such that $\phi=\partial \tilde{T} \phi+\tilde{T} \partial \phi$. Here $\tilde{T}: S_{n}(Y) \rightarrow S_{n+1}(Y)$. Now if $\phi \in S_{n}(X)$, it follows that $g_{0 \#} \phi-g_{1 \#} \phi \in S_{n}(Y)$ and so

$$
\begin{gathered}
g_{0 \#} \phi-g_{1 \#} \phi=\tilde{T} \partial\left(g_{0 \#} \phi-g_{1 \#} \phi\right)+\partial \tilde{T}\left(g_{0 \#} \phi-g_{1 \#} \phi\right) \\
\quad=\tilde{T} \partial\left(g_{0 \#}-g_{1 \#}\right) \phi+\partial \tilde{T}\left(g_{0 \#}-g_{1 \#}\right) \phi \\
\quad=\tilde{T}\left(g_{0 \#}-g_{1 \#}\right) \partial \phi+\partial \tilde{T}\left(g_{0 \#}-g_{1 \#}\right) \phi
\end{gathered}
$$

Let $\hat{T} \equiv \tilde{T}\left(g_{0 \#}-g_{1 \#}\right)$.
Definition A.3.2 Two functions $f, g: X \rightarrow Y$ are homotopic if there is a continuous function $F: X \times I \rightarrow Y$ for $I=[0,1]$ such that $f(x)=F(x, 0), g(x)=F(x, 1)$.

To begin with, consider something simpler than general homotopic maps. Let $g_{0}, g_{1}$ : $\sigma_{n} \rightarrow \sigma_{n} \times I, g_{0}(x) \equiv(x, 0)$ and let $g_{1}(x) \equiv(x, 1)$. These are obviously homotopic maps because you could just let $F(x, t)=g_{t}(x) \equiv(x, t)$ so this is essentially the simplest case. First is some notation of a technical nature. Now here is a useful lemma.

Lemma A.3.3 Let $\tau_{n}: \sigma_{n} \rightarrow \sigma_{n}$ be a singular simplex defined by

$$
\tau_{n}\left(t_{0}, \ldots, t_{n}\right)=\left(t_{0}, \ldots, t_{n}\right)
$$

Then if $\phi$ is a singular simplex in $X$, then $\phi_{\#}\left(\tau_{n}\right)=\phi$.
Proof: $\phi_{\#}\left(\tau_{n}\right) \equiv \phi \circ \tau_{n}=\phi$.
We have in mind the specific examples of $g_{0}$ and $g_{1}$ mentioned above. Using $X=\sigma_{n}$ and $Y=\sigma_{n} \times I$ we have obtained the following.

Lemma A.3.4 Let $g_{0}^{n}(x)=(x, 0), g_{1}^{n}(x)=(x, 1)$ for $x \in \sigma_{n}, n \geq 0$. Then there exist homomorphisms $\hat{T}$ mapping $S_{n}\left(\sigma_{n}\right)$ to $S_{n}\left(\sigma_{n} \times I\right)$ such that on $S_{n}\left(\sigma_{n}\right)$,

$$
\begin{equation*}
\partial \hat{T}(d)+\hat{T}(\partial d)=g_{0 \#}^{n}(d)-g_{1 \#}^{n}(d) \tag{1.2}
\end{equation*}
$$

Note how I put an $n$ on the $g_{i}$. This is because I am about to consider the case of $g_{0}(x)=(x, 0), g_{1}(x)=(x, 1)$ for $X$ an arbitrary topological space maybe not convex.

Lemma A.3.5 Let $g_{0}^{n}(x)=(x, 0), g_{1}^{n}(x)=(x, 1), x \in \sigma_{n}$, and let $g_{0}^{n}(x)=(x, 0), g_{1}^{n}(x)=$ $(x, 1)$. Then for $\phi$ a simplex in $S_{n}(X)$,

$$
\begin{equation*}
(\phi \times \mathrm{id})_{\#}\left(g_{j \#}^{n}\right)\left(\tau_{n}\right)=g_{j \#}(\phi) \tag{1.3}
\end{equation*}
$$

Proof: By definition, $g_{j \#}^{n}(\phi) \equiv g_{j}^{n} \circ \phi$. Thus

$$
g_{j \#}^{n}(\phi)(\boldsymbol{t})= \begin{cases}g_{j}^{n}(\phi(\boldsymbol{t}))=(\phi(\boldsymbol{t}), 0) & \text { if } j=0 \\ g_{j}^{n}(\phi(\boldsymbol{t}))=(\phi(\boldsymbol{t}), 1) \text { if } j=1\end{cases}
$$

Now consider the left side of 1.3 .

$$
\begin{aligned}
(\phi \times \mathrm{id})_{\#}\left(g_{j \#}^{n}\right)\left(\tau_{n}\right)(\boldsymbol{t}) & \equiv(\phi \times \mathrm{id})\left(g_{j}^{n}\left(\tau_{n}(\boldsymbol{t})\right)\right) \\
& \equiv(\phi \times \mathrm{id})\left(g_{j}^{n}(\boldsymbol{t})\right) \\
& =\left\{\begin{array}{r}
(\phi \times \mathrm{id})(\boldsymbol{t}, 0) \text { if } j=0 \\
(\phi \times \mathrm{id})(\boldsymbol{t}, 1) \text { if } j=1
\end{array}\right. \\
& =\left\{\begin{array}{r}
(\phi(\boldsymbol{t}), 0) \text { if } j=0 \\
(\phi(\boldsymbol{t}), 1) \text { if } j=1
\end{array}=g_{j \#}(\phi)(\boldsymbol{t})\right.
\end{aligned}
$$

These are the same so this proves the lemma.
I need to define $T$ according to having the following diagram commute where $\phi$ is a singular simplex. Then I can extend to make $T$ a homomorphism on all of $S_{n}(X)$ where $\hat{T}$ is the homomorphism of 1.2.

$$
\begin{array}{ccc}
S_{n}(X) & \xrightarrow{T} & S_{n+1}(X \times I) \\
\uparrow \phi_{\#} & & \uparrow(\phi \times \mathrm{id})_{\#}  \tag{1.4}\\
S_{n}\left(\sigma_{n}\right) & \xrightarrow{\hat{T}} & S_{n+1}\left(\sigma_{n} \times I\right)
\end{array}
$$

I just showed that the following diagram commutes.

$$
\begin{array}{ccc}
S_{n}(X) & \xrightarrow{g_{j \#}} & S_{n}(X \times I) \\
\uparrow \phi_{\#} & & \uparrow(\phi \times \mathrm{id})_{\#} \\
S_{n}\left(\sigma_{n}\right) & \xrightarrow{g_{j \#}^{n}} & S_{n}\left(\sigma_{n} \times I\right)
\end{array}
$$

Theorem A.3.6 Let $T$ be defined as follows for $\phi$ a simplex.

$$
T(\phi) \equiv T\left(\phi_{\#}\left(\tau_{n}\right)\right) \equiv(\phi \times \mathrm{id})_{\#} \hat{T}\left(\tau_{n}\right)
$$

where $\hat{T}$ is the homomorphism of 1.2. Then

$$
g_{0 \#}(\phi)-g_{1 \#}(\phi)=\partial T(\phi)+T \partial(\phi)
$$

Proof: From 1.4,

$$
\begin{gathered}
\partial T(\phi)+T \partial \phi=\partial\left((\phi \times \mathrm{id})_{\#} \hat{T}\left(\tau_{n}\right)\right)+T \partial \phi_{\#}\left(\tau_{n}\right) \\
=(\phi \times \mathrm{id})_{\#} \partial \hat{T}\left(\tau_{n}\right)+T \phi_{\#} \partial\left(\tau_{n}\right) \\
=(\phi \times \mathrm{id})_{\#} \partial \hat{T}\left(\tau_{n}\right)+(\phi \times \mathrm{id})_{\#} \hat{T} \partial \tau_{n} \\
=(\phi \times \mathrm{id})_{\#}\left(\partial \hat{T}\left(\tau_{n}\right)+\hat{T} \partial\left(\tau_{n}\right)\right)=(\phi \times \mathrm{id})_{\#}\left(g_{0 \#}^{n}\left(\tau_{n}\right)-g_{1 \#}^{n}\left(\tau_{n}\right)\right)
\end{gathered}
$$

From 1.3, this equals $g_{0 \#}(\phi)-g_{1 \#}(\phi)$.
Why is this significant? Suppose $\partial \phi=0$. Then from this theorem, $g_{0 \#}(\phi)-g_{1 \#}(\phi)$ equals a boundary and so $\left[g_{0 \#}(\phi)-g_{1 \#}(\phi)\right]$, the equivalence class in $H_{n}(X \times I)$ is 0 . This is called $g_{0 \#}$ and $g_{1 \#}$ are homologous and here $X$ is just a topological space.

With this preparation, it is time to consider the theorem about homotopic maps. So let $F: X \times I \rightarrow Y$ be continuous. Consider the following diagram which will help to complete the argument.


From Theorem A.3.6 we know there is a homomorphism $T$ for each $S_{n}(X)$ which maps $S_{n}(X)$ to $S_{n+1}(X \times I)$ such that

$$
\partial T(\phi)+T \partial(\phi)=g_{0 \#}(\phi)-g_{1 \#}(\phi)
$$

Thus we can do the homomorphism $F_{\#}$ to both sides. Then

$$
F_{\#} \partial T(\phi)+F_{\#} T(\partial \phi)=F_{\#} g_{0 \#}(\phi)-F_{\#} g_{1 \#}(\phi)
$$

But $F_{\#} g_{0 \#}=\left(F \circ g_{0}\right)_{\#}$ similar with $F_{\#} g_{1 \#}$. Therefore, if $F \circ g_{0}=f_{0}$ and $F \circ g_{1}=f_{1}$ so these are the two homotopic maps, it follows that

$$
\partial F_{\#} T(\phi)+F_{\#} T(\partial \phi)=f_{0 \#}(\phi)-f_{1 \#}(\phi)
$$

and this shows that for $S$ the homomorphism $F_{\#} T$,

$$
\partial S \phi+S \partial \phi=f_{0 \#}(\phi)-f_{1 \#}(\phi)
$$

This shows the main result which is the following.
Theorem A.3.7 Let $f_{0}$ and $f_{1}$ be continuous maps from $X$ to $Y$ which are homotopic. Then there exists a homomorphism $S$ mapping $S_{n}(X)$ to $S_{n+1}(Y)$ such that for all $\phi$ an $n$ simplex,

$$
\partial S \phi+S \partial \phi=f_{0 \#}(\phi)-f_{1 \#}(\phi)
$$

Since it doesn't matter which $n$ is used in this relation between $f_{0 \#}, f_{1 \#}$ in the above theorem, we say that this is a "chain homotopy". Note that if these conditions hold, then if $c$ is a cycle, then it is a boundary and so $\left[f_{0 \#}(c)-f_{1 \#}(c)\right]=0 . f_{0 \#}-f_{1 \#}$ maps cycles in $S_{n}(X)$ to boundaries in $S_{n}(Y)$.

Definition A.3.8 If $f: X \rightarrow Y$ and $g: Y \rightarrow X$ and $f \circ g$ and $g \circ f$ are both homotopic to the identity map on $Y$ and $X$ respectively, then these two are called homotopy inverses and $X, Y$ are said to have the same homotopy type.

Recall Corollary A.2.4 which said that if $X, Y$ are homeomorphic with homeomorphism $f$, then $f_{*}$ is an isomorphism of homology groups. In fact it is enough to assume less. This leads to the following theorem.

Theorem A.3.9 Let $f, g$ be homotopy inverses as just described. Then $f_{*}$ and $g_{*}$ are isomorphisms of the homology groups $H_{n}(X)$ and $H_{n}(Y)$. Recall $f_{*}[c] \equiv\left[f_{\#} c\right]$.

Proof: From Theorem A.3.7 there is a chain homotopy between $(g \circ f)_{\#}$ and $\mathrm{id}_{\#}$. Also, from the definition of \# we see that $(g \circ f)_{\#}=g_{\#} f_{\#}$ and so there is a homomorphism $S$ with $\partial S(\phi)+S \partial \phi=g_{\#} f_{\#}(\phi)-\mathrm{id}_{\#}(\phi)$ for any $n$ simplex $\phi$. Thus if $\partial c=0$ so $c$ is a cycle, then $g_{\#} f_{\#}(c)-\mathrm{id}_{\#}(c)$ is a boundary. Hence, when considered as maps on homology groups, $g_{*} f_{*}=\mathrm{id}_{*}$ on $H_{n}(X)$. Similarly $f_{*} g_{*}=\mathrm{id}_{*}$ on $H_{n}(Y)$ which shows that $f_{*}$ and $g_{*}$ are inverses as maps between homology groups and so these maps are both isomorphisms.

Here is an interesting case of the above. First is a definition. Following Vick,

Definition A.3.10 Suppose $A \subseteq X$ where $X$ is a topological space. Then $A$ is a retract of $X$ if there exists $g: X \rightarrow A$ such that $g$ is continuous and $g(x)=x$ for all $x \in A$. This function $g$ is called a retraction. Thus this can be written as $g \circ i$ is the identity map on $A$, where $i$ is the inclusion map of $A$ into $X$. If you can turn it around and have $i \circ g: X \rightarrow X$ homotopic to the identity, then $A$ is called a deformation retract of $X$.

The next lemma will be used quite a bit in considerations involving spheres.
Lemma A.3.11 Let $S^{n-1}$ denote $\sum_{k=1}^{n} x_{i}^{2}=1$. Then $S^{n-1}$ is a deformation retract of $\mathbb{R}^{n} \backslash$ $\{0\}$ and $H_{m}\left(S^{n-1}\right) \approx H_{m}\left(\mathbb{R}^{n} \backslash\{0\}\right)$. In fact, letting $g(x) \equiv \frac{x}{|x|}$ and $i$ be the identity, both $g \circ i$ and $i \circ g$ are homotopic to the identity.

Proof: Let $i$ be the inclusion map of $S^{n-1}$ into $\mathbb{R}^{n} \backslash\{0\}$ and define for $x \in \mathbb{R}^{n} \backslash\{0\}$ the unit vector $g(x) \equiv \frac{x}{|x|}$. Then $g \circ i$ is the identity on $S^{n-1}$. Consider $t(g \circ i)+(1-t)$ id, $t \in[0,1]$. This is a homotopy of $g \circ i$ and id on $S^{n-1}$. In fact their sum equals id so it maps $S^{n-1}$ to $S^{n-1}$. Next consider $t(i \circ g)+(1-t) \mathrm{id}, t \in[0,1]$. This is a homotopy of id and $(i \circ g)$ on $\mathbb{R}^{n} \backslash\{0\}$ because $(t(i \circ g)+(1-t) \mathrm{id})\left(\mathbb{R}^{n} \backslash\{0\}\right) \subseteq \mathbb{R}^{n} \backslash\{0\}$. It follows then from Theorem A.3.9 that $i_{*} g_{*}$ and $g_{*} i_{*}$ are both the identity on appropriate homology groups so $H_{m}\left(S^{n-1}\right) \approx H_{m}\left(\mathbb{R}^{n} \backslash\{0\}\right)$.

Then the following Corollary to the above theorem is obtained.
Corollary A.3.12 If $i: A \rightarrow X$ is the inclusion of a retract $A$ of $X$ then for each $n, i_{*}$ : $H_{n}(A) \rightarrow H_{n}(X)$ is a one to one homomorphism onto a direct summand. If $A$ is a deformation retract of $X$, then $i_{*}$ is an isomorphism.

Proof: We are assuming that $g \circ i$ is the identity map on $A$ where $g: X \rightarrow A$. In particular, $g \circ i$ and id are homotopic on $A$. Also $g_{*}(\gamma) \in H_{n}(A)$ because $g: X \rightarrow A$.

Therefore, from Theorem A.3.7 $g_{*} i_{*}=\mathrm{id}_{*}=$ identity on $H_{n}(A)$. Therefore, $i_{*}$ has a left inverse, namely $g_{*}$, and so it is one to one on $H_{n}(A)$. So consider $i_{*}: H_{n}(A) \rightarrow H_{n}(X)$. If $\gamma \in H_{n}(X)$, then

$$
\gamma=\stackrel{\substack{\text { image of } i_{*} \\ i_{*} g_{*} \\(\gamma)}+\left(\gamma-i_{*} g_{*}(\gamma)\right), ~}{\text { ( }}
$$

The first term is in $i_{*}\left(H_{n}(A)\right)$. In that second term,

$$
g_{*}\left(\gamma-i_{*} g_{*}(\gamma)\right)=g_{*}(\gamma)-g_{*} i_{*} g_{*}(\gamma)=g_{*}(\gamma)-g_{*}(\gamma)=0
$$

Thus every element of $H_{n}(X)$ is the sum of one in the image of $i_{*}$ and one in the $\operatorname{ker}\left(g_{*}\right)$. Suppose now that $\alpha$ is in the image of $i_{*}$ and also in $\operatorname{ker}\left(g_{*}\right)$. Then $\alpha=i_{*} \beta, g_{*}(\alpha)=0$. Then we have $0=g_{*} i_{*} \beta$, but this means $\beta=0$ since $g_{*} i_{*}=\mathrm{id}_{*}$. Now it follows that $\alpha=0$ also and so this is a direct sum $H_{n}(X)=i_{*}\left(H_{n}(A)\right) \oplus \operatorname{ker}\left(g_{*}\right)$.

In the second case, we are given from Theorem A.3.7 that $i_{*}$ is an isomorphism.

## A. 4 The Boundary Map on Geometric Simplices

Let $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]$ be a simplex in $\mathbb{R}^{n+1}$, consisting of the convex combinations of the $\boldsymbol{v}_{k}$, namely, expressions of the form $\sum_{k=0}^{n} t_{k} \boldsymbol{v}_{k}$ where $\sum_{k} t_{k}=1$ and each $t_{k} \geq 0$. This will be called a geometric $n$ simplex.

Definition A.4.1 Then there is a boundary operator defined as follows

$$
\hat{\partial}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]\right) \equiv \sum_{i=0}^{n}(-1)^{i}\left[\boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \boldsymbol{v}_{n}\right]
$$

Here this means we delete $\hat{\boldsymbol{v}}_{i}$ from the list of vertices and consider convex combinations of the rest. The i refers to position in the list from the left. Thus if you delete $\boldsymbol{v}_{i}$ the $\boldsymbol{v}_{i+1}$ which follows before deleting it is no longer in the $(i+1)^{\text {th }}$ position. It is now in the $i^{\text {th }}$. The sum means to consider these as elements of a free Abelian group. $\hat{\partial}\left(\left[\boldsymbol{v}_{0}\right]\right) \equiv 0$.

The fundamental result about $\hat{\partial}$ is that $\hat{\partial} \hat{\partial}=0$ where 0 will be the zero element in the free Abelian group just mentioned.

Proposition A.4.2 $\hat{\partial}^{2}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{p}\right]\right)=0$ whenever $p \geq 1$.
Proof: This is obvious from the definition if $p=1,0$. Suppose $p \geq 2$. Then from the definition, $\hat{\partial}^{2}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{p}\right]\right)=$

$$
\begin{aligned}
& \sum_{i=1}^{p} \sum_{j=0}^{i-1}(-1)^{i}(-1)^{j}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{j-1}, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{p}\right] \\
& +\sum_{i=0}^{p-1} \sum_{j=i+1}^{p}(-1)^{i}(-1)^{j-1}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{j-1}, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_{p}\right]
\end{aligned}
$$

There are $C(p, 2)$ pairs of deleted vectors and the two terms found in each pair occur with opposite sign.

I will generally use $\hat{\partial}$ for the boundary map on any $n$ simplex for any $n$. However, $\hat{\partial}$ will also be a homomorphism on the free Abelian group of $n$ simplices as follows:

$$
\hat{\partial}\left(\sum_{\sigma} n_{\sigma} \sigma\right) \equiv \sum_{\sigma} n_{\sigma} \hat{\partial}(\sigma)
$$

The free Abelian group will consist of $\sum_{\sigma} n_{\sigma} \sigma$ where $\sigma$ is some geometric simplex and this is a finite sum in which there are only finitely many nonzero $n_{\sigma}$.

## A. 5 The Subdivision Operation

Homology groups are of the form $Z_{n}(X) / B_{n}(X)$, cycles mod boundaries. What this section is all about is the following problem. Given $c \in Z_{n}(X)$, obtain $\hat{c}$ as $\sum_{i} n_{i} \phi_{i}$ where each $\phi_{i}$ has $\phi_{i}\left(\sigma_{p}\right)$ with diameter less than $\varepsilon$ and that also $[c]=[\hat{c}]$. First we consider geometric simplices and after that we extend to the general case.

Next is the definition of a cone which will be a new possibly higher dimensional simplex made from a single point $\boldsymbol{b}$ and the original simplex. This will lead to a triangulation of an original simplex $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$. Recall that this symbol means the convex combinations of $\left\{\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right\}$ that is, $\sum_{i=0}^{m} t_{i} \boldsymbol{v}_{i}$ where each $t_{i} \geq 0$ and $\sum_{i} t_{i}=1$. I will assume that the $\boldsymbol{v}_{j}$ are in a convex space $C$. Thus it will make perfect sense to consider things like convex combinations. Then later I will extend to arbitrary spaces.

Definition A.5.1 In general, for a simplex $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$, the cone determined by the simplex and $\boldsymbol{b}$ will just be $\left[\boldsymbol{b}, \boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$, denoted as $C_{\boldsymbol{b}}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$. The order is very important. You place $\boldsymbol{b}$ at the beginning. This is keeping track of orientation. Also make $C_{b}$ a homomorphism on $G_{n}(C)$, the free Abelian group of these geometric simplices contained in the convex set $C$. by defining $C_{b} \sum_{\sigma} m_{\sigma} \sigma \equiv \sum_{\sigma} m_{\sigma} C_{\boldsymbol{b}} \sigma$. In particular $C_{\boldsymbol{b}}\left(\hat{\partial}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]\right)\right)$ is $\sum_{i=0}^{m}(-1)^{i}\left[\boldsymbol{b}, \boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \boldsymbol{v}_{m}\right]$.

The following is a fundamental observation about this cone operation and the boundary homomorphism. It is illustrated in the following picture. If $\boldsymbol{b}$ is not in $\hat{\sigma}$, regard this picture as a view from the top


Lemma A.5.2 $\hat{\partial} C_{b} \hat{\partial}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]=\hat{\partial}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$. More generally if $c$ is a formal sum of these geometric simplices, $\hat{\partial} C_{b} \hat{\partial} c=\hat{\partial} c$.

Proof: Before proving this, here are a couple of examples. $\hat{\partial} C_{\boldsymbol{b}} \hat{\partial}\left[\boldsymbol{v}_{0}\right]=\mathbf{0}=\hat{\partial}\left[\boldsymbol{v}_{0}\right]$.
Now consider the case of a 1 simplex.

$$
\begin{aligned}
\hat{\partial} C_{\boldsymbol{b}} \hat{\partial}\left[\boldsymbol{v}_{0}, \boldsymbol{v}_{1}\right] & =\hat{\partial}\left(\left[\boldsymbol{b}, \boldsymbol{v}_{1}\right]-\left[\boldsymbol{b}, \boldsymbol{v}_{0}\right]\right)=\left[\boldsymbol{v}_{1}\right]-[\boldsymbol{b}]-\left(\left[\boldsymbol{v}_{0}\right]-[\boldsymbol{b}]\right) \\
& =\left[\boldsymbol{v}_{1}\right]-\left[\boldsymbol{v}_{0}\right]=\hat{\partial}\left[\boldsymbol{v}_{0}, \boldsymbol{v}_{1}\right]
\end{aligned}
$$

Now consider the general case. On the left is $\hat{\partial} C_{b} \sum_{i=0}^{m}(-1)^{i}\left[\boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \boldsymbol{v}_{m}\right]$ where the hat indicates $\boldsymbol{v}_{i}$ is missing. Then this equals

$$
\begin{equation*}
\sum_{i=0}^{m}(-1)^{i} \hat{\partial}\left[\boldsymbol{b}, \boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \boldsymbol{v}_{m}\right] \tag{1.5}
\end{equation*}
$$

In this last sum, when $\hat{\partial}$ is done to the inside, we get a sum of pairs

$$
(-1)^{j}(-1)^{i}\left[\boldsymbol{b}, \boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{j}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \boldsymbol{v}_{m}\right],(-1)^{j-1}(-1)^{i}\left[\boldsymbol{b}, \boldsymbol{v}_{0}, \cdots, \hat{\boldsymbol{v}}_{i}, \cdots, \hat{\boldsymbol{v}}_{j}, \cdots, \boldsymbol{v}_{m}\right]
$$

which cancel when summed. What is left of 1.5 in which the terms do not have $\boldsymbol{b}$ consists of $\hat{\partial}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{m}\right]$. The last claim follows from letting $C_{\boldsymbol{b}}$ be a homomorphism defined on the free group of geometric simplices.

Now I will define the subdivision operator. This will also be a homomorphism on the free Abelian group generated by the geometric simplices.

Definition A.5.3 $\widehat{\mathscr{S}}\left(\left[\boldsymbol{v}_{0}\right]\right) \equiv\left[\boldsymbol{v}_{0}\right]$. This defines the subdivision map on 0 simplices. We can now extend $\widehat{\mathscr{S}}$ as a homomorphism on the free group of geometric simplices. Suppose the homomorphism $\widehat{\mathscr{S}}$ has been obtained up to $n-1$. Let $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]$ be a simplex. The barycenter is defined as

$$
\boldsymbol{b} \equiv \frac{1}{n+1} \sum_{k=0}^{n} \boldsymbol{v}_{k}
$$

It is a point of $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]$ because it is a convex combination of the $\boldsymbol{v}_{k}$. It is called the barycenter because it is at the very center. Then for $c$ an $n$ simplex and $\boldsymbol{b}$ its barycenter, say $\hat{\partial} c \equiv \sum_{i=1}^{m} a_{i} c_{i}, a_{i}$ an integer $\pm 1$ and $c_{i}$ is of the form $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{n}\right]$ an $n-1$ simplex, then $\widehat{\mathscr{S}}(c) \equiv C_{b}(\widehat{\mathscr{S}}(\hat{\partial} c))$ where $\boldsymbol{b}$ is the barycenter of $c$. We extend $\widehat{\mathscr{S}}$ to make it a homomorphism on the free group of geometric simplices.

Here is how the subdivision operator and boundary interact. Here is how the subdivision operator and boundary interact

Theorem A.5.4 $\hat{\partial} \widehat{\mathscr{S}}(c)=\widehat{\mathscr{S}}(\hat{\partial} c)$.
Proof: This is clearly true for $n=0$. In case $n=1$, say $c=\left[\boldsymbol{v}_{0}, \boldsymbol{v}_{1}\right]$. Then the right is clearly $\left[\boldsymbol{v}_{1}\right]-\left[\boldsymbol{v}_{0}\right]$ and the left is

$$
\begin{aligned}
\hat{\partial}\left(C_{\boldsymbol{b}} \widehat{\mathscr{S}}\left(\hat{\partial}\left[\boldsymbol{v}_{0}, \boldsymbol{v}_{1}\right]\right)\right) & =\hat{\partial}\left(C_{\boldsymbol{b}} \widehat{\mathscr{S}}\left(\left[\boldsymbol{v}_{1}\right]-\left[\boldsymbol{v}_{0}\right]\right)\right)=\hat{\partial}\left(C_{\boldsymbol{b}}\left(\left[\boldsymbol{v}_{1}\right]-\left[\boldsymbol{v}_{0}\right]\right)\right) \\
& =\hat{\partial}\left(\left[\boldsymbol{b}, \boldsymbol{v}_{1}\right]-\left[\boldsymbol{b}, \boldsymbol{v}_{0}\right]\right)=\left[\boldsymbol{v}_{1}\right]-\left[\boldsymbol{v}_{0}\right]
\end{aligned}
$$

Assume then that $\widehat{\mathscr{S}}$ is a homomorphism defined up to $n-1$ for which the conclusion holds. Let $c$ be a geometric $n$ simplex. Then $\hat{\partial} \widehat{\mathscr{S}}(c) \equiv \hat{\partial} C_{\boldsymbol{b}} \widehat{\mathscr{S}}(\hat{\partial} c)$. Say $c=\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]$ so $\hat{\partial} c=\sum_{i=0}^{n}(-1)^{i}\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{n}\right]$, and so

$$
\widehat{\mathscr{S}} c=\sum_{i=0}^{n}(-1)^{i}\left[\boldsymbol{b}_{i}, \boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{n}\right]
$$

where $\boldsymbol{b}_{\boldsymbol{i}}$ is the barycenter of some boundary simplex. Then when $C_{\boldsymbol{b}}$ is done to this last expression, the individual terms in the resulting sum are of the form

$$
(-1)^{i}\left[\boldsymbol{b}, \boldsymbol{b}_{i}, \boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{n}\right]
$$

When $\hat{\partial}$ is done to sum of these, the terms not having $b$ reduce to

$$
\sum_{i=0}^{n}(-1)^{i}\left[\boldsymbol{b}_{i}, \boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \cdots, \boldsymbol{v}_{n}\right]
$$

which yields $\widehat{\mathscr{S}}\left(\hat{\partial}_{c}\right)$. For those which do have $\boldsymbol{b}$, it reduces to $C_{b}(\hat{\partial} \widehat{\mathscr{S}}(\hat{\partial}))$ which, by induction is the same as $C_{b}(\widehat{\mathscr{S}}(\hat{\partial} \hat{\partial} c))=0$.

Note that $\widehat{\mathscr{S}}(\hat{\partial} c)$ is composed of $n-1$ simplices which are contained in $\hat{\partial} c$ if $c$ is composed of $n$ simplices. Thus $n-1$ simplices on the "interior" disappeared.

What is accomplished by this subdivision applied to a simplex $S$ ? It yields a chain whose union is $S$ for which each simplex in the chain is small. If we do this subdivision operation enough, we can make the resulting simplices as small as desired. Here is why: If you have an $n$ simplex $\left[\boldsymbol{x}_{0}, \cdots, \boldsymbol{x}_{n}\right]$, its diameter is the maximum of $\left|\boldsymbol{x}_{k}-\boldsymbol{x}_{l}\right|$ for all $k \neq l$. Consider $\left|\boldsymbol{b}-\boldsymbol{x}_{j}\right|$. It equals

$$
\left|\sum_{i=0}^{n} \frac{1}{n+1}\left(\boldsymbol{x}_{i}-\boldsymbol{x}_{j}\right)\right|=\left|\sum_{i \neq j} \frac{1}{n+1}\left(\boldsymbol{x}_{i}-\boldsymbol{x}_{j}\right)\right| \leq \frac{n}{n+1} \operatorname{diam}(S) .
$$

The subdivision operator involves making a cone with a barycenter and each time you do it, another factor of no more than $\frac{n}{n+1}$ is introduced. Therefore, all simplices in $\mathscr{S}^{m} S$ are eventually as small as desired. By induction, $S$ is the sum of simplices which each have diameter less than $\varepsilon$. However, this delivers an order of the vertices in the geometric simplices obtained, not just their size. Recall that $\boldsymbol{b}$ always goes in the first slot. This is really why Theorem A. 5.4 holds. Note that this operator chops up all simplices of order $k$ for each $k \leq n$.

The following corollary comes from the above Theorem A.5.4.
Corollary A.5.5 For any $c$ an element of the free group of geometric p simplices,

$$
\hat{\partial} \widehat{\mathscr{S}}^{m}(c)=\widehat{\mathscr{S}}^{m}(\hat{\partial} c)
$$

Let $\sigma_{p}$ denote $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{p}\right]$ where $\left\{\boldsymbol{v}_{k}-\boldsymbol{v}_{0}\right\}_{k=1}^{p}$ are linearly independent and these vertices are at least in $\mathbb{R}^{p+1}$. For example they could be $\left(e_{1}, \cdots, e_{n+1}\right)$. I just described $\widehat{\mathscr{S}}$ as a homomorphism on the free group of oriented geometric simplices. The word "oriented" is included because the process also specifies the order of the vertices in each simplex.

Lemma A.5.6 Let $G_{n}(C)$ denote the free Abelian group of these geometric simplices. There exists a homomorphism $\hat{T}: G_{n}(C) \rightarrow G_{n+1}(C)$ such that $\hat{\partial} \hat{T}+\hat{T} \hat{\partial}=\widehat{\mathscr{S}}-\mathrm{id}$.

Proof: We prove this by induction. If $\hat{T}=0$, then this $\hat{T}$ works on $G_{0}(C)$ because $\widehat{\mathscr{S}}(c)=c$ for $c \in G_{0}(C)$. Suppose then that you have found a homomorphism which works for $k \leq n-1$. Then it is desired to have $\hat{T}$ such that $\hat{\partial} \hat{T}(c)+\hat{T}(\hat{\partial} c)=\widehat{\mathscr{S}}(c)-c$ for $c$ a geometric $n$ simplex. By induction, and Corollary A.5.5,

$$
\hat{\partial} \hat{T}(\hat{\partial} c)+\hat{T}(\hat{\partial} \hat{\partial} c)=\widehat{\mathscr{S}}(\hat{\partial} c)-\hat{\partial} c=\hat{\partial}(\widehat{\mathscr{S}}(c)-c)
$$

and so

$$
\begin{equation*}
0=\hat{\partial}((\widehat{\mathscr{S}}(c)-c)-\hat{T}(\hat{\partial} c)) \tag{1.6}
\end{equation*}
$$

Define $\hat{T}(c) \equiv C_{b(c)}((\widehat{\mathscr{S}}(c)-c)-\hat{T}(\hat{\partial} c))$. Then

$$
\partial \hat{T}(c)=\hat{\partial}\left(C_{b(c)}((\widehat{\mathscr{S}}(c)-c)-\hat{T}(\hat{\partial} c))\right)
$$

Now in the right side there are terms which have $\boldsymbol{b}(c)$ and terms which don't. As to the ones which do, we get $C_{\boldsymbol{b}(c)} \hat{\partial}((\widehat{\mathscr{S}}(c)-c)-\hat{T}(\hat{\partial} c))=0$ from 1.6 and for those which don't, we get $(\widehat{\mathscr{S}}(c)-c)-\hat{T}(\hat{\partial} c)$. Therefore, with this definition of $\hat{T}(c), \partial \hat{T}(c)=$ $(\widehat{\mathscr{S}}(c)-c)-\hat{T}(\partial c)$. Then extend $\hat{T}$ to a homomorphism on $G_{n}(C)$.

Definition A.5.7 Let $f: \sigma_{p} \rightarrow C$ where $C$ is also a convex space. Then $f$ is called affine if it does the following: $f\left(\sum_{k=0}^{m} t_{k} \boldsymbol{u}_{k}\right)=\sum_{k=0}^{m} t_{k} f\left(\boldsymbol{u}_{k}\right)$ whenever each $t_{k} \geq 0$ and $\sum_{k} t_{k}=1$.

Definition A.5.8 When one knows the ordered vertices of a geometric simplex

$$
\left[\boldsymbol{w}_{0}, \cdots . \boldsymbol{w}_{p}\right]
$$

it follows that one knows also a unique affine simplex $\phi$ having $\phi\left(\boldsymbol{v}_{i}\right)=\boldsymbol{w}_{i}$ defined by

$$
\phi\left(t_{0}, \cdots, t_{p}\right) \equiv \sum_{k=0}^{p} t_{k} \boldsymbol{\phi}\left(\boldsymbol{v}_{k}\right) \equiv \sum_{k=0}^{p} t_{k} \boldsymbol{w}_{k}
$$

whenever $\sum_{k} t_{k}=1, t_{k} \geq 0$. As to $\partial_{i} \phi\left(t_{0}, \cdots, t_{p-1}\right)$, it follows that $\partial_{i} \phi\left(\sigma_{p-1}\right)$ equals

$$
\left[\boldsymbol{w}_{0}, \cdots, \mathbf{0}, \boldsymbol{w}_{i+1}, \cdots, \boldsymbol{w}_{p}\right]
$$

which consists of ordered convex combinations of the vertices of $\left[\boldsymbol{w}_{0}, \cdots, \boldsymbol{w}_{p}\right]$ with the $\boldsymbol{w}_{i}$ replaced with $\mathbf{0}$.

Lemma A.5.9 Let $\phi$ be given in the above definition on $\sigma_{p}$ with values in a convex set $C$ which contains $\left\{\boldsymbol{w}_{0}, \cdots, \boldsymbol{w}_{p}\right\}$. Then $\phi$ is an affine map.

Proof: Let $\sum_{j=1}^{m} s_{j}=1, s_{j} \geq 0$ and consider points of $\sigma_{p} \sum_{i=0}^{p} t_{i}^{j} \boldsymbol{v}_{i}=\boldsymbol{u}_{j}$. I need to verify that $\phi\left(\sum_{j} s_{j} \boldsymbol{u}_{j}\right)=\sum_{j} s_{j} \phi\left(\boldsymbol{u}_{j}\right)$. However,

$$
\begin{aligned}
\phi\left(\sum_{j} s_{j} \boldsymbol{u}_{j}\right) & =\phi\left(\sum_{j} s_{j} \sum_{i=0}^{p} t_{i}^{j} \boldsymbol{v}_{i}\right)=\phi\left(\sum_{i=0}^{p}\left(\sum_{j} s_{j} t_{i}^{j}\right) \boldsymbol{v}_{i}\right) \\
& =\sum_{i=0}^{p}\left(\sum_{j} s_{j} t_{i}^{j}\right) \phi\left(\boldsymbol{v}_{i}\right)=\sum_{i=0}^{p}\left(\sum_{j} s_{j} t_{i}^{j}\right) \boldsymbol{w}_{i}
\end{aligned}
$$

because $\sum_{i=0}^{p}\left(\sum_{j} s_{j} t_{i}^{j}\right)=\sum_{j} s_{j} \sum_{i=0}^{p} t_{i}^{j}=1$. Also

$$
\sum_{j} s_{j} \phi\left(\boldsymbol{u}_{j}\right)=\sum_{j} s_{j} \phi\left(\sum_{i=0}^{p} t_{i}^{j} \boldsymbol{v}_{i}\right) \equiv \sum_{j} s_{j} \sum_{i=0}^{p} t_{i}^{j} \phi\left(\boldsymbol{v}_{i}\right)=\sum_{i=0}^{p}\left(\sum_{j} s_{j} t_{i}^{j}\right) \boldsymbol{w}_{i}
$$

Definition A.5.10 Denote by $A_{p}(C)$ the affine singular simplices mapping $\sigma_{p}$ to $C$.
Observation A.5.11 Because of these observations, we can regard $\hat{T}$ in Lemma A.5.6 as a homomorphism mapping $A_{n}(C)$ to $A_{n+1}(C)$. We can also regard $\widehat{\mathscr{S}}$ as a homomorphism on $A_{n}(C)$ which subdivides an affine singular simplex $\phi$ into affine singular simplices having smaller image and in terms of $A_{n}(C)$ we can replace $\hat{\partial}$ with the standard boundary operator $\partial$ in Lemma A.5.6. It amounts to nothing more than replacing each geometric simplex with a singular affine simplex, a mapping which has as its image the geometric simplex preserving order of vertices obtained by the subdivision operation.

Definition A.5.12 Define for $f$ affine, $f_{\#}: A_{n}(C) \rightarrow A_{n}(\tilde{C})$ as the extension of what was just described to all of $A_{n}(C)$.

Lemma A.5.13 Let $f, C, \tilde{C}$ be as in Definition A.5.12. Then $\hat{\partial} f_{\#}=f_{\#} \hat{\partial}$. Also if $\boldsymbol{b}$ is the barycenter of $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]$, then $f(\boldsymbol{b})$ is the barycenter of $f\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]\right)$ and $f_{\#}: A_{n}(\boldsymbol{C}) \rightarrow$ $A_{n}(\tilde{C})$. As usual, $f_{\#}$ is the homomorphism which acts on chains in $A_{n}(C)$ as before.

Proof: Note that the composition of affine maps is affine. It suffices to verify this lemma for a geometric simplex.

$$
\hat{\partial} f_{\#}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]\right) \equiv \hat{\partial}\left(\left[f\left(\boldsymbol{v}_{0}\right), \cdots, f\left(\boldsymbol{v}_{n}\right)\right]\right) \equiv \sum_{k=0}^{n}(-1)^{k}\left[f\left(\boldsymbol{v}_{0}\right), \cdots, \widehat{f\left(\boldsymbol{v}_{k}\right)}, \cdots, f\left(\boldsymbol{v}_{n}\right)\right]
$$

$$
\begin{aligned}
f_{\#} \hat{\partial}\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]\right) & \equiv f_{\#}\left(\sum_{k=0}^{n}(-1)^{k}\left[\boldsymbol{v}_{0}, \cdots, \widehat{\boldsymbol{v}_{k}}, \cdots, \boldsymbol{v}_{n}\right]\right) \\
& \equiv \sum_{k=0}^{n}(-1)^{k}\left[f\left(\boldsymbol{v}_{0}\right), \cdots, \widehat{f\left(\boldsymbol{v}_{k}\right)}, \cdots, f\left(\boldsymbol{v}_{n}\right)\right]
\end{aligned}
$$

and these are the same. As to the claim about the barycenter,

$$
f(\boldsymbol{b})=f\left(\sum_{k=0}^{n} \frac{1}{n+1} \boldsymbol{v}_{k}\right)=\sum_{k=0}^{n} \frac{1}{n+1} f\left(\boldsymbol{v}_{k}\right)
$$

which is the barycenter of $\left[f\left(\boldsymbol{v}_{0}\right), \cdots, f\left(\boldsymbol{v}_{n}\right)\right]=f\left(\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{n}\right]\right)$. Then note that, as discussed earlier, one obtains a result on $A_{n}(C)$ with $\partial$ in place of $\hat{\partial}$.

I will use $\widehat{\mathscr{S}}$ as the subdivision operator for either $\tilde{C}$ or $C$. Also I will use $\hat{T}$ as described above for either $C$ or $\tilde{C}$. Now if $f$ is an affine map $f: C \rightarrow \tilde{C}$ where $C, \tilde{C}$ are convex spaces, it is also the case that $f_{\#}$ commutes with $\hat{T}$ and with $\widehat{\mathscr{S}}$. Consider first $\widehat{\mathscr{S}}$. Let $c \in G_{n}(C)$ be a chain in $G_{n}(C)$. Then the geometric simplices corresponding to $\widehat{\mathscr{S}}(c)$ are $\sum_{i=0}^{n}(-1)^{i}\left[\boldsymbol{b}, \boldsymbol{u}_{0}, \cdots, \mathbf{0}, \cdots \boldsymbol{u}_{n}\right]$ where $\boldsymbol{b}$ is the barycenter. Then the geometric simplices corresponding to $f_{\#}(\widehat{\mathscr{S}}(c))$ would be $\sum_{i=0}^{n}(-1)^{i}\left[f(\boldsymbol{b}), f\left(\boldsymbol{u}_{0}\right), \cdots, \mathbf{0}, \cdots, f\left(\boldsymbol{u}_{n}\right)\right]$ Now consider $\widehat{\mathscr{S}}\left(f_{\#}(c)\right)=\widehat{\mathscr{S}}(f \circ c)$. The geometric simplices associated with this would involve the barycenter of $f\left(\boldsymbol{u}_{0}\right), \ldots, f\left(\boldsymbol{u}_{n}\right)$ which is $\frac{1}{n+1} \sum_{i} f\left(\boldsymbol{u}_{i}\right)=f\left(\frac{1}{n+1} \sum_{i} \boldsymbol{u}_{i}\right)=f(\boldsymbol{b})$. Thus the geometric simplices associated will be the same as the above. Hence $f_{\#} \widehat{\mathscr{S}}=\widehat{\mathscr{S}} f_{\#}$ if $f$ is affine. That $f_{\#}$ commutes with $\hat{T}$ is similar and follows from the observation that since $f$ is affine, $f$ of a barycenter of some vectors equals the barycenter of $f$ of these vectors in the same way as just noted. Now, as noted above, this leads to the same results for $A_{n}(C)$ the chains of affine singular simplices. The following is a summary.

Lemma A.5.14 If $f: C \rightarrow \tilde{C}$ convex sets, $f$ affine, the following hold:

$$
\begin{aligned}
f_{\#}\left(A_{n}(C)\right) & \subseteq A_{n}(\tilde{C}), f_{\#} \hat{T}=\hat{T} f_{\#}, f_{\#} \widehat{\mathscr{S}}=\widehat{\mathscr{S}} f_{\#} \\
\widehat{\mathscr{S}} \hat{\partial} & =\hat{\partial} \widehat{\mathscr{S}}, \hat{\partial} \hat{T}+\hat{T} \hat{\partial}=\widehat{\mathscr{S}}-\mathrm{id}
\end{aligned}
$$

Next is a specific example of an affine map from $\sigma_{p}$ to $\sigma_{p}$.
Definition A.5.15 Do the subdivision operator $\widehat{\mathscr{S}}$ on $\left[\boldsymbol{v}_{0}, \cdots, \boldsymbol{v}_{p}\right]$ multiple times to obtain a small simplices $\left\{\sigma_{k}\right\}_{k=1}^{N}$ contained in $\sigma_{p}$. Then for one of these $\sigma_{k}=\left[\boldsymbol{w}_{0}^{k}, \cdots, \boldsymbol{w}_{p}^{k}\right]$, define a map $\pi_{k}: \sigma_{p} \rightarrow \sigma_{k} \subseteq \sigma_{p}$ by $\pi_{k}\left(\sum_{i=0}^{p} t_{i} \boldsymbol{v}_{i}\right) \equiv \sum_{i=0}^{p} t_{i} \boldsymbol{w}_{i}^{k}$ where $t_{i} \geq 0, \sum_{i} t_{i}=1$. The order of the vertices in $\sigma_{k}$ is determined by the construction in the subdivision operator. Since $\pi_{k}$ is affine, we can regard $\pi_{k}$ as a homomorphism on $A_{n}\left(\sigma_{p}\right)$.

The above partition operator allows consideration of singular simplices of the form $\phi \circ \pi_{k}$ in which $\phi$ is a singular simplex, a continuous mapping defined on $\sigma_{p}$ having values in some topological space $X$, not necessarily convex. These compositions will have "small" image. It is important to use something like these to identify homology groups. One wonders whether $\phi \in S_{n}(X)$ is homologous to $\phi_{\#} \sum_{k} \pi_{k}$. Note that $\tau_{n}$, the identity map on $\sigma_{n}$ is obviously an affine map.

Definition A.5.16 For $X$ an arbitrary topological space, let $\psi$ be an n simplex. Let $\pi_{k}$ be the affine map from $\sigma_{n}$ to $\alpha_{k}$ with the $\alpha_{k}$ those simplices which result from the subdivision operator. Then we can regard $\widehat{\mathscr{S}}\left(\tau_{n}\right)$ as $\sum_{k} \pi_{k}$ with the conclusion of Lemma A.5.14 applying in terms of the $A_{n}(C)$. Then define $\mathscr{S}$ a subdivision homomorphism as follows.

$$
\begin{equation*}
\mathscr{S} \psi \equiv \psi_{\#}\left(\sum_{k} \pi_{k}\right) \equiv \psi_{\#} \widehat{\mathscr{S}}\left(\tau_{n}\right) \equiv \sum_{k} \psi \circ \pi_{k} \tag{1.7}
\end{equation*}
$$

Also define a homomorphism $T: S_{n}(X) \rightarrow S_{n+1}(X)$ as follows:

$$
\begin{equation*}
T \psi=T \psi_{\#}\left(\tau_{n}\right) \equiv \psi_{\#} \hat{T}\left(\tau_{n}\right), \text { so } T \psi \equiv \psi_{\#} \hat{T} \tag{1.8}
\end{equation*}
$$

These are well defined thanks to Lemma A.5.14 which says it holds on $A_{n}(C)$ for $C$ convex. Extend $\mathscr{S}$ and $T$ as homomorphisms on $S_{n}(X)$.

Then we get the following proposition about $T$ and $\mathscr{S}$.
Proposition A.5.17 $\partial T+T \partial=\mathscr{S}-$ id. Also $\partial \mathscr{S}=\mathscr{S} \partial$ and $\mathscr{S}^{m} \psi$ consists of a chain of simplices whose image in $X$ is $\psi(\alpha)$ for $\alpha$ as small as desired.

Proof: First,

$$
\partial \mathscr{S} \psi \equiv \partial\left(\psi_{\#} \sum_{k} \pi_{k}\right)=\psi_{\#} \partial\left(\sum_{k} \pi_{k}\right)=\psi_{\#}\left(\sum_{k} \partial \pi_{k}\right)
$$

while

$$
\mathscr{S} \partial \psi \equiv(\partial \psi)_{\#}\left(\sum_{k} \pi_{k}\right)=\partial \psi_{\#}\left(\sum_{k} \pi_{k}\right) \equiv \psi_{\#}\left(\sum_{k} \partial \pi_{k}\right)
$$

For $\partial$ the boundary operator in $S_{n}(X)$, and using the fact that $T$ is defined as a homomorphism on $S_{n}(X)$, and that $\psi_{\#} \partial=\partial \psi_{\#}$,

$$
\begin{aligned}
\partial T \psi+T \partial \psi & =(\partial T \psi+T \partial \psi)\left(\tau_{n}\right) \equiv\left(\partial \psi_{\#} \hat{T}+\partial \psi_{\#} \hat{T}\right)\left(\tau_{n}\right) \\
& =\psi_{\#}(\partial \hat{T}+\partial \hat{T})\left(\tau_{n}\right)=\psi_{\#}(\widehat{\mathscr{S}}-\mathrm{id})\left(\tau_{n}\right)=(\mathscr{S}-\mathrm{id})(\psi)
\end{aligned}
$$

and since $T$ is a homomorphism, this shows what is desired.
The last claim follows from the earlier material applied to the geometric simplices $\alpha_{k}$ obtained from successive applications of the subdivision map. Let $\psi$ be a singular $n$ simplex in $S_{n}(X)$.

Doing $\mathscr{S}$ to both sides of $\partial T+T \partial=\mathscr{S}-$ id yields $\partial \mathscr{S} T+\mathscr{S} T \partial=\mathscr{S}^{2}-\mathscr{S}$ which can now be used to see that $[\mathscr{S} \phi]=\left[\mathscr{S}^{2} \phi\right]$. Then continuing this way one sees that $\left[\mathscr{S}^{m} \phi\right]=[\phi]$ for $\phi$ a cycle.

Now here is a definition which will help to compute homology groups.
Definition A.5.18 Let $\mathscr{U}$ be a covering of $X$ and let $S_{n}^{\mathscr{U}}(X)$ consist of the subgroup of $S_{n}(X)$ generated by singular simplices $\phi$ with the property that $\phi\left(\sigma_{n}\right)$ is contained in some set $U \in \mathscr{U}$. Then $H_{n}\left(S_{n}^{\mathscr{U}}(X)\right)$ will denote the homology group obtained as before except now cycles and boundaries are with respect to $S_{n}^{\mathscr{U}}(X)$. We can do this because $\partial \phi \in$ $S_{n-1}^{\mathscr{U}}(X)$ if $\phi \in S_{n}^{\mathscr{U}}(X)$. Then $H_{n}\left(S_{n}^{\mathscr{U}}(X)\right)$ will denote the usual thing. Letting $Z_{n}^{\mathscr{U}}(X)$
consist of the free group generated by $\left\{\phi \in S_{n}^{\mathscr{U}}(X): \partial \phi=0\right\}$ and letting $B_{n}^{\mathscr{U}}(X)$ be those $c \in Z_{n}^{\mathscr{U}}(X)$ which are of the form $c=\partial d$ for some $d \in S_{n+1}(X)$,

$$
H_{n}\left(S_{n}^{\mathscr{U}}(X)\right) \equiv Z_{n}^{\mathscr{U}}(X) / B_{n}(X)
$$

Observation A.5.19 One should also observe the following. If $\mathscr{U}$ is a covering of $X$ and $\mathscr{V}$ is a covering of $Y$ and $f: X \rightarrow Y$ is continuous with the property that for each $U \in \mathscr{U}$, $f(U)$ is contained in some $V \in \mathscr{V}$ then $f_{\#}: S_{n}^{\mathscr{U}}(X) \rightarrow S_{n}^{\mathscr{V}}(Y)$ is a homomorphism.

Proposition A.5.20 Let $K$ be a compact subset in a metric space and let $\mathscr{U}$ be an open covering of $K$. Then there exists $\delta>0$ such that $B(k, \delta)$ is contained in some $U \in \mathscr{U}$ for every $k \in K$. This $\delta$ is called a Lebesgue number.

Proof: If $\delta$ does not exist, then for each $n \in \mathbb{N}$ there exists $k_{n}$ such that $B\left(k_{n}, \frac{1}{n}\right)$ is not contained in any single open set from $\mathscr{U}$. However, since $K$ is compact, there is a subsequence, still denoted as $k_{n}$ which converges to $k \in K$. Now $k \in U$ for some $U \in \mathscr{U}$ and since $U$ is open, $B(k, 2 \varepsilon) \subseteq U$ for some $\varepsilon$. Now for all $n$ large enough, $k_{n} \in B(k, \varepsilon)$ and so $B\left(k_{n}, \varepsilon\right) \subseteq U$ which is a contradiction.

Now here is the main result.
Theorem A.5.21 Let $\mathscr{U}$ be a covering of $X$ such that $\{\operatorname{int}(U): U \in \mathscr{U}\}$ which is denoted as int $(\mathscr{U})$ is also a covering of $X$. Then for $i$ the identity map, $i_{*}: H_{n}\left(S_{n}^{\mathscr{U}}(X)\right) \rightarrow H_{n}(X)$ is an isomorphism. In fact, if $c \in H_{n}(X)$ then there exists $\hat{c} \in H_{n}\left(S_{n}^{\mathscr{U}}(X)\right)$ with $[\hat{c}]=[c]$.

Proof: Let $c$ be a cycle in $S_{n}(X)$. From Proposition A.5.17, $[c]=[\mathscr{S} c]$. By induction, we have $[c]=\left[\mathscr{S}^{m} c\right]$. Say $c=\sum_{\phi} n_{\phi} \phi$ and, as observed, all the simplices $\psi$ in the chain $\mathscr{S}^{m} c$ have the property that their images in $X$ are of the form $\phi(\alpha)$ where $\phi$ is one of finitely many simplices in the cycle $c$ and $\alpha$ is a set of sufficiently small diameter that $\phi(\alpha)$ must be contained in some $\operatorname{int}(U)$ provided $m$ is sufficiently large. To see that $m$ exists, note that $\phi\left(\sigma_{n}\right)$ is compact so there exists a Lebesgue number for this compact set with the covering $\mathscr{U}$. When all the $\phi(\alpha)$ are smaller than this number they are each contained in a single int $(U)$.

From this, it appears that we can compute $H_{n}\left(S_{n}^{\mathscr{U}}(X)\right)$ and obtain $H_{n}(X)$ because the inclusion map is onto. Also note that, from the construction, the geometric simplices in $\mathscr{S}^{m+1} c$ are each contained in a simplex of $\mathscr{S}^{m} c$ which is itself the union of those in $\mathscr{S}^{m+1} c$.

## A. 6 Exact Sequences

This section is completely free of context and is pure algebra. Actually the boundary maps are on different levels so could be denoted with a subscript to indicate which level. However, I will continue to use $\partial$. It is only the algebraic properties of this map which are important.

Definition A.6.1 Let $C, D, E$ be Abelian groups. For $f, g$ the indicated homomorphisms, we say that this sequence is exact of $f(C)=\operatorname{ker}(g)$

$$
C \xrightarrow{f} D \xrightarrow{g} E
$$

One can string along more groups than three and it is called exact if every triple is like the above. When you just have three, and $f$ is one to one and $g$ is onto, the notation is as follows.

$$
0 \rightarrow C \xrightarrow{f} D \xrightarrow{g} E \rightarrow 0
$$

To express this situation that $f(C)=\operatorname{ker}(g)$ and you call it a short exact sequence. The reason for the arrow on the right is that $g$ maps onto $E$. The arrow on the left indicates that $f$ is one to one.

Now suppose you have $\partial \partial=0$

$$
\begin{array}{rllllll}
0 \rightarrow & \vdots & \dot{C}_{n+1} & \xrightarrow{f} & \vdots & \dot{D}_{n+1} & \xrightarrow{g} \\
& \downarrow \partial & & \downarrow \partial & & \downarrow \partial & \\
0 \rightarrow & C_{n} & \xrightarrow{f} & D_{n} & \xrightarrow{g} & E_{n} & \rightarrow 0  \tag{1.9}\\
& \downarrow \partial & & \downarrow \partial & & \downarrow \partial & \\
0 \rightarrow & C_{n-1} & \xrightarrow{f} & D_{n-1} & \xrightarrow{g} & E_{n-1} & \rightarrow 0 \\
& \vdots & & \vdots & & \vdots &
\end{array}
$$

In the examples of interest $f$ will be $f_{\#}, g$ will be $g_{\#}$ where $f, g$ will be continuous mappings. I will be assuming that $f$ is one to one, $g$ is onto and $\operatorname{ker}(g)=\operatorname{Im}(f)$ so that each line is a short exact sequence. However, here these maps are just homomorphisms and we call them chain maps because the rectangles commute.That is $\partial f=f \partial$, etc. Since $\partial \partial=0$ it makes sense to consider homology groups obtained from cycles (ker $\partial$ ) mod boundaries (image $\partial$ ). $f_{*}[c] \equiv\left[f_{\# c} c\right]$ where $f_{*}$ is a map on homology groups. It turns out there is something called a connecting homomorphism which leads to a long exact sequence of homology groups. I will simply use $f, g$ to denote these other mappings to save on notation. The corresponding homology groups will be $H_{C_{n}}, H_{D_{n}}, H_{E_{n}}$. The following is from Spanier [39].

Lemma A.6.2 There exists a homomorphism $\Delta: H_{E_{n+1}} \rightarrow H_{E_{n}}$ which satisfies

$$
\begin{equation*}
\Delta\left[e_{n+1}\right] \equiv\left[f^{-1} \partial g^{-1} e_{n+1}\right] \tag{1.10}
\end{equation*}
$$

This is called the connecting homomorphism. Here $\left[e_{n+1}\right] \in H_{E_{n+1}}$ so $e_{n+1}$ is a cycle.
Proof: The main problem is showing that this is well defined. Thus we need to show that the same thing is obtained with $e_{n+1}$ replaced with $e_{n+1}+\partial e_{n+2}$ on the right side independent of the choice of $e_{n+2}$. Letting the two be $e_{n+2}, \hat{e}_{n+2}$, denote with a hat as follows.

Let

$$
g\left(d_{n+1}\right)=e_{n+1}+\partial e_{n+2}, g\left(\hat{d}_{n+1}\right)=e_{n+1}+\partial \hat{e}_{n+2}
$$

so that $d_{n+1} \in g^{-1}\left(e_{n+1}+\partial e_{n+2}\right), \hat{d}_{n+1} \in g^{-1}\left(e_{n+1}+\partial \hat{e}_{n+2}\right)$. Then

$$
\begin{aligned}
& g \partial d_{n+1}=\partial\left(g\left(d_{n+1}\right)\right)=\partial\left(e_{n+1}+\partial e_{n+2}\right)=0 \\
& g \partial \hat{d}_{n+1}=\partial\left(g\left(\hat{d}_{n+1}\right)\right)=\partial\left(e_{n+1}+\partial \hat{e}_{n+2}\right)=0
\end{aligned}
$$

Thus $\partial d_{n+1} \in \operatorname{ker}(g)$ and so $\partial d_{n+1}=f\left(c_{n+1}\right)$ for some unique $c_{n+1} \in C_{n+1}$. Same with $\hat{c}_{n+1}$. Is $c_{n+1}$ a cycle?

$$
f\left(\partial c_{n+1}\right)=\partial f\left(c_{n+1}\right)=\partial \partial d_{n+1}=0
$$

so $\partial c_{n+1}=0$ since $f$ is one to one and so this is indeed a cycle and the result on the right in 1.10 is $\left[c_{n+1}\right]$. Same with $\hat{c}_{n+1}$. Will $\left[c_{n+1}-\hat{c}_{n+1}\right]=0$ ? If so, this will prove that the expression on the right in 1.10 is well defined.

$$
\begin{equation*}
f\left(c_{n+1}-\hat{c}_{n+1}\right)=\left(\partial d_{n+1}-\partial \hat{d}_{n+1}\right),\left(c_{n+1}-\hat{c}_{n+1}\right)=f^{-1}\left(\partial\left(d_{n+1}-\hat{d}_{n+1}\right)\right) \tag{1.11}
\end{equation*}
$$

Now $f\left(f^{-1}(\partial d)\right)=\partial d$ and $f\left(\partial f^{-1}(d)\right)=\partial f\left(f^{-1}(d)\right)=\partial d$ so $f^{-1}(\partial d)=\partial f^{-1}(d)$. It follows from this observation and 1.11 that $c_{n+1}-\hat{c}_{n+1}$ is a boundary and so $\left[c_{n+1}-\hat{c}_{n+1}\right]=$ 0 . Therefore, $\Delta$ is well defined.

Is $\Delta$ a homomorphism?

$$
\Delta\left(\left[e_{n+1}\right]+\left[\hat{e}_{n+1}\right]\right) \equiv \Delta\left[e_{n+1}+\hat{e}_{n+1}\right] \equiv\left[f^{-1} \partial g^{-1}\left(e_{n+1}+\hat{e}_{n+1}\right)\right]
$$

Letting $d_{n+1} \in g^{-1}\left(e_{n+1}\right), \hat{d}_{n+1} \in g^{-1}\left(\hat{e}_{n+1}\right)$, it follows that

$$
d_{n+1}+\hat{d}_{n+1} \in g^{-1}\left(e_{n+1}+\hat{e}_{n+1}\right)
$$

Similar considerations will now apply to $f^{-1}$. We can have

$$
f\left(c_{n+1}\right)=\partial d_{n+1}, f\left(\hat{c}_{n+1}\right)=\partial \hat{d}_{n+1}
$$

and the result will be $\left[c_{n+1}\right]+\left[\hat{c}_{n+1}\right]=\Delta\left[e_{n+1}\right]+\Delta\left[\hat{e}_{n+1}\right]$ so this is a homomorphism.
Theorem A.6.3 Suppose the situation of 1.9 described in Definition A.6.1. Here $\partial \partial=0$ and $\partial f=f \partial$, same with $g$. Let the homology groups be defined as before $H_{C_{n}} \equiv Z_{C_{n}} / B_{C_{n}}$ where $Z_{C_{n}} \equiv\{c \in C: \partial c=0\}$ and $B_{C_{n}} \equiv \partial c$ for some $c \in C_{n+1}$. Then we can consider $f, g$ acting on the homology groups in the natural way

$$
\begin{equation*}
f([c])=[f(c)], g([d])=[g(d)] \tag{1.12}
\end{equation*}
$$

and also there is a connecting homomorphism $\Delta$ such that

$$
\cdots \rightarrow H_{C_{n}} \xrightarrow{f} H_{D_{n}} \xrightarrow{g} H_{E_{n}} \xrightarrow{\Delta} H_{C_{n-1}} \xrightarrow{f} H_{D_{n-1}} \xrightarrow{g} \cdots
$$

is an exact sequence.
Before proving this, consider that just because $g$ is onto $E_{n}$ does not mean that $g$ is onto $H_{E_{n}}$. This is because the things in $H_{D_{n}}$ are equivalence classes of cycles.

Proof: The definitions in 1.12 are clearly true and this was shown earlier in Lemma A.2.3. Recall

$$
\begin{array}{lllllll}
0 \rightarrow & \vdots & \dot{C}_{n+1} & \xrightarrow{f} & \dot{D}_{n+1} & \xrightarrow{g} & \vdots \\
& \downarrow \partial & & \downarrow \partial & & \downarrow \partial & \\
& & & \\
0 \rightarrow & C_{n} & \xrightarrow{f} & D_{n} & \xrightarrow{g} & E_{n} & \rightarrow 0
\end{array}
$$

Is $f\left(H_{C_{n}}\right)=\operatorname{ker}(g)$ ? Letting $c$ be a cycle, consider $f(c)$. By definition of exactness, $g(f(c))=0$. Therefore, $[0]=[g(f(c))]=g[f(c)]$ showing that $f\left(H_{C_{n}}\right) \subseteq \operatorname{ker}(g)$.

For the other inclusion let $[d] \in \operatorname{ker}(g)$. In particular this assumes $d$ is a cycle. I need to show $[d]$ is in $f\left(H_{C_{n}}\right)$. We have $[0]=g([d])=[g(d)]$. It follows that $g(d)=\partial e$ for some $e \in E_{n+1}$. Since $g$ is onto, $e=g(\hat{d})$ for some $\hat{d} \in D_{n+1}$ and so $g(d)=\partial e=\partial g(\hat{d})=g(\partial \hat{d})$ so $g(d-\partial \hat{d})=0$. Therefore, $d-\partial \hat{d} \in \operatorname{ker}(g)$ and so $d-\partial \hat{d}=f(c)$ for some $c \in C_{n}$. Thus $0=f(\partial c)$ and since $f$ is one to one, it follows $\partial c=0$. Thus $[d]=f([c]) \in f\left(H_{C_{n}}\right)$. We just showed that, as mappings on homology groups, $f\left(H_{C_{n}}\right)=\operatorname{ker}(g)$.

So far we have this in terms of homology groups:

$$
H_{C_{n}} \xrightarrow{f} H_{D_{n}} \xrightarrow{g} H_{E_{n}}, \operatorname{Im}(f)=\operatorname{ker}(g)
$$

I want to get this:

$$
H_{C_{n+1}} \xrightarrow{f} H_{D_{n+1}} \xrightarrow{g} H_{E_{n+1}} \xrightarrow{\Delta} H_{C_{n}} \xrightarrow{f} H_{D_{n}} \xrightarrow{g} H_{E_{n}} \ldots
$$

where $\Delta$ is from Lemma A.6.2.
$\{\operatorname{ker} \Delta=\operatorname{Im} g\}$ Let $\left[e_{n+1}\right] \in \operatorname{ker} \Delta$. I need to show $\left[e_{n+1}\right]$ is in the image of $g$. Since $\left[e_{n+1}\right] \in \operatorname{ker} \Delta,\left[f^{-1} \partial g^{-1} e_{n+1}\right]=0$ so $f^{-1} \partial g^{-1} e_{n+1}=\partial c_{n+1}$. Thus $\partial g^{-1} e_{n+1}=f\left(\partial c_{n+1}\right)$. Letting $g\left(d_{n+1}\right)=e_{n+1}$ using the fact that $g$ is onto, we get $\partial d_{n+1}=f\left(\partial c_{n+1}\right)$. Also $\partial g\left(d_{n+1}\right)=g\left(f\left(\partial c_{n+1}\right)\right)$ so

$$
g\left(\partial d_{n+1}-f\left(\partial c_{n+1}\right)\right)=0
$$

which implies there exists $x$ for which

$$
f(x)=\partial d_{n+1}-f\left(\partial c_{n+1}\right)=\partial\left(d_{n+1}-f\left(c_{n+1}\right)\right)
$$

Therefore, $x=f^{-1} \partial\left(d_{n+1}-f\left(c_{n+1}\right)\right)=\partial f^{-1}\left(d_{n+1}-f\left(c_{n+1}\right)\right)=\partial y$ and so from the above,

$$
\begin{gathered}
\partial f(y)=\partial\left(d_{n+1}-f\left(c_{n+1}\right)\right) \\
\text { so } 0=\partial\left(d_{n+1}-\binom{\in \operatorname{ker} g}{\left.f(y)+f\left(c_{n+1}\right)\right)} . \text { But then } d_{n+1}-\left(f(y)+f\left(c_{n+1}\right)\right)\right. \text { is a cycle and } \\
g\left(d_{n+1}-\left(f(y)+f\left(c_{n+1}\right)\right)\right)=g\left(d_{n+1}\right)=e_{n+1}
\end{gathered}
$$

Therefore, $\left[e_{n+1}\right]$ is indeed in the image of $g$ as was to be shown. Thus ker $\Delta \subseteq \operatorname{Im} g$.
Now consider $\left[g\left(d_{n+1}\right)\right]$ for $d_{n+1}$ a cycle. Is this in $\operatorname{ker} \Delta$ ? From the definition of $\Delta$,

$$
\left[f^{-1} \partial g^{-1} g\left(d_{n+1}\right)\right]=\left[f^{-1} \partial \bar{d}_{n+1}^{=0}\right]=0
$$

since $f$ is one to one. Thus $\operatorname{Im}(g)=\operatorname{ker} \Delta$.
Next I need to verify exactness at $H_{C_{n}}$.
$\{\operatorname{Im} \Delta=\operatorname{ker} f\}$ First let $e_{n+1}$ be a cycle and consider $\Delta\left[e_{n+1}\right] \equiv\left[f^{-1} \partial g^{-1} e_{n+1}\right]$. Is it in $\operatorname{ker} f ?$ Is it the case that $f\left(f^{-1} \partial g^{-1} e_{n+1}\right)$ is a boundary? This expression is just $\partial g^{-1} e_{n+1}$ so this is clearly true. Thus $\operatorname{Im} \Delta \subseteq \operatorname{ker} f$.

Next suppose $f\left[c_{n}\right]=\left[f c_{n}\right]=0$ so $\left[c_{n}\right] \in \operatorname{ker} f$ for $c_{n}$ a cycle. I need to show $c_{n} \in \operatorname{Im} \Delta$. Since $\left[f\left(c_{n}\right)\right]=0$, it follows that $f\left(c_{n}\right)=\partial d_{n+1}$. By exactness, $\partial d_{n+1} \in \operatorname{ker}(g)$ because
$\partial d_{n+1}$. is in $\operatorname{Im} f$. Let $e_{n+1} \equiv g\left(d_{n+1}\right)$. Then $\partial e_{n+1}=g\left(\partial d_{n+1}\right)=0$. Thus this $e_{n+1}$ is a cycle. Then $f^{-1} \partial g^{-1} e_{n+1}=f^{-1} \partial g^{-1} g\left(d_{n+1}\right)=f^{-1} \partial d_{n+1}=c_{n}$. Therefore, $\Delta\left[e_{n+1}\right] \equiv$ $\left[f^{-1} \partial g^{-1} e_{n+1}\right]=\left[c_{n}\right]$. Therefore, $\operatorname{ker}(f) \subseteq \operatorname{Im} \Delta$. This completes the proof.

I think it may be useful to have a description of this connecting homomorphism in terms of a sequence of steps.

$$
\begin{align*}
& e_{n+1}=g\left(d_{n+1}\right) \text { since } g \text { is onto } E_{n+1} \\
& 0=g\left(\partial d_{n+1}\right) \text { since } e_{n+1} \text { is cycle } \\
& f\left(c_{n}\right)=\partial d_{n+1} \text { by } \operatorname{Im}(f)=\operatorname{ker}(g)  \tag{1.13}\\
& f\left(\partial c_{n}\right)=\partial \partial d_{n+1}=0 \Rightarrow \partial c_{n}=0 \\
& \Delta\left(\left[e_{n+1}\right]\right) \equiv\left[c_{n}\right] \in H_{C_{n}}
\end{align*}
$$

Definition A.6.4 One of the columns in 1.9 is called a chain complex. The homomorphisms $f, g$ are called chain maps. Recall that the diagram commutes so that $f(\partial c)=\partial f(c)$. In the above theorem about the connecting homomorphism, what happens to the bottom row? It is of the form

$$
0 \rightarrow C_{0} \quad \xrightarrow{f} \quad D_{0} \quad \xrightarrow{g} \quad E_{0} \quad \rightarrow 0
$$

and $\partial$ will map these Abelian groups to 0 . Thus the connecting homomorphism $\Delta$ will just be the zero map so everything just disappears after this. To save on notation, we write 1.9 as

$$
0 \rightarrow C \quad \xrightarrow{f} \quad D \quad \xrightarrow{g} \quad E \quad \rightarrow 0
$$

where $C, D, E$ symbolize the entire column. Thus, in words, a short exact sequence of chain complexes yields a long exact sequence of homology groups. This is completely algebraic. $\partial$ is just a mapping from the $n^{\text {th }}$ to the $(n-1)^{\text {st }}$ level in one of these chain complexes such that $\partial^{2}=0$ which enables the definition of the homology groups. Now suppose you have two of these short exact sequences.

$$
\begin{array}{lcccccc}
0 \rightarrow & C & \xrightarrow{f} & D & \xrightarrow{g} & E & \rightarrow 0  \tag{1.14}\\
& \downarrow \alpha & & \downarrow \beta & & \downarrow \gamma & \\
0 \rightarrow & C^{\prime} & \xrightarrow{f^{\prime}} & D^{\prime} & \xrightarrow{g^{\prime}} & E^{\prime} & \rightarrow 0
\end{array}
$$

where the mappings $\alpha, \beta, \gamma$ are homomorphisms which act between the groups $C_{n}, D_{n}, E_{n}$ and $C_{n}^{\prime}, D_{n}^{\prime}, E_{n}^{\prime}$ in such a way that the squares in the above diagram commute. That is, for $c \in C_{n}, \beta f(c)=f^{\prime} \alpha(c)$ with a similar relation holding for the next square involving $D_{n}, E_{n}, D_{n}^{\prime}, E_{n}^{\prime}$. Such mappings $\alpha, \beta, \gamma$ are called "chain homomorphisms". They are said to have degree 0 because they act on the same level.

Also we insist that for

$$
\begin{align*}
& c \in C_{n}, \partial^{\prime}(\alpha(c))=\alpha(\partial c), d \in D_{n}, \partial^{\prime}(\beta(d))=\beta(\partial(d)),  \tag{1.15}\\
& e \in E_{n}, \partial^{\prime}(\gamma(e))=\beta(\partial(e))
\end{align*}
$$

Thus we can consider the following in terms of homology groups. For c a cycle in $C$, we can say the following is well defined.

$$
[\alpha c]^{\prime}=\alpha[c],[\beta d]^{\prime}=\beta[d],[\gamma e]^{\prime}=\gamma[e]
$$

This is because of 1.15. Cycles go to cycles and boundaries go to boundaries. Thus this definition yields a homomorphism of homology groups also.

Now consider the corresponding sequence of homology groups. Then all of the rectangles commute.

$$
\begin{array}{lllllllllll}
\cdots & H_{C_{n}} & \xrightarrow{f} & H_{D_{n}} & \xrightarrow{g} & H_{E_{n}} & \xrightarrow{\Delta} & H_{C_{n-1}} & \xrightarrow{f} & H_{D_{n-1}} & \xrightarrow{g} \cdots \\
& \downarrow \alpha & & \downarrow \beta & & \downarrow \gamma & & \downarrow \alpha & & \downarrow \beta &  \tag{1.16}\\
\cdots & H_{C_{n}^{\prime}} & \xrightarrow{f^{\prime}} & H_{D_{n}^{\prime}} & \xrightarrow{g^{\prime}} & H_{E_{n}^{\prime}} & \xrightarrow{\Delta^{\prime}} & H_{C_{n-1}^{\prime}} & \xrightarrow{f^{\prime}} & H_{D_{n-1}^{\prime}} & \xrightarrow{g^{\prime}} \cdots
\end{array}
$$

This is the content of the next proposition.
Proposition A.6.5 Let $\alpha, \beta, \gamma$ be chain homomorphisms of the short exact sequences of 1.14 in which the squares commute, and let them also denote the induced homomorphisms of homology groups of 1.16. Also let $f, f^{\prime}, g, g^{\prime}$, denote the induced homomorphisms on homology groups in each level for the chain complexes and let $\Delta, \Delta^{\prime}$ be the connecting homomorphisms of Theorem A.6.3. Then the squares in 1.16 all commute.

Proof: Start with $[c] \in H_{C_{n+1}}$. Then as noted above, $f^{\prime}(\alpha[c])=f^{\prime}\left([\alpha c]^{\prime}\right)=\left[f^{\prime}(\alpha(c))\right]^{\prime}$ and $\beta(f([c]))=\beta([f(c)])=[\beta(f(c))]^{\prime}$. However, by assumption that these are chain homomorphisms, $\beta(f(c))=f^{\prime}(\alpha(c))$ and it works the same for $g, \beta, \gamma$. It remains to consider the connecting homomorphisms.

Let $e_{n}$ be a cycle.

$$
\begin{aligned}
\alpha \Delta\left[e_{n}\right] & =\alpha\left[f^{-1} \partial g^{-1} e_{n}\right]=\left[\alpha f^{-1} \partial g^{-1} e_{n}\right]^{\prime}=\left[f^{\prime-1} \beta \partial g^{-1} e_{n}\right]^{\prime} \\
& =\left[f^{\prime-1} \partial^{\prime} \beta g^{-1} e_{n}\right]^{\prime}=\left[f^{\prime-1} \partial^{\prime} g^{\prime-1} \gamma e_{n}\right]^{\prime} \equiv \Delta^{\prime}\left[\gamma e_{n}\right]^{\prime}=\Delta^{\prime} \gamma\left[e_{n}\right]
\end{aligned}
$$

Why is $\beta g^{-1}(e) \subseteq g^{-1} \gamma(e)$ ? Let $\beta d \in \beta g^{-1}(e)$ so $g(d)=e$. Then $g^{\prime} \beta d=\gamma g d$ and so $\beta d \in g^{\prime-1}(\gamma g d)=g^{\prime-1} \gamma e$ so $\beta g^{-1}(e) \subseteq g^{\prime-1} \gamma(e)$. This was what was used above. As to the interchange of $\alpha$ and $f^{-1}$ also used, if $\alpha c=\alpha f^{-1} d$, then $f^{\prime}(\alpha c)=f^{\prime}\left(\alpha f^{-1} d\right)=$ $\beta f f^{-1} d=\beta d$ and so $\alpha f^{-1} d=f^{\prime-1}(\beta d)$.

Note that if $f=f^{\prime}, g=g^{\prime}, \partial=\partial^{\prime}$, then from Lemma A.6.2 which gives a description of $\Delta$, we would have also that $\Delta^{\prime}=\Delta$.

The term used is that the connecting homomorphism is "natural". There is also notation which is used to describe the name of the maps when acting on homology groups. Recall the following notation which is to use a subscript of $*$ to denote mappings on homology groups.

Definition A.6.6 In the above situation where $f, g$ are chain maps we write $f_{*}$ and $g_{*}$ to indicate the corresponding map acting on homology groups.

## A. 7 Computing Homology Groups

This will be about finding homology groups. It is surprisingly hard, but leads to interesting theorems. Assume here that $X \subseteq \operatorname{int}(U) \cup \operatorname{int}(V)$ and consider the following:

$$
0 \rightarrow S_{n}(U \cap V) \xrightarrow{f_{\#}} S_{n}(U) \oplus S_{n}(V) \xrightarrow{g_{\#}} S_{n}^{U, V}(X) \rightarrow 0
$$

where $f_{\#}, g_{\#}$ are homomorphisms defined as $f_{\#}(c) \equiv(c,-c), g_{\#}(c, d) \equiv c+d$. Recall $S_{n}^{U, V}(X)$ is the free Abelian group of combinations of singular simplices which are supported either on $U$ or $V$. In the applications here, $U, V$ will be open. Addition is defined in
the usual way. $(c, d)+(\hat{c}, \hat{d})=c+\hat{c}+d+\hat{d}$. Then $f_{\#}$ is clearly one to one and $g_{\#}$ is onto. Also, if $g_{\#}(c, d)=0$ then $c+d=0$ and so $d=-c$ so $(c, d)=(c,-c) \in \operatorname{Im}\left(f_{\#}\right)$. Thus this is a short exact sequence. We also assume $f_{\#}, g_{\#}$ are chain maps so $\partial f_{\#}=f_{\#}(\partial \oplus \partial)$ and $\partial g_{\#}=g_{\#}(\partial \oplus \partial)$ where $(\partial \oplus \partial)$ does the obvious thing $(\partial \oplus \partial)(c, d)=(\partial c, \partial d)$. Thus this yields a short exact sequence of chain complexes. It follows from Theorem A.6.3 that there exists a long exact sequence of homology groups.

$$
\cdots \rightarrow H_{n}(U \cap V) \xrightarrow{f_{*}} H_{n}(U) \oplus H_{n}(V) \xrightarrow{g_{*}} H_{n}\left(S_{n}^{U, V}(X)\right) \xrightarrow{\Delta} H_{n-1}(U \cap V) \xrightarrow{f_{*}} \cdots
$$

This is called the Mayer Vietoris sequence.
Also notice that if $h: X \rightarrow \hat{X}$ is continuous with $h(U) \subseteq \hat{U}, h(V) \subseteq \hat{V}$ and $\hat{X}=\operatorname{int}(\hat{U}) \cup$ $\operatorname{int}(\hat{V})$ then the squares in the following diagram must commute. This is a consequence of Proposition A.6.5 and the fact that the corresponding squares in the short exact sequences of chains involving $h_{\#}$ commute. Note how $f, g$ make perfect sense independent of, $X, U, V, \hat{U}, \hat{V}, \hat{X}$ or on $h$.

$$
\begin{array}{ccccccccc}
\rightarrow & H_{n}(U \cap V) & \xrightarrow{f_{*}} & H_{n}(U) \oplus H_{n}(V) & \xrightarrow{g_{*}} & H_{n}\left(S_{n}^{U, V}(X)\right) & \xrightarrow{\Delta} & H_{n-1}(U \cap V) & \xrightarrow{f_{*}} \\
& \downarrow h_{*} & & \downarrow h_{*} \oplus h_{*} & & & \downarrow h_{*} & & \downarrow h_{*} \\
\rightarrow & H_{n}(\hat{U} \cap \hat{V}) & \xrightarrow{f_{*}} & H_{n}(\hat{U}) \oplus H_{n}(\hat{V}) & \xrightarrow{g_{*}} & H_{n}\left(S_{n}^{\hat{U}, \hat{V}}(\hat{X})\right) & \xrightarrow{\Delta} & H_{n-1}(\hat{U} \cap \hat{V}) & \xrightarrow{f_{*}}
\end{array}
$$

Lemma A.7.1 For $U, V$ open sets containing $X$ and for $h: X \rightarrow \hat{X}$ satisfying $h(U) \subseteq$ $\hat{U}, h(V) \subseteq \hat{V}$ where $\hat{X}=\operatorname{int}(\hat{U}) \cup \operatorname{int}(\hat{V})$ then the above diagram is valid in which the rectangles commute.

It is time for examples at long last. We do have a couple of good ones already. Recall that $H_{0}(X)=\mathbb{Z}$ in case $X$ is path connected. This is from Theorem A.2.5. Also recall that from Proposition A.2.8 $H_{n}(X)$ is the direct sum of homology groups of the path components of $X$. I will refer to $H_{n}\left(S_{n}^{U, V}(X)\right)$ as $H_{n}(X)$ from now on because that material on subdivisions says that if $c$ is a cycle, we can obtain that it is homologous to one in which all the simplices are supported in one of $U$ or $V$.

## A. 8 The Homology Groups of Spheres

This is done using the Mayer Vietoris sequence and induction which reduces to $S^{1}$.
Example A.8.1 $S^{1}$ is the unit circle $x^{2}+y^{2}=1$. Letting this be $X$, what are its homology groups?


It is certainly path connected so $H_{0}\left(S^{1}\right)=\mathbb{Z}$ but what of $H_{1}\left(S^{1}\right)$ ? Let $U$ be all of $S^{1}$ other than the bottom point $(0,-1)$ and let $V$ be all of $S^{1}$ other than the top point $(0,1)$. $H_{n}(U) \oplus H_{n}(V)=(0,0)$ because $U, V$ are both homeomorphic to $(-1,1)$ a convex set
whenever $n \geq 1$. Letting $n=1, \operatorname{Im}\left(g_{*}\right)=0$ which is the kernel of $\Delta$ and so $\Delta$ is one to one. Note that $H_{0}(U \cap V) \approx \mathbb{Z} \oplus \mathbb{Z}$ because $U \cap V$ consists of two path components.

$$
H_{1}(U) \oplus H_{1}(V) \quad \stackrel{0}{\rightarrow} \quad H_{1}\left(S^{1}\right) \quad \stackrel{\Delta, 1-1}{\rightarrow} \quad \underset{0}{\approx} H_{0}(U \cap \mathbb{Z} \oplus \mathbb{Z}) \quad \xrightarrow{f_{*}} \quad \underset{H_{0}(U)}{\approx \mathbb{Z}} \underset{H_{0}(V)}{\approx \mathbb{Z}}
$$

Then since the sequence is exact, $\Delta\left(H_{1}\left(S^{1}\right)\right)=\operatorname{ker}\left(f_{*}\right)$. So what is $\operatorname{ker}\left(f_{*}\right)$ ? It must be isomorphic to $m(1,-1), m \in \mathbb{Z}$ which is isomorphic to $\mathbb{Z}$.

More precisely, if $U \cap V$ is $L \cup R$ where $L, R$ are the left and right sides of $U \cap V$ in the above picture, the path components of $U \cap V . H_{0}(U \cap V)$ would be of the form $(m[c], n[d])$ because there are two path components. Here $[c],[d]$ are in $H_{0}(L), H_{0}(R)$ respectively $c, d$ being cycles. Say

$$
c=\sum_{\phi} m_{\phi} \phi, d=\sum_{\psi} n_{\psi} \psi
$$

where the $\phi$ have values in $L$ and the $\psi$ have values in $R$. Now all of the $\phi$ are homologous to each other in $L$ because $\phi-\hat{\phi}$ is indeed a boundary, so [c] can be reduced to $m[\phi]$ and similarly $[d]$ is of the form $n[\psi]$ for $\phi, \psi$ simplices. Thus we can assume $c, d$ are 0 simplices. Both $c$ and $d$ are supported in $U$ and both are supported in $V$. To be in $\operatorname{ker}\left(f_{*}\right)$ we would need $n[d]+m[c]=0$ in $H_{0}(U)$ and also in $H_{0}(V)$. This would mean that $[n d+m c]=0$ in $H_{0}(U)$ so $n d+m c$ would be a boundary in $U$. Of course this happens exactly when $n=-m$ so that you can pair the two to obtain their difference as a boundary, and so $\operatorname{ker}\left(f_{*}\right)$ is of the form $m([c]-[d])$, with $c, d$ being 0 simplices in $U \cap V, c$ in $L$ and $d$ in $R$. Note that $c-d$ is indeed a boundary in $U$ and also in $V$ but this is not a boundary in $U \cap V$. Thus $[c]-[d]$ is nonzero in $H_{0}(U \cap V)$ and

$$
\operatorname{ker}\left(f_{*}\right)=\{m([c]-[d]): m \in \mathbb{Z}\} \approx H_{1}\left(S^{1}\right) \approx \mathbb{Z}
$$

If $\left[c_{1}\right]$ generates $H_{1}\left(S^{1}\right)$, this means that $\Delta\left(\left[c_{1}\right]\right)=[c]-[d]$ where $c, d$ are two 0 cycles. This seems to be a pretty useful observation.
Lemma A.8. 2 Let $U, V$ be the open sets given above. Then for any 0 cycle $c$ in $L$ and $d$ in $R$, it follows that $H_{1}\left(S^{1}\right) \approx \mathbb{Z}$ and a generator for $H_{1}\left(S^{1}\right)$ is $\Delta^{-1}([c]-[d])$, this by Lemma A.0.2.

So what about $H_{n}\left(S^{1}\right)$ for $n>1$ ? Consider $n=2$.

$$
H_{2}(U) \oplus H_{2}(V) \xrightarrow{0} \quad H_{2}\left(S^{1}\right) \xrightarrow{g_{*}} \xrightarrow{\Delta, 1-1} H_{1}(U \cap V) \quad \xrightarrow{f_{*}} \quad H_{1}(U) \stackrel{0 \oplus 0}{\oplus} H_{1}(V)
$$

As just noted, $H_{2}(U) \oplus H_{2}(V)=(0,0)$ so $\Delta$ is one to one since ker $\Delta=0$. However, ker $f_{*}=$ $\operatorname{Im}(\Delta)=H_{1}(U \cap V)$ because $f_{*}$ maps to $0 \oplus 0=0$. Now $H_{1}(U \cap V)=0$. It is the direct sum of homology groups of the two path components of $U \cap V$ each of which is 0 . Therefore, $\Delta\left(H_{2}\left(S^{1}\right)\right)$ and consequently $H_{2}\left(S^{1}\right)$ is 0 . Similarly the other $H_{n}\left(S^{1}\right)=0$ for $n \geq 2$. Just replace 2 with $n$ and repeat.

Now it is time to find the homology groups of spheres in any dimension. The case of $S^{1}$ was just done. In particular, $H_{1}\left(S^{1}\right)$ is of the form $m([c]-[d])$ for $c, d 0$ simplices. Therefore, it is assumed that $n>1$ in what follows.


Then $S^{n-1}$ is the intersection of $\mathbb{R}^{n}$ with $S^{n}$. Also $U$ will be all of $S^{n}$ except for the top point $t$ while $V$ will be all of $S^{n}$ except the bottom point $b$. The line illustrates how $\mathbb{R}^{n}$ is homeomorphic to $U$ and similarly homeomorphic to $V$. Thus, from the picture, $U \cap V$ is homeomorphic to $\mathbb{R}^{n} \backslash \mathbf{0}$. From Lemma A.3.11

$$
H_{m}(U \cap V) \approx H_{m}\left(\mathbb{R}^{n} \backslash \mathbf{0}\right) \approx H_{m}\left(S^{n-1}\right)
$$

Then from the Mayer Vietoris sequence above and letting $X=S^{n}=U \cup V$,

$$
\begin{gather*}
\begin{array}{l}
0=H_{m}(U) \quad \begin{array}{c}
0=H_{m}(V) \\
H_{m}\left(\mathbb{R}^{n}\right) \oplus
\end{array} H_{m}\left(\mathbb{R}^{n}\right) \xrightarrow{g_{*}} H_{m}\left(S^{n}\right) \xrightarrow{\Delta} \begin{array}{c}
H_{m-1}(U \cap V) \\
H_{m-1}\left(S^{n-1}\right)
\end{array} \\
\stackrel{f_{*}}{\rightarrow} H_{m-1}\left(\mathbb{R}^{n}\right) \oplus H_{m-1}\left(\mathbb{R}^{n}\right) \xrightarrow{g_{*}}
\end{array}
\end{gather*}
$$

Proposition A.8.3 For $m \geq 1, H_{m}\left(S^{m}\right)=\mathbb{Z}$ and if $m \neq n$, then $H_{m}\left(S^{n}\right)=0$. Also, for any $n \geq 1, H_{0}\left(S^{n}\right)=\mathbb{Z}$.

Proof: The last claim follows because $S^{n}$ is path connected. The first claim was shown above in case $n=1$. So suppose the claim is true for $n-1$. Consider $n$ and the case where $m=n$. Then from 1.17 , the left side is 0 because $\mathbb{R}^{n}$ is convex, so $\operatorname{Im}\left(g_{*}\right)=$ $0=\operatorname{ker}(\Delta)$ which shows that $\Delta$ is one to one. Also $H_{m-1}\left(\mathbb{R}^{n}\right) \oplus H_{m-1}\left(\mathbb{R}^{n}\right)=0$ and so $\operatorname{ker} f_{*}=\operatorname{Im}(\Delta)=H_{m-1}\left(S^{n-1}\right)$ which shows that $\Delta$ is an isomorphism. Hence by induction $H_{n}\left(S^{n}\right) \approx H_{n-1}\left(S^{n-1}\right) \approx \mathbb{Z}$.

Next consider the case that $m<n$. By induction, $H_{m-1}\left(S^{n-1}\right)=0$ but $\Delta$ in 1.17 is still one to one. Hence $H_{m}\left(S^{n}\right)=0$ since otherwise $\Delta$ would fail to be one to one.

Next consider the case that $m>n$. In this case $\operatorname{Im}\left(g_{*}\right)$ is still 0 and so $\Delta$ is still one to one. Again, by induction, we have $H_{m-1}\left(S^{n-1}\right)=0$ so again $H_{m}\left(S^{n}\right)=0$ since otherwise $\Delta$ would fail to be one to one. This proves the proposition.

## A. 9 Brouwer Fixed Points

Corollary A.9.1 $S^{n}$ and $S^{m}$ are not homeomorphic if $n \neq m$.
Proof: If they were homeomorphic, they would have the same homology groups and they don't.

Note that if $\mathbb{R}^{n}$ and $\mathbb{R}^{m}$ for $n \neq m$ were homeomorphic, then, their one point compactifications would be homeomorphic and hence, using steriographic projection $S^{n}$ and $S^{m}$ would also be homeomorphic which they are not. Steriographic projection involves adding a point at $\infty$ with the understanding that neighborhoods of this point are complements of compact sets so $\left\{\boldsymbol{x}_{n}\right\}$ converges to $\infty$ means that $\lim _{n \rightarrow \infty}\left|\boldsymbol{x}_{n}\right|=\infty$ in the usual manner from calculus. The picture illustrating the idea is as follows. The point at $\infty$ maps to the top point of the sphere.


Corollary A.9.2 If $m \neq n$, then $\mathbb{R}^{n}$ is not homeomorphic to $\mathbb{R}^{m}$.

Corollary A.9.3 Let $D_{n}$ be the closed ball of radius 1 centered at 0 in $\mathbb{R}^{n}, n \geq 1$. Then there does not exist a function $g: D_{n} \rightarrow S^{n-1}$ which is continuous and leaves all points of $S^{n-1}$ unchanged.

Proof: Suppose there were such a map. Then letting $i$ be the inclusion map of $S^{n-1}$ into $D_{n}$ it follows that $g \circ i=\mathrm{id}$ on $S^{n-1}$ and also $i \circ g=\mathrm{id}$ on $D^{n}$. Also for $t \in[0,1]$,

$$
\begin{aligned}
& t(\mathrm{id})(x)+(1-t)(i \circ g)(x) \quad D_{n} \\
& t(\mathrm{id})(x)+(1-t)(g \circ i)(x)=x \in S^{n-1}
\end{aligned}
$$

This is by convexity of $D_{n}$. Therefore, $g \circ i, i \circ g$ are both homotopy inverses on $S^{n-1}$ and $D_{n}$ respectively, and so by Theorem A.3.9, it follows that $H_{n-1}\left(D_{n}\right)$ and $H_{n-1}\left(S^{n-1}\right)$ are isomorphic, but this is certainly not the case because the first is 0 or $\mathbb{Z}$ depending on whether $n>1$ or $n=1$. The second is $\mathbb{Z}$ if $n>1$ and if $n=1$, it is $H_{0}\left(S^{0}\right)=\mathbb{Z} \oplus \mathbb{Z}$ because in the last case, $S^{0}$ has two path components.

With this, it is easy to prove the Brouwer fixed point theorem.
Corollary A.9.4 Let $D_{n}$ be the closed unit ball, $n \geq 1$ and let $\boldsymbol{h}: D_{n} \rightarrow D_{n}$ be continuous. Then $\boldsymbol{h}$ has a fixed point.

Proof: If $\boldsymbol{h}$ has no fixed point, consider the mapping $\boldsymbol{g}$ in the following picture which would deliver a retraction onto $S^{n-1}$ the boundary of $D_{n}$.


Definition A.9.5 If a set $A$ has the property that whenever $f: A \rightarrow A$ is continuous, there is a fixed point, then we say that $A$ has the fixed point property.

Note that if two sets are homeomorphic and one has the fixed point property, then so does the other. Letting $f: A \rightarrow \hat{A}$ be a homeomorphism with $A$ having the fixed point property and letting $g: \hat{A} \rightarrow \hat{A}$ be continuous, then $f^{-1} \circ g \circ f: A \rightarrow A$ and is continuous so it has a fixed point $x$. Then $g(f(x))=f(x)$ and so $g$ also has a fixed point.

Corollary A.9.6 If $C$ is any compact convex subset of $\mathbb{R}^{n}$ for $n \geq 1$, and if $f: C \rightarrow C$ is continuous, then $f$ has a fixed point.

Proof: Let $P$ be the continuous projection map onto $C$. Then take $B$ a large ball which contains $C$. Consider $f \circ P: B \rightarrow B$. It has a fixed point $x$ and so $f(P(x))=x$. Since $f$ maps to $C$, it follows that $x \in C$ and so $P(x)=x$. Hence $f(x)=x$.

An examination of the argument used shows the following.
Corollary A.9.7 Suppose $K$ is a continuous retraction of $C$ where $C$ has the fixed point property. Then so does $K$.

## A. 10 Topological Degree on Spheres

Degree theory, as presented here is all based on Proposition A.6.5 and Lemma A.7.1. It might be a good idea to have a quick review of this.

Definition A.10.1 Suppose for $n \geq 1$ we have $f: S^{n} \rightarrow S^{n}$ a continuous function. We know that $H_{n}\left(S^{n}\right)=\mathbb{Z}$ and so there is a generator of $H_{n}\left(S^{n}\right)$ called $[c]$. Thus $f_{*}([c]) \in H_{*}\left(S^{n}\right)$ and so there is an integer $d$ such that $f_{*}([c])=d[c]$. This $d$ is called the degree of $f$, denoted as $d(f)$. Obviously $d(\mathrm{id})=1$.

It is a good idea to find the degree of some other mappings besides the identity.
Lemma A.10.2 Let $f: S^{n} \rightarrow S^{n}$ be continuous, $n \geq 1$ and defined as

$$
f\left(x_{0}, \ldots, x_{n}\right)=\left(-x_{0}, x_{2}, \ldots, x_{n}\right)
$$

Then $d(f)=-1$.
Proof: First let $n=1$ and let $U$ be all of $S^{1}$ except the bottom point and $V$ is all of $S^{1}$ except the top point. Thus $U \cap V$ has two components and $U \cup V=S^{1}$. Then we have the following Mayer Vietoris sequence in which $f_{*}$ is an isomorphism because clearly $f$ is a homeomorphism.

$$
\begin{aligned}
& H_{1}(U) \oplus H_{1}(V) \xrightarrow{g_{*}} H_{1}\left(S^{1}\right) \xrightarrow{\Delta} H_{0}(U \cap V) \rightarrow H_{0}(U) \oplus H_{0}(V)
\end{aligned}
$$

From the diagram, $\Delta$ is one to one. What is $\operatorname{ker}\left(h_{*}\right)$ ? If we know this, we will know the image of $\Delta$. As earlier when homology of spheres was presented, $\operatorname{ker}\left(h_{*}\right)=\operatorname{Im} \Delta$ will be $\{m([c]-[d]): m \in \mathbb{Z}\}$ where here $[c],[d]$ are 0 simplices in the left side of $U \cap V$ and the right side of $U \cap V$ respectively. Any pair will work. Recall that $c-d$ is a boundary in $U$. Thus $([c]-[d])$ generates $\operatorname{Im}(\Delta)$. Now

$$
\begin{aligned}
f_{\#}(c-d)(\boldsymbol{t}) & \equiv f\left(c_{1}(\boldsymbol{t}), c_{2}(\boldsymbol{t})\right)-f\left(d_{1}(\boldsymbol{t}), d_{2}(\boldsymbol{t})\right) \\
& =\left(-c_{1}(\boldsymbol{t}), c_{2}(\boldsymbol{t})\right)-\left(-d_{1}(\boldsymbol{t}), d_{2}(\boldsymbol{t})\right) \\
& =\left(d_{1}(\boldsymbol{t}),-d_{2}(\boldsymbol{t})\right)-\left(c_{1}(\boldsymbol{t}),-c_{2}(\boldsymbol{t})\right)
\end{aligned}
$$

These points on the left and right sides of $U \cap V$ were arbitrary, so $[d]=[\hat{d}]$ where $\hat{d}(\boldsymbol{t})=$ $\left(d_{1}(\boldsymbol{t}),-d_{2}(\boldsymbol{t})\right)$, a similar thing holding for $c$. Thus

$$
f_{*}([c]-[d])=[d]-[c]
$$

Now, since $\Delta$ is an isomorphism onto $\operatorname{Im}(\Delta)$, a generator for $H_{1}\left(S^{1}\right)$ is $\Delta^{-1}([c]-[d])$ and so

$$
\Delta f_{*} \Delta^{-1}([c]-[d])=f_{*} \Delta\left(\Delta^{-1}([c]-[d])\right)=[d]-[c]=\Delta\left(\Delta^{-1}([d]-[c])\right)
$$

and so $f_{*} \Delta^{-1}([c]-[d])=-\Delta^{-1}([c]-[d])$ showing that $d(f)=-1$.
Now assume this is true for $S^{n-1}$ and consider $S^{n}, n>1$. Let $U$ and $V$ be as described above. Thus $U, V$ are homeomorphic to $\mathbb{R}^{n}$ and $U \cap V$ is homeomorphic to $\mathbb{R}^{n} \backslash\{0\}$ and by

Lemma A.3.11 $H_{m}\left(S^{n-1}\right)$ is the isomorphic to $H_{m}\left(\mathbb{R}^{n} \backslash\{0\}\right)$ while $U \cup V$ is $S^{n}$. Thus the Mayer Vietoris sequence is

Then $\operatorname{Im}\left(g_{*}\right)=0$ and so $\Delta$ is one to one. However, $\operatorname{ker}\left(h_{*}\right)=\operatorname{Im}(\Delta)=H_{n-1}\left(S^{n-1}\right)$ because $h_{*}$ sends everything to 0 . Thus $\Delta$ is an isomorphism. Then a generator of $H_{n-1}\left(S^{n-1}\right)$ is just $\Delta([\hat{c}])$ where $[\hat{c}]$ is a generator of $H_{n}\left(S^{n}\right)$. Thus by induction and Proposition A.6.5, $\Delta f_{*}([\hat{c}])=f_{*}(\Delta[\hat{c}])=-\Delta([\hat{c}])$ and so $f_{*}[\hat{c}]=-[\hat{c}]$.

Now here is another important result about the degree. It will be helpful to consider the following picture.


Lemma A.10.3 Let $n \geq 1$ and $f: S^{n} \rightarrow S^{n}$ defined as

$$
f\left(x_{1}, \ldots x_{i}, \ldots, x_{j}, \ldots, x_{n}\right)=\left(x_{1}, \ldots x_{j}, \ldots, x_{i}, \ldots, x_{n}\right)
$$

Then $d(f)=-1$.
Proof: First consider the case when $n=1$. Let $U$ be everything except the point at the South West circle and let $V$ be everything except the point at the North East circle in the above picture. Let $C_{1}$ and $C_{2}$ be the connected components of $U \cap V$ as above. Thus $U \cup V=S^{1}$ and as before, we have the following Mayer Vietoris sequence.

$$
\begin{align*}
& \begin{array}{cc}
=H_{1}(\mathbb{R})=0 & =H_{1}(\mathbb{R})=0 \\
H_{1}(U) \oplus & =\mathbb{Z} \\
H_{1}(V) \xrightarrow{g_{*}}(V) \\
H_{1}\left(S^{1}\right) \xrightarrow{\Delta} H_{0}(U \cap \mathbb{Z} & \xrightarrow{h_{*}}=\mathbb{Z}=H_{0}(\mathbb{R}) \\
H_{0}(U) & =\mathbb{Z}=H_{0}(\mathbb{R}) \\
H_{0}(V)
\end{array} \tag{1.18}
\end{align*}
$$

$$
\begin{aligned}
& H_{1}(U) \oplus H_{1}(V) \xrightarrow{g_{*}} H_{1}\left(S^{1}\right) \xrightarrow{\Delta} H_{0}(U \cap V) \rightarrow \stackrel{H_{0}(U) \oplus H_{0}(V)}{ }
\end{aligned}
$$

As discussed earlier, $\operatorname{ker}\left(h_{*}\right)=\{m([c]-[d]): m \in \mathbb{Z}\}=\Delta\left(H_{1}\left(S^{1}\right)\right)$ where $c, d$ are 0 simplices which have values at the indicated points. It didn't really matter which points we picked in $C_{1}$ and $C_{2}$ since any two in $C_{1}$ and any two in $C_{2}$ will have difference a boundary so they will lead to homologous 0 simplices. The ones I picked in the picture are convenient because when the components of the two points are switched the two points $c, d$ switch position. Thus $f_{*}([c]-[d])=([d]-[c])=-([c]-[d])$. Also, since $\Delta$ is an isomorphism, a generator for $H_{1}\left(S^{1}\right)$ will be $\Delta^{-1}([c]-[d])$. Then

$$
\begin{aligned}
\Delta f_{*} \Delta^{-1}([c]-[d]) & =f_{*} \Delta \Delta^{-1}([c]-[d])=f_{*}([c]-[d]) \\
& =-([c]-[d])=\Delta\left(-\Delta^{-1}([c]-[d])\right)
\end{aligned}
$$

and so $f_{*} \Delta^{-1}([c]-[d])=-\Delta^{-1}([c]-[d])$ showing that in this case the degree is -1 . Now in general, let the two circles on North East and South West be the points on $S^{n}$ which are on the line $\mathbf{0}+t\left(\boldsymbol{e}_{i}+\boldsymbol{e}_{j}\right)$. Let $U$ be all but the point on the South West and $V$ be all of $S^{n}$ but the North East point. Then the Mayer Vietoris sequence is the following for $n>1$

$$
\begin{gathered}
=H_{n}\left(\mathbb{R}^{n}\right)=0=H_{n}\left(\mathbb{R}^{n}\right)=0 \\
\left.H_{n}(U) \oplus \stackrel{H_{n}}{H_{n}(V)} \xrightarrow{H_{n-1}(U \cap V)} H_{n}\left(S^{n}\right) \xrightarrow{\Delta} H_{n-1}\left(\mathbb{R}^{n-1}\right)=0 \quad H_{n-1}\left(S^{n-1}\right) \xrightarrow{R_{n}-1}\right)=0 \\
H_{n-1}(U) \oplus \quad H_{n-1}(V)
\end{gathered}
$$

Then $\operatorname{Im}\left(g_{*}\right)=0$ and so $\Delta$ is one to one. However, $\operatorname{ker}\left(h_{*}\right)=\operatorname{Im}(\Delta)=H_{n-1}\left(S^{n-1}\right)$ because $h_{*}$ sends everything to 0 . Thus $\Delta$ is an isomorphism. Then a generator of $H_{n-1}\left(S^{n-1}\right)$ is just $\Delta([c])$ where $[c]$ is a generator of $H_{n}\left(S^{n}\right)$. Thus by induction and Proposition A.6.5, $\Delta f_{*}([c])=f_{*}(\Delta[c])=-\Delta([c])$ and so $f_{*}[c]=-[c]$.

Lemma A.10.4 Let $f, g: S^{n} \rightarrow S^{n}$ for any $n \geq 1$. Then $d(f \circ g)=d(f) d(g)$.
Proof: This follows from the observation that $(f \circ g)_{*}=f_{*} g_{*}$ and so if $[c]$ is a generator of $H_{n}\left(S^{n}\right)$, then

$$
(f \circ g)_{*}[c]=f_{*} g_{*}[c]=f_{*} d(g)[c]=d(g) f_{*}[c]=d(g) d(f)[c]
$$

From these lemmas, we obtain the following theorem about the degree of -id .
Theorem A.10.5 Let $-\mathrm{id}: S^{n} \rightarrow S^{n}$ be the antipodal map which takes $x$ to $-x$. Then $d(-\mathrm{id})=(-1)^{n}$.

Proof: Let $f\left(x_{1}, \ldots, x_{n}\right) \equiv\left(-x_{1}, \ldots, x_{n}\right)$ and let $g_{j}$ be the map which switches the first component and the $j^{\text {th }}$ component. Then $-\mathrm{id}=g_{n} \circ f \circ g_{n} \circ f \cdots g_{3} \circ f \circ g_{3} \circ g_{2} \circ f \circ g_{2} \circ f$. For example,

$$
\begin{aligned}
\left(x_{1}, x_{2}, x_{3}\right) & \rightarrow\left(-x_{1}, x_{2}, x_{3}\right) \rightarrow\left(x_{2},-x_{1}, x_{3}\right) \rightarrow\left(-x_{2},-x_{1}, x_{3}\right) \\
& \rightarrow\left(-x_{1},-x_{2}, x_{3}\right) \rightarrow\left(x_{3},-x_{1},-x_{2}\right) \\
& \rightarrow\left(-x_{3},-x_{1},-x_{2}\right) \rightarrow\left(-x_{1},-x_{2},-x_{3}\right)
\end{aligned}
$$

By Lemma A.10.4 $d(-\mathrm{id})=(-1)^{n}$ since the switching maps occur an even number of times.

If you had $C$ homeomorphic to $S^{n}$, and a continuous function $f$ mapping $S^{n}$ to itself, then you could use the homeomorphism and the degree of $f$ to get a generalization of the winding number from complex analysis.

## A. 11 Functions Defined on a Subset of $S^{n}$

The approach to defining the degree from using the homology groups of spheres is in Hocking and Young [23] although they appear to be using a somewhat different approach to homology. I am trying to include the standard theorems which are usually obtained from degree theory presented from an analytical point of view. I have tried to avoid mistakes, but I have made many revisions of this material when I found mistakes. Sometimes I wonder whether I have caught them all.

An important conclusion about approximation is next.
We think of the Tietze extension theorem as a way to extend a real valued function keeping its values in a given interval which contains $f(C)$ for $C$ a closed set. However, using spheres we can consider $\bar{f}$ the extended function in terms of keeping the values of this extended function away from a particular point. In the following picture, suppose $f(C)$ is contained in lower part of $S^{n}$. The top point $(\overrightarrow{0}, 2)$ will be denoted as $q$ for simplicity. Consider the following picture.


In the picture $p$ is the center of a small ball in $S^{n}$ and $b$ is the point opposite to $p$. In the picture $V$ is a translate of an $n$ dimensional subspace of $\mathbb{R}^{n+1}$ which is perpendicular to the vector $\overrightarrow{b p}$. Also $L \equiv S^{n} \backslash A$ and $f$ is a continuous function $f: C \rightarrow L, C$ a closed proper subset of $\mathbb{R}^{n}$. $\theta$ is the map illustrated by $y \rightarrow x$. We let $\infty \equiv \theta^{-1}(p)$ and make $V$ into a metric space by the rule $d(y, \hat{y}) \equiv|\theta(y)-\theta(\hat{y})|$ so $\theta$ is a homeomorphism of $S^{n}$ and $V \cup\{\infty\}$. Now $\theta^{-1}(L)$ is a closed ball in $V$ denoted as $\overline{B(b, R)}$. Let $V=b+Q\left(\mathbb{R}^{n}\right) \equiv$ $\alpha\left(\mathbb{R}^{n}\right)$ where $Q$ is an orthogonal transformation preserving distances with determinant 1. Thus $\alpha^{-1} \circ \theta^{-1}(L)=\overline{B(0, R)}$. Then $\alpha^{-1} \circ \theta^{-1} \circ f: C \rightarrow \overline{B(0, R)} \subseteq[-R, R]^{n}$. Using the Tietze extension theorem on the components of $\alpha^{-1} \circ \theta^{-1} \circ f$ there is an extension $\tilde{F}: S^{n} \rightarrow[-R, R]^{n}$ which agrees with $\alpha^{-1} \circ \theta^{-1} \circ f$ on $C$. Now let $P$ be the projection onto $\overline{B(0, R)}$. Then $P \circ \tilde{F}$ agrees with $\alpha^{-1} \circ \theta^{-1} \circ f$ on $C$ but is defined on all of $S^{n}$ and maps everything into $\overline{B(0, R)}$. Now $\alpha \circ P \circ \tilde{F}$ maps all of $S^{n}$ into $L$ and agrees with $f$ on $C$. This yields the following proposition.

Proposition A.11.1 Let $A$ be the intersection of a ball in $\mathbb{R}^{n+1}$ of small radius which is intersected with $S^{n}$ having center $p \in A$. Suppose $f(C) \cap \bar{A}=\emptyset, f$ is continuous where $C$ is a closed proper subset of $S^{n}$. Then there is a continuous extension of $f$ called $\bar{f}$ such that $\bar{f}: S^{n} \rightarrow S^{n}$ which maps all of $S^{n}$ to $L \equiv S^{n} \backslash A$.

Next is the topological degree. For a set $S \neq \emptyset, \partial S$ will denote those points $x$ where every ball containing $x$ contains points of $S$ and points of $S^{C} \equiv \mathbb{R}^{n} \backslash S$. Note that for an open set $\Omega$ in a metric space, $\partial \Omega=\bar{\Omega} \backslash \Omega$.

Proposition A.11.2 Let $\bar{\Omega} \subseteq S^{n}$ be a proper subset of $S^{n}$ with $\Omega$ an open set and let $p \notin$ $f(\partial \Omega)$ where $f: \bar{\Omega} \rightarrow S^{n}$ is continuous. Then there exists an extension of $f$ to all of $S^{n}$ called $\hat{f}$ such that $\hat{f}^{-1}(\bar{A}) \subseteq \Omega$ for $A$ an open ball containing $p$.

Proof: Since $p \notin f(\partial \Omega)$, there is a small ball $A$ centered at $p$ such that $\bar{A} \cap f(\partial \Omega)=\emptyset$. Using Proposition A.11.1, there is $\bar{f}: S^{n} \rightarrow S^{n} \backslash A$ extending $f$ off of $\partial \Omega$ to all of $S^{n}$ such that $\bar{f}\left(S^{n}\right) \subseteq L \equiv S^{n} \backslash A$. Then define $\hat{f}: S^{n} \rightarrow S^{n}$ to equal $f$ on $\bar{\Omega}$ and $\bar{f}$ on $S^{n} \backslash \bar{\Omega}$.Thus $\hat{f}^{-1}(\bar{A}) \subseteq \Omega$.

From now on $\hat{f}$ will refer to such an extension of $f$. Now here is the definition of the degree.

Definition A.11.3 Let $\Omega$ be a nonempty open set in $S^{n}$ with $\bar{\Omega}$ a proper subset of $S^{n}$ and let $p \notin f(\partial \Omega)$. Then $d(f, \Omega, p)$ is the integer such that $\hat{f}_{*}[\hat{c}]=d(f, \Omega, p)[\hat{c}]$ where $[\hat{c}]$ is a generator of $H_{n}\left(S^{n}\right)$.

I need to show this degree is well defined, but first, assuming this, it is easy to see that $p \rightarrow d(f, \Omega, p)$ is locally constant. First note that $p \notin f(\partial \Omega)$ is equivalent to saying that there exists an open ball $A$ such that $\bar{A} \cap f(\partial \Omega)=\emptyset$.

Lemma A.11.4 There is an open ball containing $p$ such that $d(f, \Omega, \hat{p})$ is the same for all $\hat{p}$ in this ball.

Proof: By definition of $\hat{f}$, there is an open ball $A$ containing $p$ such that $\hat{f}^{-1}(\bar{A}) \subseteq \Omega$. Then if $\hat{p} \in A$, there is a smaller open ball $\hat{A}$ centered at $\hat{p}$ with $\hat{f}^{-1}(\hat{A}) \subseteq \Omega$ and so, since this is the same $\hat{f}$, the $d(f, \Omega, p)=d(f, \Omega, \hat{p})$ because $d(f, \Omega, p)$ is defined in terms of $\hat{f}_{*}$ and the generator of $H_{n}\left(S^{n}\right)$.

Say $[c]$ generates $H_{n}\left(S^{n}\right)$ and let $\alpha$ be the isomorphism of $H_{n}\left(S^{n}\right)$ and $\mathbb{Z}$. Then $[c]$ would be either $\alpha(1)$ or else $\alpha(-1)=[c]$ and so either $[c]$ or $-[c]$ will work. But then $d$ still would be the same because $\hat{f}_{*}(-[c])=-\hat{f}_{*}([c])=-d[c]=d(-[c])$. Thus the definition of the degree does not change if $[c]$ is switched to $-[c]$ the other generator of $H_{n}\left(S^{n}\right)$. The hard issue is whether two different extensions satisfying the above definition give the same degree.This is discussed in the following proposition and shows that $d(f, \Omega, p)$ is well defined.

The proper open balls $A$ on $S^{n}, n \geq 1$, are just intersections of open balls in $\mathbb{R}^{n+1}$ with $S^{n}$ and so they are connected and open sets. Each is homeomorphic to a convex subset of $\mathbb{R}^{n}$. Say $\alpha$ is the name of this homeomorphism so that $C=\alpha A$ where $C$ is convex. Then if $c$ is a cycle with values in $A$ we have $\alpha_{*}[c]_{A}=\left[\alpha_{\#} c\right]_{C}=0$ because $C$ is convex so cycles and boundaries are the same. Since $\alpha_{*}$ is an isomorphism, this implies $[c]_{A}=0$.
Proposition A.11.5 Let $\Omega, f, \hat{f}$ be as defined in Definition A.11.3 with respect to some open ball $A$ for which $\hat{f}^{-1}(\bar{A}) \subseteq \Omega$. If $\hat{f}, \tilde{f}$ are two extensions of $f$ from Definition A.11.3 and $[\hat{c}]$ generating $H_{n}\left(S^{n}\right), \hat{f}_{*}[\hat{c}]=\tilde{f}_{*}[\hat{c}]$ and so, for any $p \in A, d(f, \Omega, p)$ is well defined. If $f^{-1}(\bar{A})=\emptyset$ then for $\hat{f}$ such an extension, $\hat{f}_{*}[\hat{c}]=0$ and so for $p \in A, d(f, \Omega, p)=0$.

Proof: Letting $L, \theta$ be as in Proposition A.11.1, one can consider the following homotopy of $\hat{f}, \tilde{f}$

$$
\theta\left(t \theta^{-1} \hat{f}+(1-t) \theta^{-1} \tilde{f}\right), t \in[0,1]
$$

The two functions are homotopic and so by Theorem A.3.7 $\hat{f}_{*}[\hat{c}]=\tilde{f}[\hat{c}]$. In particular, the definition of the degree is well defined.

If $f^{-1}(\bar{A})=\emptyset$ then for each simplex $\phi$ in $\hat{c}, \hat{f}_{\#} \phi$ has values in $S^{n} \backslash A$ which is homeomorphic to a convex subset of $\mathbb{R}^{n}$ and so $\left[\hat{f}_{\#} \hat{c}\right]=\hat{f}_{*}[\hat{c}]=0$.

Another claim which is easy to get is the following which deals with homeomophisms of $S^{n}$.

Lemma A.11.6 From the definition, if $\alpha$ is a homeomorphism on $S^{n}$, then $d(f, \Omega, p)=$ $\pm d(\alpha f, \alpha \Omega, \alpha p)$.

Proof: Let $\hat{f}$ and $A$ go with the definition for $d(f, \Omega, p)$. Then, since $\alpha$ is a homeomorphism, $\alpha(A)$ contains $A^{\prime}$ a ball centered at $\alpha p$ and $\alpha \hat{f}$ will serve for the definition of $d(\alpha f, \alpha \Omega, \alpha p)$. Then for $[\hat{c}]$ a generator of $H_{n}\left(S^{n}\right), \alpha_{*}$ is an isomorphism of homology groups and so $\alpha_{*}[\hat{c}]$ is also a generator of $H_{n}\left(S^{n}\right)$.

$$
d(\alpha f, \alpha \Omega, \alpha p)[\hat{c}] \equiv(\alpha \hat{f})_{*}[\hat{c}]=\alpha_{*} \hat{f}_{*}[\hat{c}]=\alpha_{*} d(f, \Omega, p)[\hat{c}]=d(f, \Omega, p) \alpha_{*}[\hat{c}]
$$

But, since $\alpha_{*}$ is an isomorphism, $\alpha_{*}[\hat{c}]$ is also a generator of $H_{n}\left(S^{n}\right) \approx \mathbb{Z}$ so $\alpha_{*}[\hat{c}]$ is either $[\hat{c}]$ or $-[\hat{c}]$.

There is also a generalization of Proposition A. 11.5 which replaces $\tilde{f}$ with $\hat{g}$ where $g$ is sufficiently close to $f$ using the observation that all that mattered in the above argument was $\tilde{f}(\partial \Omega) \cap \bar{A}=\emptyset$.

Corollary A.11.7 In the context of the above proposition, let $\bar{\Omega}, f, \hat{f}$ be as defined in Definition A.11.3 with respect to some open ball $A, f(\partial \Omega) \cap \bar{A}=\emptyset$. Then if $p \in A$, and $\|f-g\|_{\partial \Omega}$ is small enough, $d(f, \Omega, p)=d(g, \Omega, p)$.

Proof: Letting $\|f-g\|_{\partial \Omega}$ be small enough, we can assume that both $\hat{g}^{-1}(\bar{A})$ and $\hat{f}^{-1}(\bar{A})$ are contained in $\Omega$. Indeed, we choose $\|f-g\|_{\partial \Omega}$ so small that $g(\partial \Omega) \cap \bar{A}=\emptyset$. Then extend $g$ to $\bar{g}$ off $\partial \Omega$ to lie in $S^{n} \backslash A$ and let $\hat{g}=\bar{g}$ off $\bar{\Omega}$ and $\hat{g}=g$ on $\Omega$ similar to $\hat{f}$. Then the argument is similar to the above with $\hat{g}$ in place of $\hat{f}$ and having $\hat{g}$ and $\hat{f}$ homotopic. $\hat{g}_{*}[\hat{c}]=\hat{f}_{*}[\hat{c}]$ and both these are usable in the definition of the degree.

Next I want to consider a variation on what is in the above proposition. First observe that if $\Omega$ is a nonempty open set in a metric space, and for $n \in \mathbb{N}$,

$$
\Omega_{n} \equiv\left\{x \in \Omega: \operatorname{dist}\left(x, \Omega^{C}\right)>\frac{1}{n}\right\},
$$

then $\cup_{n} \Omega_{n}=\Omega$ and for all $n, \bar{\Omega}_{n} \subseteq \Omega_{n+1}$. I will use this observation in the proof of the following.

Corollary A.11.8 Suppose in the situation of Proposition A.11.5, $\Omega=\cup_{i} \Omega_{i}$ where the $\Omega_{i}$ are disjoint open sets. Let $\hat{f}_{i}$ be the kind of extension described above where $\hat{f}_{i}^{-1}(\bar{A}) \subseteq \Omega_{i}$, and $\hat{f}_{i}=f$ on $\Omega_{i}$. Then for $[\hat{c}]$ generating $H_{n}\left(S_{n}\right), n \geq 1$, it follows $\hat{f}_{*}[\hat{c}]=\sum_{i} \hat{f}_{i *}[\hat{c}]$ and $d(f, \Omega, p)=\sum_{i} d\left(f_{i}, \Omega, p\right)$.

Proof: Let $[\hat{c}]$ generate $H_{n}\left(S^{n}\right)$. Then the support of $\hat{c}, K$ is compact as is $\hat{f}^{-1}(\bar{A})$. Thus $K \cap \hat{f}^{-1}(\bar{A})$ is assumed contained in the union of the $\Omega_{i}$ and so this compact set is contained in finitely many. The $i$ in what follows will refer to these finitely many. There exists open $\Omega_{i 0} \subseteq \bar{\Omega}_{i 0} \subseteq \Omega_{i}$ such that $\hat{f}_{i}^{-1}(\bar{A}) \subseteq \Omega_{i 0}$ since $\hat{f}_{i}^{-1}(\bar{A})$ is a compact subset of $\Omega_{i}$. Similarly we can assume $\hat{f}^{-1}(\bar{A}) \subseteq \cup_{i} \Omega_{i 0}$ because the compact set $\hat{f}^{-1}(\bar{A})$ is contained in finitely many of the $\Omega_{i}$. Also, we can let $\Omega_{i 0}=\emptyset$ if $\hat{f}^{-1}(\bar{A}) \cap \Omega_{i}=\emptyset$. Then we can assume by Theorem A.5.21 that all the simplices in $\hat{c}$ have support in some $\Omega_{i}$ or in $S^{n} \backslash \cup_{i} \bar{\Omega}_{i 0}$. If $c$ has all simplices in $S^{n} \backslash \cup_{i} \bar{\Omega}_{i 0}$ then of course $\left[\hat{f}_{i \# c}\right]=0$ because $H_{n}\left(S^{n} \backslash A\right)=0$ and similarly $\left[\hat{f}_{\# c} c\right]=0$ because $S^{n} \backslash A$ is homeomorphic to a convex set. Next suppose one of the cycles of $\hat{c}$ called $c$ has all simplices in $\Omega_{i}$. Then, from the construction, $\hat{f}_{j \#} c, j \neq i$ is a cycle with support in $S^{n} \backslash A$ since $\hat{f}_{j}^{-1}(\bar{A}) \subseteq \Omega_{j}$ which has empty intersection with $\Omega_{i}$, and so $\hat{f}_{j *}[c] \in H_{n}\left(S^{n} \backslash A\right)=0$. However, $\hat{f}_{i \#} c=\hat{f}_{\#} c$ and so $\hat{f}_{*}[c]=\hat{f}_{i *}[c]$ which shows $\hat{f}_{*}[\hat{c}]=\sum_{i} \hat{f}_{i *}[\hat{c}]$ and so $d(f, \Omega, p)=\sum_{i} d\left(f_{i}, \Omega, p\right)$.

Summarizing the above, is the following proposition. Always we assume $p \notin f(\partial \Omega)$.
Proposition A.11.9 The definition of the degree is well defined and also $p \rightarrow d(f, \Omega, p)$ is locally constant. If $p \notin f(\Omega)$, then $d(f, \Omega, p)=0$. Thus if $d(f, \Omega, p) \neq 0$, then $p \in f(\Omega)$. Also if $\Omega$ is an open subset of $S^{n}$ equal to the union of disjoint open sets $\Omega_{i}$, then for $f: \bar{\Omega} \rightarrow S^{n}$ continuous and $p \notin f(\partial \Omega), d(f, \Omega, p)=\sum_{i} d\left(f, \Omega_{i}, p\right)$. The sum will be finite.

Proof: That $d(f, \Omega, p)$ is well defined follows from Proposition A.11.5. Also from the proposition is the claim that if $p \notin f(\Omega)$ then $d(f, \Omega, p)=0$. That $p \rightarrow d(f, \Omega, p)$ is locally constant follows from the observation that we get the same degree for any $p \in A$ from this definition. For the other claim, Corollary A.11.8 says that if we consider the extension $\hat{f_{i}}$ which goes with $\Omega_{i}$ to define $d\left(f_{i}, \Omega_{i}, p\right)$ where $p \in A$ as described there, then

$$
d(f, \Omega, p)[\hat{c}] \equiv \hat{f}_{*}[\hat{c}]=\sum_{i} \hat{f}_{i *}[\hat{c}]=\sum_{i} d\left(f_{i}, \Omega_{i}, p\right)[\hat{c}]
$$

Proposition A.11.5 and Corollary A.11.5 also imply the following corollary because it says that if $\|f-g\|_{\bar{\Omega}}$ is small enough, then $d(f, \Omega, p)=d(g, \Omega, p)$.
Corollary A.11.10 In the situation of Proposition A.11.9, given $p \in S^{n} \backslash f(\partial \Omega)$,

$$
d(f, \Omega, p)=d(g, \Omega, q)
$$

whenever $\|f-g\|_{\bar{\Omega}}+|p-q|$ is sufficiently small.
A similar result is the following:
Corollary A.11.11 If $\Omega \supseteq \hat{\Omega}$ where $\Omega, \hat{\Omega}$ are open and if $p \notin f(\bar{\Omega} \backslash \hat{\Omega})$, then

$$
d(f, \hat{\Omega}, p)=d(f, \Omega, p)
$$

Proof: This follows from using the closed set $\bar{\Omega} \backslash \hat{\Omega}$ rather than $\partial \Omega$ in Definition A.11.3 to get an extension which works for both $\Omega$ and $\hat{\Omega}$.

From Corollary A.11.10 there is a convenient result about homotopy.
Lemma A.11.12 Let continuous $h:[0,1] \times \Omega \rightarrow S^{n}$ and let $t \rightarrow p(t)$ also be continuous with $p(t) \notin h(t, \partial \Omega)$ for each $t \in[0,1]$. Then $t \rightarrow d(h(t, \cdot), \Omega, p(t))$ is constant. Also $p \rightarrow d(f, \Omega, p)$ is constant on each component of $f(\partial \Omega)^{C}$.

Proof: If $d(h(t, \cdot), \Omega, p(t))$ were not constant, then by Corollary A.11.10 we could separate $[0,1]$ by considering $t$ associated to different integer values achieved by the degree, which is not possible. For the last claim, letting $p, q$ be in a connected component $U$ of $(f(\partial \Omega))^{C}$, there exists $r:[0,1] \rightarrow U$ such that $r(0)=p, r(1)=q$. Then from the first part, $d(f, \Omega, r(t))$ is constant for $t \in[0,1]$.

Lemma A.11.13 $d(\mathrm{id}, \Omega, x)=1$ if $x \in \Omega$ and 0 if $x \notin \Omega$. Also if $d(f, \Omega, p) \neq 0$ then $f(x)=$ pfor some $x \in \Omega$. If $f=g$ on $\partial \Omega, p \notin f(\partial \Omega)$, then $d(f, \Omega, p)=d(g, \Omega, p)$.

Proof: We can extend id to be id on all of $S^{n}$. Then from the definition $d=1$.
Next suppose $d(f, \Omega, p) \neq 0$. Then by Proposition A.11.9 there is a point $x \in \Omega$ such that $f(x)=p$.

As to the last claim, one can consider $d(f+t(g-f), \Omega, p)$ which must be constant for $t \in[0,1]$ because $g=f$ on $\partial \Omega$ so $p \notin(f+t(g-f))(\partial \Omega)$.

The following simple lemma holds in $S^{n}$ or in $\mathbb{R}^{n}$ or more generally for a metric space in which open balls are connected.

Lemma A.11.14 Let $\left\{K_{i}\right\}_{i=1}^{N}, N \leq \infty$ be the connected components of $S^{n} \backslash C$ where $C$ is a closed set. Then $\partial K_{i} \subseteq C$.

Proof: Since $K_{i}$ is a connected component of an open set, it is itself open. Recall that this happens because open balls are connected. Thus $\partial K_{i}$ consists of all limit points of $K_{i}$ which are not in $K_{i}$. Let $y$ be such a point. If it is not in $C$ then it must be in some other $K_{j}$ which is impossible because these are disjoint open sets. Thus if $x$ is a point in $U$ it cannot be a limit point of $V$ for $V$ disjoint from $U$.

I want to consider $d(g \circ f, \Omega, p)$. Here this assumes that $g: S^{n} \rightarrow S^{n}$ and also requires $p \notin g(f(\partial \Omega))$ so $g^{-1}(p) \in(f(\partial \Omega))^{C}$. Let $K_{i}$ be the components of $(f(\partial \Omega))^{C}$. Pick $q_{i} \in$ $K_{i}$ so $d\left(f, \Omega, K_{i}\right)=d\left(f, \Omega, q_{i}\right)$. Since $g^{-1}(p)$ is a compact set, which lies in $(f(\partial \Omega))^{C}$, it can have empty intersection with only finitely many of the components $K_{i}$.

Lemma A.11.15 $g^{-1}(p)$ has empty intersection with all but finitely many of the $K_{i}$.
The next theorem is a major result called the product formula. Recall that by Propostion A.11.12, $d(f, \Omega, p)$ is constant for $p \in K_{i}$ and so we denote this common value as $d\left(f, \Omega, K_{i}\right)$.

Theorem A.11.16 Let $g: S^{n} \rightarrow S^{n}$ and let $f: \Omega \rightarrow S^{n}$. Suppose $p \notin g(f(\partial \Omega))$. Then for $K_{i}$ the components of $f(\partial \Omega)^{C}$,

$$
\begin{equation*}
d(g \circ f, \Omega, p)=\sum_{i} d\left(f, \Omega, K_{i}\right) d\left(g, K_{i}, p\right) \tag{1.19}
\end{equation*}
$$

and the sum is finite.
Proof: By the above Lemma A.11.15, we will consider only finitely many $i$ for which $g^{-1}(p)$ intersects $K_{i}$. Otherwise $d\left(g, K_{i}, p\right)=0$ and the term in the sum is 0 . Consider $f^{-1}\left(K_{i}\right) \cap \Omega$, disjoint open sets because the $K_{i}$ are disjoint. Also we are assuming that $g^{-1}(p) \subseteq \cup_{i} K_{i}$ so

$$
(g \circ f)^{-1}(p) \subseteq \cup_{i} f^{-1}\left(K_{i}\right) \cap \Omega \subseteq \Omega
$$

and $p \notin(g \circ f)\left(\bar{\Omega} \backslash \cup_{i} f^{-1}\left(K_{i}\right)\right)$. It follows from Corollary A.11.11 and Proposition A.11.9 that

$$
d(g \circ f, \Omega, p)=\sum_{i} d\left(g \circ f, f^{-1}\left(K_{i}\right) \cap \Omega, p\right)
$$

Now letting $\hat{f_{i}}$ be an appropriate extension of $f$, on $f^{-1}\left(K_{i}\right) \cap \Omega$ and $[c]$ a generator of $H_{n}\left(S^{n}\right), d\left(f, f^{-1}\left(K_{i}\right) \cap \Omega, q_{i}\right)[c]=\hat{f}_{i *}([c])$. Therefore,

$$
\begin{aligned}
& d\left(g \circ f, f^{-1}\left(K_{i}\right) \cap \Omega, p\right)[c]=g_{*} \hat{f}_{i *}([c])=g_{*} d\left(f, f^{-1}\left(K_{i}\right) \cap \Omega, q_{i}\right)[c] \\
& \quad=g_{*} d\left(f, \Omega, q_{i}\right)[c]=d\left(f, \Omega, K_{i}\right) g_{*}[c]=d\left(f, \Omega, K_{i}\right) d\left(g, K_{i}, p\right)[c]
\end{aligned}
$$

Note that on the top line, $f$ restricted to $f^{-1}\left(K_{i}\right) \cap \Omega$ has the properties that $q_{i}$ is not in $f\left(\bar{\Omega} \backslash f^{-1}\left(K_{i}\right) \cap \Omega\right)$ and so Corollary A.11.11 applies. Thus we obtain the product formula 1.19.

## A. 12 The Degree on Open Subsets of $\mathbb{R}^{n}$

I am not all that interested in spheres. I am much more interested in what can be said about open bounded subsets of $\mathbb{R}^{n}$. However, this is essentially included in the above. It will be based on the following mapping $\theta$ illustrated in this picture.


Note that if $\Omega$ is an open bounded subset of $\mathbb{R}^{n}$, then $\theta \bar{\Omega}=\overline{\theta \Omega}$ is a proper subset of $S^{n}$ which does not contain $q$. Letting $p \notin f(\partial \Omega) d(f, \Omega, p) \equiv d\left(\theta \circ f \circ \theta^{-1}, \theta \Omega, \theta p\right)$.

Definition A.12.1 Let $\Omega$ be a bounded open set in $\mathbb{R}^{n}$ and $f: \bar{\Omega} \rightarrow \mathbb{R}^{n}$ is continuous. Let $p \notin f(\partial \Omega)$. Then for $\theta$ the homeomorphism which maps $\mathbb{R}^{n} \cup\{\infty\}$ to all of $S^{n}, p=\theta(\infty)$,
$\theta \Omega$ is an open set in $S^{n}$. Then $\left(\theta \circ f \circ \theta^{-1}\right): \theta \Omega \rightarrow S^{n}$ does not include $\theta p$ along with a closed ball $\bar{A}$ and we can define $d(f, \Omega, p) \equiv d\left(\theta \circ f \circ \theta^{-1}, \theta \Omega, \theta p\right)$ as in Definition A.11.3. This will be $d_{\hat{f}}$ where $\hat{f}$ is the extention of $\theta \circ f \circ \theta^{-1}$ off of $\theta \Omega$ described by letting $\hat{f}=\theta \circ \bar{f} \circ \theta^{-1}$ off $\bar{\Omega}$ and $\theta \circ f \circ \theta^{-1}$ on $\theta \bar{\Omega}$.

The earlier material on spheres yield the following proposition.
Proposition A.12.2 The degree has the following properties:

1. Let continuous $h:[0,1] \times \Omega \rightarrow \mathbb{R}^{n}$ and let $t \rightarrow p(t)$ also be continuous with $p(t) \notin$ $h(t, \partial \Omega)$ for each $t \in[0,1]$. Then $t \rightarrow d(h(t, \cdot), \Omega, p(t))$ is constant. Also $p \rightarrow$ $d(f, \Omega, p)$ is constant on each component of $f(\partial \Omega)^{C}$.
2. The identity map id, satisfies $d(\mathrm{id}, \Omega, x)=1$ if $x \in \Omega$ and $d(\mathrm{id}, \Omega, x)=0$ if $x \notin \Omega$. If $\Omega$ is a ball centered at 0 , then $d(-\mathrm{id}, \Omega, x)=(-1)^{n}$ for all $x \in \Omega$.
3. If $p \notin f(\bar{\Omega} \backslash \hat{\Omega})$ for $\hat{\Omega}$ an open subset of $\Omega$, then $d(f, \hat{\Omega}, p)=d(f, \Omega, p)$.
4. Also if $\Omega$ is an open bounded subset of $\mathbb{R}^{n}$ equal to the union of disjoint open sets $\Omega_{i}$, then for $f: \bar{\Omega} \rightarrow \mathbb{R}^{n}$ continuous, $p \notin f(\partial \Omega), d(f, \Omega, p)=\sum_{i} d\left(f, \Omega_{i}, p\right)$. The sum will be finite.
5. If $f=g$ on $\partial \Omega, p \notin f(\partial \Omega)$, then $d(f, \Omega, p)=d(g, \Omega, p)$.
6. If $\Omega$ is an open ball centered at 0 in $\mathbb{R}^{n}$ and $f: \Omega \rightarrow \mathbb{R}^{n}$ is given by $f\left(x_{1}, \ldots, x_{n}\right)=$ $\left(k_{1} x_{1}, \ldots, k_{n} x_{n}\right)$ where none of the $k_{i}$ are 0 , then $d(f, \Omega, 0)=(-1)^{m}$ where $m$ is the number of negative constants $k_{i}$.
7. If $\Omega$ is an open ball centered at 0 in $\mathbb{R}^{n}$ and $f: \Omega \rightarrow \mathbb{R}^{n}$ is given by $f\left(x_{1}, \ldots, x_{n}\right)=A x$ where $A^{-1}$ exists, then $d(f, \Omega, 0)=\operatorname{sgn}(\operatorname{det}(A))$.

Proof: Consider 2. the one about - id. The points on $S^{n},\left(x_{1}, \ldots, x_{n}, x_{n+1}\right)$ with the sum of the squares of the components equal to 1 are obtained by letting $\hat{f}$ correspond to the following on the sphere $S^{n}:\left(x_{1}, \ldots, x_{n}, x_{n+1}\right) \rightarrow\left(-x_{1}, \ldots,-x_{n}, x_{n+1}\right)$ and so from Theorem A.10.5 and those leading to this theorem, this would be $(-1)^{n}$ because we changed sign in $n$ components.

For part 6., consider first the case where all the $k_{i}>0$. From 1., $d(\lambda f+(1-\lambda)$ id, $B, 0)$ is constant in $\lambda \in[0,1]$ since $0 \notin(\lambda f+(1-\lambda) \mathrm{id})(\partial B)$. When $\lambda=0$ this is 1 and so when $\lambda=1$ it is also 1 , this by part 2 . Now if you compose such an $f$ with a map which changes the sign of $m$ of the entries, then similar to 2., the result will be $(-1)^{m}$ because of Lemma A.10.4 about the degree of a composition being the product of the degrees.

For part 7. I will consider other elementary operations. Also regard each of these invertible elementary operations $E$ as taking $\infty$ to $\infty$. Thus $x_{n} \rightarrow \infty$ if and only if $E x_{n} \rightarrow \infty$, similar for an invertible $A$. Let $L_{i j}(\boldsymbol{x}) \equiv\left(x_{1}, \ldots, x_{i-1}, x_{i}+x_{j}, x_{i+1}, \ldots, x_{n}\right)^{T}$ which comes from an elementary matrix in which the $i^{\text {th }}$ row is replaced with the $j^{t h}$ row added to the $i^{\text {th }}$ row. Then for $\boldsymbol{x} \in \partial \Omega$ and $t \in[0,1], 0 \notin\left(t L_{i j}+(1-t) \mathrm{id}\right) \partial \Omega$. Indeed, the matrix for the mapping $t L_{i j}+(1-t)$ id is just the identity with $t$ added into the $i j^{t h}$ position which is invertible and so cannot send anything in $\partial B$ to 0 . Therefore, from 1 ., $d\left(L_{i j}, \Omega, 0\right)=d(\mathrm{id}, \Omega, 0)=1$. Since $A^{-1}$ exists, $A$ is the composition of finitely many elementary operations. For the operation which switches two components, the degree is -1 as is the determinant of the elementary
matrix which produces this operation. The case of multiplication of a row by a nonzero constant is covered in 6. Also, for invertible $E, d(E, \Omega, 0) \equiv d\left(\theta \circ E \circ \theta^{-1}, \theta \Omega, \theta 0\right)$ and we can regard $\hat{E}$ as simply $\theta \circ E \circ \theta^{-1}$ on $S^{n}$. Thus, from Lemma A.10.4, for $A$ invertible, $\theta \circ A \circ \theta^{-1}=\prod_{i} \theta \circ E_{i} \circ \theta^{-1}, E_{i}$ an elementary operation and so

$$
d(A, \Omega, 0)=d_{\theta \circ A \circ \theta^{-1}}=\prod_{i} d_{\theta \circ E_{i} \circ \theta^{-1}}=\prod_{i} \operatorname{sgn}\left(\operatorname{det}\left(E_{i}\right)\right)=\operatorname{sgn}(\operatorname{det}(A))
$$

Now here is a very easy proof of the Brouwer fixed point theorem.
Proposition A.12.3 Let $f: B \rightarrow B$ where $B$ is a closed ball of radius $R$ and center 0 . Then there exists a fixed point for $f$.

Proof: This is from Proposition A.12.2. If there is no fixed point for $f$, then for $|x|=$ $R, 0 \notin x-t f(x)$ for all $t \in[0,1]$ hence $d(\mathrm{id}-t f, B, 0)$ is constant for $t \in[0,1]$. This degree is 1 when $t=0$ and so it is also 1 when $t=1$ but this means there must exist $x$ such that $x-f(x)=0$ and this is the fixed point.

The next is the cow lick theorem.
Theorem A.12.4 Let $n$ be odd and let $\Omega$ be an open bounded set in $\mathbb{R}^{n}$ with $0 \in \Omega$. Suppose $f: \partial \Omega \rightarrow \mathbb{R}^{n} \backslash\{0\}$ is continuous. Then for some $x \in \partial \Omega$ and $\lambda \neq 0, f(x)=\lambda x$.

Proof: Using the Tietze extension theorem, extend $f$ to all of $\mathbb{R}^{n}$. Also denote the extended function by $f$. Suppose for all $x \in \partial \Omega, f(x) \neq \lambda x$ for all $\lambda \in \mathbb{R}$. Then

$$
\begin{aligned}
& 0 \notin t f(x)+(1-t) x, \quad(x, t) \in \partial \Omega \times[0,1] \\
& 0 \notin t f(x)-(1-t) x, \quad(x, t) \in \partial \Omega \times[0,1]
\end{aligned}
$$

Thus there exists a homotopy of $f$ and id and a homotopy of $f$ and -id . Then by the homotopy invariance of degree, $d(f, \Omega, 0)=d(\mathrm{id}, \Omega, 0), d(f, \Omega, 0)=d(-\mathrm{id}, \Omega, 0)$. But this is impossible because $d(\mathrm{id}, \Omega, 0)=1$ but $d(-\mathrm{id}, \Omega, 0)=(-1)^{n}=-1$.

The product formula is from using the homeomophism $\theta$ of $\mathbb{R}^{n} \cup\{\infty\}$ and $S^{n}$.
Theorem A.12.5 Let $g: f(\bar{\Omega}) \rightarrow \mathbb{R}^{n}$ be continuous and let $f: \bar{\Omega} \rightarrow \mathbb{R}^{n}$ be continuous. Suppose $p \notin g(f(\partial \Omega)), p \in \mathbb{R}^{n}$. Then for $K_{i}$ the components of $f(\partial \Omega)^{C}$ which contain a point of $g^{-1}(p)$,

$$
\begin{equation*}
d(g \circ f, \Omega, p)=\sum_{i} d\left(f, \Omega, K_{i}\right) d\left(g, K_{i}, p\right) \tag{1.20}
\end{equation*}
$$

and the sum is finite. None of these components $K_{i}$ contain $\infty$ so these are all bounded components.

Proof: Extend $g$ off $f(\partial \Omega)$ so that $\lim _{x \rightarrow \infty} g(x)=0$ and let $g(\infty)=0$ so $g$ is continuous on $\mathbb{R}^{n} \cup\{\infty\}$. Then $g \circ f$ is unchanged on $\partial \Omega$. Also $g^{-1}(\infty)=\emptyset$. If $d(g, K, p) \neq 0$ then $K$ cannot be the unbounded component of $f(\partial \Omega)^{C}$ because that one has $\infty, g$ fails to equal $\infty$ and the degree is constant on components. Thus we get 1.20 for $K_{i}$ the bounded components of $f(\partial \Omega)^{C}$. It only remains to verify the sum is finite. However, this follows because $g^{-1}(p)$ is a compact set contained in $\mathbb{R}^{n}$.

## A. 13 Jordan Separation Theorem

Recall that if a function $f$ is continuous and one to one on a compact set $K$, then $f$ is a homeomorphism of $K$ and $f(K)$. Also recall that if $U$ is a nonempty open set, the boundary of $U$, denoted as $\partial U$ and meaning those points $x$ with the property that for all $r>0, B(x, r)$ intersects both $U$ and $U^{C}$, is $\bar{U} \backslash U$. Note that it is not possible for a compact set $H$ in $\mathbb{R}$ to have $H^{C}$ posess only one connected component. Thus the next proposition considers the case $n \geq 2$.

Proposition A.13.1 Let $H$ be a compact set and let $f: H \rightarrow \mathbb{R}^{n}, n \geq 2$ be one to one and continuous so that $H$ and $f(H) \equiv C$ are homeomorphic. Suppose $\bar{H}^{C}$ has only one connected component so $H^{C}$ is connected. Then $C^{C}$ also has only one component.

Proof: Extend $f$, using the Tietze extension theorem on its components to all of $\mathbb{R}^{n}$ and let $g$ be an extension of $f^{-1}$ to all of $\mathbb{R}^{n}$. Suppose $K$ is a bounded component of $C^{C}$. Then by Lemma A.11.14 $\partial K \subseteq C$. Hence $g(\partial K) \subseteq g(C)=H$. If $Q$ is a bounded component of $g(\partial K)^{C}$ then if $Q$ contains a point of the connected set $H^{C}$ then $Q$ would need to contain all of $H^{C}$ and $Q$ is not bounded after all. Therefore, there are no bounded components of $g(\partial K)^{C}$. But by the product formula, Theorem A.12.5, for $\mathscr{Q}$ the set of bounded components of $g(\partial K)^{C}, d(f \circ g, K, z)=\sum_{Q \in \mathscr{Q}} d(g, K, Q) d(f, Q, z)=0$ because $\mathscr{Q}=\emptyset$. However, $d(f \circ g, K, z)=d(\mathrm{id}, K, z)$ because $f \circ g=\mathrm{id}$ on $\partial K \subseteq C$. See Proposition A.12.2 which comes from the earlier development of the degree on spheres. Thus there is no bounded component of $C^{C}$ so $C^{C}$ has only one component just as $H^{C}$.

This says that if a compact set $H$ fails to separate $\mathbb{R}^{n}$ for $n \geq 2$ and if $f: H \rightarrow \mathbb{R}^{n}$ is continuous and one to one, then also $f(H)$ fails to separate $\mathbb{R}^{n}$.

It is obvious that the unit sphere $S^{n-1}$ divides $\mathbb{R}^{n}$ into two disjoint open sets, the inside and the outside, this for $n \geq 2$. The following shows that this also holds for any homeomorphic image of $S^{n-1}$.

Proposition A.13.2 Let $B$ be the ball $B(\mathbf{0}, 1)$ with $S^{p-1}$ its boundary, $p \geq 2$. Suppose $f: S^{p-1} \rightarrow C \equiv f\left(S^{p-1}\right) \subseteq \mathbb{R}^{p}$ is a homeomorphism. Then $C^{C}$ also has exactly two components, one bounded and one unbounded.

Proof: Let $f$ denote the extension of $f$ to all of $\mathbb{R}^{p}$ and let $g=f^{-1}$ on $f(\partial B)$ where $g$ is also extended using the Tietze extension theorem to all of $\mathbb{R}^{p}$. Let $H$ be the unbounded component of $\mathbb{R}^{p} \backslash S^{p-1}$. Assuming there exists $K$ a bounded component of $f(\partial B)^{C}$, then from Lemma ??, $\partial K \subseteq f(\partial B)$ so $g(\partial K) \subseteq \partial B$. Also, $f \circ g(\partial K) \subseteq f \circ g(f(\partial B))=f(\partial B)$. Recall that $K$ has no points in $f(\partial B)$ so if $p \in K$, then $p$ cannot be in $f(\partial B)$ and consequently $p$ cannot be in $f \circ g(\partial K)$ either. Summarizing this,

$$
\partial K \subseteq f(\partial B), g(\partial K) \subseteq \partial B, f \circ g(\partial K) \cap K=\emptyset
$$

Then picking $p \in K$, by the product rule,

$$
1=d(i d, K, p)=d(f \circ g, K, p)=\sum_{i} d\left(g, K, Q_{i}\right) d\left(f, Q_{i}, p\right)
$$

where here the $Q_{i}$ are the bounded components of $(g(\partial K))^{C}$. These are maximal open connected sets in $\mathbb{R}^{p}$. Recall $g(\partial K) \subseteq \partial B$. If $Q_{i}$ has a point of $H$, then $H$ would be connected and contain no points of $g(\partial K)$ and so $H$ would be contained in $Q_{i}$ which does
not happen because $Q_{i}$ is bounded. Thus $Q_{i} \subseteq \bar{B}$ but also $Q_{i}$ is open and so it must be contained in $B$. Now $B$ is connected and open and contains no points of $g(\partial K)$ because it contains no points of $\partial B$ which is a larger set than $g(\partial K)$ and so in fact $Q_{i}=B$ and there is only one term in the above sum. Thus, from properties of the degree,

$$
\begin{aligned}
1 & =d(i d, K, p)=d(f \circ g, K, p)=d(g, K, B) d(f, B, p) \\
& =d(g, K, \mathbf{0}) d(f, B, K)=d(g \circ f, B, \mathbf{0})
\end{aligned}
$$

so by the product rule there is no more than one bounded component of $f(\partial B)^{C}$ the $K$ just mentioned. However, there is at least one bounded component and one term in the sum for the product rule because if there were no bounded components, this would contradict Proposition ?? since it is clear that $\left(S^{p-1}\right)^{C}$ is not connected. If you had other components of $f(\partial B)^{C}$ called $K_{i}, i \leq m \leq \infty$ you could repeat the above argument and obtain $1=d\left(g, K_{i}, 0\right) d\left(f, B, K_{i}\right)=d(g \circ f, B, 0)$, but then, by the product rule, you would have for $K \equiv K_{0}, 1=d(g \circ f, B, 0)=\sum_{k=0}^{m} d\left(g, K_{i}, 0\right) d\left(f, B, K_{i}\right)=m+1$. Thus there is exactly one bounded component of $f(\partial B)^{C}$.

Proposition A.13.3 Let $B$ be the ball $B(0,1)$ with $S^{n-1}$ its boundary, $n \geq 2$. Suppose $f: S^{n-1} \rightarrow C \equiv f\left(S^{n-1}\right) \subseteq \mathbb{R}^{n}$ is a homeomorphism. Then $C^{C}$ also has exactly two components, one bounded and one unbounded.

Proof: Let $f$ denote the extension of $f$ to all of $\mathbb{R}^{n}$ and let $g=f^{-1}$ on $f(\partial B)$ where $g$ is also extended using the Tietze extension theorem to all of $\mathbb{R}^{n}$. Thus $0 \notin g(f(\partial B))$ because if $x \in \partial B$, then $g(f(x))=x \neq 0$. By Proposition A.12.2 and the product formula,

$$
\begin{equation*}
1=d(\mathrm{id}, B, 0)=d(g \circ f, B, 0)=\sum_{i} d\left(f, B, K_{i}\right) d\left(g, K_{i}, 0\right) \tag{1.21}
\end{equation*}
$$

where the $K_{i}$ are finitely many of the bounded components of $f(\partial B)^{C}$. Thus, from Lemma A.11.14 $\partial K_{i} \subseteq f(\partial B)$ and so $g\left(\partial K_{i}\right) \subseteq g(f(\partial B))=\partial B$. It follows that $0 \notin g\left(\partial K_{i}\right) \subseteq \partial B$. Pick a term in the sum which is nonzero. Let it involve $K_{i}$. Letting $y_{i} \in K_{i}$,

$$
\begin{equation*}
d\left(f, B, K_{i}\right) d\left(g, K_{i}, 0\right)=d\left(f, B, y_{i}\right) d\left(g, K_{i}, B\right) \tag{1.22}
\end{equation*}
$$

Indeed, $B$ is connected and contains no points of $g\left(\partial K_{i}\right) \subseteq \partial B$ so $z \rightarrow d\left(g, K_{i}, z\right)$ is constant on $B$. I want to argue that $B$ is the only bounded component of $g\left(\partial K_{i}\right)^{C}$.

Let $H$ be a bounded component of $g\left(\partial K_{i}\right)^{C}$. Then $H$ cannot have any points of $g\left(\partial K_{i}\right)$. If $H$ has any points of $U$, the unbounded component of $(\partial B)^{C}$, then $U$ is a connected set in $g\left(\partial K_{i}\right)^{C}$ and intersects $H$ so the component determined by a point of intersection must be $H$ and contain $U$. Hence $H$ is not bounded after all. Thus the only bounded components of $g\left(\partial K_{i}\right)^{C}$ are contained in $B$. Since $B$ is connected and has no points of $g\left(\partial K_{i}\right) \subseteq \partial B, B$ must be contained in exactly one bounded component of $g\left(\partial K_{i}\right)^{C}$. Therefore this bounded component and $B$ are equal and the only bounded component of $g\left(\partial K_{i}\right)^{C}$ is $B$. Therefore in 1.22, the right side is just $d\left(f \circ g, K_{i}, y_{i}\right)=d\left(\mathrm{id}, K_{i}, y_{i}\right)=1$. This is because $f \circ g=\mathrm{id}$ on $\partial K_{i} \subseteq f(\partial B)$ and when the two functions coincide on the boundary, they have the same degree by Proposition A.12.2. It follows from 1.21 that, since each term in the sum is 1 or 0 , there is exactly one term in the sum and hence exactly one bounded component of $f(\partial B)^{C}$.

A repeat of the above proof yields the following corollary. Replace $B$ with $\Omega$.

Corollary A.13.4 Let $\Omega \subseteq \mathbb{R}^{n}, n \geq 2$ be a bounded open connected set such that $\partial \Omega^{C}$ has two components, a bounded and an unbounded component. Suppose $f: \partial \Omega \rightarrow C \equiv$ $f(\partial \Omega) \subseteq \mathbb{R}^{n}$ is a homeomorphism. Then $C^{C}$ also has exactly two components, one bounded and one unbounded.

As an application, here is a very interesting result about orientation and the invariance of domain theorem which says that a one to one continuous function maps open sets to open sets.

Proposition A.13.5 Let $\Omega$ be an open connected bounded set in $\mathbb{R}^{n}, n \geq 2$ such that $\mathbb{R}^{n} \backslash$ $\partial \Omega$ consists of two connected components. Let $f \in C\left(\bar{\Omega} ; \mathbb{R}^{n}\right)$ be continuous and one to one. Then $f(\Omega)$ is the bounded component of $\mathbb{R}^{n} \backslash f(\partial \Omega)$ and for $y \in f(\Omega), d(f, \Omega, y)$ either equals 1 or -1 .

Proof: By the Jordan separation theorem, Corollary A.13.4, $\mathbb{R}^{n} \backslash f(\partial \Omega)$ consists of two components, a bounded component $B$ and an unbounded component $U$. Using the Tietze extention theorem, there exists $g$ defined on $\mathbb{R}^{n}$ such that $g=f^{-1}$ on $f(\bar{\Omega})$. Thus on $\partial \Omega, g \circ f=\mathrm{id}$. It follows from this and the product formula that

$$
1=d(\mathrm{id}, \Omega, g(y))=d(g \circ f, \Omega, g(y))=d(g, B, g(y)) d(f, \Omega, B)
$$

Therefore, $d(f, \Omega, B) \neq 0$ and so for every $z \in B$, it follows $z \in f(\Omega)$. Thus $B \subseteq f(\Omega)$. On the other hand, $f(\Omega)$ cannot have points in both $U$ and $B$ because it is a connected set. Therefore $f(\Omega) \subseteq B$ and this shows $B=f(\Omega)$. Thus $d(f, \Omega, B)=d(f, \Omega, y)$ for each $y \in B$ and the above formula shows this equals either 1 or -1 because the degree is an integer.

This shows how to generalize orientation. It is just the degree. One could use this to describe an orientable manifold without any direct reference to differentiability.

In the case of $f\left(S^{n-1}\right)$ for $f$ one to one and continuous, one wants to verify that this is the boundary of both components, the bounded one and the unbounded one.

Theorem A.13.6 Let $S^{n-1}$ be the unit sphere in $\mathbb{R}^{n}, n \geq 2$. Suppose $\gamma: S^{n-1} \rightarrow \Gamma \subseteq \mathbb{R}^{n}$ is one to one onto and continuous. Then $\mathbb{R}^{n} \backslash \Gamma$ consists of two components, a bounded component (called the inside) $U_{i}$ and an unbounded component (called the outside), $U_{o}$. Also the boundary of each of these two components of $\mathbb{R}^{n} \backslash \Gamma$ is $\Gamma$ and $\Gamma$ has empty interior.

Proof: $\gamma^{-1}$ is continuous since $S^{n-1}$ is compact and $\gamma$ is one to one. By the Jordan separation theorem, $\mathbb{R}^{n} \backslash \Gamma=U_{o} \cup U_{i}$ where these on the right are the connected components of the set on the left, both open sets. Only $U_{i}$ is bounded. Thus $\Gamma \cup U_{i} \cup U_{o}=\mathbb{R}^{n}$. Since both $U_{i}, U_{o}$ are open, $\partial U \equiv \bar{U} \backslash U$ for $U$ either $U_{o}$ or $U_{i}$. If $x \in \Gamma$, and is not a limit point of $U_{i}$, then there is $B(x, r)$ which contains no points of $U_{i}$. Let $S$ be those points $x$ of $\Gamma$ for which, $B(x, r)$ contains no points of $U_{i}$ for some $r>0$. This $S$ is open in $\Gamma$. Let $\hat{\Gamma}$ be $\Gamma \backslash S$. Then if $\hat{C}=\gamma^{-1}(\hat{\Gamma})$, it follows that $\hat{C}$ is a closed set in $S^{n-1}$ and is a proper subset of $S^{n-1}$. It is obvious that taking a relatively open set from $S^{n-1}$ results in a compact set whose complement in $\mathbb{R}^{n}$ is an open connected set. By Proposition A.13.1, $\mathbb{R}^{n} \backslash \hat{\Gamma}$ is also an open connected set. Start with $x \in U_{i}$ and consider a continuous curve which goes from $x$ to $y \in U_{o}$ which is contained in $\mathbb{R}^{n} \backslash \hat{\Gamma}$. Thus the curve contains no points of $\hat{\Gamma}$. However, it must contain points of $\Gamma$ which can only be in $S$. The first point of $\Gamma$ intersected by this curve is a point in $\overline{U_{i}}$ and so this point of intersection is not in $S$ after all because every ball containing it must contain points of $U_{i}$. Thus $S=\emptyset$ and every point of $\Gamma$ is in $\overline{U_{i}}$. Similarly, every point of $\Gamma$ is in $\overline{U_{o}}$. Thus $\Gamma \subseteq \overline{U_{i}} \backslash U_{i}$ and $\Gamma \subseteq \overline{U_{o}} \backslash U_{o}$. However, if $x \in \overline{U_{i}} \backslash U_{i}$, then
$x \notin U_{o}$ because it is a limit point of $U_{i}$ and so $x \in \Gamma$. It is similar with $U_{o}$. Thus $\Gamma=\overline{U_{i}} \backslash U_{i}$ and $\Gamma=\overline{U_{o}} \backslash U_{o}$. This could not happen if $\Gamma$ had an interior point. Such a point would be in $\Gamma$ but would fail to be in either $\partial U_{i}$ or $\partial U_{o}$.

When $n=2$, this theorem is called the Jordan curve theorem.
Corollary A.13.7 Let $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ be one to one, continuous and let $\lim _{|x| \rightarrow \infty}|f(x)|=\infty$. Then $f$ is onto.

Proof: From the invariance of domain, Proposition A.13.5, $f\left(\mathbb{R}^{n}\right)$ is open. However, if $f\left(x_{n}\right) \rightarrow y$, then $\left\{x_{n}\right\}$ must be bounded and so there is a convergent subsequence $x_{n_{k}} \rightarrow x$. Therefore, by continuity of $f$ it follows that $f\left(x_{n_{k}}\right) \rightarrow y=f(x)$ and so $f\left(\mathbb{R}^{n}\right)$ is also closed. But $\mathbb{R}^{n}=\left(\mathbb{R}^{n} \backslash f\left(\mathbb{R}^{n}\right)\right) \cup f\left(\mathbb{R}^{n}\right)$, the union of disjoint open sets. Hence one must be empty since otherwise $\mathbb{R}^{n}$ would not be connected.

Corollary A.13.8 If $n>m$ then there is no continuous one to one function $f$ which maps $\mathbb{R}^{n}$ onto $\mathbb{R}^{m}$. Thus $\mathbb{R}^{n}, \mathbb{R}^{m}$ are not homeomorphic.

Proof: If there were, then you could let $\theta: \mathbb{R}^{m} \rightarrow \mathbb{R}^{n}$ be given by $x \rightarrow(x, 0)$ where $0=(0,0, \ldots, 0)$. and we would have $\theta \circ f$ is one to one and continuous mapping $\mathbb{R}^{n}$ to $\mathbb{R}^{n}$ but does not take open sets to open sets.

Theorem A.13.9 Let $B$ be a ball in $\mathbb{R}^{n}, n \geq 1$ and let continuous $f: \bar{B} \rightarrow \mathbb{R}^{n}$ be odd meaning that $f(x)=-f(-x)$. Suppose $0 \notin f(\partial B)$. Then $d(f, B, 0) \neq 0$.

Proof: From the above construction of $\hat{f}$, this reduces to showing that $\hat{f}([\hat{c}]) \neq 0$ where $[\hat{c}]$ generates $H_{n}\left(S^{n}\right)$.

First consider the case where $n=1$ so the ball is just the interval $(-R, R)$. Let $g$ denote the function of the form $g(x)=k x$ where $g(R)=f(R), g(-R)=f(-R)$. Then $d(t f+(1-t) g,(-R, R), 0)$ is constant by the earlier properties of the degree because for all $t \in[0,1], 0 \notin(t f+(1-t) g)(\partial(-R, R))$. However, $d(g,(-R, R), 0)= \pm 1$ by Proposition A.12.2. Thus the theorem holds in case $n=1$. Thus $\hat{f}([\hat{c}]) \neq 0$.

This shows what is desired in case $n=1$. Assume that $d(f, B, 0) \neq 0$ for $f$ odd for $n-1, n-1 \geq 1$ so $\hat{f}_{*}[\hat{c}] \neq 0$ for $[\hat{c}]$ generating the homology group in dimension $n-1$. Recall that for $U$ and $V$ missing the top and bottom of $S^{n}, H_{n-1}(U \cap V) \approx H_{n-1}\left(S^{n-1}\right)$ and so we have the following Mayer Vietoris sequence in which $\Delta$ is an isomorphism:

$$
\left.\begin{array}{ccccccc}
=H_{n}\left(\mathbb{R}^{n}\right)=0 & =H_{n}\left(\mathbb{R}^{n}\right)=0 \\
H_{n}(U) \oplus \oplus H_{n}(V) & \xrightarrow{g_{*}} & H_{n}\left(S^{n}\right) & \xrightarrow{\Delta} & H_{n-1}\left(S^{n-1}\right) & \xrightarrow{h_{*}} & H_{n-1}(U) \oplus H_{n-1}(V) \\
\downarrow \hat{f}_{*} \oplus \hat{f}_{*} & & \downarrow \hat{f}_{*} & & \downarrow \hat{f}_{*} & & \\
H_{n}(\hat{f}(U)) \oplus H_{n}(\hat{f}(V)) & \xrightarrow{g_{*}} & H_{n}\left(\hat{f}\left(S^{n}\right)\right) & \xrightarrow{\Delta} & H_{n-1}\left(\hat{f}\left(S^{n-1}\right)\right) & \xrightarrow{h_{*}} & H_{n-1}(\hat{f}(U)) \oplus \hat{f}_{*}(\hat{f}
\end{array}\right)
$$

In the above, $g_{*}, h_{*}$ mapping to 0 in the bottom line comes from the above observation that $\hat{f}(U) \subseteq U$ and $\hat{f}(V) \subseteq V$. Then this implies that both connecting homomorphisms $\Delta$ are isomorphisms. Thus, using Lemma A.7.1, the above Mayer Vietoris sequence commutes and the following holds for $n>1$.

If $[\hat{c}]$ generates $H_{n}\left(S^{n}\right)$, then $\Delta[\hat{c}]$ generates $H_{n-1}(U \cap V) \approx H_{n-1}\left(S^{n-1}\right)$ and by induction and Lemma A.7.1, $0 \neq \hat{f}_{*} \Delta[\hat{c}]=\Delta \hat{f}_{*}[\hat{c}]$ and so $\hat{f}_{*}[\hat{c}] \neq 0$, so $d(f, B, 0) \neq 0$.

Since we know that $d(f, B, 0) \neq 0$ for $f$ an odd mapping, this leads to the Borsuk Ulam theorem.

Theorem A.13.10 Let $B$ be a bounded open ball in $\mathbb{R}^{n}$ centered at 0 and let $f: \partial B \rightarrow V$ be continuous where $V$ is an $m$ dimensional subspace of $\mathbb{R}^{n}, m \leq n-1$. Then $f(-x)=f(x)$ for some $x \in \partial B$.

Proof: We can assume $V$ is $\mathbb{R}^{m}$. Suppose the conclusion of the theorem is not so. Using the Tietze extension theorem on components of the function, extend $f$ to all of $\mathbb{R}^{n}$, $f(\bar{B}) \subseteq V$. (Here the extended function is also denoted by $f$.) Let $g(x)=f(x)-f(-x)$. Thus $g$ is odd, maps into $V$ and assuming the theorem is not true, $0 \notin g(\partial B)$ and so for some $r>0, B(0, r) \subseteq \mathbb{R}^{n} \backslash g(\partial B)$. For $z \in B(0, r), d(g, B(0, r), z)=d(g, B(0, r), 0) \neq 0$ because $B(0, r)$ is contained in a component of $\mathbb{R}^{n} \backslash g(\partial B)$. Hence $V \supseteq g(B(0, r)) \supseteq B(0, r)$ and this is a contradiction because $V$ is $m$ dimensional.

One can also show the invariance of domain theorem with the above theorem about the degree.

Lemma A.13.11 Let $g: \overline{B(0, r)} \rightarrow \mathbb{R}^{p}$ be one to one and continuous where here $B(0, r)$ is the ball centered at 0 of radius $r$ in $\mathbb{R}^{p}$. Then there exists $\delta>0$ such that

$$
g(0)+B(0, \delta) \subseteq g(B(0, r))
$$

The symbol on the left means: $\{g(0)+x: x \in B(0, \delta)\}$.
Proof: For $t \in[0,1]$, let $h(x, t) \equiv g\left(\frac{x}{1+t}\right)-g\left(\frac{-t x}{1+t}\right)$. Then for $x \in \partial B(0, r), h(x, t) \neq 0$ because if this were so, the fact $g$ is one to one implies $\frac{x}{1+t}=\frac{-t x}{1+t}$ and this requires $x=0$, not the case since $\|x\|=r$. Then $d(h(\cdot, t), B(0, r), 0)$ is constant. Hence it is nonzero for all $t$ thanks to Theorem A.14.6, because $h(\cdot, 1)$ is odd. Now let $B(0, \delta)$ be such that $B(0, \delta) \cap h(\partial \Omega, 0)=\emptyset$. Then $d(h(\cdot, 0), B(0, r), 0)=d(h(\cdot, 0), B(0, r), z)$ for $z \in B(0, \delta)$ because the degree is constant on connected components of $\mathbb{R}^{p} \backslash h(\partial \Omega, 0)$. Hence $z=$ $h(x, 0)=g(x)-g(0)$ for some $x \in B(0, r)$. Thus $g(B(0, r)) \supseteq g(0)+B(0, \delta)$

Here is another proof of invariance of domain.
Theorem A.13.12 (invariance of domain)Let $\Omega$ be any open subset of $\mathbb{R}^{p}$ and let $f: \Omega \rightarrow$ $\mathbb{R}^{p}$ be continuous and one to one. Then $f$ maps open subsets of $\Omega$ to open sets in $\mathbb{R}^{p}$.

Proof: Let $\overline{B\left(x_{0}, r\right)} \subseteq \Omega$ where $f$ is one to one on $\overline{B\left(x_{0}, r\right)}$. Let $g$ be defined on $\overline{B(0, r)}$ given by $g(x) \equiv f\left(x+x_{0}\right)$. Then $g$ satisfies the conditions of Lemma A.13.11, being one to one and continuous. It follows from that lemma there exists $\delta>0$ such that

$$
\begin{gathered}
f(\Omega) \supseteq f\left(B\left(x_{0}, r\right)\right)=f\left(x_{0}+B(0, r)\right) \\
=g(B(0, r)) \supseteq g(0)+B(0, \delta)=f\left(x_{0}\right)+B(0, \delta)=B\left(f\left(x_{0}\right), \delta\right)
\end{gathered}
$$

This shows that for any $x_{0} \in \Omega, f\left(x_{0}\right)$ is an interior point of $f(\Omega)$ which shows $f(\Omega)$ is open.

## A. 14 Analysis and the Degree

The degree can be presented in a different way using more linear algebra and analysis.
Lemma A.14.1 Let $y \notin f(\partial \Omega)$. Then $d(f, \Omega, y)=d(f-y, \Omega, 0)$. Also,

$$
d(f, \Omega, y)=d(f((\cdot)+z), \Omega-z, y) .
$$

Proof: Consider $d(t f+(1-t)(f-y), \Omega, t y)$. If $x \in \partial \Omega, t \in[0,1]$,

$$
\begin{aligned}
(t f+(1-t)(f-y))(x)-t y & =t f(x)+(1-t) f(x)-(1-t) y-t y \\
& =f(x)-y \neq 0
\end{aligned}
$$

When $t=0, d(t f+(1-t)(f-y), \Omega, t y)=d(f-y, \Omega, 0)$. When $t=1$, it is $d(f, \Omega, y)$ so by Proposition A.12.2 the two are equal. Now consider the second claim.

Let $\alpha^{z}(x)=x+z$. Then the claim is that $d(f, \Omega, y)=d\left(f \circ \alpha^{z}, \alpha^{-z} \Omega, y\right)$. Let $\hat{f}$ go with $f$ and $\widehat{f \circ \alpha^{z}}$ go with $f \circ \alpha^{z}$. Then $\widehat{f \circ \alpha^{z}}=\hat{f} \circ \theta \circ \alpha^{z} \circ \theta^{-1}$ since the latter does what $\widehat{f \circ \alpha^{z}}$ is supposed to do:

$$
\left(\hat{f} \circ \theta \circ \alpha^{z} \circ \theta^{-1}\right)^{-1}(y)=\theta \circ \alpha^{-z} \circ \theta^{-1} \circ \hat{f}^{-1}(y) \in \theta\left(\alpha^{-z} \Omega\right)
$$

Now $\theta \circ \alpha^{z} \circ \theta^{-1}$ is a homeomorphism on $S^{n}$ and so if $[\hat{c}]$ generates $H_{n}\left(S^{n}\right)$, then

$$
\left(\theta \circ \alpha^{z} \circ \theta^{-1}\right)_{*}[\hat{c}]= \pm[\hat{c}] .
$$

However, letting $t \in[0,1],\left(\theta \circ \alpha^{t z} \circ \theta^{-1}\right)$ is a homotopy of id and $\left(\theta \circ \alpha^{z} \circ \theta^{-1}\right)$ so by Theorem A.3.7 $\left(\theta \circ \alpha^{z} \circ \theta^{-1}\right)_{*}[\hat{c}]=\operatorname{id}_{*}[\hat{c}]=[\hat{c}]$ and so

$$
\begin{aligned}
d\left(f \circ \alpha^{z}, \alpha^{-z} \Omega, y\right)[\hat{c}] & \equiv\left(\hat{f} \circ \theta \circ \alpha^{z} \circ \theta^{-1}\right)_{*}[\hat{c}] \\
& =\hat{f}_{*}\left(\theta \circ \alpha^{z} \circ \theta^{-1}\right)_{*}[\hat{c}]=\hat{f}_{*}[\hat{c}] \equiv d(f, \Omega, y)[\hat{c}]
\end{aligned}
$$

Suppose $f$ is one to one on $\overline{B(w, r)}$ and we want to consider the case where $D f(w)$ is invertible. From the above lemma,

$$
\begin{align*}
d(f, B(w, r), f(w)) & =d(f-f(w), B(w, r), 0) \\
& =d(f((\cdot)+w)-f(w), B(0, r), 0) \tag{1.23}
\end{align*}
$$

Lemma A.14.2 Let $f: \overline{B(w, R)} \rightarrow \mathbb{R}^{n}$ be such that $f^{-1}(f(w))=\{w\}$ and suppose $D f(w)$ is invertible. Then $d(f, B(w, R), f(w))=\operatorname{sgn}(\operatorname{det}(D f(w)))$.

Proof: Referring to 1.23 , let $g(x) \equiv f(x+w)-f(w)$. Thus $D g(0)=D f(w)$. Denote by $B_{r}$ the ball with center at 0 and radius $r$. By Corollary A.11.11 and Lemma 1.23,

$$
d(f, B(w, R), f(w))=d(g, B(0, R), 0)=d\left(g, B_{r}, 0\right)
$$

for all $0<r<R$ where $B_{r}$ is centered at 0 with radius $r$. Thus it suffices to consider $d\left(g, B_{r}, 0\right)$. For $x \in \partial B_{r}$, consider $\operatorname{tg}(x)+(1-t) D g(0) x$ for $t \in[0,1]$.

$$
\operatorname{tg}(x)+(1-t) D g(0) x=t(g(x)-D g(0) x)+D g(0) x
$$

Thus, if for some $x_{r} \in \partial B_{r}$ and $t \in[0,1]$, the above is 0 , then

$$
\left|D g(0) \frac{x_{r}}{r}\right| \leq \frac{\left|g\left(x_{r}\right)-D g(0) x_{r}\right|}{r} .
$$

Now $\lim _{r \rightarrow 0} \frac{\left|g\left(x_{r}\right)-D g(0) x\right|}{r}=0$ by differentiability and so, since $\frac{x_{r}}{r}$ is a unit vector, there is a subsequence converging to $y$ another unit vector as $r \rightarrow 0$. Therefore, $|D g(0) y|=0$ contrary to the assumption that $D g(0)$ is invertible. It follows that there exists some $r \leq R$ such that $0 \notin(t g+(1-t) D g(0))\left(\partial B_{r}\right)$ and now it follows from Proposition A.12.2 that $d\left(g, B_{r}, 0\right)=d\left(D g(0), B_{r}, 0\right)$ which is the sign of the determinant of $D g(0)$, either 1 or -1 because $\operatorname{Dg}(0)$ is the product of the elementary matrices described in that proposition.

Lemma A.14.3 Let $\Omega$ be a bounded open set and let $y \notin f(\partial \Omega)$ where $f: \bar{\Omega} \rightarrow \mathbb{R}^{n}$ is continuous and $f$ is differentiable on $\Omega$ with $\operatorname{det}(D f(x)) \neq 0$ for each $x \in \Omega$. Then there are finitely many $x_{i} \in f^{-1}(y)$ each the center of a ball $B_{i}$ where $\bar{B}_{i} \cap \bar{B}_{j}=\emptyset$ and $f$ is one to one on $\bar{B}_{i}$. Then

$$
d(f, \Omega, y)=\sum_{i} d\left(f, B_{i}, y\right)=\sum_{i} \operatorname{sgn}\left(\operatorname{det}\left(D f\left(x_{i}\right)\right)\right)
$$

Proof: That a collection of points $x_{i} \in f^{-1}(y)$ exists together with disjoint balls $B_{i}$ centered at $x_{i}$ having disjoint closures on which $f$ is one to one follows from the inverse function theorem. If there were infinitely many of these $x_{i}$ then a subsequence of distinct points would converge to a point $z$ of $\bar{\Omega}$ which must satisfy $f(z)=y$ and so by assumption, $y \notin \partial \Omega$. But now, by the inverse function theorem, there would be a ball $B_{z}$ containing $z$ on which $f$ is one to one which is a contradiction since this ball must contain some $x_{i} \neq z$. Thus there are only finitely many. The formula follows from Lemma A.14.2 and Proposition A.12.2 part 3.

From Lemma A.14.2, we know $d(f, B(w, R), f(w))$ whenever $f$ is one to one on $B(w, R)$. It is just $\operatorname{sgn}(\operatorname{det}(D f(w)))$. See my book "Real and Abstract Analysis" for a treatment of the degree starting with this.

If $f: U \subseteq \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ is differentiable, and if $S$ is the set of $x$ where $\operatorname{det}(D f(x))=0$ then Sard's lemma says that $f(S)$ has measure zero. I will use this fact in the proof of the following lemma. This is in both of my books in the chapter on change of variables theorems "Analysis of Functions of Complex and Many Variables" or "Real and Abstract Analysis".

Lemma A.14.4 Let $h_{\eta}$ be a polynomial meaning each component is a polynomial, with $\operatorname{det}\left(D h_{\eta}(0)\right) \neq 0, h_{\eta}(0)=0$, and let $\eta>0$. Then there are vectors $y^{k}$ each with $\left|y^{k}\right|<\eta$ such that $h(x) \equiv h_{\eta}(x)-\sum_{k=1}^{n} x_{k}^{3} y^{k}$ has 0 as a regular value meaning that if $h(x)=0$, then $D h(x)$ is invertible.

Proof: Let $V_{k}$ consist of all $x \in \mathbb{R}^{n}$ such that $x_{k} \neq 0$. Consider $f(x) \equiv \frac{h_{\eta}(x)}{x_{1}^{3}}$. Then $x_{1}^{3} D f(x)+3 x_{1}^{2} f(x) J_{1}=D h_{\eta}(x)$ where $J_{1}$ has the first column ones, the others 0 . So

$$
\begin{equation*}
D f(x)=x_{1}^{-3}\left(D h_{\eta}(x)-3 x_{1}^{2} f(x) J_{1}\right) \tag{1.24}
\end{equation*}
$$

Now the singular set for $f$ in $V_{1}$ where $\operatorname{det}(D f(x))=0$ called $S$, has $f(S)$ with measure zero. In particular there exists $y^{1} \in B(0, \eta)$ a regular value for $f$ on $V_{k}$. Then $\hat{h}(x) \equiv h_{\eta}(x)-$ $x_{1}^{3} y^{1}=x_{1}^{3}\left(f(x)-y^{1}\right)$ sends 0 to 0 and $D \hat{h}(0)$ is invertible while at the other points $x$ where $\hat{h}(x)=0$, we have $f(x)=y^{1}$ and so $D f(x)$ is invertible. Thus $D \hat{h}(x)=3 x_{1}^{2}\left(f(x)-y^{1}\right) I+$ $x_{1}^{3} D f(x)=x_{1}^{3} D f(x)$ which is invertible for any $x \in V_{1}$ for which $\hat{h}(x)=0$. What if $x_{1}=0$ but $D h_{\eta}(x)$ is invertible, does this continue to hold for $\hat{h}$ ? From the above, $D \hat{h}(x)=$ $D h_{\eta}(x)-3 x_{1}^{2} y^{1} J$ where all columns of $J$ are zero except the first, so at $x_{1}=0, D \hat{h}(x)=$ $D h_{\eta}(x)$ which is invertible.

Now let $\hat{h}$ be the new $h_{\eta}$ and do the same construction with $V_{2}$. Iterate this process eventually getting $h(x)=h_{\eta}(x)-\sum_{k=1}^{n} x_{k}^{3} y^{k}$ such that 0 is a regular value for this function on $\{0\} \cup V_{1} \cup V_{2} \cdots \cup V_{n}=\mathbb{R}^{n}$.

Lemma A.14.5 Let $f: \bar{B} \rightarrow \mathbb{R}^{n}$ where $B$ is a ball of radius $R$ centered at 0 and $f$ is odd and has each component function a polynomial and $0 \notin f(\partial B)$. Then $d(f, B, 0)$ is an odd integer.

Proof: Clearly $f(0)=0$ because $f$ is odd. First assume that $D f(x)$ is invertible for all $x \in f^{-1}(0)$. Then from Lemma A.14.3 there are finitely many open balls, $B_{0}$ centered at 0 and $B_{i}$ centered at $x_{i} \in f^{-1}(0)$ such that $f^{-1}(0)$ is contained in the union of these balls, their closures are disjoint, and if $B_{i}$ is one of these balls with center at $x_{i}, i=0,1, \ldots, 2 m$ then $B_{i} \cap f^{-1}(0)=\left\{x_{i}\right\}$. These disjoint open balls come in pairs $B,-B$ since $f$ is odd. Now from Lemma A.14.2 and the fact that $f$ is odd, $d\left(f, B\left(x_{i}, r_{i}\right), 0\right)=d\left(f,-B\left(x_{i}, r_{i}\right), 0\right)=$ $\operatorname{sgn}\left(\operatorname{det}\left(D f\left(x_{i}\right)\right)\right)$ for $i>0$. Thus $d(f, B, 0)$ is an odd integer.

If this is not so that $D f(x)$ is invertible for all $x \in f^{-1}(0)$, pick $\eta \neq 0$ small enough that $h_{\eta} \equiv f+\eta I$ has $D h_{\eta}(0)$ invertible (Problem 11 on Page 190) and also that $d\left(h_{\eta}, B, 0\right)=$ $d(f, B, 0)$. Then if $\eta$ is still smaller so that $h$ in the above Lemma A.14.4 has $d(h, B, 0)=$ $d(f, B, 0)$ the first part applies to $h$ and proves the desired result.

Actually there is an easier way to prove a simpler version of Lemma A.14.4 since $h_{\eta}$ is a polynomial and the iterates in the process are also polynomials which will suffice in the above. Note that the singular sets $S$ all have measure zero because the set where $\operatorname{det}(D f(x))$ equals 0 is the set where $\operatorname{det}(D g(x))=0$ for some $g$ a polynomial in each step of the iteration, and $x_{1} \rightarrow \operatorname{det}(D g(x))$ has only finitely many zeros, being a polynomial. It is also relatively easy in comparison to proving Sard's lemma, to verify that if $f$ has one continuous derivative on an open set $V$ containing $S$ a closed set of measure zero, then $f(S)$ has measure zero. In particular $f(S)$ does not contain any ball, making the iteration in that lemma possible.

Next we need to reduce to the case of this lemma for odd continuous functions. Say $f$ is odd and defined on $\bar{B}$. Then $\frac{1}{2}(f(x)+(-f(-x)))=f(x)$. Using the Weierstrass approximation theorem of Chapter 16, there is a polynomial $g$ which is close enough to $f$ on $\bar{B}$ that $d(f, B, 0)=d(g, B, 0)$. Let $\hat{f}(x)=\frac{1}{2}(g(x)+(-g(-x)))$. Then $\hat{f}(-x)=$ $\frac{1}{2}(g(-x)+(-g(x)))=-\frac{1}{2}(g(x)+(-g(-x)))=-\hat{f}(x)$. Thus $\hat{f}$ is also a polynomial, is odd, and is just as close to $f$ as is $g$, so by Lemma A.14.5 $d(f, B, 0)=d(g, B, 0)=$ $d(\hat{f}, B, 0)$ an odd integer. Thus we get the following theorem.

Theorem A.14.6 Let $B$ be an open ball centered at 0 and let $f: \bar{B} \rightarrow \mathbb{R}^{n}$ be odd with $0 \notin f(\partial B)$. Then $d(f, B, 0)$ is an odd integer. In particular it is not zero.

Note that you could replace the ball with a symmetric open set $\Omega$ which means $x \in \Omega$ if and only if $-x \in \Omega$. There would be no change in the argument.

## Appendix B

## The Hausdorff Maximal Theorem

First is the definition of what is meant by a partial order.
Definition B.0.1 A nonempty set $\mathscr{F}$ is called a partially ordered set if it has a partial order denoted by $\prec$. This means it satisfies the following. If $x \prec y$ and $y \prec z$, then $x \prec z$. Also $x \prec x$. It is like $\subseteq$ on the set of all subsets of a given set. It is not the case that given two elements of $\mathscr{F}$ that they are related. In other words, you cannot conclude that either $x \prec y$ or $y \prec x$. A chain, denoted by $\mathscr{C} \subseteq \mathscr{F}$ has the property that it is totally ordered meaning that if $x, y \in \mathscr{C}$, either $x \prec y$ or $y \prec x$. A maximal chain is a chain $\mathscr{C}$ which has the property that there is no strictly larger chain. In other words, if $x \in \mathscr{F} \backslash \cup \mathscr{C}$, then $\mathscr{C} \cup\{x\}$ is no longer a chain so $x$ fails to be related to something in $\mathscr{C}$.

Here is the Hausdorff maximal theorem. The proof is a proof by contradiction. We assume there is no maximal chain and then show this cannot happen. The axiom of choice is used in choosing the $x_{\mathscr{C}}$ right at the beginning of the argument.

Theorem B.0. 2 Let $\mathscr{F}$ be a nonempty partially ordered set with order $\prec$. Then there exists a maximal chain.

Proof: For $\mathscr{C}$ a chain, let $\theta \mathscr{C}$ denote $\mathscr{C} \cup\left\{x_{\mathscr{C}}\right\}$. Thus for $\mathscr{C}$ a chain, $\theta \mathscr{C}$ is a larger chain which has exactly one more element of $\mathscr{F}$. Since $\mathscr{F} \neq \emptyset$, pick $x_{0} \in \mathscr{F}$. Note that $\left\{x_{0}\right\}$ is a chain. Let $\mathscr{X}$ be the set of all chains $\mathscr{C}$ such that $x_{0} \in \cup \mathscr{C}$. Thus $\mathscr{X}$ contains $\left\{x_{0}\right\}$. Call two chains comparable if one is a subset of the other. Also, if $\mathscr{S}$ is a nonempty subset of $\mathscr{F}$ in which all chains are comparable, then $\cup \mathscr{S}$ is also a chain. From now on $\mathscr{S}$ will always refer to a nonempty set of chains in which any pair are comparable. Then summarizing,

1. $x_{0} \in \cup \mathscr{C}$ for all $\mathscr{C} \in \mathscr{X}$.
2. $\left\{x_{0}\right\} \in \mathscr{X}$
3. If $\mathscr{C} \in \mathscr{X}$ then $\theta \mathscr{C} \in \mathscr{X}$.
4. If $\mathscr{S} \subseteq \mathscr{X}$ then $\cup \mathscr{S} \in \mathscr{X}$.

A subset $\mathscr{Y}$ of $\mathscr{X}$ will be called a "tower" if $\mathscr{Y}$ satisfies 1.) - 4.). Let $\mathscr{Y}_{0}$ be the intersection of all towers. Then $\mathscr{Y}_{0}$ is also a tower, the smallest one. Then the next claim might seem to be so because if not, $\mathscr{Y}_{0}$ would not be the smallest tower.

Claim 1: If $\mathscr{C}_{0} \in \mathscr{Y}_{0}$ is comparable to every chain $\mathscr{C} \in \mathscr{Y}_{0}$, then if $\mathscr{C}_{0} \subsetneq \mathscr{C}$, it must be the case that $\theta \mathscr{C}_{0} \subseteq \mathscr{C}$. In other words, $x_{\mathscr{C}_{0}} \in \cup \mathscr{C}$. The symbol $\subsetneq$ indicates proper subset.

This is done by considering a set $\mathscr{B} \subseteq \mathscr{Y}_{0}$ consisting of $\mathscr{D}$ which acts like $\mathscr{C}$ in the above and showing that it actually equals $\mathscr{Y}_{0}$ because it is a tower.

Proof of Claim 1: Consider $\mathscr{B} \equiv\left\{\mathscr{D} \in \mathscr{Y}_{0}: \mathscr{D} \subseteq \mathscr{C}_{0}\right.$ or $\left.x_{\mathscr{C}_{0}} \in \cup \mathscr{D}\right\}$. Let $\mathscr{Y}_{1} \equiv \mathscr{Y}_{0} \cap \mathscr{B}$. I want to argue that $\mathscr{Y}_{1}$ is a tower. By definition all chains of $\mathscr{Y}_{1}$ contain $x_{0}$ in their unions. If $\mathscr{D} \in \mathscr{Y}_{1}$, is $\theta \mathscr{D} \in \mathscr{Y}_{1}$ ? If $\mathscr{S} \subseteq \mathscr{Y}$, is $\cup \mathscr{S} \in \mathscr{Y}_{1}$ ? Is $\left\{x_{0}\right\} \in \mathscr{B}$ ?
$\left\{x_{0}\right\}$ cannot properly contain $\mathscr{C}_{0}$ since $x_{0} \in \cup \mathscr{C}_{0}$. Therefore, $\mathscr{C}_{0} \supseteq\left\{x_{0}\right\}$ so $\left\{x_{0}\right\} \in \mathscr{B}$.
If $\mathscr{S} \subseteq \mathscr{Y}_{1}$, and $\mathscr{D} \equiv \cup \mathscr{S}$, is $\mathscr{D} \in \mathscr{Y}_{1}$ ? Since $\mathscr{Y}_{0}$ is a tower, $\mathscr{D}$ is comparable to $\mathscr{C}_{0}$. If $\mathscr{D} \subseteq \mathscr{C}_{0}$, then $\mathscr{D}$ is in $\mathscr{B}$. Otherwise $\mathscr{D} \supseteq \mathscr{C}_{0}$ and in this case, why is $\mathscr{D}$ in $\mathscr{B}$ ? Why is $x_{\mathscr{C}_{0}} \in \cup \mathscr{D}$ ? The chains of $\mathscr{S}$ are in $\mathscr{B}$ so one of them, called $\tilde{\mathscr{C}}$ must properly contain $\mathscr{C}_{0}$
and so $x_{\mathscr{C}_{0}} \in \cup \tilde{C} \subseteq \cup \mathscr{D}$. Therefore, $\mathscr{D} \in \mathscr{B} \cap \mathscr{Y}_{0}=\mathscr{Y}_{1}$. 4.) holds. Two cases remain, to show that $\mathscr{Y}_{1}$ satisfies 3.).
case 1: $\mathscr{D} \supseteq \mathscr{C}_{0}$. Then by definition of $\mathscr{B}, x_{\mathscr{C}_{0}} \in \cup \mathscr{D}$ and so $x_{\mathscr{C}_{0}} \in \cup \theta \mathscr{D}$ so $\theta \mathscr{D} \in \mathscr{Y}_{1}$.
case 2: $\mathscr{D} \subseteq \mathscr{C}_{0} . \theta \mathscr{D} \in \mathscr{Y}_{0}$ so $\theta \mathscr{D}$ is comparable to $\mathscr{C}_{0}$. First suppose $\theta \mathscr{D} \supsetneq \mathscr{C}_{0}$. Thus $\mathscr{D} \subseteq \mathscr{C}_{0} \varsubsetneqq \mathscr{D} \cup\left\{x_{\mathscr{D}}\right\}$. If $x \in \mathscr{C}_{0}$ and $x$ is not in $\mathscr{D}$ then $\mathscr{D} \cup\{x\} \subseteq \mathscr{C}_{0} \subsetneq \mathscr{D} \cup\left\{x_{\mathscr{D}}\right\}$. This is impossible. Consider $x$. Thus in this case that $\theta \mathscr{D} \supsetneq \mathscr{C}_{0}, \mathscr{D}=\mathscr{C}_{0}$. It follows that $x_{\mathscr{D}}=x_{\mathscr{C}_{0}} \in \cup \theta \mathscr{C}_{0}=\cup \theta \mathscr{D}$ and so $\theta \mathscr{D} \in \mathscr{Y}_{1}$. The other case is that $\theta \mathscr{D} \subseteq \mathscr{C}_{0}$ so $\theta \mathscr{D} \in \mathscr{B}$ by definition. This shows 3.) so $\mathscr{Y}_{1}$ is a tower and must equal $\mathscr{Y}_{0}$.

Claim 2: Any two chains in $\mathscr{Y}_{0}$ are comparable.
Proof of Claim 2: Let $\mathscr{Y}_{1}$ consist of all chains of $\mathscr{Y}_{0}$ which are comparable to every chain of $\mathscr{Y}_{0} .\left\{x_{0}\right\}$ is in $\mathscr{Y}_{1}$ by definition. All chains of $\mathscr{Y}_{0}$ have $x_{0}$ in their union. If $\mathscr{S} \subseteq \mathscr{Y}_{1}$, is $\cup \mathscr{S} \in \mathscr{Y}_{1}$ ? Given $\mathscr{D} \in \mathscr{Y}_{0}$ either every chain of $\mathscr{S}$ is contained in $\mathscr{D}$ or at least one contains $\mathscr{D}$. Either way $\mathscr{D}$ is comparable to $\cup \mathscr{S}$ so $\cup \mathscr{S} \in \mathscr{Y}_{1}$. It remains to show 3.). Let $\mathscr{C} \in \mathscr{Y}_{1}$ and $\mathscr{D} \in \mathscr{Y}_{0}$. Since $\mathscr{C}$ is comparable to all chains in $\mathscr{Y}_{0}$, it follows from Claim 1 either $\mathscr{C} \subsetneq \mathscr{D}$ when $x_{\mathscr{C}} \in \cup \mathscr{D}$ and $\theta \mathscr{C} \subseteq \mathscr{D}$ or $\mathscr{C} \supseteq \mathscr{D}$ when $\theta \mathscr{C} \supseteq \mathscr{D}$. Hence $\mathscr{Y}_{1}=\mathscr{Y}_{0}$ because $\mathscr{T}_{0}$ is as small as possible.

Since every pair of chains in $\mathscr{Y}_{0}$ are comparable and $\mathscr{Y}_{0}$ is a tower, it follows that $\cup \mathscr{Y}_{0} \in \mathscr{Y}_{0}$ so $\cup \mathscr{Y}_{0}$ is a chain. However, $\theta \cup \mathscr{Y}_{0}$ is a chain which properly contains $\cup \mathscr{Y}_{0}$ and since $\mathscr{Y}_{0}$ is a tower, $\theta \cup \mathscr{Y}_{0} \in \mathscr{Y}_{0}$. Thus $\cup\left(\theta \cup \mathscr{Y}_{0}\right) \supseteq \cup\left(\cup \mathscr{Y}_{0}\right) \supseteq \cup\left(\theta \cup \mathscr{Y}_{0}\right)$ which is a contradiction. Therefore, for some chain $\mathscr{C}$ it is impossible to obtain the $x_{C}$ described above and so, this $\mathscr{C}$ is a maximal chain.

If $X$ is a nonempty set, $\leq$ is an order on $X$ if

$$
x \leq x
$$

and if $x, y \in X$, then

$$
\text { either } x \leq y \text { or } y \leq x
$$

and

$$
\text { if } x \leq y \text { and } y \leq z \text { then } x \leq z
$$

$\leq$ is a well order and say that $(X, \leq)$ is a well-ordered set if every nonempty subset of $X$ has a smallest element. More precisely, if $S \neq \emptyset$ and $S \subseteq X$ then there exists an $x \in S$ such that $x \leq y$ for all $y \in S$. A familiar example of a well-ordered set is the natural numbers.

Lemma B.0.3 The Hausdorff maximal principle implies every nonempty set can be wellordered.

Proof: Let $X$ be a nonempty set and let $a \in X$. Then $\{a\}$ is a well-ordered subset of $X$. Let

$$
\mathscr{F}=\{S \subseteq X: \text { there exists a well order for } S\}
$$

Thus $\mathscr{F} \neq \emptyset$. For $S_{1}, S_{2} \in \mathscr{F}$, define $S_{1} \prec S_{2}$ if $S_{1} \subseteq S_{2}$ and there exists a well order for $S_{2}$, $\leq_{2}$ such that

$$
\left(S_{2}, \leq_{2}\right) \text { is well-ordered }
$$

and if

$$
y \in S_{2} \backslash S_{1} \text { then } x \leq_{2} y \text { for all } x \in S_{1},
$$

and if $\leq_{1}$ is the well order of $S_{1}$ then the two orders are consistent on $S_{1}$. Then observe that $\prec$ is a partial order on $\mathscr{F}$. By the Hausdorff maximal principle, let $\mathscr{C}$ be a maximal chain in $\mathscr{F}$ and let

$$
X_{\infty} \equiv \cup \mathscr{C}
$$

Define an order, $\leq$, on $X_{\infty}$ as follows. If $x, y$ are elements of $X_{\infty}$, pick $S \in \mathscr{C}$ such that $x, y$ are both in $S$. Then if $\leq_{S}$ is the order on $S$, let $x \leq y$ if and only if $x \leq_{S} y$. This definition is well defined because of the definition of the order, $\prec$. Now let $U$ be any nonempty subset of $X_{\infty}$. Then $S \cap U \neq \emptyset$ for some $S \in \mathscr{C}$. Because of the definition of $\leq$, if $y \in S_{2} \backslash S_{1}, S_{i} \in \mathscr{C}$, then $x \leq y$ for all $x \in S_{1}$. Thus, if $y \in X_{\infty} \backslash S$ then $x \leq y$ for all $x \in S$ and so the smallest element of $S \cap U$ exists and is the smallest element in $U$. Therefore $X_{\infty}$ is well-ordered. Now suppose there exists $z \in X \backslash X_{\infty}$. Define the following order, $\leq_{1}$, on $X_{\infty} \cup\{z\}$.

$$
\begin{aligned}
& x \leq_{1} y \text { if and only if } x \leq y \text { whenever } x, y \in X_{\infty} \\
& \qquad x \leq_{1} z \text { whenever } x \in X_{\infty} .
\end{aligned}
$$

Then let

$$
\tilde{\mathscr{C}}=\left\{S \in \mathscr{C} \text { or } X_{\infty} \cup\{z\}\right\}
$$

Then $\tilde{\mathscr{C}}$ is a strictly larger chain than $\mathscr{C}$ contradicting maximality of $\mathscr{C}$. Thus $X \backslash X_{\infty}=\emptyset$ and this shows $X$ is well-ordered by $\leq$.

With these two lemmas the main result follows.
Theorem B.0.4 The following are equivalent.
The axiom of choice
The Hausdorff maximal principle
The well-ordering principle.
Proof: It only remains to prove that the well-ordering principle implies the axiom of choice. Let $I$ be a nonempty set and let $X_{i}$ be a nonempty set for each $i \in I$. Let $X=\cup\left\{X_{i}\right.$ : $i \in I\}$ and well order $X$. Let $f(i)$ be the smallest element of $X_{i}$. Then $f \in \prod_{i \in I} X_{i}$.

## B. 1 The Hamel Basis

A Hamel basis is nothing more than the correct generalization of the notion of a basis for a finite dimensional vector space to vector spaces which are possibly not of finite dimension.

Definition B.1. 1 Let $X$ be a vector space. A Hamel basis is a subset of $X, \Lambda$ such that every vector of $X$ can be written as a finite linear combination of vectors of $\Lambda$ and the vectors of $\Lambda$ are linearly independent in the sense that if $\left\{x_{1}, \cdots, x_{n}\right\} \subseteq \Lambda$ and

$$
\sum_{k=1}^{n} c_{k} x_{k}=0
$$

then each $c_{k}=0$.
The main result is the following theorem.
Theorem B.1.2 Let X be a nonzero vector space. Then it has a Hamel basis.

Proof: Let $x_{1} \in X$ and $x_{1} \neq 0$. Let $\mathscr{F}$ denote the collection of subsets of $X, \Lambda$ containing $x_{1}$ with the property that the vectors of $\Lambda$ are linearly independent as described in Definition B.1.1 partially ordered by set inclusion. By the Hausdorff maximal theorem, there exists a maximal chain, $\mathscr{C}$ Let $\Lambda=\cup \mathscr{C}$. Since $\mathscr{C}$ is a chain, it follows that if $\left\{x_{1}, \cdots, x_{n}\right\} \subseteq \mathscr{C}$ then there exists a single $\Lambda^{\prime} \in \mathbb{C}$ containing all these vectors. Therefore, if

$$
\sum_{k=1}^{n} c_{k} x_{k}=0
$$

it follows each $c_{k}=0$. Thus the vectors of $\Lambda$ are linearly independent. Is every vector of $X$ a finite linear combination of vectors of $\Lambda$ ?

Suppose not. Then there exists $z$ which is not equal to a finite linear combination of vectors of $\Lambda$. Consider $\Lambda \cup\{z\}$. If

$$
c z+\sum_{k=1}^{m} c_{k} x_{k}=0
$$

where the $x_{k}$ are vectors of $\Lambda$, then if $c \neq 0$ this contradicts the condition that $z$ is not a finite linear combination of vectors of $\Lambda$. Therefore, $c=0$ and now all the $c_{k}$ must equal zero because it was just shown $\Lambda$ is linearly independent. It follows $\mathscr{C} \cup\{\Lambda \cup\{z\}\}$ is a strictly larger chain than $\mathscr{C}$ and this is a contradiction. Therefore, $\Lambda$ is a Hamel basis as claimed.

## B. 2 Exercises

1. Zorn's lemma states that in a nonempty partially ordered set, if every chain has an upper bound, there exists a maximal element, $x$ in the partially ordered set. $x$ is maximal, means that if $x \prec y$, it follows $y=x$. Show Zorn's lemma is equivalent to the Hausdorff maximal theorem.
2. Show that if $Y, Y_{1}$ are two Hamel bases of $X$, then there exists a one to one and onto map from $Y$ to $Y_{1}$. Thus any two Hamel bases are of the same size.

## Bibliography

[1] Apostol T. Calculus Volume II Second edition, Wiley 1969.
[2] Apostol, T. Mathematical Analysis, Addison Wesley Publishing Co., 1974.
[3] Artin M., Algebra, Pearson 2011.
[4] Baker, Roger, Linear Algebra, Rinton Press 2001.
[5] Baker, A. Transcendental Number Theory, Cambridge University Press 1975.
[6] Baker, Roger, and Kuttler, K. Linear Algebra With Applications, with Roger Baker, World Scientific March (2014). (311 pages)
[7] Birkhoff G. and Maclane S., Algebra, Macmillan 1965.
[8] Davis H. and Snider A., Vector Analysis Wm. C. Brown 1995.
[9] Diestal J. and Uhl J., Vector Measures, American Math. Society, Providence, R.I., 1977.
[10] Edwards C.H. Advanced Calculus of several Variables, Dover 1994.
[11] Gaal L. Classical Galois Theory, Chelsea publishing company, New York 1973
[12] Chahal J.S., Historical Perspective of Mathematics 2000 B.C. - 2000 A.D. Kendrick Press, Inc. (2007)
[13] Cheney, E. W., Introduction To Approximation Theory, McGraw Hill 1966.
[14] Donal O'Regan, Yeol Je Cho, and Yu-Qing Chen, Topological Degree Theory and Applications, Chapman and Hall/CRC 2006.
[15] Golub, G. and Van Loan, C.,Matrix Computations, Johns Hopkins University Press, 1996.
[16] Greenberg M.D. Advanced Engineering Mathematics Prentice Hall 1998 Second edition.
[17] Friedberg S. Insel A. and Spence L., Linear Algebra, Prentice Hall, 2003.
[18] Gurtin M. An introduction to continuum mechanics, Academic press 1981.
[19] Hardy G. A Course Of Pure Mathematics, Tenth edition, Cambridge University Press 1992.
[20] Herstein I. N., Topics In Algebra, Xerox, 1964.
[21] Hewitt E. and Stromberg K. Real and Abstract Analysis, Springer-Verlag, New York, 1965.
[22] Hobson E.W., The Theory of functions of a Real Variable and the Theory of Fourier's Series V. 1, Dover 1957.
[23] Hocking J. and Young G., Topology, Addison-Wesley Series in Mathematics, 1961.
[24] Hofman K. and Kunze R., Linear Algebra, Prentice Hall, 1971.
[25] Horn R. and Johnson C. matrix Analysis, Cambridge University Press, 1985.
[26] Jacobsen N. Basic Algebra Freeman 1974.
[27] Karlin S. and Taylor H., A First Course in Stochastic Processes, Academic Press, 1975.
[28] Kreyszig E. Introductory Functional Analysis With applications, Wiley 1978.
[29] Kuttler K.L., Modern Analysis CRC Press 1998.
[30] Kuttler K., Linear Algebra On web page. Linear Algebra
[31] Kuttler K., Calculus Theory and Applications Vol. 2 World Scientific 2011.
[32] Marsden J. E. and Hoffman J. M., Elementary Classical Analysis, Freeman, 1993.
[33] Marcus M., and Minc H., A Survey Of Matrix Theory and Matrix Inequalities, Allyn and Bacon, INc. Boston, 1964
[34] McShane E. J. Integration, Princeton University Press, Princeton, N.J. 1944.
[35] Nobel B. and Daniel J. Applied Linear Algebra, Prentice Hall, 1977.
[36] Rudin W. Principles of Mathematical Analysis, McGraw Hill, 1976.
[37] Rudin W. Functional Analysis, second edition, McGraw-Hill, 1991.
[38] Salas S. and Hille E., Calculus One and Several Variables, Wiley 1990.
[39] Spanier E., Algebraic Topology, McGraw Hill 1966.
[40] Vick, J., homology theory, An Introduction to Algebraic Topology, Academic Press 1973.
[41] Strang Gilbert, Linear Algebra and its Applications, Harcourt Brace Jovanovich 1980.
[42] Wilkinson, J.H., The Algebraic Eigenvalue Problem, Clarendon Press Oxford 1965.
[43] Yosida K., Functional Analysis, Springer Verlag, 1978.

## Index

$\cap, 1$
$\cup, 1$
$\varepsilon$ net, 253
A close to B
eigenvalues, 336
Abel's formula, 189, 194
Abelian group, 51
free, 455
absolute convergence
convergence, 394
absolute value
complex number, 10
accumulation point, 246
adjoint, 329
of matrix, 275
adjugate, 180
affine maps, 470
algebraic number
minimum polynomial, 70
algebraic numbers, 69
field, 71
alternating group, 228
3 cycles, 229
analytic function of matrix, 302
arcwise connected, 261
ascending chains, 153
at most countable, 4
automorphism, 211
axiom of choice, 4
barycenter, 469
basis, 262
existence, 55
basis of eigenvectors
diagonalizable, 136
basis of vector space, 54
Bernstein polynomial
approximation of derivative, 444
Binet Cauchy
volumes, 326
Binet Cauchy formula, 177
binomial theorem, 32
block diagonal matrices
direct sum of subspaces, 120
block diagonal matrix, 120
block matrix, 88
block multiplication, 88,89
Borsuk Ulam theorem, 497
boundary operator, 457
bounded linear transformations, 274
Brouwer
fixed point theorem, 482, 483, 493
Cauchy interlacing theorem, 348, 350
Cauchy Schwarz inequality, 264
Cauchy sequence, 248, 324
Cayley Hamilton theorem, 186, 191, 192, 384
chain, 503
chain complex, 478
chain homomorphisms, 478
chain homotopy, 465
characteristic polynomial, 185
Cholesky factorization, 387
closed set, 246
closed sets
limit points, 246
closure of a set, 250
cofactor, 178
column rank, 183
commutative ring, 67, 195
Noetherian, 169
commutative ring with unity, 28
commutator, 232, 348
commutator subgroup, 232
compact
sequentially compact, 283
compact set, 253
compactness
closed interval, 249
equivalent conditions, 253
companion matrix, 419
complete, 399
complex conjugate, 10
complex numbers, 8
complex numbers
arithmetic, 8
roots, 12
triangle inequality, 10
components of a vector, 262
composition of linear transformations, 111
condition number, 390
cone, 468
conjugate
of a product, 33
conjugate fields, 237
conjugate linear, 314
connected, 259
connected component, 260
connected components, 260
equivalence class, 260
equivalence relation, 260
open sets, 260
connected sets
intersection, 259
intervals, 260
real line, 260
connecting sequence, 475
consistent, 44
continuous function, 251
continuous functions, 284
equivalent conditions, 251
contraction map, 255
fixed point theorem, 255
convex combination, 116, 280
convex hull, 116, 280
compactness, 116
convex set
homology, 461
Coordinates, 53
countable, 4
counting zeros, 336
Courant Fischer theorem, 369
cow lick, 493
Cramer's rule, 181
cyclic basis, 148
cyclic set, 147
De Moivre's theorem, 12
degree
antipodal map, 486
Euclidean space, 491
odd function, 501
on spheres, 484
orientation, 496
product formula, 491
properties, 492
regular value, 500
switching variables, 485
well defined, 488
derivative, 284
determinant
definition, 174
estimate for Hermitian matrix, 388
expansion along row, column, 179
Hadamard inequality, 388
matrix inverse, 179
partial derivative, cofactor, 190
permutation of rows, 174
product, 176
product of eigenvalues, 341
row, column operations, 175
summary of properties, 185
symmetric definition, 175
transpose, 175
diagonal matrix, 135
diagonalizability, 135
diagonalizable, 135, 357
formal derivative, 140
minimal polynomial and its derivative, 140
differential equations
first order systems, 343
dimension of a vector space, 262
dimension of vector space, 54
direct sum, 119, 158, 460
minimum polynomial splits, 162
notation, 119
discrete Fourier transform, 384
distance, 245
to a subspace, 313
distance to a nonempty set, 247
distinct eigenvalues, 147
distinct roots
polynomial and its derivative, 239
dot product, 263
dyadics, 103
echelon form, 41
eigen-pair, 132
eigenvalue, 132
existence, 133
eigenvalues, 186, 336
AB and BA, 187
eigenvector, 132
existence, 133
eigenvectors
distinct eigenvalues, 147
independent, 147
elementary matrices, 91
product, 95
elementary matrix
inverse, 95
properties, 95
elementary operations, 39
elementary symmetric polynomials, 196
empty set, 1
equivalence class, $6,63,107$
of polynomials, 64
equivalence relation, $6,63,107$
Euclidean algorithm, 19
exchange theorem, 54
existence of a fixed point, 401
factorization of matrix
Euclidean domain, 151
field axioms, 8
field extension
dimension, 67
finite, 67
field extensions, 67
Field of scalars, 51
fields
characteristic, 240
perfect, 241
fields
perfect, 240
finite dimensional vector space, 54
finite fields, 27
fixed field, 219
fixed fields and subgroups, 224
fixed point property, 483
formal derivative, 140
Fourier series, 323
Fredholm alternative, 318
Frobenius
inner product, 347
Frobenius norm, 275, 374
singular value decomposition, 374
Frobinius norm, 383
function, 3
fundamental theorem of algebra, 13, 15, 210
fundamental theorem of algebra
plausibility argument, 15
rigorous proof, 16
fundamental theorem of arithmetic, 20
fundamental theorem of Galois theory, 226
Galois group
size, 217
Gauss Elimination, 44
Gauss elimination, 40

Gauss Jordan method for inverses, 82
Gauss Seidel method, 402
geometric simplices boundary, 466
Gerschgorin's theorem, 334
Gram Schmidt process, 271, 309
Grammian matrix, 312, 319 invertible, 312
greatest common divisor, 19, 22
characterization, 19
description, 22
Gronwall's inequality, 407
group
definition, 217
group
solvable, 232
Hadamard
inequality, 388
Hamel basis, 505
Hausdorff
maximal principle, 503
Hermitian, 331
orthonormal basis eigenvectors, 367
positive definite, 371
Hermitian matrix
factorization, 387
positive part, 382
positive part, Lipschitz continuous, 382
Hermitian operator, 315
largest, smallest, eigenvalues, 368
Hilbert space, 307
Holder's inequality, 268
homeomorphism, 455
homogeneous coordinates, 115
homology
convex spaces, 461
of spheres, 480
path components, 460
pathconnected, 459
spheres, 480
homology group, 458
homomorphism, 211, 456
boundary, 456
connecting, 476
homotopic, 463
maps, 465
homotopy
inverses, 465
type, 465
ideal, 67, 169
maximal, 169
inconsistent, 44
initial value problem
uniqueness, 408
inner product, 263
inner product space
adjoint operator, 314
integers modulo a prime, 28
integral
continuous function, 447
operator valued function, 407
vector valued function, 406
integral domain, 28, 169
integrals
iterated, 450
interior point, 245
intermediate value theorem, 260
intersection, 1
intervals
notation, 1
invariance of domain, 498
invariant subspaces
direct sum, block diagonal matrix, 120
inverse image, 3
inverses and determinants, 180
invertible, 81
irreducible, 22
relatively prime, 23
isomorphism, 152, 211, 459
extensions, 213
iterated integrals, 450
iterative methods
alternate proof of convergence, 404
diagonally dominant, 405
proof of convergence, 402
Jacobi method, 402
Jordan
separation theorem, 494
Jordan canonical form, 144, 162
Jordan curve theorem, 496
ker, 62
kernel of a product
direct sum decomposition, 126

Kirchoff's law, 49, 50

Laplace expansion, 178
leading entry, 41
least squares, 317
lim inf, 29
properties, 32
lim sup, 29
properties, 32
limiit point, 246
limit
continuity, 278
infinite limits, 276
limit of a function, 276
limit of a sequence, 246
well defined, 246
limit point, 276
limits
combinations of functions, 276
existence of limits, 30
limits and continuity, 278
Lindeloff property, 252
Lindemann Weierstrass theorem, 206
Lindemannn Weierstrass theorem, 199
linear combination, 53, 176
linear independence, 262
linear transformation, 62, 101
defined on a basis, 102
dimension of vector space, 102
kernel, 124
matrix, 101
linear transformations
a vector space, 101
commuting, 126
composition, matrices, 111
sum, 101
linearly independent, 53
linearly independent set
enlarging to a basis, 262
Lipschitz continuous, 255
Markov matrix, 285
limit, 288
regular, 288
steady state, 285,288
math induction, 6
mathematical induction, 6
matrices
block diagonal, 91
block multiplication, 89
commuting, 360
invertible, 95
notation, 77
rotation, 111
transpose, 80
matrix, 77
differentiation operator, 105
inverse, 81
left inverse, 180
linear transformation, 104
lower triangular, 181
main diagonal, 135
Markov, 285
polynomial, 191
right inverse, 180
right, left inverse, 180
row, column, determinant rank, 183
stochastic, 285
upper triangular, 181
matrix
positive definite, 386
matrix exponential, 405
matrix multiplication, 78
properties, 80
maximal chain, 503
maximal ideal, 67
maximum likelihood estimates
covariance, 356
mean, 356
Mayer Vietoris
sequence, 480
mean value theorem
Cauchy, 447
metric, 245
properties, 245
metric space, 245
compact sets, 253
complete, 248
completely separable, 252
open set, 245
separable, 252
metric tensor, 319
migration matrix, 289
minimal polynomial
finding it, 165
minimum polynomial, 69, 127
algebraic number, 69
direct sum, 160
finding it, 129
minor, 178
module, 152
cyclical, 153
direct sum, 154
Noetherian, 153
monomorphism, 152, 211
Moore Penrose inverse, 377
least squares, 378
uniqueness, 385
morphism, 152
multivariate normal distribution, 355
Muntz theorems, 453
negative definite, 370
Neuman
series, 410
nilpotent, 144
non solvable group, 234
nondefective, 135
norm
p norm, 268
strictly convex, 403
uniformly convex, 403
normal closure, 237
normal extension, 221
normal matrix, 332
normal subgroup, 222, 232
null and rank, 325
open ball, 245
open set, 245
open cover, 252
open set, 245
open sets
countable basis, 252
operator norm, 274
order, 17
ordered
partial, 503
totally ordered, 503
orthonormal, 270
orthonormal basis
existence, 309
orthonormal polynomials, 322
parallelepiped
volume, 319
partial fractions, 25
decomposition, 26
unique, 26
partial order, 503
partially ordered set, 503
partitioned matrix, 88
path components, 460
Penrose conditions, 379
permutation, 173
permutation matrices, 91, 227
permutations
cycle, 228
Perron's theorem, 293
piecewise continuous, 451
pointwise convergence, 258
polar form complex number, 11
Polish space, 252
polynomial, 21
addition, 21
degree, 21
divides, 22
division, 21
equality, 21
greatest common divisor, 22
greatest common divisor, uniqueness, 22
irreducible, 22
irreducible factorization, 23
multiplication, 21
relatively prime, 22
polynomial
leading term, 21
matrix coefficients, 191
monic, 21
polynomials
coefficients in a field, 63
factoring, 13
factorization, 24, 126
relatively prime?, 44
polynomials in finitely many algebraic numbers, 70
positive, 363
positive definite
postitive eigenvalues, 371
principle minors, 371
positive definite matrix, 386
positive self adjoint
products, 363
roots, 364
postitive definite, 370
power method, 413
powers of a matrix
existence of a limit, 285
Jordan form, 285
stochastic matrix, 285
prime number, 19
principal submatrix, 350
principle ideal domain, 169
principle minors, 371
projection map
convex set, 324
QR algorithm, 339, 423
convergence, 426
convergence theorem, 426
non convergence, 340,430
QR factorization, 316
quadratic form, 333
quadratic formula, 13
quotient group, 222
quotient module, 153
quotient space, 64, 74
quotient vector space, 75
rank
number of pivot columns, 87
rank of a matrix, 87,183
rank one transformations, 103
rational canonical form, 161
Rayleigh quotient, 420
how close?, 420
regression line, 317
regular Sturm Liouville problem, 322
relatively prime, 19
residue class
integers, 27
modulo a prime, 28
retract, 466
retraction, 466
deformation, 466
Riesz representation theorem, 313
right polar factorization, 351, 352
ring
including 1,170
row operations, 41, 91
row rank, 183
row reduced echelon form, 41
description, 86
unique, 87
scalars, 77
Schroder Bernstein theorem, 4
Schur's theorem, 329
self adjoint, 315,363
self adjoint nonnegative
roots, 354,362
separable
polynomial, 223
separable metric space
Lindeloff property, 252
separated sets, 259
sequence, 246
Cauchy, 248
connecting, 475
exact, 474
subsequence, 247
sequential compactness, 283
sequentially compact, 283
sequentially compact set, 253
set notation, 1
sgn, 171
uniqueness, 173
shifted inverse power method, 415
complex eigenvalues, 418
short and long exact sequences, 479
short exact sequence, 154
sign of a permutation, 173
similar
matrix and its transpose, 167
similar matrices, 107, 189
similarity
characteristic polynomial, 189
determinant, 189
trace, 189
similarity transformation, 107
simple field extension, 72
simple groups, 231
simplex
singular, 456
simultaneously diagonalizable, 359
commuting family, 361
singular simplex
boundary, 456
singular value decomposition, 372
singular values, 372
skew symmetric, 81
solution set, 39
solvable by radicals, 235
solvable group, 232
span, 53, 176
spectral mapping theorem, 304
spectral norm, 381
spectral radius, 390
splitting field, 67, 68
dimension, 68
splitting fields
isomorphic, 216
normal extension, 221
stochastic matrix, 285
subdivision and boundary, 469
subdivision map, 467
submodule, 152
subsequence, 247
subspace, 56
complementary, 192
vector space, 56
subspaces
direct sum, 119
direct sum, basis, 119
substituting matrix into polynomial identity, 191
Sylvester
law of inertia, 346
dimention of kernel of product, 124
Sylvester's equation, 326
symmetric, 81
symmetric polynomial theorem, 196
symmetric polynomials, 195
the space $\mathrm{AU}, 326$
Tietze extension theorem, 268, 446
totally bounded, 253
totally ordered, 503
trace, 117, 167
eigenvalues, 117, 167
product, 117, 167
similar matrices, 117, 167
sum of eigenvalues, 341
transpose, 80
properties, 80
transposition, 228
triangle inequality, 266
complex numbers, 10
trichotomy, 17
uniform convergence, 258
uniform convergence and continuity, 258
union, 1
uniqueness of limits, 276
unitary, 316, 329
Unitary matrix
representation, 410
upper Hessenberg matrix, 435
Vandermonde determinant, 190
variation of constants formula, 345
variational inequality, 324
vector space, 51
axioms, 78
dimension, 262
vector space axioms, 51
vector valued function
limit theorems, 276
vectors, 53,78
volume
parallelepiped, 319
Weierstrass approximation
estimate, 441
well ordered, 6
well ordered sets, 504
well ordering, 6
Wilson's theorem, 35
Wronskian, 189, 344
Wronskian alternative, 344


[^0]:    ${ }^{1}$ In this context $t$ is called a parameter.

[^1]:    ${ }^{1}$ I grew up calling this and similar things the minimal polynomial, but $I$ think it is better to call it the minimum polynomial because it is unique. If you see minimal polynomial, this is what it is.

[^2]:    ${ }^{2}$ Gilbert, the librettist of the Savoy operas, may have heard about this great achievement. In Princess Ida which opened in 1884 he has the following lines. "As for fashion they forswear it, so they say - so they say; and the circle - they will square it some fine day some fine day." Of course it had been proved impossible to do this a couple of years before.

[^3]:    ${ }^{1}$ More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, which means possibly more than two rows are switched.

[^4]:    ${ }^{1}$ Note that this is the standard way of defining the sum of two functions.

[^5]:    ${ }^{1}$ This word has 9 syllables! Such words belong in Iceland. Eyjafjallajökull actually only has seven syllables.

[^6]:    ${ }^{1}$ A special case was first proved by Hamilton in 1853. The general case was announced by Cayley some time later and a proof was given by Frobenius in 1878.

[^7]:    ${ }^{1}$ This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in "The sixth shiek's sixth sheep is sick". This is the reason that bases is used instead of basiss.

[^8]:    ${ }^{2}$ Sometimes people put the conjugate on the components of the first entry. It doesn't matter a lot, but it is good to be consistent. I have chosen to place the conjugate on the components of the second entry.

[^9]:    ${ }^{1}$ This notation is just about the most abominable thing imaginable because it is the same notation but entirely different meaning than the norm. However, it saves space in the presentation of this theory of positive matrices and avoids the use of new symbols. Please forget about it when you leave this section.

[^10]:    ${ }^{1}$ If you haven't studied the theory of a complex variable, you should skip this section because you won't understand any of it.

