Analysis of Functions of Many Variables

Kenneth Kuttler klkuttler@gmail.com

February 12, 2024

Contents

Preface

This book is directed to people who have a good understanding of the concepts of one variable calculus including the notions of limit of a sequence and completeness of $\mathbb{R}$. It develops real analysis for functions of many real variables. It is intended to follow my book on real analysis of functions of one variable. The emphasis is on basic concepts from topology, the derivative and the integral. It does not go into functional analysis.

In order to do multivariable calculus correctly, you must first understand some linear algebra. One cannot escape the fact that the derivative is a linear transformation, for example. Therefore, a condensed course in linear algebra is presented first, emphasizing those topics in linear algebra which are useful in analysis, not those topics which are primarily dependent on row operations. It is best to have had a good linear algebra course before attempting a book like this one, however.

I chose to feature the Lebesgue integral because I have gone through the theory of the Riemann integral for a function of n variables and ended up thinking it was too fussy and that the extra abstraction of the Lebesgue integral was worthwhile in order to avoid this fussiness and to also get much better theorems. Also, it seemed to me that this book should be in some sense "more advanced" than my Engineering Math book which has a development of the Riemann integral as an appendix.

Chapter 1

Review of Some Linear Algebra

This material can be referred to as needed. It is here in order to make the book self contained.

1.1 The Matrix of a Linear Map

Recall the definition of a linear map. First of all, these need to be defined on a linear space and have values in a linear space.

Definition 1.1.1 *Let $T : V \to W$ be a function. Here V and W are linear spaces. Then $T \in \mathscr{L}(V,W)$ and is a linear map means that for α, β scalars and v_1, v_2 vectors,*

$$T(\alpha v_1 + \beta v_2) = \alpha T v_1 + \beta T v_2$$

Also recall from linear algebra that if you have $T \in \mathscr{L}(\mathbb{F}^n, \mathbb{F}^m)$ it can always be understood in terms of a matrix. That is, there exists an $m \times n$ matrix A such that for all $\boldsymbol{x} \in \mathbb{F}^n$,

$$A\boldsymbol{x} = T\boldsymbol{x}$$

Recall that, from the way we multiply matrices,

$$A = \left(\begin{array}{ccc} T\boldsymbol{e}_1 & \cdots & T\boldsymbol{e}_n \end{array} \right)$$

That is, the i^{th} column is just $T\boldsymbol{e}_i$.

1.2 Block Multiplication of Matrices

Consider the following problem

$$\left(\begin{array}{cc} A & B \\ C & D \end{array} \right) \left(\begin{array}{cc} E & F \\ G & H \end{array} \right).$$

You know how to do this. You get

$$\left(\begin{array}{cc} AE + BG & AF + BH \\ CE + DG & CF + DH \end{array} \right).$$

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?

Suppose A is a matrix of the form

$$A = \left(\begin{array}{ccc} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{array} \right) \tag{1.1}$$

where A_{ij} 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_{ij}? Here is how for A an $m \times n$ matrix:

$$\overbrace{\begin{pmatrix} \mathbf{0} & I_{s_i \times s_i} & \mathbf{0} \end{pmatrix}}^{s_i \times m} A \overbrace{\begin{pmatrix} \mathbf{0} \\ I_{p_j \times p_j} \\ \mathbf{0} \end{pmatrix}}^{n \times p_j}. \tag{1.2}$$

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_{ij} 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_{ij} 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. Thus the block A_{ij}, 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

$$\begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_m \times p_m} \end{pmatrix},$$

where these little identity matrices don't overlap. A similar conclusion follows from consideration of the matrices $I_{s_i \times s_i}$. Note that in (1.2), the matrix on the right is a block column matrix for the above block diagonal matrix, and the matrix on the left in (1.2) is a block row matrix taken from a similar block diagonal matrix consisting of the $I_{s_i \times s_i}$.

Next consider the question of multiplication of two block matrices. Let B be a block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{1.3}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{1.4}$$

such that for all i, j, it makes sense to multiply $B_{is}A_{sj}$ for all $s \in \{1, \cdots, p\}$. (That is the two matrices B_{is} and A_{sj} are conformable.) and that for fixed ij, it follows that $B_{is}A_{sj}$ is the same size for each s so that it makes sense to write $\sum_s B_{is}A_{sj}$.

The following theorem says essentially that when you take the product of two matrices, you can partition both matrices, formally multiply the blocks to get another block matrix and this one will be BA partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

Lemma 1.2.1 *Consider the following product.*

$$\begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I & \mathbf{0} \end{pmatrix}$$

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

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Proof: From the definition of matrix multiplication, the product is

$$\left(\begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \mathbf{0} \quad \cdots \quad \begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \mathbf{e}_1 \quad \cdots \quad \begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \mathbf{e}_r \quad \cdots \quad \begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \mathbf{0} \right)$$

which yields the claimed result. In the formula $\mathbf{e}_j$ refers to the column vector of length r which has a 1 in the j^{th} position. This proves the lemma. ■

Theorem 1.2.2 *Let B be a $q \times p$ block matrix as in (1.3) and let A be a $p \times n$ block matrix as in (1.4) such that B_{is} is conformable with A_{sj} and each product, $B_{is}A_{sj}$ for $s = 1, \cdots, p$ is of the same size, so that they can be added. Then BA can be obtained as a block matrix such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{1.5}$$

Proof: From (1.2)

$$B_{is}A_{sj} = \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} B \begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix}$$

where here it is assumed B_{is} is $r_i \times p_s$ and A_{sj} is $p_s \times q_j$. The product involves the s^{th} block in the i^{th} row of blocks for B and the s^{th} block in the j^{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 1.2.1

$$\begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Since the blocks of small identity matrices do not overlap,

$$\sum_s \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_p \times p_p} \end{pmatrix} = I,$$

and so,

$$\sum_s B_{is}A_{sj} = \sum_s \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} B \begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0}\end{pmatrix} B \sum_s \begin{pmatrix}\mathbf{0}\\ I_{p_s\times p_s}\\ \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{0} & I_{p_s\times p_s} & \mathbf{0}\end{pmatrix} A \begin{pmatrix}\mathbf{0}\\ I_{q_j\times q_j}\\ \mathbf{0}\end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0}\end{pmatrix} BIA \begin{pmatrix}\mathbf{0}\\ I_{q_j\times q_j}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0}\end{pmatrix} BA \begin{pmatrix}\mathbf{0}\\ I_{q_j\times q_j}\\ \mathbf{0}\end{pmatrix}$$

which equals the ij^{th} block of BA. Hence the ij^{th} block of BA equals the formal multiplication according to matrix multiplication,

$$\sum_s B_{is}A_{sj}.$$

This proves the theorem. ■

Example 1.2.3 *Multiply the following pair of partitioned matrices using the above theorem by multiplying the blocks as described above and then in the conventional manner.*

$$\left(\begin{array}{|cc|c|}\hline 1 & 2 & 3\\ \hline -1 & 2 & 3\\ 3 & -2 & 1\\ \hline\end{array}\right)\left(\begin{array}{|c|cc|}\hline 1 & -1 & 2\\ 2 & 3 & 0\\ \hline -2 & 2 & 1\\ \hline\end{array}\right)$$

Doing it in terms of the blocks, this yields, after the indicated multiplications of the blocks,

$$\left(\begin{array}{|c|c|}\hline 5+(-6) & \begin{pmatrix}5 & 2\end{pmatrix}+3\begin{pmatrix}2 & 1\end{pmatrix}\\ \hline \begin{pmatrix}3\\ -1\end{pmatrix}+\begin{pmatrix}3\\ 1\end{pmatrix}(-2) & \begin{pmatrix}7 & -2\\ -9 & 6\end{pmatrix}+\begin{pmatrix}6 & 3\\ 2 & 1\end{pmatrix}\\ \hline\end{array}\right)$$

This is

$$\left(\begin{array}{|c|c|}\hline -1 & \begin{pmatrix}11 & 5\end{pmatrix}\\ \hline \begin{pmatrix}-3\\ -3\end{pmatrix} & \begin{pmatrix}13 & 1\\ -7 & 7\end{pmatrix}\\ \hline\end{array}\right)$$

Multiplying it out the usual way, you have

$$\begin{pmatrix}1 & 2 & 3\\ -1 & 2 & 3\\ 3 & -2 & 1\end{pmatrix}\begin{pmatrix}1 & -1 & 2\\ 2 & 3 & 0\\ -2 & 2 & 1\end{pmatrix} = \begin{pmatrix}-1 & 11 & 5\\ -3 & 13 & 1\\ -3 & -7 & 7\end{pmatrix}$$

you see this is the same thing without the partition lines.

1.3 Schur's Theorem

For some reason, not understood by me, Schur's theorem is often neglected in beginning linear algebra. This is too bad because it is one of the best theorems in linear algebra. Here $|\cdot|$ denotes the usual norm in $\mathbb{C}^n$ given by

$$|\boldsymbol{x}|^2 \equiv \sum_{j=1}^{n} |x_j|^2$$

Definition 1.3.1 *A complex $n \times n$ matrix U is said to be unitary if $U^*U = I$. Here U^* is the transpose of the conjugate of U. The matrix is unitary if and only if its columns form an orthonormal set in $\mathbb{C}^n$. This follows from the way we multiply matrices in which the ij^{th} entry of U^*U is obtained by taking the conjugate of the i^{th} row of U times the j^{th} column of U.*

Theorem 1.3.2 *(Schur) Let A be a complex $n \times n$ matrix. Then there exists a unitary matrix U such that*

$$U^*AU = T, \tag{1.6}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal, listed with multiplicity[1].

Proof: The theorem is clearly true if A is a 1×1 matrix. Just let $U = 1$, the 1×1 matrix which has entry 1. Suppose it is true for $(n-1) \times (n-1)$ matrices and let A be an $n \times n$ matrix. Then let $\boldsymbol{v}_1$ be a unit eigenvector for A. Then there exists λ_1 such that

$$A\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1,\ |\boldsymbol{v}_1| = 1.$$

Extend $\{\boldsymbol{v}_1\}$ to a basis and then use the Gram - Schmidt process to obtain

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$$

an orthonormal basis of $\mathbb{C}^n$. Let U_0 be a matrix whose i^{th} column is $\boldsymbol{v}_i$. Then from the definition of a unitary matrix Definition 1.3.1, it follows that U_0 is unitary. Consider $U_0^*AU_0$.

$$U_0^*AU_0 = \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_n^* \end{pmatrix} \begin{pmatrix} A\boldsymbol{v}_1 & \cdots & A\boldsymbol{v}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_n^* \end{pmatrix} \begin{pmatrix} \lambda_1 \boldsymbol{v}_1 & \cdots & A\boldsymbol{v}_n \end{pmatrix}$$

Thus $U_0^*AU_0$ is of the form

$$\begin{pmatrix} \lambda_1 & \boldsymbol{a} \\ \mathbf{0} & A_1 \end{pmatrix}$$

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

$$U_1 \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}.$$

An application of block multiplication shows that U_1 is a unitary matrix and also that

$$U_1^* U_0^* A U_0 U_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & * \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & T_{n-1} \end{pmatrix} = T$$

where T is upper triangular. Then let $U = U_0 U_1$. Since $(U_0 U_1)^* = U_1^* U_0^*$, it follows that A is similar to T and that $U_0 U_1$ is unitary. Hence A and T have the same characteristic

[1] 'Listed with multiplicity' means that the diagonal entries are repeated according to their multiplicity as roots of the characteristic equation.

polynomials, and since the eigenvalues of T are the diagonal entries listed with multiplicity, this proves the theorem. ■

The same argument yields the following corollary in the case where A has real entries. The only difference is the use of the real inner product instead of the complex inner product.

Corollary 1.3.3 *Let A be a real $n \times n$ matrix which has only real eigenvalues. Then there exists a real orthogonal matrix Q such that*

$$Q^T AQ = T$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal, listed with multiplicity.

Proof: This follows by observing that if all eigenvalues are real, then corresponding to each real eigenvalue, there exists a real eigenvector. Thus the argument of the above theorem applies with the real inner product in $\mathbb{R}^n$. ■

1.4 Hermitian and Symmetric Matrices

A complex $n \times n$ matrix A with $A^* = A$ is said to be **Hermitian**. A real $n \times n$ matrix A with $A^T = A$ is said to be **symmetric**. In either case, note that for $\langle \cdot, \cdot \rangle$ the inner product in $\mathbb{C}^n$,

$$\langle A\boldsymbol{u}, \boldsymbol{v} \rangle = (A\boldsymbol{u})^T \bar{\boldsymbol{v}} = \boldsymbol{u}^T A^T \bar{\boldsymbol{v}} = \boldsymbol{u}^T \overline{A} \bar{\boldsymbol{v}} = \langle \boldsymbol{u}, A\boldsymbol{v} \rangle.$$

Thus, as a numerical example, the matrix

$$\begin{pmatrix} 1 & 1-i \\ 1+i & 2 \end{pmatrix}$$

is Hermitian, while

$$\begin{pmatrix} 1 & -1 & -2 \\ -1 & 2 & 4 \\ -2 & 4 & 3 \end{pmatrix}$$

is symmetric. Hermitian matrices are named in honor of the French mathematician Charles Hermite (1822–1901).

With Schur's theorem, the theorem on diagonalization of a Hermitian matrix follows.

Theorem 1.4.1 *Let A be Hermitian. Then the eigenvalues of A are all real, and there exists a unitary matrix U such that*

$$U^* AU = D,$$

a diagonal matrix whose diagonal entries are the eigenvalues of A listed with multiplicity. In case A is symmetric, U may be taken to be an orthogonal matrix. The columns of U form an orthonormal basis of eigenvectors of A.

Proof: By Schur's theorem and the assumption that A is Hermitian, there exists a triangular matrix T, whose diagonal entries are the eigenvalues of A listed with multiplicity, and a unitary matrix U such that

$$T = U^* AU = U^* A^* U = (U^* AU)^* = T^*.$$

It follows from this that T is a diagonal matrix and has all real entries down the main diagonal. Hence the eigenvalues of A are real. If A is symmetric (real and Hermitian) it follows from Corollary 1.3.3 that U may be taken to be orthogonal (The columns are an orthonormal set in the inner product of $\mathbb{R}^n$).

That the columns of U form an orthonormal basis of eigenvectors of A, follows right away from the definition of matrix multiplication which implies that if $\boldsymbol{u}_i$ is a column of U, then $A\boldsymbol{u}_i =$ column i of $(UD) = \lambda_i \boldsymbol{u}_i$. ■

1.5 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. 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. 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 [20].

Lemma 1.5.1 *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 $A = U^*DU$. This is from Theorem 1.4.1 above. 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

$$\left(A^{1/2}\right)^2 = U^*D^{1/2}UU^*D^{1/2}U = U^*DU = A.$$

Since $D^{1/2}$ is real,

$$\left(U^*D^{1/2}U\right)^* = U^*\left(D^{1/2}\right)^*(U^*)^* = U^*D^{1/2}U$$

so $A^{1/2}$ is Hermitian. ■

Next it is helpful to recall the Gram Schmidt algorithm and observe a certain property stated in the next lemma.

Lemma 1.5.2 *Suppose $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_p\}$ is a linearly independent set of vectors such that $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ 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 $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}$ be the orthonormal set delivered by the Gram Schmidt process. Then $\boldsymbol{u}_1 = \boldsymbol{w}_1$ because by definition, $\boldsymbol{u}_1 \equiv \boldsymbol{w}_1/|\boldsymbol{w}_1| = \boldsymbol{w}_1$. Now suppose $\boldsymbol{u}_j = \boldsymbol{w}_j$ for all $j \leq k \leq r$. Then if $k < r$, consider the definition of $\boldsymbol{u}_{k+1}$.

$$\boldsymbol{u}_{k+1} \equiv \frac{\boldsymbol{w}_{k+1} - \sum_{j=1}^{k+1}(\boldsymbol{w}_{k+1}, \boldsymbol{u}_j)\boldsymbol{u}_j}{\left|\boldsymbol{w}_{k+1} - \sum_{j=1}^{k+1}(\boldsymbol{w}_{k+1}, \boldsymbol{u}_j)\boldsymbol{u}_j\right|}$$

By induction, $\boldsymbol{u}_j = \boldsymbol{w}_j$ and so this reduces to $\boldsymbol{w}_{k+1}/|\boldsymbol{w}_{k+1}| = \boldsymbol{w}_{k+1}$. ■

This lemma immediately implies the following lemma.

Lemma 1.5.3 *Let V be a subspace of dimension p and let $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ be an orthonormal set of vectors in V. Then this orthonormal set of vectors may be extended to an orthonormal basis for V,*

$$\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_p\}$$

Proof: First extend the given linearly independent set $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ to a basis for V and then apply the Gram Schmidt theorem to the resulting basis. Since $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ is orthonormal it follows from Lemma 1.5.2 the result is of the desired form, an orthonormal basis extending $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$. ■

Here is another lemma about preserving distance.

Lemma 1.5.4 *Suppose R is an $m \times n$ matrix with $m \geq n$ and R preserves distances. Then $R^*R = I$. Also, if R takes an orthonormal basis to an orthonormal set, then R must preserve distances.*

Proof: Since R preserves distances, $|R\boldsymbol{x}| = |\boldsymbol{x}|$ for every $\boldsymbol{x}$. Therefore from the axioms of the dot product,

$$\begin{aligned} & |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 + (\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{y},\boldsymbol{x}) = |\boldsymbol{x}+\boldsymbol{y}|^2 = (R(\boldsymbol{x}+\boldsymbol{y}), R(\boldsymbol{x}+\boldsymbol{y})) \\ = & (R\boldsymbol{x},R\boldsymbol{x}) + (R\boldsymbol{y},R\boldsymbol{y}) + (R\boldsymbol{x},R\boldsymbol{y}) + (R\boldsymbol{y},R\boldsymbol{x}) \\ = & |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 + (R^*R\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{y},R^*R\boldsymbol{x}) \end{aligned}$$

and so for all $\boldsymbol{x},\boldsymbol{y}$,

$$(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{y},R^*R\boldsymbol{x}-\boldsymbol{x}) = 0$$

Hence for all $\boldsymbol{x},\boldsymbol{y}$, $\operatorname{Re}(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y}) = 0$. Now for a $\boldsymbol{x},\boldsymbol{y}$ given, choose $\alpha \in \mathbb{C}$ such that

$$\alpha(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y}) = |(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y})|$$

Then

$$0 = \operatorname{Re}(R^*R\boldsymbol{x}-\boldsymbol{x},\overline{\alpha}\boldsymbol{y}) = \operatorname{Re}\alpha(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y}) = |(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y})|$$

Thus $|(R^*R\boldsymbol{x}-\boldsymbol{x},\boldsymbol{y})| = 0$ for all $\boldsymbol{x},\boldsymbol{y}$ because the given $\boldsymbol{x},\boldsymbol{y}$ were arbitrary. Let $\boldsymbol{y} = R^*R\boldsymbol{x}-\boldsymbol{x}$ to conclude that for all $\boldsymbol{x}$,

$$R^*R\boldsymbol{x}-\boldsymbol{x} = \boldsymbol{0}$$

which says $R^*R = I$ since $\boldsymbol{x}$ is arbitrary.

Consider the last claim. Let $R : \mathbb{F}^n \to \mathbb{F}^m$ such that $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$ is an orthonormal basis for $\mathbb{F}^n$ and $\{R\boldsymbol{u}_1, \cdots, R\boldsymbol{u}_n\}$ is also an orthormal set, then

$$\left|R\left(\sum_i x_i\boldsymbol{u}_i\right)\right|^2 = \left|\sum_i x_i R\boldsymbol{u}_i\right|^2 = \sum_i |x_i|^2 = \left|\sum_i x_i\boldsymbol{u}_i\right|^2 \blacksquare$$

With this preparation, here is the big theorem about the right polar factorization.

Theorem 1.5.5 *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 satisfies $R^*R = I$ such that $F = RU$.*

Proof: Consider F^*F. This is a Hermitian matrix because

$$(F^*F)^* = F^*(F^*)^* = F^*F$$

Also the eigenvalues of the $n \times n$ matrix F^*F are all nonnegative. This is because if $\boldsymbol{x}$ is an eigenvalue,

$$\lambda(\boldsymbol{x},\boldsymbol{x}) = (F^*F\boldsymbol{x},\boldsymbol{x}) = (F\boldsymbol{x},F\boldsymbol{x}) \geq 0.$$

Therefore, by Lemma 1.5.1, there exists an $n \times n$ Hermitian matrix U having all nonnegative eigenvalues such that

$$U^2 = F^*F.$$

Consider the subspace $U(\mathbb{F}^n)$. Let $\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r\}$ be an orthonormal basis for

$$U(\mathbb{F}^n) \subseteq \mathbb{F}^n.$$

Note that $U(\mathbb{F}^n)$ might not be all of $\mathbb{F}^n$. Using Lemma 1.5.3, extend to an orthonormal basis for all of $\mathbb{F}^n$,

$$\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_n\}.$$

Next observe that $\{F\boldsymbol{x}_1, \cdots, F\boldsymbol{x}_r\}$ is also an orthonormal set of vectors in $\mathbb{F}^m$. This is because

$$\begin{aligned}(F\boldsymbol{x}_k, F\boldsymbol{x}_j) &= (F^*F\boldsymbol{x}_k, \boldsymbol{x}_j) = \left(U^2\boldsymbol{x}_k, \boldsymbol{x}_j\right) \\ &= (U\boldsymbol{x}_k, U^*\boldsymbol{x}_j) = (U\boldsymbol{x}_k, U\boldsymbol{x}_j) = \delta_{jk}\end{aligned}$$

Therefore, from Lemma 1.5.3 again, this orthonormal set of vectors can be extended to an orthonormal basis for $\mathbb{F}^m$,

$$\{F\boldsymbol{x}_1, \cdots, F\boldsymbol{x}_r, \boldsymbol{z}_{r+1}, \cdots, \boldsymbol{z}_m\}$$

Thus there are at least as many $\boldsymbol{z}_k$ as there are $\boldsymbol{y}_j$. Now for $\boldsymbol{x} \in \mathbb{F}^n$, since

$$\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_n\}$$

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

$$R\boldsymbol{x} \equiv \sum_{k=1}^{r} c_k F\boldsymbol{x}_k + \sum_{k=r+1}^{n} d_k \boldsymbol{z}_k \tag{1.7}$$

Then also there exist scalars b_k such that

$$U\,\boldsymbol{x} = \sum_{k=1}^{r} b_k U\boldsymbol{x}_k$$

and so from 1.7,

$$RU\,\boldsymbol{x} = \sum_{k=1}^{r} b_k F\boldsymbol{x}_k = F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right)$$

Is $F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) = F(\boldsymbol{x})$?

$$\left(F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) - F(\boldsymbol{x}), F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) - F(\boldsymbol{x})\right)$$

$$\begin{aligned}
&= \left((F^*F)\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), \left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), \left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), U\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U\boldsymbol{x}_k - U\boldsymbol{x}, \sum_{k=1}^{r} b_k U\boldsymbol{x}_k - U\boldsymbol{x}\right) = 0
\end{aligned}$$

Therefore, $F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) = F(\boldsymbol{x})$ and this shows $RU\boldsymbol{x} = F\boldsymbol{x}$. From 1.7 it follows that R maps an orthonormal set to an orthonormal set and so R preserves distances. Therefore, by Lemma 1.5.4 $R^*R = I$. ■

1.6 Elementary matrices

The elementary matrices result from doing a row operation to the identity matrix.

As before, everything will apply to matrices having coefficients in an arbitrary field of scalars, although we will mainly feature the real numbers in the examples.

Definition 1.6.1 *The row operations consist of the following*

1. *Switch two rows.*
2. *Multiply a row by a nonzero number.*
3. *Replace a row by the same row added to a multiple of another row.*

We refer to these as the row operations of type 1,2, and 3 respectively.

The elementary matrices are given in the following definition.

Definition 1.6.2 *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. More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.*

As an example of why these elementary matrices are interesting, consider the following. Letting $\boldsymbol{r}_i$ be the row vector of all zeros except for a 1 in the i^{th} slot,

$$\begin{pmatrix} \boldsymbol{r}_2 \\ \boldsymbol{r}_1 \\ \boldsymbol{r}_3 \end{pmatrix} \begin{pmatrix} a & b & c & d \\ x & y & z & w \\ f & g & h & i \end{pmatrix} = \begin{pmatrix} x & y & z & w \\ a & b & c & d \\ f & g & h & i \end{pmatrix}.$$

A 3×4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to switching the first two rows of the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. They are obtained from switching a couple of rows of the identity matrix. First P_{ij}, which involves switching row i and row j of the identity where Let $i < j$. Then, as above, $P^{ij} =$

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}$$

where

$$\boldsymbol{r}_j = (0 \cdots 1 \cdots 0)$$

with the 1 in the j^{th} position from the left.

For P^{ij} this matrix which involves switching the i and j rows of the identity. Now consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_j \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix}.$$

Now we try multiplication of a matrix on the left by this elementary matrix P^{ij}. Thus,

$$P^{ij} \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}.$$

has the indicated columns listed in order:

$$\left(P^{ij}\begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{n1} \end{pmatrix}, P^{ij}\begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, P^{ij}\begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \left(\begin{pmatrix} a_{11} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{n1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1p} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

and so the resulting matrix is

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}.$$

This has established the following lemma.

Lemma 1.6.3 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows of I. Then if P^{ij}, A are conformable, we have*

$$P^{ij}A = B$$

where B is obtained from A by switching the i^{th} and the j^{th} rows.

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c. We write

$$I = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}$$

where

$$\boldsymbol{r}_j = (0 \cdots 1 \cdots 0)$$

with the 1 in the j^{th} position from the left. The elementary matrix which results from applying this operation to the i^{th} row of the identity matrix is of the form

$$E(c,i) = \begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ c\boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}.$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ c\boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ cv_i \\ \vdots \\ v_n \end{pmatrix}.$$

Denote by $E(c,i)$ this elementary matrix which multiplies the i^{th} row of the identity by the nonzero constant, c. Then from what was just discussed and the way matrices are multiplied,

$$E(c,i) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

equals a matrix having the columns indicated below.

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ ca_{i1} & ca_{i2} & \cdots & ca_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}.$$

This proves the following lemma.

Lemma 1.6.4 *Let $E(c,i)$ denote the elementary matrix corresponding to the row operation in which the i^{th} row is multiplied by the nonzero constant c. Thus $E(c,i)$ involves multiplying the i^{th} row of the identity matrix by c. Then*

$$E(c,i)A = B$$

where B is obtained from A by multiplying the i^{th} row of A by c.

Finally consider the third of these row operations. Letting $\boldsymbol{r}_j$ be the j^{th} row of the identity matrix, denote by $E(c \times i + j)$ the elementary matrix obtained from the identity matrix by replacing $\boldsymbol{r}_j$ with $\boldsymbol{r}_j + c\boldsymbol{r}_i$. In case $i < j$ this will be of the form

$$P^{ij} = \begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ c\boldsymbol{r}_i + \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}.$$

Consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ c\boldsymbol{r}_i + \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ cv_i + v_j \\ \vdots \\ v_n \end{pmatrix}.$$

From this and the way matrices are multiplied,

$$E\left(c \times i + j\right) \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix having the indicated columns listed in order.

$$\left(E\left(c \times i + j\right) \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n1} \end{pmatrix}, E\left(c \times i + j\right) \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots E\left(c \times i + j\right) \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{j2}+ca_{i1} & a_{j2}+ca_{i2} & \cdots & a_{jp}+ca_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}.$$

The case where $i > j$ is similar. This proves the following lemma in which, as above, the i^{th} row of the identity is $\boldsymbol{r}_i$.

Lemma 1.6.5 *Let $E(c \times i + j)$ denote the elementary matrix obtained from I by replacing the j^{th} row of the identity $\boldsymbol{r}_j$ with $c\boldsymbol{r}_i + \boldsymbol{r}_j$. Letting the k^{th} row of A be $\boldsymbol{a}_k$,*

$$E(c \times i + j) A = B$$

where B has the same rows as A except the j^{th} row of B is $c\boldsymbol{a}_i + \boldsymbol{a}_j$.

The above lemmas are summarized in the following theorem.

Theorem 1.6.6 *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, EA. In addition to this, the following identities hold for the elementary matrices described above.*

$$E(c \times i + j) E(-c \times i + j) = E(-c \times i + j) E(c \times i + j) = I. \tag{1.8}$$

$$E(c, i) E\left(c^{-1}, i\right) = E\left(c^{-1}, i\right) E(c, i) = I. \tag{1.9}$$

$$P^{ij} P^{ij} = I. \tag{1.10}$$

Proof: Consider (1.8). Starting with I and taking $-c$ times the i^{th} row added to the j^{th} yields $E(-c \times i + j)$ which differs from I only in the j^{th} row. Now multiplying on the left by $E(c \times i + j)$ takes c times the i^{th} row and adds to the j^{th} thus restoring the j^{th} row to its original state. Thus $E(c \times i + j) E(-c \times i + j) = I$. Similarly $E(-c \times i + j) E(c \times i + j) = I$. The reasoning is similar for (1.9) and (1.10). ■

Each of these elementary matrices has a significant geometric significance. The effect of doing $E\left(\frac{1}{2} \times 3 + 1\right)$ shears the box in one direction. Of course there would be corresponding shears in the other directions also. Note that this does not change the volume. You should think about the geometric effect of the other elementary matrices on a box.

Definition 1.6.7 *For an $n \times n$ matrix A, an $n \times n$ matrix B which has the property that $AB = BA = I$ is denoted by A^{-1}. Such a matrix is called an* ***inverse.*** *When A has an inverse, it is called* ***invertible.***

The following lemma says that if a matrix acts like an inverse, then it is **the** inverse. Also, the product of invertible matrices is invertible.

Lemma 1.6.8 *If B,C are both inverses of A, then $B = C$. That is, there exists at most one inverse of a matrix. If $A_1, \cdots, A_m$ are each invertible $m \times m$ matrices, then the product $A_1 A_2 \cdots A_m$ is also invertible and*

$$(A_1 A_2 \cdots A_m)^{-1} = A_m^{-1} A_{m-1}^{-1} \cdots A_1^{-1}.$$

Proof. From the definition and associative law of matrix multiplication,

$$B = BI = B(AC) = (BA)C = IC = C.$$

This proves the uniqueness of the inverse.

Next suppose A, B are invertible. Then

$$AB\left(B^{-1}A^{-1}\right) = A\left(BB^{-1}\right)A^{-1} = AIA^{-1} = AA^{-1} = I$$

and also

$$\left(B^{-1}A^{-1}\right)AB = B^{-1}\left(A^{-1}A\right)B = B^{-1}IB = B^{-1}B = I.$$

It follows from Definition 1.6.7 that AB has an inverse and it is $B^{-1}A^{-1}$. Thus the case of $m = 1, 2$ in the claim of the lemma is true. Suppose this claim is true for k. Then

$$A_1 A_2 \cdots A_k A_{k+1} = (A_1 A_2 \cdots A_k) A_{k+1}.$$

By induction, the two matrices $(A_1 A_2 \cdots A_k), A_{k+1}$ are both invertible and

$$(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} \cdots A_2^{-1} A_1^{-1}.$$

By the case of the product of two invertible matrices shown above,

$$((A_1 A_2 \cdots A_k) A_{k+1})^{-1} = A_{k+1}^{-1}(A_1 A_2 \cdots A_k)^{-1} = A_{k+1}^{-1} A_k^{-1} \cdots A_2^{-1} A_1^{-1}.$$

This proves the lemma. ■

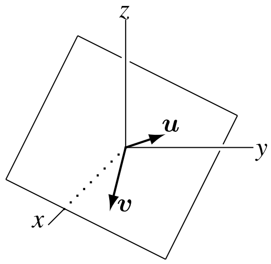
We will discuss methods for finding the inverse later. For now, observe that Theorem 1.6.6 says that elementary matrices are invertible and that the inverse of such a matrix is also an elementary matrix. The major conclusion of the above Lemma and Theorem is the following lemma about linear relationships.

Definition 1.6.9 *Let $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k, \boldsymbol{u}$ be vectors. Then $\boldsymbol{u}$ is said to be a **linear combination** of the vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ if there exist scalars $c_1, \cdots, c_k$ such that*

$$\boldsymbol{u} = \sum_{i=1}^{k} c_i \boldsymbol{v}_i.$$

*We also say that when the above holds for some scalars $c_1, \cdots, c_k$, there exists a **linear relationship** between the vector $\boldsymbol{u}$ and the vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$.*

We will discuss this more later, but the following picture illustrates the geometric significance of the vectors which have a linear relationship with two vectors $\boldsymbol{u}, \boldsymbol{v}$ pointing in different directions.



The following lemma states that linear relationships between columns in a matrix are preserved by row operations. This simple lemma is the main result in understanding all the major questions related to the row reduced echelon form as well as many other topics.

Lemma 1.6.10 *Let A and B be two $m\times n$ matrices and suppose B results from a row operation applied to A. Then the k^{th} column of B is a linear combination of the $i_1,\cdots,i_r$ columns of B if and only if the k^{th} column of A is a linear combination of the $i_1,\cdots,i_r$ columns of A. Furthermore, the scalars in the linear combinations are the same. (The linear relationship between the k^{th} column of A and the $i_1,\cdots,i_r$ columns of A is the same as the linear relationship between the k^{th} column of B and the $i_1,\cdots,i_r$ columns of B.)*

Proof: Let A be the following matrix in which the $\mathbf{a}_k$ are the columns

$$\begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}$$

and let B be the following matrix in which the columns are given by the $\mathbf{b}_k$

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{pmatrix}.$$

Then by Theorem 1.6.6 on Page 23, $\mathbf{b}_k = E\mathbf{a}_k$ where E is an elementary matrix. Suppose then that one of the columns of A is a linear combination of some other columns of A. Say

$$\mathbf{a}_k = c_1\mathbf{a}_{i_1} + \cdots + c_r\mathbf{a}_{i_r}.$$

Then multiplying by E,

$$\mathbf{b}_k = E\mathbf{a}_k = c_1E\mathbf{a}_{i_1} + \cdots + c_rE\mathbf{a}_{i_r} = c_1\mathbf{b}_{i_1} + \cdots + c_r\mathbf{b}_{i_r}.$$

This proves the lemma. ■

Example 1.6.11 Find linear relationships between the columns of the matrix

$$A = \begin{pmatrix} 1 & 3 & 11 & 10 & 36 \\ 1 & 2 & 8 & 9 & 23 \\ 1 & 1 & 5 & 8 & 10 \end{pmatrix}.$$

It is not clear what the relationships are, so we do row operations to this matrix. Lemma 1.6.10 says that all the linear relationships between columns are preserved, so the idea is to do row operations until a matrix results which has the property that the linear relationships are obvious. First take -1 times the top row and add to the two bottom rows. This yields

$$\begin{pmatrix} 1 & 3 & 11 & 10 & 36 \\ 0 & -1 & -3 & -1 & -13 \\ 0 & -2 & -6 & -2 & -26 \end{pmatrix}$$

Next take -2 times the middle row and add to the bottom row followed by multiplying the middle row by -1 :

$$\left(\begin{array}{ccccc} 1 & 3 & 11 & 10 & 36 \\ 0 & 1 & 3 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Next take -3 times the middle row added to the top:

$$\left(\begin{array}{ccccc} 1 & 0 & 2 & 7 & -3 \\ 0 & 1 & 3 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right). \tag{1.11}$$

At this point it is clear that the last column is -3 times the first column added to 13 times the second. By Lemma 1.6.10, the same is true of the corresponding columns in the original matrix A. As a check,

$$-3\left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right)+13\left(\begin{array}{c} 3 \\ 2 \\ 1 \end{array}\right)=\left(\begin{array}{c} 36 \\ 23 \\ 10 \end{array}\right).$$

You should notice that other linear relationships are also easily seen from (1.11). For example the fourth column is 7 times the first added to the second. This is obvious from (1.11) and Lemma 1.6.10 says the same relationship holds for A.

This is really just an extension of the technique for finding solutions to a linear system of equations. In solving a system of equations earlier, row operations were used to exhibit the last column of an augmented matrix as a linear combination of the preceding columns. The **row reduced echelon form** just extends this by making obvious the linear relationships between **every** column, not just the last, and those columns preceding it. The matrix in 1.11 is in row reduced echelon form. The row reduced echelon form is the topic of the next section.

1.7 The Row Reduced Echelon Form Of A Matrix

When you do row operations on a matrix, there is an ultimate conclusion. It is called the **row reduced echelon form**. We show here that every matrix has such a row reduced echelon form and that this row reduced echelon form is unique. The significance is that it becomes possible to use the definite article in referring to **the** row reduced echelon form. Hence important conclusions about the original matrix may be logically deduced from an examination of its unique row reduced echelon form. First we need the following definition.

Definition 1.7.1 *Define special column vectors $\mathbf{e}_i$ as follows.*

$$\mathbf{e}_i = \left(\begin{array}{ccccc} 0 & \cdots & 1 & \cdots & 0 \end{array}\right)^T.$$

Recall that ^{T} says to take the transpose. Thus $\mathbf{e}_i$ is the column vector which has all zero entries except for a 1 in the i^{th} position down from the top.

Now here is the description of the row reduced echelon form.

Definition 1.7.2 *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*

$\mathbf{e}_1$ and if, in viewing the columns of the matrix from left to right, you have encountered $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$, the next column is either $\mathbf{e}_{k+1}$ or this next column is a linear combination of the vectors, $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$.

Example 1.7.3 The following matrices are in row reduced echelon form.

$$\begin{pmatrix} 1 & 0 & 4 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 0 & 7 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -5 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Definition 1.7.4 *Given a matrix A, row reduction produces one and only one row reduced matrix B with $A \sim B$. See Corollary 1.7.9. We call B **the** row reduced echelon form of A.*

Theorem 1.7.5 *Let A be an $m \times n$ matrix. Then A has a row reduced echelon form determined by a simple process.*

Proof. Viewing the columns of A from left to right, take the first nonzero column. Pick a nonzero entry in this column and switch the row containing this entry with the top row of A. Now divide this new top row by the value of this nonzero entry to get a 1 in this position and then use row operations to make all entries below this equal to zero. Thus the first nonzero column is now $\mathbf{e}_1$. Denote the resulting matrix by A_1. Consider the sub-matrix of A_1 to the right of this column and below the first row. Do exactly the same thing for this sub-matrix that was done for A. This time the $\mathbf{e}_1$ will refer to F^{m-1}. Use the first 1 obtained by the above process which is in the top row of this sub-matrix and row operations, to produce a zero in place of every entry above it and below it. Call the resulting matrix A_2. Thus A_2 satisfies the conditions of the above definition up to the column just encountered. Continue this way till every column has been dealt with and the result must be in row reduced echelon form. ■

Here is some terminology about pivot columns.

Definition 1.7.6 *The first **pivot column** of A is the first nonzero column of A which becomes $\mathbf{e}_1$ in the row reduced echelon form. The next pivot column is the first column after this which becomes $\mathbf{e}_2$ in the row reduced echelon form. The third is the next column which becomes $\mathbf{e}_3$ in the row reduced echelon form and so forth.*

The algorithm just described for obtaining a row reduced echelon form shows that these columns are well defined, but we will deal with this issue more carefully in Corollary 1.7.9 where we show that every matrix corresponds to exactly one row reduced echelon form.

Definition 1.7.7 *Two matrices A, B are said to be **row equivalent** if B can be obtained from A by a sequence of row operations. When A is row equivalent to B, we write $A \sim B$.*

Proposition 1.7.8 *In the notation of Definition 1.7.7. $A \sim A$. If $A \sim B$, then $B \sim A$. If $A \sim B$ and $B \sim C$, then $A \sim C$.*

Proof. That $A \sim A$ is obvious. Consider the second claim. By Theorem 1.6.6, there exist elementary matrices $E_1, E_2, \cdots, E_m$ such that

$$B = E_1 E_2 \cdots E_m A.$$

It follows from Lemma 1.6.8 that $(E_1 E_2 \cdots E_m)^{-1}$ exists and equals the product of the inverses of these matrices in the reverse order. Thus

$$E_m^{-1} E_{m-1}^{-1} \cdots E_1^{-1} B = (E_1 E_2 \cdots E_m)^{-1} B$$

$$= (E_1 E_2 \cdots E_m)^{-1} (E_1 E_2 \cdots E_m) A = A.$$

By Theorem 1.6.6, each E_k^{-1} is an elementary matrix. By Theorem 1.6.6 again, the above shows that A results from a sequence of row operations applied to B. The last claim is left for an exercise. This proves the proposition. ■

There are three choices for row operations at each step in Theorem 1.7.5. A natural question is whether the same row reduced echelon matrix always results in the end from following any sequence of row operations.

We have already made use of the following observation in finding a linear relationship between the columns of the matrix A, but here it is stated more formally.

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \mathbf{e}_1 + \cdots + x_n \mathbf{e}_n,$$

so to say two column vectors are equal, is to say the column vectors are the same linear combination of the special vectors $\mathbf{e}_j$.

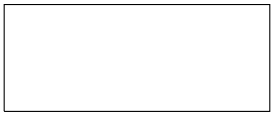
Corollary 1.7.9 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are obtained from A by a sequence of row operations, then $B = C$.*

Proof. Suppose B and C are both row reduced echelon forms for the matrix A. It follows that B and C have zero columns in the same positions because row operations do not affect zero columns. By Proposition 1.7.8, B and C are row equivalent. In reading from left to right in B, suppose $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in positions $i_1, \cdots, i_r$ respectively. Then from the description of the row reduced echelon form, each of these columns of B, in positions $i_1, \cdots, i_r$, is not a linear combination of the preceding columns. Since C is row equivalent to B, it follows from Lemma 1.6.10, that each column of C in positions $i_1, \cdots, i_r$ is not a linear combination of the preceding columns of C. By the description of the row reduced echelon form, $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in C, in positions $i_1, \cdots, i_r$ respectively. Therefore, both B and C have the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occurring first (reading from left to right) in the positions, $i_1, i_2, \cdots, i_r$. Since these matrices are row equivalent, it follows from Lemma 1.6.10, that the columns between the i_k and i_{k+1} position in the two matrices are linear combinations involving the same scalars, of the columns in the $i_1, \cdots, i_k$ position. Similarly, the columns after the i_r position are linear combinations of the columns in the $i_1, \cdots, i_r$ positions involving the same scalars in both matrices. This is equivalent to the assertion that each of these columns is identical in B and C. ■

Now with the above corollary, here is a very fundamental observation. The number of nonzero rows in the row reduced echelon form is the same as the number of pivot columns.

Namely, this number is r in both cases where $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_r$ are the pivot columns in the row reduced echelon form. This number r is called the **rank** of the matrix. This is discussed more later, but first here are some other applications.

Consider a matrix which looks like this: (More columns than rows.)



Corollary 1.7.10 *Suppose A is an $m \times n$ matrix and that $m < n$. That is, the number of rows is less than the number of columns. Then one of the columns of A is a linear combination of the preceding columns of A. Also, there exists $\boldsymbol{x} \in F^n$ such that $\boldsymbol{x} \neq \mathbf{0}$ and $A\boldsymbol{x} = \mathbf{0}$.*

Proof: Since $m < n$, not all the columns of A can be pivot columns. In reading from left to right, pick the first one which is not a pivot column. Then from the description of the row reduced echelon form, this column is a linear combination of the preceding columns. Say

$$\boldsymbol{a}_j = x_1\boldsymbol{a}_1 + \cdots + x_{j-1}\boldsymbol{a}_{j-1}.$$

Therefore, from the way we multiply a matrix times a vector,

$$A\begin{pmatrix} x_1 \\ \vdots \\ x_{j-1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1 \cdots \boldsymbol{a}_{j-1} \boldsymbol{a}_j \cdots \boldsymbol{a}_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{j-1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}. \blacksquare$$

1.8 Finding the Inverse of a Matrix

Recall that the inverse of an $n \times n$ matrix A is a matrix B such that

$$AB = BA = I$$

where I is the identity matrix. It was shown that an elementary matrix is invertible and that its inverse is also an elementary matrix. Also the product of invertible matrices is invertible and its inverse is the product of the inverses in the reverse order. In this section, we consider the problem of finding an inverse for a given $n \times n$ matrix.

Example 1.8.1 Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. Show that $\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ is the inverse of A.

To check this, multiply

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

showing that this matrix is indeed the inverse of A.

In the last example, how would you find A^{-1}? You wish to find a matrix $\begin{pmatrix} x & z \\ y & w \end{pmatrix}$ such that

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x & z \\ y & w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This requires the solution of the systems of equations,

$$x+y=1, x+2y=0$$

and

$$z+w=0, z+2w=1.$$

Writing the augmented matrix for these two systems gives

$$\left(\begin{array}{ccc|c} 1 & 1 & | & 1 \\ 1 & 2 & | & 0 \end{array}\right) \tag{1.12}$$

for the first system and

$$\left(\begin{array}{cccc} 1 & 1 & | & 0 \\ 1 & 2 & | & 1 \end{array}\right) \tag{1.13}$$

for the second. Let's solve the first system. Take (-1) times the first row and add to the second to get

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

Now take (-1) times the second row and add to the first to get

$$\left(\begin{array}{cccc} 1 & 0 & | & 2 \\ 0 & 1 & | & -1 \end{array}\right).$$

Putting in the variables, this says $x=2$ and $y=-1$.

Now solve the second system, (1.13) to find z and w. Take (-1) times the first row and add to the second to get

$$\left(\begin{array}{cccc} 1 & 1 & | & 0 \\ 0 & 1 & | & 1 \end{array}\right).$$

Now take (-1) times the second row and add to the first to get

$$\left(\begin{array}{cccc} 1 & 0 & | & -1 \\ 0 & 1 & | & 1 \end{array}\right).$$

Putting in the variables, this says $z=-1$ and $w=1$. Therefore, the inverse is

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Didn't the above seem rather repetitive? Exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|\mathbf{v})$ where I is the identity and $\boldsymbol{v}$ gave a column of the inverse. In the above $\begin{pmatrix} x \\ y \end{pmatrix}$, the first column of the inverse was obtained first and then the second column $\begin{pmatrix} z \\ w \end{pmatrix}$.

To simplify this procedure, you could have written

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

and row reduced till you obtained

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

Then you could have read off the inverse as the 2×2 matrix on the right side. You should be able to see that it is valid by adapting the argument used in the simple case above.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the **Gauss-Jordan procedure**.

Procedure 1.8.2 *Suppose A is an $n \times n$ matrix. To find A^{-1} if it exists, form the augmented $n \times 2n$ matrix*

$$(A|I)$$

and then if possible, do row operations until you obtain an $n \times 2n$ matrix of the form

$$(I|B). \tag{1.14}$$

When this has been done, $B = A^{-1}$. If it is impossible to row reduce to a matrix of the form $(I|B)$, then A has no inverse.

The procedure just described along with the preceding explanation shows that this procedure actually yields a **right inverse**. This is a matrix B such that $AB = I$. We will show in Theorem 1.8.4 that this right inverse is really **the** inverse. This is a stronger result than that of Lemma 1.6.8 about the uniqueness of the inverse. For now, here is an example.

Example 1.8.3 Let $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. Find A^{-1} if it exists.

Set up the augmented matrix $(A|I)$:

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

Next take (-1) times the first row and add to the second followed by (-3) times the first row added to the last. This yields

$$\left(\begin{array}{ccccccc} 1 & 2 & 2 & | & 1 & 0 & 0 \\ 0 & -2 & 0 & | & -1 & 1 & 0 \\ 0 & -5 & -7 & | & -3 & 0 & 1 \end{array}\right).$$

Then take 5 times the second row and add to -2 times the last row.

$$\begin{pmatrix} 1 & 2 & 2 & | & 1 & 0 & 0 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{pmatrix}$$

Next take the last row and add to (-7) times the top row. This yields

$$\begin{pmatrix} -7 & -14 & 0 & | & -6 & 5 & -2 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{pmatrix}.$$

Now take $(-7/5)$ times the second row and add to the top.

$$\begin{pmatrix} -7 & 0 & 0 & | & 1 & -2 & -2 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{pmatrix}.$$

Finally divide the top row by -7, the second row by -10 and the bottom row by 14, which yields

$$\begin{pmatrix} 1 & 0 & 0 & | & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & | & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & | & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}.$$

Therefore, the inverse is

$$\begin{pmatrix} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}.$$

What you have really found in the above algorithm is a **right inverse.** Is this right inverse matrix, which we have called the inverse, really **the** inverse, the matrix which when multiplied on both sides gives the identity?

Theorem 1.8.4 *Suppose A,B are $n \times n$ matrices and $AB = I$. Then it follows that $BA = I$ also, and so $B = A^{-1}$. For $n \times n$ matrices, the left inverse, right inverse and inverse are all the same thing.*

Proof. If $AB = I$ for A,B $n \times n$ matrices, is $BA = I$? If $AB = I$, there exists a unique solution $\boldsymbol{x}$ to the equation

$$B\boldsymbol{x} = \boldsymbol{y}$$

for any choice of $\boldsymbol{y}$. In fact,

$$\boldsymbol{x} = A(B\boldsymbol{x}) = A\boldsymbol{y}.$$

This means the row reduced echelon form of B must be I. Thus every column is a pivot column. Otherwise, there exists a free variable and the solution, if it exists, would not be unique, contrary to what was just shown must happen if $AB = I$. It follows that a right inverse B^{-1} for B exists. The above procedure yields

$$\left(\begin{array}{cc} B & I \end{array}\right) \rightarrow \left(\begin{array}{cc} I & B^{-1} \end{array}\right).$$

Now multiply both sides of the equation $AB = I$ on the right by B^{-1}. Then

$$A = A\left(BB^{-1}\right) = (AB)B^{-1} = B^{-1}.$$

Thus A is the right inverse of B, and so $BA = I$. This shows that if $AB = I$, then $BA = I$ also. Exchanging roles of A and B, we see that if $BA = I$, then $AB = I$. This proves the theorem. ■

This has shown that in the context of $n \times n$ matrices, right inverses, left inverses and inverses are all the same and this matrix is called A^{-1}.

The following corollary is also of interest.

Corollary 1.8.5 *An $n \times n$ matrix A has an inverse if and only if the row reduced echelon form of A is I.*

Proof. First suppose the row reduced echelon form of A is I. Then Procedure 1.8.2 yields a right inverse for A. By Theorem 1.8.4 this is **the** inverse. Next suppose A has an inverse. Then there exists a unique solution $\boldsymbol{x}$ to the equation $A\boldsymbol{x} = \boldsymbol{y}$. given by $\boldsymbol{x} = A^{-1}\boldsymbol{y}$. It follows that in the augmented matrix $(A|\mathbf{0})$ there are no free variables, and so every column to the left of the zero column is a pivot column. Therefore, the row reduced echelon form of A is I. ■

1.9 The Mathematical Theory of Determinants

It is easiest to give a definition of the determinant which is clearly well defined and then prove the Laplace expansion gives the same thing. Let $(i_1, \cdots, i_n)$ 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 [38]. A recent book which also has a good introduction is Baker [4]

1.9.1 The Function sgn

The following Lemma will be essential in the definition of the determinant.

Lemma 1.9.1 *There exists a function, 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.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{1.15}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{1.16}$$

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 $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ so that every number from $\{1, \cdots, n\}$ appears in the ordered list, $(i_1, \cdots, i_n)$,

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta} \operatorname{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{1.17}$$

where $n = i_\theta$ in the ordered list, $(i_1, \cdots, i_n)$.

Proof: Define $\text{sign}(x) = 1$ 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 $\text{sgn}_1(1) \equiv 1$. For the general case where $n > 1$, simply define

$$\text{sgn}_n(i_1, \cdots, i_n) \equiv \text{sign}\left(\prod_{r<s}(i_s - i_r)\right)$$

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 $(i_1, \cdots, i_n)$ are switched. Denote the new ordered list of numbers as $(j_1, \cdots, j_n)$. 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 1	i_2 2	$\cdots$	i_p p	$\cdots$	i_q q	$\cdots$	i_n n
i_1 1	i_2 2	$\cdots$	i_q p	$\cdots$	i_p q	$\cdots$	i_n n
j_1 1	j_2 2	$\cdots$	j_p p	$\cdots$	j_q q	$\cdots$	j_n n

Then

$$\text{sgn}_n(j_1, \cdots, j_n) \equiv \text{sign}\left(\prod_{r<s}(j_s - j_r)\right)$$

$$= \text{sign}\left(\overset{\text{both } p,q}{(i_p - i_q)} \prod_{p<j<q} (i_j - i_q) \overset{\text{one of } p,q}{} \prod_{p<j<q} (i_p - i_j) \overset{\text{neither } p \text{ nor } q}{\prod_{r<s,r,s\notin\{p,q\}}} (i_s - i_r)\right)$$

The last product consists of the product of terms which were in $\prod_{r<s}(i_s - i_r)$ 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 $\text{sgn}_n(i_1, \cdots, i_n)$. Therefore, this switch introduced a minus sign and

$$\text{sgn}_n(j_1, \cdots, j_n) = -\text{sgn}_n(i_1, \cdots, i_n)$$

Now consider the last claim. In computing $\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n)$ there will be the product of $n - \theta$ negative terms

$$(i_{\theta+1} - n) \cdots (i_n - n)$$

and the other terms in the product for computing $\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n)$ are those which are required to compute $\text{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n)$ multiplied by terms of the form $(n - i_j)$ which are nonnegative. It follows that

$$\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) = (-1)^{n-\theta}\, \text{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n)$$

It is obvious that if there are repeats in the list the function gives 0. ■

Lemma 1.9.2 *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, 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 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 sgn_n is unique, if there exist two functions, f and g both satisfying 1.15 and 1.16, you could start with $f(1, \cdots, n) = g(1, \cdots, n) = 1$ and applying the same sequence of switches, eventually arrive at $f(i_1, \cdots, i_n) = g(i_1, \cdots, i_n)$. If any numbers are repeated, then 1.16 gives both functions are equal to zero for that ordered list. ■

Definition 1.9.3 *Given an ordered list of distinct numbers from* $\{1, 2, \cdots, n\}$, *say*

$$(i_1, \cdots, i_n),$$

this ordered list is called a permutation. The symbol for all such permutations is S_n. *The number* $\text{sgn}_n(i_1, \cdots, i_n)$ *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 sgn_n because the context supplies the appropriate n.

1.9.2 The Definition of the Determinant

Definition 1.9.4 *Let f be a real valued function which has the set of ordered lists of numbers from* $\{1, \cdots, n\}$ *as its domain. Define*

$$\sum_{(k_1, \cdots, k_n)} f(k_1 \cdots k_n)$$

to be the sum of all the $f(k_1 \cdots k_n)$ *for all possible choices of ordered lists* $(k_1, \cdots, k_n)$ *of numbers of* $\{1, \cdots, n\}$. *For example,*

$$\sum_{(k_1, k_2)} f(k_1, k_2) = f(1, 2) + f(2, 1) + f(1, 1) + f(2, 2).$$

Definition 1.9.5 *Let* $(a_{ij}) = A$ *denote an $n \times n$ matrix. The determinant of A, denoted by* $\det(A)$ *is defined by*

$$\det(A) \equiv \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots a_{nk_n}$$

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,* $\text{sgn}(k_1, \cdots, k_n) = 0$ *and so that term contributes 0 to the sum.*

Let A be an $n\times n$ matrix $A=(a_{ij})$ and let $(r_1,\cdots,r_n)$ denote an ordered list of n numbers from $\{1,\cdots,n\}$. Let $A(r_1,\cdots,r_n)$ denote the matrix whose k^{th} row is the r_k row of the matrix A. Thus

$$\det(A(r_1,\cdots,r_n))=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n} \tag{1.18}$$

and $A(1,\cdots,n)=A$.

Proposition 1.9.6 *Let $(r_1,\cdots,r_n)$ be an ordered list of numbers from*

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

Then

$$\begin{aligned}\operatorname{sgn}(r_1,\cdots,r_n)\det(A) &= \sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n} && (1.19)\\ &= \det(A(r_1,\cdots,r_n)). && (1.20)\end{aligned}$$

Proof: Let $(1,\cdots,n)=(1,\cdots,r,\cdots s,\cdots,n)$ so $r<s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n))= \tag{1.21}$$

$$\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\,a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s,k_r and k_r, k_s, this equals

$$=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\,a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$=\sum_{(k_1,\cdots,k_n)}-\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right)a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$=-\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{1.22}$$

Consequently,

$$\det(A(1,\cdots,s,\cdots,r,\cdots,n))=-\det(A(1,\cdots,r,\cdots,s,\cdots,n))=-\det(A)$$

Now letting $A(1,\cdots,s,\cdots,r,\cdots,n)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det(A(r_1,\cdots,r_n))=(-1)^p\det(A)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Lemma 1.9.1, this implies

$$\det(A(r_1,\cdots,r_n))=(-1)^p\det(A)=\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1,\cdots,r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 1.21 -1.22 shows that det$(A(r_1,\cdots,r_n))=0$ and also $\operatorname{sgn}(r_1,\cdots,r_n)=0$ so the formula holds in this case also. ■

Observation 1.9.7 *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.

1.9.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A)=\det\left(A^T\right)$.

Corollary 1.9.8 *The following formula for* $\det(A)$ *is valid.*

$$\det(A)=\frac{1}{n!}\cdot\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}. \tag{1.23}$$

And also $\det\left(A^T\right)=\det(A)$ *where A^T is the transpose of A. (Recall that for $A^T=\left(a_{ij}^T\right)$, $a_{ij}^T=a_{ji}$.)*

Proof: From Proposition 1.9.6, if the r_i are distinct,

$$\det(A)=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}(r_1,\cdots,r_n)=0$ and so there is no contribution to the sum.)

$$n!\det(A)=\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T. ■

Corollary 1.9.9 *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* $\det(A)=0$. *Suppose the i^{th} row of A equals*

$$(xa_1+yb_1,\cdots,xa_n+yb_n)$$

Then

$$\det(A)=x\det(A_1)+y\det(A_2)$$

where the i^{th} row of A_1 is $(a_1,\cdots,a_n)$ and the i^{th} row of A_2 is $(b_1,\cdots,b_n)$, 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 1.9.6 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 1.9.8 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,

$$\det(A) = \det\left(A^T\right) = -\det\left(A_1^T\right) = -\det(A_1).$$

If A has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, $\det(A) = -\det(A)$ and so $\det(A) = 0$.

It remains to verify the last assertion.

$$\det(A) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots\left(xa_{rk_i}+yb_{rk_i}\right)\cdots a_{nk_n}$$

$$= x\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_i}\cdots a_{nk_n}$$

$$+y\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots b_{rk_i}\cdots a_{nk_n} \equiv x\det(A_1)+y\det(A_2).$$

The same is true of columns because $\det\left(A^T\right) = \det(A)$ and the rows of A^T are the columns of A. ■

1.9.4 Basic Properties of the Determinant

Definition 1.9.10 *A vector, $\boldsymbol{w}$, is a linear combination $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r\}$ if there exist scalars $c_1,\cdots c_r$ such that $\boldsymbol{w} = \sum_{k=1}^r c_k\boldsymbol{v}_k$. This is the same as saying*

$$\boldsymbol{w} \in \operatorname{span}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r).$$

The following corollary is also of great use.

Corollary 1.9.11 *Suppose A is an $n\times n$ matrix and some column (row) is a linear combination of r other columns (rows). Then* $\det(A) = 0$.

Proof: Let $A = \begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \end{pmatrix}$ 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 1.9.9, $\det(A) =$

$$\det\begin{pmatrix} \boldsymbol{a}_1 & \cdots & \sum_{j\neq i} c_j\boldsymbol{a}_j & \cdots & \boldsymbol{a}_n \end{pmatrix} = \sum_{j\neq i} c_j \det\begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_j & \cdots & \boldsymbol{a}_n \end{pmatrix} = 0$$

because each of these determinants in the sum has two equal rows. ■

Recall the following definition of matrix multiplication.

Definition 1.9.12 *If A and B are $n\times n$ matrices, $A = (a_{ij})$ and $B = (b_{ij})$, $AB = (c_{ij})$ where $c_{ij} \equiv \sum_{k=1}^n a_{ik}b_{kj}$.*

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

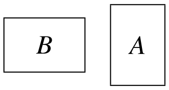
Theorem 1.9.13 *Let A and B be $n\times n$ matrices. Then* $\det(AB)=\det(A)\det(B)$.

Proof: Let c_{ij} be the ij^{th} entry of AB. Then by Proposition 1.9.6,

$$\begin{aligned}\det(AB) &= \sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,c_{1k_1}\cdots c_{nk_n}\\ &= \sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\left(\sum_{r_1}a_{1r_1}b_{r_1k_1}\right)\cdots\left(\sum_{r_n}a_{nr_n}b_{r_nk_n}\right)\end{aligned}$$

$$\begin{aligned}&= \sum_{(r_1\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,b_{r_1k_1}\cdots b_{r_nk_n}\,(a_{1r_1}\cdots a_{nr_n})\\ &= \sum_{(r_1\cdots,r_n)}\operatorname{sgn}(r_1\cdots r_n)\,a_{1r_1}\cdots a_{nr_n}\det(B)=\det(A)\det(B).\blacksquare\end{aligned}$$

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 1.9.14 *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*

$$\det(BA)=\sum_{k=1}^{C(n,m)}\det(B_k)\det(A_k)$$

Proof: This follows from a computation. By Corollary 1.9.8 on Page 37, $\det(BA)=$

$$\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)}\operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)(BA)_{i_1j_1}(BA)_{i_2j_2}\cdots(BA)_{i_mj_m}=$$

$$\begin{gathered}\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)}\operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot\\ \sum_{r_1=1}^{n}B_{i_1r_1}A_{r_1j_1}\sum_{r_2=1}^{n}B_{i_2r_2}A_{r_2j_2}\cdots\sum_{r_m=1}^{n}B_{i_mr_m}A_{r_mj_m}\end{gathered}$$

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.

$$\begin{aligned}=\sum_{k=1}^{C(n,m)}\sum_{\{r_1,\cdots,r_m\}=I_k}\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)}\operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot\\ B_{i_1r_1}A_{r_1j_1}B_{i_2r_2}A_{r_2j_2}\cdots B_{i_mr_m}A_{r_mj_m}\end{aligned}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1,\cdots,r_m\}=I_k} \frac{1}{m!} \sum_{(i_1\cdots i_m)} \operatorname{sgn}(i_1\cdots i_m) B_{i_1r_1}B_{i_2r_2}\cdots B_{i_mr_m} \cdot$$
$$\sum_{(j_1\cdots j_m)} \operatorname{sgn}(j_1\cdots j_m) A_{r_1j_1}A_{r_2j_2}\cdots A_{r_mj_m}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1,\cdots,r_m\}=I_k} \frac{1}{m!} \operatorname{sgn}(r_1\cdots r_m)^2 \det(B_k)\det(A_k) = \sum_{k=1}^{C(n,m)} \det(B_k)\det(A_k)$$

since there are $m!$ ways of arranging the indices $\{r_1,\cdots,r_m\}$. ■

1.9.5 Expansion Using Cofactors

Lemma 1.9.15 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \text{ or } \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{1.24}$$

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 $\mathbf{0}$ denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then $\det(M) = a\det(A)$.

Proof: Denote M by (m_{ij}). Thus in the first case, $m_{nn} = a$ and $m_{ni} = 0$ if $i \neq n$ while in the second case, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$\det(M) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}_n(k_1,\cdots,k_n) m_{1k_1}\cdots m_{nk_n}$$

Letting θ denote the position of n in the ordered list, $(k_1,\cdots,k_n)$ then using the earlier conventions used to prove Lemma 1.9.1, $\det(M)$ equals

$$\sum_{(k_1,\cdots,k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_1,\cdots,k_{\theta-1},\overset{\theta}{k_{\theta+1}},\cdots,\overset{n-1}{k_n}\right) m_{1k_1}\cdots m_{nk_n}$$

Now suppose the second case. Then if $k_n \neq n$, the term involving m_{nk_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

$$a \sum_{(k_1,\cdots,k_{n-1})} \operatorname{sgn}_{n-1}(k_1,\cdots k_{n-1}) m_{1k_1}\cdots m_{(n-1)k_{n-1}} = a\det(A).$$

To get the assertion in the first case, use Corollary 1.9.8 to write

$$\det(M) = \det\left(M^T\right) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det\left(A^T\right) = a\det(A).■$$

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 1.9.16 *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix $\operatorname{cof}(A)$ is defined by $\operatorname{cof}(A) = (c_{ij})$ where to obtain c_{ij} delete the i^{th} row and the j^{th} column of A, take the determinant of the $(n-1) \times (n-1)$ matrix which results, (This is called the ij^{th} minor of A.) and then multiply this number by $(-1)^{i+j}$. To make the formulas easier to remember, $\operatorname{cof}(A)_{ij}$ will denote the ij^{th} entry of the cofactor matrix.*

The following is the main result. Earlier this was given as a definition and the outrageous totally unjustified assertion was made that the same number would be obtained by expanding the determinant along any row or column. The following theorem proves this assertion.

Theorem 1.9.17 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{1.25}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Proof: Let $(a_{i1}, \cdots, a_{in})$ be the i^{th} row of A. Let B_j be the matrix obtained from A by leaving every row the same except the i^{th} row which in B_j equals $(0, \cdots, 0, a_{ij}, 0, \cdots, 0)$. Then by Corollary 1.9.9,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

For example if

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ h & i & j \end{pmatrix}$$

and $i = 2$, then

$$B_1 = \begin{pmatrix} a & b & c \\ d & 0 & 0 \\ h & i & j \end{pmatrix}, B_2 = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ h & i & j \end{pmatrix}, B_3 = \begin{pmatrix} a & b & c \\ 0 & 0 & f \\ h & i & j \end{pmatrix}$$

Denote by A^{ij} the $(n-1) \times (n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of A. Thus $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j} \det(A^{ij})$. At this point, recall that from Proposition 1.9.6, 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 1.9.15,

$$\begin{aligned} \det(B_j) &= (-1)^{n-j} (-1)^{n-i} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) \\ &= (-1)^{i+j} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) = a_{ij} \operatorname{cof}(A)_{ij}. \end{aligned}$$

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the i^{th} row. Also,

$$\det(A) = \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} = \sum_{j=1}^{n} a_{ji} \operatorname{cof}(A)_{ji}$$

which is the formula for expanding $\det(A)$ along the i^{th} column. ■

1.9.6 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. Recall the definition of the inverse of a matrix in Definition 1.6.7 on Page 23.

Theorem 1.9.18 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} \operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ the ij^{th} cofactor of A.

Proof: By Theorem 1.9.17 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ir} \det(A)^{-1} = \det(A)\det(A)^{-1} = 1.$$

Now in the matrix A, replace the k^{th} column with the r^{th} column and then expand along the k^{th} column. This yields for $k \neq r$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = 0$$

because there are two equal columns by Corollary 1.9.9. Summarizing,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 1.9.17, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then A^{-1} exists with $A^{-1} = \left(a_{ij}^{-1}\right)$, where

$$a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose A^{-1} exists. Then by Theorem 1.9.13,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A)\det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. ■

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 1.9.19 *Let A be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix B such that $BA = I$. Then A^{-1} exists and $A^{-1} = B$. Also, if there exists C an $n \times n$ matrix such that $AC = I$, then A^{-1} exists and $A^{-1} = C$.*

Proof: Since $BA = I$, Theorem 1.9.13 implies $\det B \det A = 1$ and so $\det A \neq 0$. Therefore from Theorem 1.9.18, A^{-1} exists. Therefore,

$$A^{-1} = (BA)A^{-1} = B\left(AA^{-1}\right) = BI = B.$$

The case where $CA = I$ 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 1.9.18 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.

1.9.7 Cramer's Rule

In case you are solving a system of equations, $A\boldsymbol{x} = \boldsymbol{y}$ for $\boldsymbol{x}$, it follows that if A^{-1} exists,

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

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Definition 1.9.20 *A matrix M, is upper triangular if $M_{ij} = 0$ whenever $i > j$. Thus such a matrix equals zero below the main diagonal, the entries of the form M_{ii} as shown.*

$$\begin{pmatrix} * & * & \cdots & * \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & * \end{pmatrix}$$

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 1.9.17.

Corollary 1.9.21 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

1.9.8 Rank of a Matrix

Definition 1.9.22 *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 1.9.23 *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 $A = (a_{ij})$ 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^{th} row and let p be an index between 1 and n. Form the following $(r+1) \times (r+1)$ matrix

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} & a_{i_1 p} \\ \vdots & & \vdots & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} & a_{i_r p} \\ a_{l j_1} & \cdots & a_{l j_r} & a_{lp} \end{pmatrix}$$

Of course you can assume $l \notin \{i_1, \cdots, i_r\}$ because there is nothing to prove if the l^{th} row is one of the chosen ones. The above matrix has determinant 0. This is because if $p \notin \{j_1, \cdots, j_r\}$ then the above would be a submatrix of A which is too large to have non zero determinant. On the other hand, if $p \in \{j_1, \cdots, j_r\}$ 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_{lp}. This is given to be nonzero, it being the determinant of the matrix $r \times r$ matrix in the upper left corner. Thus $0 = a_{lp}C + \sum_{k=1}^r C_k a_{i_k p}$ which implies $a_{lp} = \sum_{k=1}^r \frac{-C_k}{C} a_{i_k p} \equiv \sum_{k=1}^r m_k a_{i_k p}$ Since this is true for every p and since m_k does not depend on p, this has shown the l^{th} row is a linear combination of the $i_1, i_2, \cdots, i_r$ rows. ■

Corollary 1.9.24 *The determinant rank equals the row rank.*

Proof: From Theorem 1.9.23, 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 1.9.23 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 4.2.3, 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 1.9.11 the determinant would be 0, a contradiction. ■

Corollary 1.9.25 *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 1.9.24, 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 1.9.26 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. $\det(A) = 0$.

2. *A, A^T are not one to one.*

3. *A is not onto.*

Proof: Suppose $\det(A) = 0$. 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 1.9.23. In particular, it follows that for some m, the m^{th} column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m & \cdots & \boldsymbol{a}_n \end{pmatrix}$ where the columns are denoted by $\boldsymbol{a}_i$, there exists scalars α_i such that $\boldsymbol{a}_m = \sum_{k \neq m} \alpha_k \boldsymbol{a}_k$. Now consider the column vector, $\boldsymbol{x} \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \alpha_n \end{pmatrix}^T$. 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 $|\boldsymbol{x}|^2 = \boldsymbol{x}^T (A\boldsymbol{y}) = (\boldsymbol{x}^T A)\, \boldsymbol{y} = \mathbf{0}\boldsymbol{y} = 0$ 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 $\det(A) \neq 0$ but then from Theorem 1.9.18 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 1.9.27 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. *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.

1.9.9 An Identity of Cauchy

Theorem 1.9.28 *Both the left and the right sides in the following yield the same polynomial in the variables a_i, b_i for $i \leq n$.*

$$\prod_{i,j}(a_i+b_j)\begin{vmatrix} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & & \vdots \\ \frac{1}{a_n+b_1} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix} = \prod_{j<i}(a_i-a_j)(b_i-b_j). \tag{1.26}$$

Proof: The theorem is true if $n=2$. This follows from some computations. Suppose it is true for $n-1$, $n \geq 3$.

$$\begin{vmatrix} \frac{1}{a_1+b_1} & \frac{1}{a_1+b_2} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{a_{n-1}+b_1} & \frac{1}{a_{n-1}+b_2} & & \frac{1}{a_{n-1}+b_n} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix}$$

$$= \begin{vmatrix} \frac{a_n-a_1}{(a_1+b_1)(b_1+a_n)} & \frac{a_n-a_1}{(a_1+b_2)(b_2+a_n)} & \cdots & \frac{a_n-a_1}{(a_1+b_n)(a_n+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{a_n-a_{n-1}}{(a_{n-1}+b_1)(a_n+b_1)} & \frac{a_n-a_{n-1}}{(b_2+a_n)(b_2+a_{n-1})} & & \frac{a_n-a_{n-1}}{(a_n+b_n)(b_n+a_{n-1})} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix}$$

Continuing to use the multilinear properties of determinants, this equals

$$\begin{vmatrix} \frac{1}{(a_1+b_1)(b_1+a_n)} & \frac{1}{(a_1+b_2)(b_2+a_n)} & \cdots & \frac{1}{(a_1+b_n)(a_n+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(a_{n-1}+b_1)(a_n+b_1)} & \frac{1}{(b_2+a_n)(b_2+a_{n-1})} & & \frac{1}{(a_n+b_n)(b_n+a_{n-1})} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix} \prod_{k=1}^{n-1}(a_n-a_k)$$

and this equals

$$\begin{vmatrix} \frac{1}{(a_1+b_1)} & \frac{1}{(a_1+b_2)} & \cdots & \frac{1}{(a_1+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(a_{n-1}+b_1)} & \frac{1}{(b_2+a_{n-1})} & & \frac{1}{(b_n+a_{n-1})} \\ 1 & 1 & \cdots & 1 \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)}$$

Now take -1 times the last column and add to each previous column. Thus it equals

$$\begin{vmatrix} \frac{b_n-b_1}{(a_1+b_1)(a_1+b_n)} & \frac{b_n-b_2}{(a_1+b_2)(a_1+b_n)} & \cdots & \frac{1}{(a_1+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{b_n-b_1}{(b_1+a_{n-1})(b_n+a_{n-1})} & \frac{b_n-b_2}{(b_2+a_{n-1})(b_n+a_{n-1})} & & \frac{1}{(a_{n-1}+b_n)} \\ 0 & 0 & \cdots & 1 \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)}$$

Now continue simplifying using the multilinear property of the determinant.

$$\left|\begin{array}{cccc} \frac{1}{(a_1+b_1)} & \frac{1}{(a_1+b_2)} & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(b_1+a_{n-1})} & \frac{1}{(b_2+a_{n-1})} & & 1 \\ 0 & 0 & \cdots & 1 \end{array}\right| \frac{\prod_{k=1}^{n-1}(a_n-a_k)\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n}(a_n+b_k)\prod_{k=1}^{n-1}(a_k+b_n)}$$

Expanding along the bottom row, what has just resulted is

$$\left|\begin{array}{ccc} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_{n-1}} \\ \vdots & \cdots & \vdots \\ \frac{1}{a_{n-1}+b_1} & \cdots & \frac{1}{a_{n-1}+b_{n-1}} \end{array}\right| \frac{\prod_{k=1}^{n-1}(a_n-a_k)\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n}(a_n+b_k)\prod_{k=1}^{n-1}(a_k+b_n)}$$

By induction this equals

$$\frac{\prod_{j<i\leq n-1}(a_i-a_j)(b_i-b_j)}{\prod_{i,j\leq n-1}(a_i+b_j)} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)} \frac{\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n-1}(a_k+b_n)}$$

$$= \frac{\prod_{j<i\leq n}(a_i-a_j)(b_i-b_j)}{\prod_{i,j\leq n}(a_i+b_j)} \blacksquare$$

1.10 The Cayley Hamilton Theorem

Definition 1.10.1 *Let A be an $n\times n$ matrix. The characteristic polynomial is defined as*

$$q_A(t) \equiv \det(tI-A)$$

and the solutions to $q_A(t)=0$ are called eigenvalues. For A a matrix and $p(t)=t^n+a_{n-1}t^{n-1}+\cdots+a_1t+a_0$, denote by $p(A)$ the matrix defined by

$$p(A) \equiv A^n + a_{n-1}A^{n-1} + \cdots + a_1A + a_0I.$$

The explanation for the last term is that A^0 is interpreted as I, the identity matrix.

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[2]. 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 1.10.2 *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 $A_i=0$.

[2] 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.

Proof: Suppose some $A_i \neq 0$. Let p be the largest index of those which are non zero. Then multiply by λ^{-p}.

$$A_0\lambda^{-p} + A_1\lambda^{-p+1} + \cdots + A_{p-1}\lambda^{-1} + A_p = 0$$

Now let $\lambda \to \infty$. Thus $A_p = 0$ after all. Hence each $A_i = 0$. ■

With the lemma, here is a simple corollary.

Corollary 1.10.3 *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 λ 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 1.10.4 *Let A be an $n \times n$ matrix and let $q(\lambda) \equiv \det(\lambda I - A)$ 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 λ cannot be in the finite list of eigenvalues of A and so for such λ, $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 1.9.18

$$C(\lambda) = q(\lambda)(\lambda I - A)^{-1}.$$

Say

$$q(\lambda) = a_0 + a_1\lambda + \cdots + \lambda^n$$

Note that each entry in $C(\lambda)$ is a polynomial in λ having degree no more than $n-1$. For example, you might have something like

$$C(\lambda) = \begin{pmatrix} \lambda^2 - 6\lambda + 9 & 3 - \lambda & 0 \\ 2\lambda - 6 & \lambda^2 - 3\lambda & 0 \\ \lambda - 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2 \end{pmatrix}$$

$$= \begin{pmatrix} 9 & 3 & 0 \\ -6 & 0 & 0 \\ -1 & -1 & 2 \end{pmatrix} + \lambda \begin{pmatrix} -6 & -1 & 0 \\ 2 & -3 & 0 \\ 1 & 1 & -3 \end{pmatrix} + \lambda^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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,

$$\begin{aligned} a_0 I + a_1 I\lambda + \cdots + I\lambda^n &= C_0\lambda + C_1\lambda^2 + \cdots + C_{n-1}\lambda^n \\ &\quad - \left[C_0 A + C_1 A\lambda + \cdots + C_{n-1}A\lambda^{n-1}\right] \\ &= -C_0 A + (C_0 - C_1 A)\lambda + (C_1 - C_2 A)\lambda^2 + \cdots + (C_{n-2} - C_{n-1}A)\lambda^{n-1} + C_{n-1}\lambda^n \end{aligned}$$

Then, using Corollary 1.10.3, one can replace λ 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. ■

Chapter 2

Some Basic Topics

This chapter contains basic definitions and a few fundamental theorems which will be used throughout the book whenever convenient.

2.1 Basic Definitions

A set is a collection of things called elements of the set. For example, the set of integers, the collection of signed whole numbers such as 1,2,-4, etc. This set whose existence will be assumed is denoted by $\mathbb{Z}$. Other sets could be the set of people in a family or the set of donuts in a display case at the store. Sometimes parentheses, $\{ \}$ specify a set by listing the things which are in the set between the parentheses. For example the set of integers between -1 and 2, including these numbers could be denoted as $\{-1,0,1,2\}$. The notation signifying x is an element of a set S, is written as $x \in S$. Thus, $1 \in \{-1,0,1,2,3\}$. Here are some axioms about sets. Axioms are statements which are accepted, not proved.

Axiom 2.1.1 *Two sets are equal if and only if they have the same elements.*

Axiom 2.1.2 *To every set, A, and to every condition $S(x)$ there corresponds a set, B, whose elements are exactly those elements x of A for which $S(x)$ holds.*

Axiom 2.1.3 *For every collection of sets there exists a set that contains all the elements that belong to at least one set of the given collection. (You can take the union of a bunch of sets.)*

Axiom 2.1.4 *The Cartesian product of a nonempty family of nonempty sets is nonempty.*

Axiom 2.1.5 *If A is a set there exists a set, $\mathscr{P}(A)$ such that $\mathscr{P}(A)$ is the set of all subsets of A. This is called the power set.*

These axioms are referred to as the axiom of extension, axiom of specification, axiom of unions, axiom of choice, and axiom of powers respectively.

It seems fairly clear you should want to believe in the axiom of extension. It is merely saying, for example, that $\{1,2,3\} = \{2,3,1\}$ since these two sets have the same elements in them. Similarly, it would seem you should be able to specify a new set from a given set using some "condition" which can be used as a test to determine whether the element in question is in the set. For example, the set of all integers which are multiples of 2. This set could be specified as follows.

$$\{x \in \mathbb{Z} : x = 2y \text{ for some } y \in \mathbb{Z}\}.$$

In this notation, the colon is read as "such that" and in this case the condition is being a multiple of 2.

Another example of political interest, could be the set of all judges who are not judicial activists. I think you can see this last is not a very precise condition since there is no way to determine to everyone's satisfaction whether a given judge is an activist. Also, **just because something is grammatically correct does not mean it makes any sense.** For example consider the following nonsense.

$$S = \{x \in \text{set of dogs : it is colder in the mountains than in the winter}\}.$$

So what is a condition?

We will leave these sorts of considerations and assume our conditions make sense, whatever that means. The axiom of unions states that for any collection of sets, there is a set consisting of all the elements in each of the sets in the collection. Of course this is also open to further consideration. What is a collection? Maybe it would be better to say "set of sets" or, given a set whose elements are sets there exists a set whose elements consist of exactly those things which are elements of at least one of these sets. If $\mathscr{S}$ is such a set whose elements are sets,

$$\cup\{A : A \in \mathscr{S}\} \text{ or } \cup\mathscr{S}$$

signify this union.

Something is in the Cartesian product of a set or "family" of 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.

A is a subset of B, written $A \subseteq B$, if every element of A is also an element of B. This can also be written as $B \supseteq A$. A is a proper subset of B, written $A \subset B$ or $B \supset A$ if A is a subset of B but A is not equal to $B, A \neq B$. $A \cap B$ denotes the intersection of the two sets, A and B and it means the set of elements of A which are also elements of B. The axiom of specification shows this is a set. The empty set is the set which has no elements in it, denoted as $\emptyset$. $A \cup B$ denotes the union of the two sets, A and B and it means the set of all elements which are in either of the sets. It is a set because of the axiom of unions.

The complement of a set, (the set of things which are not in the given set) must be taken with respect to a given set called the universal set which is a set which contains the one whose complement is being taken. Thus, the complement of A, denoted as A^C (or more precisely as $X \setminus A$) is a set obtained from using the axiom of specification to write

$$A^C \equiv \{x \in X : x \notin A\}$$

The symbol $\notin$ means: "is not an element of". Note the axiom of specification takes place relative to a given set. Without this universal set it makes no sense to use the axiom of specification to obtain the complement.

Words such as "all" or "there exists" are called quantifiers and they must be understood relative to some given set. For example, the set of all integers larger than 3. Or there exists an integer larger than 7. Such statements have to do with a given set, in this case the integers. Failure to have a reference set when quantifiers are used turns out to be illogical even though such usage may be grammatically correct. Quantifiers are used often enough that there are symbols for them. The symbol $\forall$ is read as "for all" or "for every" and the symbol $\exists$ is read as "there exists". Thus $\forall\forall\exists\exists$ could mean for every upside down A there exists a backwards E.

DeMorgan's laws are very useful in mathematics. Let $\mathscr{S}$ be a set of sets each of which is contained in some universal set, U. Then

$$\cup\left\{A^C : A\in\mathscr{S}\right\} = \left(\cap\left\{A : A\in\mathscr{S}\right\}\right)^C$$

and

$$\cap\left\{A^C : A\in\mathscr{S}\right\} = \left(\cup\left\{A : A\in\mathscr{S}\right\}\right)^C.$$

These laws follow directly from the definitions. Also following directly from the definitions are:

Let $\mathscr{S}$ be a set of sets then

$$B\cup\cup\left\{A : A\in\mathscr{S}\right\} = \cup\left\{B\cup A : A\in\mathscr{S}\right\}.$$

and: Let $\mathscr{S}$ be a set of sets show

$$B\cap\cup\left\{A : A\in\mathscr{S}\right\} = \cup\left\{B\cap A : A\in\mathscr{S}\right\}.$$

Unfortunately, there is no single universal set which can be used for all sets. Here is why: Suppose there were. Call it S. Then you could consider A the set of all elements of S which are not elements of themselves, this from the axiom of specification. If A is an element of itself, then it fails to qualify for inclusion in A. Therefore, it must not be an element of itself. However, if this is so, it qualifies for inclusion in A so it is an element of itself and so this can't be true either. Thus the most basic of conditions you could imagine, that of being an element of, is meaningless and so allowing such a set causes the whole theory to be meaningless. The solution is to not allow a universal set. As mentioned by Halmos in Naive set theory, "Nothing contains everything". Always beware of statements involving quantifiers wherever they occur, even this one. This little observation described above is due to Bertrand Russell and is called Russell's paradox.

2.2 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 2.2.1 *Let X and Y be sets.*

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

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 (x,y_1) 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 $f : D(f)\to Y$. 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.

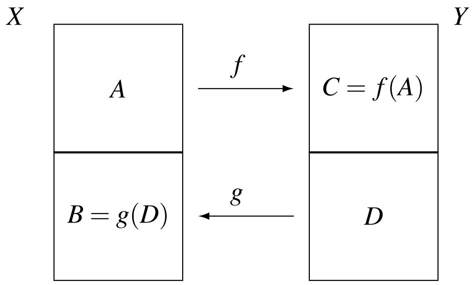
The following theorem which is interesting for its own sake will be used to prove the Schroder Bernstein theorem.

Theorem 2.2.2 *Let $f : X \to Y$ and $g : Y \to X$ be two functions. Then there exist sets A, B, C, D, such that*

$$A \cup B = X,\ C \cup D = Y,\ A \cap B = \emptyset,\ C \cap D = \emptyset,$$

$$f(A) = C,\ g(D) = B.$$

The following picture illustrates the conclusion of this theorem.



Proof: Consider the empty set, $\emptyset \subseteq X$. If $y \in Y \setminus f(\emptyset)$, 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 $y \in Y \setminus f(A_0)$, $g(y) \notin A_0$.

$$\mathscr{A} \equiv \{A_0 \subseteq X : A_0 \text{ satisfies } \mathscr{P}\}.$$

Let $A = \cup\mathscr{A}$. If $y \in Y \setminus f(A)$, then for each $A_0 \in \mathscr{A}$, $y \in Y \setminus f(A_0)$ 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

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

It only remains to verify that $g(D) = B$. It was just shown that $g(D) \subseteq B$.

Suppose $x \in B = X \setminus A$. Then $A \cup \{x\}$ does not satisfy $\mathscr{P}$ and so there exists $y \in Y \setminus f(A \cup \{x\}) \subseteq D$ 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 2.2.3 *(Schroder Bernstein) If $f : X \to Y$ and $g : Y \to X$ are one to one, then there exists $h : X \to Y$ which is one to one and onto.*

Proof: Let A,B,C,D be the sets of Theorem2.2.2 and define

$$h(x) \equiv \begin{cases} f(x) & \text{if } x \in A \\ g^{-1}(x) & \text{if } x \in B \end{cases}$$

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 2.2.4 *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 2.2.5 *If $f : X \to Y$ is onto and $g : Y \to X$ is onto, then there exists $h : X \to Y$ 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 $f_0^{-1}(y_1) = f_0^{-1}(y_2)$, then $y_1 = f\left(f_0^{-1}(y_1)\right) = f\left(f_0^{-1}(y_2)\right) = y_2$. Similarly g_0^{-1} is one to one. Therefore, by the Schroder Bernstein theorem, there exists $h : X \to Y$ which is one to one and onto. ■

Definition 2.2.6 *A set S, is finite if there exists a natural number n and a map θ 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 θ mapping $\mathbb{N}$ one to one and onto S.(When θ maps a set A to a set B, this will be written as $\theta : A \to B$ in the future.) Here $\mathbb{N} \equiv \{1, 2, \cdots\}$, the natural numbers. S is at most countable if there exists a map $\theta : \mathbb{N} \to S$ 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 2.2.7 *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 $\eta : \mathbb{N} \to X$ which is onto. Define $\eta(i) \equiv x_i$ and consider X as the set $\{x_1, x_2, x_3, \cdots\}$. Similarly, consider Y as the set $\{y_1, y_2, y_3, \cdots\}$. It follows the elements of $X \times Y$ are included in the following rectangular array.

$$\begin{array}{lllll} (x_1,y_1) & (x_1,y_2) & (x_1,y_3) & \cdots & \leftarrow \text{Those which have } x_1 \text{ in first slot.} \\ (x_2,y_1) & (x_2,y_2) & (x_2,y_3) & \cdots & \leftarrow \text{Those which have } x_2 \text{ in first slot.} \\ (x_3,y_1) & (x_3,y_2) & (x_3,y_3) & \cdots & \leftarrow \text{Those which have } x_3 \text{ in first slot.} \;\cdot \\ \vdots & \vdots & \vdots & & \qquad\qquad\vdots \end{array}$$

Follow a path through this array as follows.

$$\begin{tikzcd}
(x_1,y_1) \arrow[r] & (x_1,y_2) \arrow[dl] & (x_1,y_3) \arrow[r] & {} \\
(x_2,y_1) \arrow[d] & (x_2,y_2) \arrow[ur] & & \\
(x_3,y_1) \arrow[ur] & & &
\end{tikzcd}$$

Thus the first element of $X \times Y$ is (x_1, y_1), the second element of $X \times Y$ is (x_1, y_2), the third element of $X \times Y$ is (x_2, y_1) 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 $\alpha : \mathbb{N} \to X$ which is one to one and onto. Let $\beta : X \times Y \to \mathbb{N}$ be defined by $\beta((x,y)) \equiv \alpha^{-1}(x)$. Thus β is onto $\mathbb{N}$. By the first part there exists a function from $\mathbb{N}$ onto $X \times Y$. Therefore, by Corollary 2.2.5, there exists a one to one and onto mapping from $X \times Y$ to $\mathbb{N}$. ■

Theorem 2.2.8 *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,

$$X = \{x_1, x_2, x_3, \cdots\}$$

and

$$Y = \{y_1, y_2, y_3, \cdots\}.$$

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

$$\begin{tikzcd}
x_1 \arrow[r] & x_2 \arrow[dl] & x_3 \arrow[r] & {} \\
y_1 \arrow[r] & y_2 \arrow[ur] & &
\end{tikzcd}$$

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 $\alpha : \mathbb{N} \to X$ 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, β 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 2.2.5 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.

2.3 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 weighed 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 2.3.1 *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 2.3.2 $[x]$ *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 2.3.3 *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 $[x] \cap [y] = \emptyset$.*

2.4 sup and inf

It is assumed in all that is done that $\mathbb{R}$ is complete. There are two ways to describe completeness of $\mathbb{R}$. One is to say that every bounded set has a least upper bound and a greatest lower bound. The other is to say that every Cauchy sequence converges. These two equivalent notions of completeness will be taken as given. Cauchy sequences are discussed a little later.

The symbol, $\mathbb{F}$ will mean either $\mathbb{R}$ or $\mathbb{C}$. The symbol $[-\infty,\infty]$ will mean all real numbers along with $+\infty$ and $-\infty$ which are points which we pretend are at the right and left ends of the real line respectively. The inclusion of these make believe points makes the statement of certain theorems less trouble.

Definition 2.4.1 *For $A \subseteq [-\infty,\infty]$, $A \neq \emptyset$ $\sup A$ is defined as the least upper bound in case A is bounded above by a real number and equals ∞ if A is not bounded above. Similarly $\inf A$ is defined to equal the greatest lower bound in case A is bounded below by a real number and equals $-\infty$ in case A is not bounded below.*

Lemma 2.4.2 *If $\{A_n\}$ is an increasing sequence in $[-\infty,\infty]$, then*

$$\sup\{A_n : n \in \mathbb{N}\} = \lim_{n\to\infty} A_n.$$

Similarly, if $\{A_n\}$ is decreasing, then

$$\inf\{A_n : n \in \mathbb{N}\} = \lim_{n\to\infty} A_n.$$

Proof: Let $\sup(\{A_n : n \in \mathbb{N}\}) = r$. In the first case, suppose $r < \infty$. Then letting $\varepsilon > 0$ be given, there exists n such that $A_n \in (r-\varepsilon, r]$. Since $\{A_n\}$ is increasing, it follows if $m > n$, then $r - \varepsilon < A_n \leq A_m \leq r$ and so $\lim_{n\to\infty} A_n = r$ as claimed. In the case where $r = \infty$, then if a is a real number, there exists n such that $A_n > a$. Since $\{A_k\}$ is increasing, it follows that if $m > n$, $A_m > a$. But this is what is meant by $\lim_{n\to\infty} A_n = \infty$. The other case is that $r = -\infty$. But in this case, $A_n = -\infty$ for all n and so $\lim_{n\to\infty} A_n = -\infty$. The case where A_n is decreasing is entirely similar. ■

2.5 Double Series

Double series are of the form $\sum_{k=m}^{\infty}\sum_{j=m}^{\infty}a_{jk} \equiv \sum_{k=m}^{\infty}\left(\sum_{j=m}^{\infty}a_{jk}\right)$. In other words, first sum on j yielding something which depends on k and then sum these. The major consideration for these double series is the question of when

$$\sum_{k=m}^{\infty}\sum_{j=m}^{\infty}a_{jk} = \sum_{j=m}^{\infty}\sum_{k=m}^{\infty}a_{jk}$$

In other words, when does it make no difference which subscript is summed over first? In the case of finite sums there is no issue here. You can always write $\sum_{k=m}^{M}\sum_{j=m}^{N}a_{jk} = \sum_{j=m}^{N}\sum_{k=m}^{M}a_{jk}$ because addition is commutative. However, there are limits involved with infinite sums and the interchange in order of summation involves taking limits in a different order. Therefore, it is not always true that it is permissible to interchange the two sums. A general rule of thumb is this: If something involves changing the order in which two limits are taken, you may not do it without agonizing over the question. In general, limits foul up algebra and also introduce things which are counter intuitive. Here is an example. This example is a little technical. It is placed here just to prove conclusively there is a question which needs to be considered.

Example 2.5.1 *Consider the following picture which depicts some of the ordered pairs (m,n) where m,n are positive integers.*

$$\begin{array}{cccccc}
 & & \vdots & & & \\
0 & 0 & c & 0 & -c & \\
0 & c & 0 & -c & 0 & \cdots \\
b & 0 & -c & 0 & 0 & \\
0 & a & 0 & 0 & 0 &
\end{array}$$

The a,b,c are the values of a_{mn}. Thus $a_{nn} = 0$ for all $n \geq 1$, $a_{21} = a, a_{12} = b, a_{m(m+1)} = -c$ whenever $m > 1$, and $a_{m(m-1)} = c$ whenever $m > 2$. The numbers next to the point are the values of a_{mn}. You see $a_{nn} = 0$ for all n, $a_{21} = a, a_{12} = b, a_{mn} = c$ for (m,n) on the line $y = 1 + x$ whenever $m > 1$, and $a_{mn} = -c$ for all (m,n) on the line $y = x - 1$ whenever $m > 2$.

Then $\sum_{m=1}^{\infty}a_{mn} = a$ if $n = 1$, $\sum_{m=1}^{\infty}a_{mn} = b - c$ if $n = 2$ and if $n > 2, \sum_{m=1}^{\infty}a_{mn} = 0$. Therefore, $\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}a_{mn} = a + b - c$. Next observe that $\sum_{n=1}^{\infty}a_{mn} = b$ if $m = 1, \sum_{n=1}^{\infty}a_{mn} = a + c$ if $m = 2$, and $\sum_{n=1}^{\infty}a_{mn} = 0$ if $m > 2$. Therefore, $\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}a_{mn} = b + a + c$ and so the two sums are different. Moreover, you can see that by assigning different values of a, b, and c, you can get an example for any two different numbers desired.

It turns out that if $a_{ij} \geq 0$ for all i, j, then you can always interchange the order of summation. This is shown next and is based on the following lemma. First, some notation should be discussed.

Definition 2.5.2 *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, ∞, 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 ∞. The symbol, $-\infty$ is interpreted similarly. Then* $\sup_{a\in A} f(a,b)$ *means* $\sup(S_b)$ *where* $S_b \equiv \{f(a,b) : a \in A\}$.

Unlike limits, you can take the sup in different orders.

Lemma 2.5.3 *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. ■

Theorem 2.5.4 *Let $a_{ij} \geq 0$. Then $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$.*

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

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij}$$

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

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij} = \sup_n \lim_{m\to\infty} \sum_{j=r}^{m}\sum_{i=r}^{n} a_{ij}$$

$$= \sup_n \lim_{m\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{m} a_{ij} = \sup_n \sum_{i=r}^{n} \lim_{m\to\infty} \sum_{j=r}^{m} a_{ij}$$

$$= \sup_n \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \lim_{n\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty}\sum_{j=r}^{\infty} a_{ij}$$

Interchanging the i and j in the above argument proves the theorem. ■

2.6 lim sup and lim inf

Sometimes the limit of a sequence does not exist. For example, if $a_n = (-1)^n$, then $\lim_{n\to\infty} a_n$ does not exist. This is because the terms of the sequence are a distance of 1 apart. Therefore there can't exist a single number such that all the terms of the sequence are ultimately within $1/4$ of that number. The nice thing about lim sup and lim inf is that they always exist. First here is a simple lemma and definition. First review the definition of inf and sup on Page 55 along with the simple properties of these things.

Definition 2.6.1 *Denote by $[-\infty,\infty]$ the real line along with symbols ∞ and $-\infty$. It is understood that ∞ is larger than every real number and $-\infty$ is smaller than every real number. Then if $\{A_n\}$ is an increasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals ∞ if the only upper bound of the set $\{A_n\}$ is ∞. If $\{A_n\}$ is bounded above by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the least upper bound of $\{A_n\}$. If $\{A_n\}$ is a decreasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals $-\infty$ if the only lower bound of the sequence $\{A_n\}$ is $-\infty$. If $\{A_n\}$ is bounded below by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the greatest lower bound of $\{A_n\}$. More simply, if $\{A_n\}$ is increasing,$\lim_{n\to\infty} A_n \equiv \sup\{A_n\}$ and if $\{A_n\}$ is decreasing then $\lim_{n\to\infty} A_n \equiv \inf\{A_n\}$.*

Lemma 2.6.2 *Let $\{a_n\}$ be a sequence of real numbers and let $U_n \equiv \sup\{a_k : k \geq n\}$. Then $\{U_n\}$ is a decreasing sequence. Also if $L_n \equiv \inf\{a_k : k \geq n\}$, then $\{L_n\}$ is an increasing sequence. Therefore, $\lim_{n\to\infty} L_n$ and $\lim_{n\to\infty} U_n$ both exist.*

Proof: Let W_n be an upper bound for $\{a_k : k \geq n\}$. Then since these sets are getting smaller, it follows that for $m < n$, W_m is an upper bound for $\{a_k : k \geq n\}$. In particular if $W_m = U_m$, then U_m is an upper bound for $\{a_k : k \geq n\}$ and so U_m is at least as large as U_n, the least upper bound for $\{a_k : k \geq n\}$. The claim that $\{L_n\}$ is decreasing is similar. ■

From the lemma, the following definition makes sense.

Definition 2.6.3 *Let $\{a_n\}$ be any sequence of points of $[-\infty,\infty]$*

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\}$$

$$\liminf_{n\to\infty} a_n \equiv \lim_{n\to\infty} \inf\{a_k : k \geq n\}.$$

Theorem 2.6.4 *Suppose $\{a_n\}$ is a sequence of real numbers and also that*

$$\limsup_{n\to\infty} a_n, \liminf_{n\to\infty} a_n$$

are both real numbers. Then $\lim_{n\to\infty} a_n$ exists if and only if the two numbers are equal and in this case,

$$\lim_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n.$$

Proof: First note that $\sup\{a_k : k \geq n\} \geq \inf\{a_k : k \geq n\}$ and so,

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\} \geq \lim_{n\to\infty} \inf\{a_k : k \geq n\} \equiv \liminf_{n\to\infty} a_n.$$

Suppose first that $\lim_{n\to\infty} a_n$ 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

$$|a_n - a_m| \leq |a_n - a| + |a - a_n| < \frac{\varepsilon}{6} + \frac{\varepsilon}{6} = \frac{\varepsilon}{3}.$$

From the definition of $\sup\{a_k : k \geq N\}$, there exists $n_1 \geq N$ such that $\sup\{a_k : k \geq N\} \leq a_{n_1} + \varepsilon/3$. Similarly, there exists $n_2 \geq N$ such that $\inf\{a_k : k \geq N\} \geq a_{n_2} - \varepsilon/3$. It follows that

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} \leq |a_{n_1} - a_{n_2}| + \frac{2\varepsilon}{3} < \varepsilon.$$

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

$$0 \leq \lim_{N\to\infty} \sup\{a_k : k \geq N\} - \lim_{N\to\infty} \inf\{a_k : k \geq N\} \leq \varepsilon$$

Since ε is arbitrary, this shows

$$\lim_{N\to\infty} \sup\{a_k : k \geq N\} = \lim_{N\to\infty} \inf\{a_k : k \geq N\} \tag{2.1}$$

Next suppose 2.1 and both equal $a \in \mathbb{R}$. Then

$$\lim_{N\to\infty} \left(\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\}\right) = 0$$

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

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} < \varepsilon,$$

and for every N,

$$\inf\{a_k : k \geq N\} \leq a \leq \sup\{a_k : k \geq N\}$$

Thus if $n \geq N$, $|a - a_n| < \varepsilon$ which implies that $\lim_{n\to\infty} a_n = a$. In case

$$a = \infty = \lim_{N\to\infty} \sup\{a_k : k \geq N\} = \lim_{N\to\infty} \inf\{a_k : k \geq N\}$$

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

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

Theorem 2.6.5 *Suppose $\{a_n\}$ is a sequence of points of $[-\infty,\infty]$. Also define $\lambda = \limsup_{n\to\infty} a_n$. 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 $\gamma = \liminf_{n\to\infty} a_n$.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 2.6.6 *Let* $\lim_{n\to\infty} a_n = a > 0$. *Then* $\limsup_{n\to\infty} a_n b_n = a \limsup_{n\to\infty} b_n$.

Proof: This follows from the definition. Let $\lambda_n = \sup\{a_k b_k : k \geq n\}$. For all n large enough, $a_n > a - \varepsilon$ where ε is small enough that $a - \varepsilon > 0$. Therefore,

$$\lambda_n \geq \sup\{b_k : k \geq n\}(a - \varepsilon)$$

for all n large enough. Then

$$\begin{aligned}
\limsup_{n\to\infty} a_n b_n &= \lim_{n\to\infty} \lambda_n \equiv \limsup_{n\to\infty} a_n b_n \geq \lim_{n\to\infty} \left(\sup\{b_k : k \geq n\}(a - \varepsilon)\right) \\
&= (a - \varepsilon) \limsup_{n\to\infty} b_n
\end{aligned}$$

Similar reasoning shows $\limsup_{n\to\infty} a_n b_n \leq (a + \varepsilon) \limsup_{n\to\infty} b_n$. Now since $\varepsilon > 0$ is arbitrary, the conclusion follows. ■

2.7 Nested Interval Lemma

The nested interval lemma is a simple and important lemma which is used later quite a bit.

Lemma 2.7.1 *Let $[a_k, b_k] \supseteq [a_{k+1}, b_{k+1}]$ for all $k = 1, 2, 3, \cdots$. Then there exists a point p in $\cap_{k=1}^{\infty} [a_k, b_k]$. If $\lim_{k\to\infty}(b_k - a_k) = 0$, then there is only one such point*

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 $x \in [\sup_k a_k, \inf_l b_l]$. Then for every $k, a_k \leq x \leq b_k$. Hence $x \in \cap_{k=1}^{\infty} [a_k, b_k]$.

To see the last claim, if q is another point in all the intervals, then both p and q are in $[a_k, b_k]$ and so $|p-q| \leq (b_k - a_k) < \varepsilon$ if k is large enough. Since ε is arbitrary, $p = q$. ■

2.8 The Hausdorff Maximal Theorem

This major theorem, or something like it is either absolutely essential or extremely convenient. First is the definition of what is meant by a partial order.

Definition 2.8.1 *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 $x \in \mathscr{F} \setminus \cup \mathscr{C}$, then $\mathscr{C} \cup \{x\}$ is no longer a chain.*

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 2.8.2 *Let $\mathscr{F}$ be a nonempty partially ordered set with order $\prec$. Then there exists a maximal chain.*

Proof: Suppose not. Then for $\mathscr{C}$ a chain, let $\theta\mathscr{C}$ denote $\mathscr{C} \cup \{x_{\mathscr{C}}\}$. 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 $\{x_0\}$ 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 $\{x_0\}$. 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. $\{x_0\} \in \mathscr{X}$

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 $\mathscr{B} \equiv \left\{\mathscr{D} \in \mathscr{Y}_0 : \mathscr{D} \subseteq \mathscr{C}_0 \text{ or } x_{\mathscr{C}_0} \in \cup\mathscr{D}\right\}$. 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 $\{x_0\} \in \mathscr{B}$?

$\{x_0\}$ cannot properly contain $\mathscr{C}_0$ since $x_0 \in \cup\mathscr{C}_0$. Therefore, $\mathscr{C}_0 \supseteq \{x_0\}$ so $\{x_0\} \in \mathscr{B}$.

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 $\mathscr{D} \supsetneq \mathscr{C}_0$ 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 $x_{\mathscr{C}_0} \in \cup\tilde{\mathscr{C}} \subseteq \cup\mathscr{D}$. 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: $\mathscr{D} \supsetneq \mathscr{C}_0$. 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 $\mathscr{D} \subseteq \mathscr{C}_0 \subsetneq \mathscr{D} \cup \{x_{\mathscr{D}}\}$. If $x \in \mathscr{C}_0$ and x is not in $\mathscr{D}$ then $\mathscr{D} \cup \{x\} \subseteq \mathscr{C}_0 \subsetneq \mathscr{D} \cup \{x_{\mathscr{D}}\}$. 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 $\mathscr{Y}_0$. $\{x_0\}$ 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 $\mathscr{Y}_0$ 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 $\cup(\theta \cup \mathscr{Y}_0) \supsetneq \cup(\cup\mathscr{Y}_0) \supseteq \cup(\theta \cup \mathscr{Y}_0)$ 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

$$\begin{gathered} x \leq x, \\ \text{either } x \leq y \text{ or } y \leq x \\ \text{if } x \leq y \text{ and } y \leq z \text{ then } x \leq z. \end{gathered}$$

and $\leq$ is a well order if $(X, \leq)$ 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 2.8.3 *The Hausdorff maximal principle implies every nonempty set can be well-ordered.*

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 $(S_2, \leq_2)$ is well-ordered and if $y \in S_2 \setminus S_1$ then $x \leq_2 y$ 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_∞ as follows. If $x,\ y$ are elements of X_∞, 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_∞. Then $S \cap U \neq \emptyset$ for some $S \in \mathscr{C}$. Because of the definition of $\leq$, if $y \in S_2 \setminus S_1,\ S_i \in \mathscr{C}$, then $x \leq y$ for all $x \in S_1$. Thus, if $y \in X_\infty \setminus S$ 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_∞ is well-ordered. Now suppose there exists $z \in X \setminus X_\infty$. Define the following order, $\leq_1$, on $X_\infty \cup \{z\}$.

$$x \leq_1 y \text{ if and only if } x \leq y \text{ whenever } x, y \in X_\infty$$

$$x \leq_1 z \text{ whenever } x \in X_\infty.$$

Then let $\widetilde{\mathscr{C}} = \{S \in \mathscr{C} \text{ or } X_\infty \cup \{z\}\}$. Then $\widetilde{\mathscr{C}}$ is a strictly larger chain than $\mathscr{C}$ contradicting maximality of $\mathscr{C}$. Thus $X \setminus X_\infty = \emptyset$ and this shows X is well-ordered by $\leq$. ■

With these two lemmas the main result follows.

Theorem 2.8.4 *The following are equivalent.*

The axiom of choice

The Hausdorff maximal principle

The well-ordering principle.

Proof: It remains to show 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 $X = \cup\{X_i : i \in I\}$ and well order X. Let $f(i)$ be the smallest element of X_i. Then $f \in \prod_{i \in I} X_i$. ■

The book by Hewitt and Stromberg [23] has more equivalences.

2.9 Exercises

1. Zorn's lemma says that if you have a nonempty partially ordered set $\mathscr{F}$ and every chain $\mathscr{C}$ has an upper bound, then there is a maximal element in $\mathscr{F}$, some x such that if $x \prec y$ then $x = y$. Show this is equivalent to the Hausdorff maximal principle.

2. A Hamel basis is a set of vectors B in a vector space X such that every element of X can be written in a unique way as a **finite** linear combination of vectors of B. Show every vector space has a Hamel basis. In fact, these are not used much outside of finite dimensional settings because it can be shown that in every complete normed linear space which is not finite dimensional, the Hamel basis must be uncountable but it is nice to know they exist.

Chapter 3

Metric Spaces

3.1 Open and Closed Sets, Sequences, Limit Points

It is most efficient to discus things in terms of abstract metric spaces to begin with.

Definition 3.1.1 *A non empty set X is called a metric space if there is a function $d : X \times X \to [0,\infty)$ 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 also.

Definition 3.1.2 *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 3.1.3 *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 3.1.4 *Let S be a nonempty subset of a metric space. Then p is a limit point (accumulation point) of S if for 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'.*

The following proposition is fairly obvious from the above definition and will be used whenever convenient. It is equivalent to the above definition and so it can take the place of the above definition if desired.

Proposition 3.1.5 *A point x is a limit point of the nonempty set A if and only if every $B(x,r)$ contains infinitely many points of A.*

Proof: $\Leftarrow$ is obvious. Consider $\Rightarrow$. Let x be a limit point. Let $r_1 = 1$. Then $B(x,r_1)$ contains $a_1 \neq x$. If $\{a_1, \cdots, a_n\}$ have been chosen none equal to x and with no repeats in the list, let $0 < r_n < \min\left(\frac{1}{n}, \min\{d(a_i,x), i = 1,2,\cdots n\}\right)$. Then let $a_{n+1} \in B(x,r_n)$. Thus every $B(x,r)$ contains $B(x,r_n)$ for all n large enough and hence it contains a_k for $k \geq n$ where the a_k are distinct, none equal to x. ■

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 $\{x_n\}$ or $\{x_n\}_{n=1}^{\infty}$ if we want to emphasize the values of n. Then the following definition is what it means for a sequence to converge.

Definition 3.1.6 *We say that $x = \lim_{n\to\infty} x_n$ when for every $\varepsilon > 0$ there exists N such that if $n \geq N$, then*

$$d(x,x_n) < \varepsilon$$

Often we write $x_n \to x$ for short. This is equivalent to saying

$$\lim_{n\to\infty} d(x,x_n) = 0.$$

Proposition 3.1.7 *The limit is well defined. That is, if x, x' are both limits of a sequence, then $x = x'$.*

Proof: From the definition, there exist N, N' such that if $n \geq N$, then $d(x,x_n) < \varepsilon/2$ and if $n \geq N'$, then $d(x,x_n) < \varepsilon/2$. Then let $M \geq \max(N,N')$. Let $n > M$. Then

$$d\left(x,x'\right) \leq d(x,x_n) + d\left(x_n,x'\right) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Since ε is arbitrary, this shows that $x = x'$ because $d(x,x') = 0$. ■

Next there is an important theorem about limit points and convergent sequences.

Theorem 3.1.8 *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, $\{x_n\}$ none of which equal p such that* $\lim_{n\to\infty} x_n = p$.

Proof: $\Longrightarrow$ Suppose p is a limit point. Why does there exist the promissed 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 $B(p, \delta_{n+1}) \cap S$ where

$$\delta_{n+1} = \min\left\{\frac{1}{n+1}, d(x_i,p), i = 1,2,\cdots,n\right\}.$$

Then this constructs the necessary convergent sequence.

$\Longleftarrow$ Conversely, if such a sequence $\{x_n\}$ 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 3.1.9 *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 relationship between closed sets and limit points.

Theorem 3.1.10 *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 3.1.11 *A set H is closed if and only if whenever $\{h_n\}$ 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 $h_n \to x$. 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 because none of the h_n are equal to x. Hence $x \in H$ after all.

$\Longleftarrow$ Suppose the limit condition holds, why is H closed? Let $x \in H'$ the set of limit points of H. By Theorem 3.1.8 there exists a sequence of points of H, $\{h_n\}$ such that $h_n \to x$. Then by assumption, $x \in H$. Thus H contains all of its limit points and so it is closed by Theorem 3.1.10. ■

Next is the important concept of a subsequence.

Definition 3.1.12 *Let $\{x_n\}_{n=1}^{\infty}$ 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 $\{x_n\}_{n=1}^{\infty}$.*

The really important thing about subsequences is that they preserve convergence.

Theorem 3.1.13 *Let $\left\{x_{n_k}\right\}$ be a subsequence of a convergent sequence $\{x_n\}$ where $x_n \to x$. Then $\lim_{k\to\infty} x_{n_k} = x$ also.*

Proof: Let $\varepsilon > 0$ be given. Then there exists N such that $d(x_n,x) < \varepsilon$ 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 $\lim_{k\to\infty} x_{n_k} = x$. ■

3.2 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 3.2.1 *$\{x_n\}$ is called a Cauchy sequence if for every $\varepsilon > 0$ there exists N such that if $m,n \geq N$, then $d(x_n,x_m) < \varepsilon$.*

Now the major theorem about this is the following.

Theorem 3.2.2 *Let $\{x_n\}$ 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 $\{x_n\}$ is a Cauchy sequence and $\lim_{k\to\infty} x_{n_k} = x$. 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 $d(x_m,x_n) < \varepsilon/2$. Let $N \geq \max(N_1,N_2)$. Then if $k \geq N$, then $n_k \geq N$ and so $d(x,x_k) \leq d\left(x,x_{n_k}\right) + d\left(x_{n_k},x_k\right) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. It follows from the definition that $\lim_{k\to\infty} x_k = x$. ■

Definition 3.2.3 *A metric space is said to be* ***complete*** *if every Cauchy sequence converges.*

There certainly are metric spaces which are not complete. For example, if you consider $\mathbb{Q}$ with $d(x,y) \equiv |x-y|$, this will not be complete because you can get a sequence which is obtained as x_n defined as the n decimal place description of $\sqrt{2}$. However, if a sequence converges, then it must be Cauchy.

Lemma 3.2.4 *If $x_n \to x$, then $\{x_n\}$ is a Cauchy sequence.*

Proof: Let $\varepsilon > 0$. Then there exists n_ε such that if $m \geq n_\varepsilon$, then $d(x,x_m) < \varepsilon/2$. If $m,k \geq n_\varepsilon$, then by the triangle inequality, $d(x_m,x_k) \leq d(x_m,x) + d(x,x_k) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ showing that the convergent sequence is indeed a Cauchy sequence as claimed. ■

Another nice thing to note is this.

Proposition 3.2.5 *If $\{x_n\}$ is a sequence and if p is a limit point of the set $S = \cup_{n=1}^{\infty} \{x_n\}$, then there is a subsequence $\{x_{n_k}\}$ such that* $\lim_{k\to\infty} x_{n_k} = x$.

Proof: By Theorem 3.1.8, there exists a sequence of distinct points of S denoted as $\{y_k\}$ such that none of them equal p and $\lim_{k\to\infty} y_k = p$. 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 $x_{n_i} \in B(p,\delta_i)$. Then let

$$\delta_{k+1} < \min \left\{ \frac{1}{2^{k+1}}, d\left(p, x_{n_j}\right), \delta_j, j = 1,2\cdots,k \right\}$$

Let $x_{n_{k+1}} \in B(p,\delta_{k+1})$ where n_{k+1} is the first index such that $x_{n_{k+1}}$ is contained $B(p,\delta_{k+1})$. Then $\lim_{k\to\infty} x_{n_k} = p$. ■

Another useful result is the following.

Lemma 3.2.6 *Suppose $x_n \to x$ and $y_n \to y$. Then $d(x_n,y_n) \to d(x,y)$.*

Proof: Consider the following.

$$d(x,y) \leq d(x,x_n) + d(x_n,y) \leq d(x,x_n) + d(x_n,y_n) + d(y_n,y)$$

so $d(x,y) - d(x_n,y_n) \leq d(x,x_n) + d(y_n,y)$. Similar reasoning to what was just used shows that $d(x_n,y_n) - d(x,y) \leq d(x,x_n) + d(y_n,y)$, so $|d(x_n,y_n) - d(x,y)| \leq d(x,x_n) + d(y_n,y)$ and the right side converges to 0 as $n \to \infty$. ■

3.3 Closure of a Set

Next is the topic of the closure of a set.

Definition 3.3.1 *Let A be a nonempty subset of (X,d) a metric space. Then $\overline{A}$ 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 3.3.2 *Let A be a nonempty set in (X,d). Then $\overline{A}$ is a closed set and $\overline{A} = A \cup A'$where A' 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 $\overline{A} = \cap \mathscr{C}$ and this will be closed if its complement is open. However, $\overline{A}^C = \cup \left\{ H^C : H \in \mathscr{C} \right\}$. 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 $\overline{A}$ is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \overline{A}$ so if $x \in A$, then $x \in \overline{A}$. Now consider $y \in A'$ but $y \notin A$. If $y \notin \overline{A}$, a closed set, then there exists $B(y,r) \subseteq \overline{A}^C$. Thus y cannot be a limit point of A, a contradiction. Therefore, $A \cup A' \subseteq \overline{A}$.

Next suppose $x \in \overline{A}$ 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 $\overline{A}$ which is contrary to the assertion that $x \in \overline{A}$. Hence if $x \notin A$, then $x \in A'$ and so $A \cup A' \supseteq \overline{A}$ ■

3.4 Separable Metric Spaces

Definition 3.4.1 *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 3.4.2 *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 $\mathscr{P} \equiv \{p_B : B \in \mathscr{B}\}$. 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 $\mathscr{B}$. ■

Corollary 3.4.3 *If (X,d) is a metric space and S is a nonempty subset of X, then S is also separable.*

Proof: Let $\mathscr{B}$ be a countable basis for (X,d). Say $\mathscr{B}_S$ be those sets of $\mathscr{B}$ which have nonempty intersections with S. By axiom of choice, there is a point in each of these intersections. The resulting countable selection of points must be dense in S. Indeed, if $x \in S$, then $B(x,r)$ is the union of sets of $\mathscr{B}$ and so some point just described is in $B(x,r)$. ■

Definition 3.4.4 *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 3.4.5 *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 3.4.6 *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 3.4.2. 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 3.4.7 *A Polish space is a complete separable metric space. These things turn out to be very useful in probability theory and in other areas.*

3.5 Compact Sets

As usual, we are not worrying about empty sets. Fussing over these is usually a waste of time. Thus if a set is mentioned, the default is that it is nonempty.

Definition 3.5.1 *A metric space K is compact if whenever $\mathscr{C}$ is an open cover of K, meaning $K \subseteq \cup\mathscr{C}$, there exists a finite subset of $\mathscr{C}$ $\{U_1, \cdots, U_n\}$ such that $K \subseteq \cup_{k=1}^n U_k$. In words, every open cover admits a finite sub-cover.*

Directly from this definition is the following proposition.

Proposition 3.5.2 *If K is a closed, nonempty subset of a nonempty compact set H, then K is compact.*

Proof: Let $\mathscr{C}$ be an open cover for K. Then $\mathscr{C} \cup \left\{K^C\right\}$ is an open cover for H. Thus there are finitely many sets from this last collection of open sets, $U_1, \cdots, U_m$ which covers H. Include only those which are in $\mathscr{C}$. These cover K because K^C covers no points of K. ■

This is the real definition given above. However, in metric spaces, it is equivalent to another definition called sequentially compact.

Definition 3.5.3 *A metric space K is sequentially compact means that whenever $\{x_n\} \subseteq K$, there exists a subsequence $\{x_{n_k}\}$ such that $\lim_{k\to\infty} x_{n_k} = x \in K$ for some point x. In words, every sequence has a subsequence which converges to a point in the set.*

There is a fundamental property possessed by a sequentially compact set in a metric space which is described in the following proposition. The special number described is called a Lebesgue number.

Proposition 3.5.4 *Let K be a sequentially compact set in a metric space and let $\mathscr{C}$ be an open cover of K. Then there exists a number $\delta > 0$ such that whenever $x \in K$, it follows that $B(x,\delta)$ is contained in some set of $\mathscr{C}$.*

Proof: If $\mathscr{C}$ is an open cover of K and has no Lebesgue number, then for each $n \in \mathbb{N}$, $\frac{1}{n}$ is not a Lebesgue number. Hence there exists $x_n \in K$ such that $B\left(x_n, \frac{1}{n}\right)$ is not contained in any set of $\mathscr{C}$. By sequential compactness, there is a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \to x \in K$. Now there is $r > 0$ such that $B(x,r) \subseteq U \in \mathscr{C}$. Let k be large enough that $\frac{1}{n_k} < \frac{r}{2}$ and also large enough that $x_{n_k} \in B\left(x, \frac{r}{2}\right)$. Then $B\left(x_{n_k}, \frac{1}{n_k}\right) \subseteq B\left(x_{n_k}, \frac{r}{2}\right) \subseteq B(x,r)$ contrary to the requirement that $B\left(x_{n_k}, \frac{1}{n_k}\right)$ is not contained in any set of $\mathscr{C}$. ■

In any metric space, these two definitions of compactness are equivalent.

Theorem 3.5.5 *Let K be a nonempty subset of a metric space (X,d). Then it is compact if and only if it is sequentially compact.*

Proof: $\Leftarrow$ Suppose K is sequentially compact. Let $\mathscr{C}$ be an open cover of K. By Proposition 3.5.4 there is a Lebesgue number $\delta > 0$. Let $x_1 \in K$. If $B(x_1,\delta)$ covers K, then pick a set of $\mathscr{C}$ containing this ball and this set will be a finite subset of $\mathscr{C}$ which covers K. If $B(x_1,\delta)$ does not cover K, let $x_2 \notin B(x_1,\delta)$. Continue this way obtaining x_k such that $d(x_k,x_j) \geq \delta$ whenever $k \neq j$. Thus eventually $\{B(x_i,\delta)\}_{i=1}^n$ must cover K because if not, you could get a sequence $\{x_k\}$ which has every pair of points further apart than δ and hence it has no Cauchy subsequence. Therefore, by Lemma 3.2.4, it would have no convergent subsequence. This would contradict K is sequentially compact. Now let $U_i \in \mathscr{C}$ with $U_i \supseteq B(x_i,\delta)$. Then $\cup_{i=1}^n U_i \supseteq K$.

$\Rightarrow$ Now suppose K is compact. If it is not sequentially compact, then there exists a sequence $\{x_n\}$ which has no convergent subsequence to a point of K. In particular, no point of this sequence is repeated infinitely often. By Proposition 3.2.5 the set of points $\cup_n \{x_n\}$ has no limit point in K. (If it did, you would have a subsequence converging to this point since every ball containing this point would contain infinitely many points of $\cup_n \{x_n\}$.) Now consider the sets $H_n \equiv \cup_{k\geq n} \{x_k\} \cup H'$ where H' denotes all limit points of $\cup_n \{x_n\}$ in X which is the same as the limit points of $\cup_{k\geq n} \{x_k\}$. Therefore, each H_n is closed thanks to Lemma 3.3.2. Now let $U_n \equiv H_n^C$. This is an increasing sequence of open sets whose union contains K thanks to the fact that there is no constant subsequence. However, none of these open sets covers K because U_n is missing x_n, violating the definition of compactness. Next is an alternate argument.

$\Rightarrow$ Now suppose K is compact. If it is not sequentially compact, then there exists a sequence $\{x_n\}$ which has no convergent subsequence to a point of K. If $x \in K$, then there exists $B(x,r_x)$ which contains x_n for only finitely many n. This is because x is not the limit of a subsequence. Then $\{B(x_i,r_i)\}_{i=1}^N$ is a finite sub-cover of K. If p is the largest index for any x_k contained in $\cup_{i=1}^N B(x_i,r_i)$, let $n > p$ and consider x_n. It is a point in K but it can't be in any of the sets covering K. ■

Definition 3.5.6 *X be a metric space. Then a finite set of points $\{x_1,\cdots,x_n\}$ is called an ε net if $X\subseteq\cup_{k=1}^{n}B(x_k,\varepsilon)$. If, for every $\varepsilon>0$ a metric space has an ε net, then we say that the metric space is totally bounded.*

Lemma 3.5.7 *If a metric space (K,d) is sequentially compact, then it is separable and totally bounded.*

Proof: Pick $x_1\in K$. If $B(x_1,\varepsilon)\supseteq K$, then stop. Otherwise, pick $x_2\notin B(x_1,\varepsilon)$. Continue this way. If $\{x_1,\cdots,x_n\}$ have been chosen, either $K\subseteq\cup_{k=1}^{n}B(x_k,\varepsilon)$ in which case, you have found an ε net or this does not happen in which case, you can pick $x_{n+1}\notin\cup_{k=1}^{n}B(x_k,\varepsilon)$. 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 ε. See Lemma 3.2.4. Thus for every $\varepsilon>0$, there is an ε net. Thus the metric space is totally bounded. Let N_ε denote an ε 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 3.5.8 *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: By Theorem 3.5.5, the first two conditions are equivalent.

2.$\Longrightarrow$ 3. If (X,d) is sequentially compact, then by Lemma 3.5.7, it is totally bounded. If $\{x_n\}$ is a Cauchy sequence, then there is a subsequence which converges to $x\in X$ by assumption. However, from Theorem 3.2.2 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 3.4.6 has the Lindeloff property. Hence, if X is not compact, there is a countable set of open sets $\{U_i\}_{i=1}^{\infty}$ which covers X but no finite subset does. Consider the nonempty closed sets F_n and pick $x_n\in F_n$ where

$$X\setminus\cup_{i=1}^{n}U_i\equiv X\cap\left(\cup_{i=1}^{n}U_i\right)^C\equiv F_n$$

Let $\{x_m^k\}_{m=1}^{M_k}$ 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 out of the finitely many $\left\{x_m^{k+1}\right\}$ where $B\left(x_m^{k+1},1/2^{k+1}\right)$ 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 $p < q$,

$$d\left(x^p_{m_p}, x^q_{m_q}\right) \leq \sum_{k=p}^{q-1} d\left(x^k_{m_k}, x^{k+1}_{m_{k+1}}\right) < \sum_{k=p}^{\infty} \frac{1}{2^{k-1}} = \frac{1}{2^{p-2}}$$

Now take a subsequence $x_{n_k} \in B\left(x^k_{m_k}, 2^{-k}\right)$ so it follows that $\lim_{k\to\infty} x_{n_k} = \lim_{k\to\infty} x^k_{m_k} = x \in X$. However, $x \in F_n$ for each n since each F_n is closed and these sets are nested. Thus $x \in \cap_n F_n$ contrary to the claim that $\{U_i\}_{i=1}^{\infty}$ covers X. ■

For the sake of another point of view, here is another argument, this time that 3.)⇒ 2.). This will illustrate something called the Cantor diagonalization process.

Assume 3.). Suppose $\{x_k\}$ is a sequence in X. By assumption there are finitely many open balls of radius $1/n$ covering X. This for each $n \in \mathbb{N}$. Therefore, for $n = 1$, there is one of the balls, having radius 1 which contains x_k for infinitely many k. Therefore, there is a subsequence with every term contained in this ball of radius 1. Now do for this subsequence what was just done for $\{x_k\}$. There is a further subsequence contained in a ball of radius 1/2. Continue this way. Denote the i^{th} subsequence as $\{x_{ki}\}_{k=1}^{\infty}$. Arrange them as shown

$$\begin{array}{c} x_{11}, x_{21}, x_{31}, x_{41} \cdots \\ x_{12}, x_{22}, x_{32}, x_{42} \cdots \\ x_{13}, x_{23}, x_{33}, x_{43} \cdots \\ \vdots \end{array}$$

Thus all terms of $\{x_{ki}\}_{k=1}^{\infty}$ are contained in a ball of radius $1/i$. Consider now the diagonal sequence defined as $y_k \equiv x_{kk}$. Given n, each y_k is contained in a ball of radius $1/n$ whenever $k \geq n$. Thus $\{y_k\}$ is a subsequence of the original sequence and $\{y_k\}$ is a Cauchy sequence. By completeness of X, this converges to some $x \in X$ which shows that every sequence in X has a convergent subsequence. This shows 3.)⇒ 2.). ■

Lemma 3.5.9 *The closed interval $[a,b]$ in $\mathbb{R}$ is compact and every Cauchy sequence in $\mathbb{R}$ converges.*

Proof: To show this, suppose it is not. Then there is an open cover $\mathscr{C}$ which admits no finite subcover for $[a,b] \equiv I_0$. Consider the two intervals $\left[a, \frac{a+b}{2}\right], \left[\frac{a+b}{2}, b\right]$. One of these, maybe both cannot be covered with finitely many sets of $\mathscr{C}$ since otherwise, there would be a finite collection of sets from $\mathscr{C}$ covering $[a,b]$. Let I_1 be the interval which has no finite subcover. Now do for it what was done for I_0. Split it in half and pick the half which has no finite covering of sets of $\mathscr{C}$. Thus there is a "nested" sequence of closed intervals $I_0 \supseteq I_1 \supseteq I_2 \cdots$, each being half of the preceding interval. Say $I_n = [a_n, b_n]$. By the nested interval Lemma, Lemma 2.7.1, there is a point x in all these intervals. The point is unique because the lengths of the intervals converge to 0. This point is in some $O \in \mathscr{C}$. Thus for some $\delta > 0, [x-\delta, x+\delta]$, having length 2δ, is contained in O. For k large enough, the interval $[a_k, b_k]$ has length less than δ but contains x. Therefore, it is contained in $[x-\delta, x+\delta]$ and so must be contained in a single set of $\mathscr{C}$ contrary to the construction. This contradiction shows that in fact $[a,b]$ is compact.

Now if $\{x_n\}$ is a Cauchy sequence, then it is contained in some interval $[a,b]$ which is compact. Hence there is a subsequence which converges to some $x \in [a,b]$. By Theorem 3.2.2 the original Cauchy sequence converges to x. ■

3.6 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 3.6.1 *Let $f : X \to Y$ be a function where (X,d) and (Y,ρ) 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 $d(\hat{x},x) < \delta$, 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 3.6.2 *Let $f : X \to Y$ where (X,d) and (Y,ρ) are metric spaces. Then the following two are equivalent.*

a f is continuous at x.

b Whenever $x_n \to x$, it follows that $f(x_n) \to f(x)$.

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 $f^{-1}(H) \equiv \{x : f(x) \in H\}$ is closed in X.

Proof: a $\Longrightarrow$ b: Let f be continuous at x and suppose $x_n \to x$. 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 $x_n \to x$, it follows that there exists N such that if $n \geq N$, then $d(x_n,x) < \delta$ and so, if $n \geq N$, it follows that $\rho(f(x_n), f(x)) < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it follows that $f(x_n) \to f(x)$.

b $\Longrightarrow$ 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 $d(x_n,x) < 1/n$ but $\rho(f(x_n), f(x)) \geq \varepsilon$. Now this is a contradiction because by assumption, the fact that $x_n \to x$ implies that $f(x_n) \to f(x)$. In particular, for large enough n, $\rho(f(x_n), f(x)) < \varepsilon$ contrary to the construction.

c $\Longrightarrow$ 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 $f^{-1}(V)$ is an interior point which implies $f^{-1}(V)$ is open.

d $\Longrightarrow$ e: Let H be closed in Y. Then $f^{-1}(H)^C = f^{-1}(H^C)$ which is open by assumption. Hence $f^{-1}(H)$ is closed because its complement is open.

e $\Longrightarrow$ d: Let V be open in Y. Then $f^{-1}(V)^C = f^{-1}(V^C)$ which is assumed to be closed. This is because the complement of an open set is a closed set.

d $\Longrightarrow$ 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, $f(B(x,\delta)) \subseteq B(f(x),\varepsilon)$ so $\rho(f(\hat{x}), f(x)) < \varepsilon$. Thus f is continuous at x for every x. ■

Example 3.6.3 *$x \to d(x,y)$ is a continuous function from the metric space to the metric space of nonnegative real numbers.*

This follows from Lemma 3.2.6. You can also define a metric on a Cartesian product of metric spaces.

Proposition 3.6.4 *Let (X,d) be a metric space and consider $(X \times X, \rho)$ where*

$$\rho((x,\tilde{x}),(y,\tilde{y})) \equiv d(x,y) + d(\tilde{x},\tilde{y}).$$

Then this is also a metric space.

Proof: The only condition not obvious is the triangle inequality. However,

$$\rho((x,\tilde{x}),(y,\tilde{y})) + \rho((y,\tilde{y}),(z,\tilde{z})) \equiv d(x,y) + d(\tilde{x},\tilde{y}) + d(y,z) + d(\tilde{y},\tilde{z})$$

$$\geq d(x,z) + d(\tilde{x},\tilde{z}) = \rho((x,\tilde{x}),(z,\tilde{z})) \blacksquare$$

Definition 3.6.5 *If you have two metric spaces (X,d) and (Y,ρ), a function $f : X \to Y$ is called a homeomorphism if and only if it is continuous, one to one, onto, and its inverse is also continuous.*

Here is a useful proposition.

Proposition 3.6.6 *Let (X,d) be a metric space and let S be a nonempty subset of X. Define*

$$\text{dist}(x,S) \equiv \inf\{d(x,s) : s \in S\}$$

Then $|\text{dist}(x,S) - \text{dist}(y,S)| \leq d(x,y)$ so $x \to \text{dist}(x,S)$ is continuous.

Proof: Say $\text{dist}(x,S) \geq \text{dist}(y,S)$. Then there is $s \in S$ such that $\text{dist}(y,S) + \varepsilon > d(y,s)$. Then

$$|\text{dist}(x,S) - \text{dist}(y,S)| = \text{dist}(x,S) - \text{dist}(y,S) \leq d(x,s) - (d(y,s) - \varepsilon)$$

$$\leq d(x,y) + d(y,s) - (d(y,s) - \varepsilon) = d(x,y) + \varepsilon$$

Since $\varepsilon > 0$ is arbitrary, this shows the claimed result. If $\text{dist}(x,S) \leq \text{dist}(y,S)$, repeat switching roles of x and y. $\blacksquare$

3.7 Continuity and Compactness

How does compactness relate to continuity? It turns out that the continuous image of a compact set is always compact. This is an easy consequence of the above major theorem.

Theorem 3.7.1 *Let $f : X \to Y$ where (X,d) and (Y,ρ) are metric spaces and f is continuous on X. Then if $K \subseteq X$ is compact, it follows that $f(K)$ is compact in (Y,ρ).*

Proof: Let $\mathscr{C}$ be an open cover of $f(K)$. Denote by $f^{-1}(\mathscr{C})$ the sets of the form $\{f^{-1}(U) : U \in \mathscr{C}\}$. Then $f^{-1}(\mathscr{C})$ is an open cover of K. It follows there are finitely many sets of the form $\{f^{-1}(U_1), \cdots, f^{-1}(U_n)\}$ which covers K. It follows that $\{U_1, \cdots, U_n\}$ is an open cover for $f(K)$. $\blacksquare$

The following is the important extreme values theorem for a real valued function defined on a compact set.

Theorem 3.7.2 *Let K be a compact metric space and suppose $f : K \to \mathbb{R}$ is a continuous function. That is, $\mathbb{R}$ is the metric space where the metric is given by $d(x,y) = |x-y|$. Then f achieves its maximum and minimum values on K.*

Proof: Let $\lambda \equiv \sup\{f(x) : x \in K\}$. Then from the definition of sup, you have the existence of a sequence $\{x_n\} \subseteq K$ such that $\lim_{n\to\infty} f(x_n) = \lambda$. There is a subsequence still called $\{x_n\}$ which converges to some $x \in K$. From continuity, $\lambda = \lim_{n\to\infty} f(x_n) = f(x)$ and so f achieves its maximum value at x. Similar reasoning shows that it achieves its minimum value on K. ■

Definition 3.7.3 *Let $f : (X,d) \to (Y,\rho)$ be a function. Then it is said to be uniformly continuous on X if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that whenever $x, \hat{x}$ are two points of X with $d(x,\hat{x}) < \delta$, it follows that $\rho(f(x), f(\hat{x})) < \varepsilon$.*

Note the difference between this and continuity. With continuity, the δ could depend on x but here it works for any pair of points in X.

There is a remarkable result concerning compactness and uniform continuity.

Theorem 3.7.4 *Let $f : (X,d) \to (Y,\rho)$ be a continuous function and let K be a compact subset of X. Then the restriction of f to K is uniformly continuous.*

Proof: First of all, K is a metric space and f restricted to K is continuous. Now suppose it fails to be uniformly continuous. Then there exists $\varepsilon > 0$ and pairs of points $x_n, \hat{x}_n$ such that $d(x_n, \hat{x}_n) < 1/n$ but $\rho(f(x_n), f(\hat{x}_n)) \geq \varepsilon$. Since K is compact, it is sequentially compact and so there exists a subsequence, still denoted as $\{x_n\}$ such that $x_n \to x \in K$. Then also $\hat{x}_n \to x$ also and so by Lemma 3.2.6, $\rho(f(x), f(x)) = \lim_{n\to\infty} \rho(f(x_n), f(\hat{x}_n)) \geq \varepsilon$ which is a contradiction. ■

3.8 Lipschitz Continuity and Contraction Maps

The following may be 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 3.8.1 *Let $f : X \to Y$ where (X,d) and (Y,ρ) 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 3.8.2 *Let $f : (X,d) \to (X,d)$ 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*

$$d(x, x_0) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Also, for each n,

$$d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1-r},$$

and $x = \lim_{n\to\infty} f^n(x_0)$.

Proof: Pick $x_0 \in X$ and consider the sequence of the iterates of the map$f$ given by $x_0, f(x_0), f^2(x_0), \cdots$. We argue that this is a Cauchy sequence. For $m < n$, it follows from the triangle inequality,

$$d(f^m(x_0), f^n(x_0)) \leq \sum_{k=m}^{n-1} d\left(f^{k+1}(x_0), f^k(x_0)\right) \leq \sum_{k=m}^{\infty} r^k d(f(x_0), x_0)$$

The reason for this last is as follows.

$$d\left(f^2(x_0), f(x_0)\right) \leq rd(f(x_0), x_0)$$

$$d\left(f^3(x_0), f^2(x_0)\right) \leq rd\left(f^2(x_0), f(x_0)\right) \leq r^2 d(f(x_0), x_0)$$

and so forth. Therefore, by the triangle inequality,

$$\begin{aligned} d(f^m(x_0), f^n(x_0)) &\leq \sum_{k=m}^{n-1} d\left(f^{k+1}(x_0), f^k(x_0)\right) \\ &\leq \sum_{k=m}^{\infty} r^k d(f(x_0), x_0) \leq d(f(x_0), x_0) \frac{r^m}{1-r} \end{aligned} \tag{3.1}$$

which shows that this is indeed a Cauchy sequence. Therefore, there exists x such that $\lim_{n\to\infty} f^n(x_0) = x$. By continuity, $f(x) = f(\lim_{n\to\infty} f^n(x_0)) = \lim_{n\to\infty} f^{n+1}(x_0) = x$.

Also note that, letting $m = 0$ in 3.1, this estimate yields

$$d(x_0, f^n(x_0)) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Now $d(x_0, x) \leq d(x_0, f^n(x_0)) + d(f^n(x_0), x)$ and so

$$d(x_0, x) - d(f^n(x_0), x) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Letting $n \to \infty$, it follows that $d(x_0, x) \leq \frac{d(x_0, f(x_0))}{1-r}$because $\lim_{n\to\infty} d(f^n(x_0), x) = d(x, x) = 0$ by Lemma 3.2.6.

It only remains to verify that there is only one fixed point. Suppose then that x, x' are two. Then

$$d\left(x, x'\right) = d\left(f(x), f\left(x'\right)\right) \leq rd\left(x', x\right)$$

and so $d(x, x') = 0$ because $r < 1$. ■

The above is the usual formulation of this important theorem, but we actually proved a better result.

Corollary 3.8.3 *Let B be a closed subset of the complete metric space (X, d) and let $f : B \to X$ be a contraction map*

$$d(f(x), f(\hat{x})) \leq rd(x, \hat{x}),\ r < 1.$$

Also suppose ***there exists*** *$x_0 \in B$ such that the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ 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 $B = \overline{B(x_0, \delta)}$, the sequence of iterates satisfies the inequality*

$$d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1-r}$$

and so it will remain in B if $\frac{d(x_0, f(x_0))}{1-r} < \delta$.

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,

$$x = \lim_{n\to\infty} f^{n}\left(x_0\right) = \lim_{n\to\infty} f^{n+1}\left(x_0\right) = f\left(\lim_{n\to\infty} f^{n}\left(x_0\right)\right) = f\left(x\right)$$

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 δ, then the iterates remain in B. As to the fixed point being unique, it is as before. If x, x' are both fixed points in B, then $d\left(x, x'\right) = d\left(f\left(x\right), f\left(x'\right)\right) \leq rd\left(x, x'\right)$ and so $x = x'$. ■

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 3.8.4 *Let $f : (X, d) \to (X, d)$ 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 $\{f^{n}\left(x_0\right)\}_{n=1}^{\infty}$ also converges to the fixed point.*

Proof: From Theorem 3.8.2 there is a unique fixed point for f^n. Thus $f^n\left(x\right) = x$ Then

$$f^{n}\left(f\left(x\right)\right) = f^{n+1}\left(x\right) = f\left(x\right)$$

By uniqueness, $f\left(x\right) = x$.

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 3.8.2, as $k \to \infty, f^{p_k n}\left(f^{r}\left(x_0\right)\right) \to x$ which contradicts the above inequality. Hence the sequence of iterates converges to x, as it did for f a contraction map. ■

3.9 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 3.9.1 *Let $f_n : X \to Y$ where $(X, d), (Y, \rho)$ are two metric spaces. Then $\{f_n\}$ is said to converge pointwise to a function $f : X \to Y$ if for every $x \in X$, $\lim_{n\to\infty} f_n\left(x\right) = f\left(x\right)$. $\{f_n\}$ is said to converge uniformly if for all $\varepsilon > 0$, there exists N such that if $n \geq N$, then $\sup_{x\in X} \rho\left(f_n\left(x\right), f\left(x\right)\right) < \varepsilon$*

Here is a well known example illustrating the difference between pointwise and uniform convergence.

Example 3.9.2 *Let $f_n(x) = x^n$ on the metric space $[0,1]$. Then this function converges pointwise to*

$$f(x) = \begin{cases} 0 \text{ on } [0,1) \\ 1 \text{ at } 1 \end{cases}$$

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 3.9.3 *Let $f_n : X \to Y$ 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

$$\rho(f(x), f(\hat{x})) \leq \rho(f(x), f_n(x)) + \rho(f_n(x), f_n(\hat{x})) + \rho(f_n(\hat{x}), f(\hat{x}))$$

By uniform convergence, there exists N such that both $\rho(f(x), f_n(x)), \rho(f_n(\hat{x}), f(\hat{x}))$ are less than $\varepsilon/3$ provided $n \geq N$. Thus picking such an n

$$\rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x}))$$

From the continuity of f_n, there exists a positive number $\delta > 0$ such that if $d(x,\hat{x}) < \delta$, then $\rho(f_n(x), f_n(\hat{x})) < \varepsilon/3$. Hence, if $d(x,\hat{x}) < \delta$, then

$$\rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x})) < \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

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 $n, \rho(f(x), f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x), f_n(\hat{x}))$. By uniform continuity of f_n there exists δ 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. ■

3.10 Compactness in $C(X,Y)$ Ascoli Arzela Theorem

This will use the characterization of compact metric spaces to give a proof of a general version of the Arzella Ascoli theorem. See Naylor and Sell [36] which is where I saw this general formulation.

Definition 3.10.1 *Let (X,d_X) be a compact metric space. Let (Y,d_Y) be another complete metric space. Then $C(X,Y)$ will denote the continuous functions which map X to Y. Then ρ is a metric on $C(X,Y)$ defined by $\rho(f,g) \equiv \sup_{x\in X} d_Y(f(x), g(x))$.*

Theorem 3.10.2 *$(C(X,Y),\rho)$ is a complete metric space where (X,d_X) is a compact metric space*

Proof: It is first necessary to show that ρ is well defined. In this argument, I will just write d rather than d_X or d_Y. To show this, note that from Lemma 3.2.6, if $x_n \to x$, and $y_n \to y$, then $d(x_n,y_n) \to d(x,y)$. Therefore, if f,g are continuous, and $x_n \to x$ so $f(x_n) \to f(x)$ and $g(x_n) \to g(x)$, $d(f(x_n),g(x_n)) \to d(f(x),g(x))$ and so, $\rho(f,g)$ is just the maximum of a continuous function defined on a compact set. By Theorem 3.7.2, the extreme values theorem, this maximum exists.

Clearly $\rho(f,g) = \rho(g,f)$ and

$$\begin{aligned}
\rho(f,g)+\rho(g,h) &= \sup_{x\in X} d(f(x),g(x)) + \sup_{x\in X} d(g(x),h(x)) \\
&\geq \sup_{x\in X} (d(f(x),g(x)) + d(g(x),h(x))) \\
&\geq \sup_{x\in X} (d(f(x),h(x))) = \rho(f,h)
\end{aligned}$$

so the triangle inequality holds.

It remains to check completeness. Let $\{f_n\}$ be a Cauchy sequence. Then from the definition, $\{f_n(x)\}$ is a Cauchy sequence in Y and so it converges to something called $f(x)$. By Theorem 3.9.3, f is continuous. It remains to show that $\rho(f_n,f) \to 0$. Let $x \in X$. Then from what was just noted,

$$d(f_n(x),f(x)) = \lim_{m\to\infty} d(f_n(x),f_m(x)) \leq \lim\sup_{m\to\infty} \rho(f_n,f_m)$$

since $\{f_n\}$ is given to be a Cauchy sequence, there exists N such that if $m,n > N$, then $\rho(f_n,f_m) < \varepsilon$. Therefore, if $n > N, d(f_n(x),f(x)) \leq \lim\sup_{m\to\infty} \rho(f_n,f_m) \leq \varepsilon$. Since x is arbitrary, it follows that $\rho(f_n,f) \leq \varepsilon$, if $n \geq N$. ■

Here is a useful lemma.

Lemma 3.10.3 *Let S be a totally bounded subset of (X,d) a metric space. Then $\overline{S}$ is also totally bounded.*

Proof: Suppose not. Then there exists a sequence $\{p_n\} \subseteq \overline{S}$ such that

$$d(p_m,p_n) \geq \varepsilon$$

for all $m \neq n$. Now let $q_n \in B\left(p_n,\frac{\varepsilon}{8}\right) \cap S$. Then it follows that

$$\frac{\varepsilon}{8} + d(q_n,q_m) + \frac{\varepsilon}{8} \geq d(p_n,q_n) + d(q_n,q_m) + d(q_m,p_m) \geq d(p_n,q_m) \geq \varepsilon$$

and so $d(q_n,q_m) > \frac{\varepsilon}{2}$. This contradicts total boundedness of S. ■

Next, here is an important definition.

Definition 3.10.4 *Let $\mathscr{A} \subseteq C(X,Y)$ where (X,d_X) and (Y,d_Y) are metric spaces. Thus $\mathscr{A}$ is a set of continuous functions mapping X to Y. Then $\mathscr{A}$ is said to be equicontinuous if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $d_X(x_1,x_2) < \delta$ then for all $f \in \mathscr{A}$, $d_Y(f(x_1),f(x_2)) < \varepsilon$. (This is uniform continuity which is uniform in $\mathscr{A}$.) $\mathscr{A}$ is said to be pointwise compact if $\{f(x) : f \in \mathscr{A}\}$ has compact closure in Y.*

Here is the Ascoli Arzela theorem.

Theorem 3.10.5 *Let (X,d_X) be a compact metric space and let (Y,d_Y) be a complete metric space. Thus $(C(X,Y),\rho)$ is a complete metric space. Let $\mathscr{A}\subseteq C(X,Y)$ be pointwise compact and equicontinuous. Then $\overline{\mathscr{A}}$ is compact. Here the closure is taken in $(C(X,Y),\rho)$.*

Proof: The more useful direction is that the two conditions imply compactness of $\overline{\mathscr{A}}$. I prove this first. Since $\overline{\mathscr{A}}$ is a closed subset of a complete space, it follows from Theorem 3.5.8, that $\overline{\mathscr{A}}$ will be compact if it is totally bounded. In showing this, it follows from Lemma 3.10.3 that it suffices to verify that $\mathscr{A}$ is totally bounded. Suppose this is not so. Then there exists $\varepsilon>0$ and a sequence of points of $\mathscr{A}$, $\{f_n\}$ such that $\rho(f_n,f_m)\geq\varepsilon$ whenever $n\neq m$.

By equicontinuity, there exists $\delta>0$ such that if

$$d(x,y)<\delta,$$

then $d_Y(f(x),f(y))<\frac{\varepsilon}{8}$ for all $f\in\mathscr{A}$. Let $\{x_i\}_{i=1}^p$ be a δ net for X. Since there are only finitely many x_i, it follows from pointwise compactness that there exists a subsequence, still denoted by $\{f_n\}$ which converges at each x_i. Now let $x\in X$ be arbitrary. There exists N such that for each x_i in that δ net,

$$d_Y(f_n(x_i),f_m(x_i))<\varepsilon/8 \text{ whenever } n,m\geq N$$

Then for $m,n\geq N$,

$$\begin{aligned}
& d_Y(f_n(x),d_{Ym}(x)) \\
\leq\; & d_Y(f_n(x),f_n(x_i))+d_Y(f_n(x_i),f_m(x_i))+d_Y(f_m(x_i),f_m(x)) \\
<\; & d_Y(f_n(x),f_n(x_i))+\varepsilon/8+d_Y(f_m(x_i),f_m(x))
\end{aligned}$$

Pick x_i such that $d(x,x_i)<\delta$. $\{x_i\}_{i=1}^p$ is a δ net and so this is surely possible. Then by equicontinuity, the two ends are each less than $\varepsilon/8$ and so for $m,n\geq N$,

$$d_Y(f_n(x),f_m(x))\leq\frac{3\varepsilon}{8}$$

Since x is arbitrary, it follows that $\rho(f_n,f_m)\leq 3\varepsilon/8<\varepsilon$ which is a contradiction. It follows that $\mathscr{A}$ and hence $\overline{\mathscr{A}}$ is totally bounded. This proves the more important direction.

Next suppose $\overline{\mathscr{A}}$ is compact. Why must $\mathscr{A}$ be pointwise compact and equicontinuous? If it fails to be pointwise compact, then there exists $x\in X$ such that $\{f(x):f\in\mathscr{A}\}$ is not contained in a compact set of Y. Thus there exists $\varepsilon>0$ and a sequence of functions in $\mathscr{A}$ $\{f_n\}$ such that $d(f_n(x),f_m(x))\geq\varepsilon$. But this implies $\rho(f_m,f_n)\geq\varepsilon$ and so $\overline{\mathscr{A}}$ fails to be totally bounded, a contradiction. Thus $\mathscr{A}$ must be pointwise compact. Now why must it be equicontinuous? If it is not, then for each $n\in\mathbb{N}$ there exists $\varepsilon>0$ and $x_n,y_n\in X$ such that $d(x_n,y_n)<1/n$ but for some $f_n\in\mathscr{A}$, $d(f_n(x_n),f_n(y_n))\geq\varepsilon$. However, by compactness, there exists a subsequence $\{f_{n_k}\}$ such that $\lim_{k\to\infty}\rho(f_{n_k},f)=0$ and also that $x_{n_k},y_{n_k}\to x\in X$. Hence

$$\begin{aligned}
\varepsilon \;\leq\; & d\left(f_{n_k}(x_{n_k}),f_{n_k}(y_{n_k})\right)\leq d\left(f_{n_k}(x_{n_k}),f(x_{n_k})\right) \\
& +d\left(f(x_{n_k}),f(y_{n_k})\right)+d\left(f(y_{n_k}),f_{n_k}(y_{n_k})\right) \\
& \leq\rho\left(f_{n_k},f\right)+d\left(f(x_{n_k}),f(y_{n_k})\right)+\rho\left(f,f_{n_k}\right)
\end{aligned}$$

and now this is a contradiction because each term on the right converges to 0. The middle term converges to 0 because $f(x_{n_k}),f(y_{n_k})\to f(x)$. See Lemma 3.2.6. ■

3.11 Connected Sets

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 3.11.1 *A set, S in a metric space, is separated if there exist sets A,B such that*

$$S = A\cup B,\ A,B \neq \emptyset,\ \text{and}\ \overline{A}\cap B = \overline{B}\cap A = \emptyset.$$

In this case, the sets A and B are said to separate S. A set is connected if it is not separated. Remember $\overline{A}$ 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 3.11.2 *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: The argument is dependent on Lemma 3.3.2. Suppose

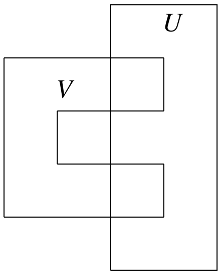
$$K = A\cup B$$

where $\bar{A}\cap B = \bar{B}\cap A = \emptyset, A\neq\emptyset, B\neq\emptyset$. Then p is in one of these sets. Say $p\in A$. Then if $U\in\mathscr{U}$, it must be the case that $U\subseteq A$ since if not, you would have

$$U = (A\cap U)\cup(B\cap U)$$

and the limit points of $A\cap U$ cannot be in B hence not in $B\cap U$ while the limit points of $B\cap U$ cannot be in A hence not in $A\cap U$. Thus $B=\emptyset$. It follows that K cannot be separated and so it is connected. ■

The intersection of connected sets is not necessarily connected as is shown by the following picture.



Theorem 3.11.3 *Let $\boldsymbol{f} : X\to Y$ 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 $\boldsymbol{f}(X)$. Then consider the sets $\boldsymbol{f}^{-1}(A)$ and $\boldsymbol{f}^{-1}(B)$. 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 $\boldsymbol{f}(\boldsymbol{z})$ such that $U \cap A = \emptyset$. But then, the continuity of $\boldsymbol{f}$ and Theorem 3.6.2 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 3.11.4 *Let S be a set and let $\mathbf{p} \in S$. Denote by $C_{\mathbf{p}}$ the union of all connected subsets of S which contain $\mathbf{p}$. This is called the connected component determined by $\mathbf{p}$.*

Theorem 3.11.5 *Let $C_{\mathbf{p}}$ be a connected component of a set S in a metric space. Then $C_{\mathbf{p}}$ is a connected set and if $C_{\mathbf{p}} \cap C_{\mathbf{q}} \neq \emptyset$, then $C_{\mathbf{p}} = C_{\mathbf{q}}$.*

Proof: Let $\mathscr{C}$ denote the connected subsets of S which contain $\mathbf{p}$. By Theorem 3.11.2, $\cup\mathscr{C} = C_{\mathbf{p}}$ is connected. If $\boldsymbol{x} \in C_{\mathbf{p}} \cap C_{\mathbf{q}}$, then from Theorem 3.11.2, $C_{\mathbf{p}} \supseteq C_{\mathbf{p}} \cup C_{\mathbf{q}}$ and so $C_{\mathbf{p}} \supseteq C_{\mathbf{q}}$. 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 3.11.6 *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 $x \in (p, q) \setminus C$, 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 $l \in \overline{A}$ so $l \notin B$ which implies $l \in A$. But if $l \notin \overline{B}$, then for some $\delta > 0, (l, l + \delta) \cap B = \emptyset$ contradicting the definition of l as an upper bound for S. Therefore, $l \in \overline{B}$ 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 3.11.7 *Let E be a connected set in a metric space and suppose $f : E \to \mathbb{R}$ and that $y \in (f(e_1), f(e_2))$ where $e_i \in E$. Then there exists $e \in E$ such that $f(e) = y$.*

Proof: From Theorem 3.11.3, $f(E)$ is a connected subset of $\mathbb{R}$. By Theorem 3.11.6 $f(E)$ 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 3.11.8 *Let U be an open set in $\mathbb{R}$. Then there exist countably many disjoint open sets $\{(a_i, b_i)\}_{i=1}^{\infty}$ such that $U = \cup_{i=1}^{\infty} (a_i, b_i)$.*

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 3.11.2 that

$$(z-\delta, z+\delta) \subseteq C_p.$$

This shows C_p is open. By Theorem 3.11.6, 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 $\{(a_i, b_i)\}_{i=1}^{\infty}$ the set of these connected components. ■

Definition 3.11.9 *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, $\boldsymbol{\gamma} : [a,b] \to E$ such that $\boldsymbol{\gamma}(a) = p$ and $\boldsymbol{\gamma}(b) = q$.*

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.

$$\left\{\left(x, \sin\frac{1}{x}\right) : x \in (0,1]\right\} \cup \{(0,y) : y \in [-1,1]\} \tag{3.2}$$

You can verify that this set of points in the normed vector space $\mathbb{R}^2$ is not arcwise connected but is connected.

Lemma 3.11.10 *In $\mathbb{R}^p$, $B(\boldsymbol{z}, r)$ is arcwise connected.*

Proof: This is easy from the convexity of the set. If $\boldsymbol{x}, \boldsymbol{y} \in B(\boldsymbol{z}, r)$, then let $\boldsymbol{\gamma}(t) = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$ for $t \in [0,1]$.

$$\begin{aligned} \|\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{z}\| &= \|(1-t)(\boldsymbol{x} - \boldsymbol{z}) + t(\boldsymbol{y} - \boldsymbol{z})\| \\ &\leq (1-t)\|\boldsymbol{x} - \boldsymbol{z}\| + t\|\boldsymbol{y} - \boldsymbol{z}\| \\ &< (1-t)r + tr = r \end{aligned}$$

showing $\boldsymbol{\gamma}(t)$ stays in $B(\boldsymbol{z}, r)$.■

Proposition 3.11.11 *If $X \neq \emptyset$ is arcwise connected, then it is connected.*

Proof: Let $p \in X$. Then by assumption, for any $x \in X$, there is an arc joining p and x. This arc is connected because it is the continuous image of an interval which is connected. Since x is arbitrary, every x is in a connected subset of X which contains p. Hence $C_p = X$ and so X is connected. ■

Theorem 3.11.12 *Let U be an open subset of $\mathbb{R}^p$. Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.*

Proof: By Proposition 3.11.11 it is only necessary to verify that if U is connected and open, then U is arcwise connected. Pick $\boldsymbol{p} \in U$. Say $\boldsymbol{x} \in U$ satisfies $\mathscr{P}$ if there exists a continuous function, $\boldsymbol{\gamma} : [a,b] \to U$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}$ and $\boldsymbol{\gamma}(\boldsymbol{b}) = \boldsymbol{x}$.

$$A \equiv \{\boldsymbol{x} \in U \text{ such that } \boldsymbol{x} \text{ satisfies } \mathscr{P}.\}$$

If $\boldsymbol{x} \in A$, then Lemma 3.11.10 implies $B(\boldsymbol{x},r) \subseteq U$ is arcwise connected for small enough r. Thus letting $\boldsymbol{y} \in B(\boldsymbol{x},r)$, there exist intervals, $[a,b]$ and $[c,d]$ and continuous functions having values in U, $\boldsymbol{\gamma},\boldsymbol{\eta}$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}, \boldsymbol{\gamma}(b) = \boldsymbol{x}, \boldsymbol{\eta}(c) = \boldsymbol{x}$, and $\boldsymbol{\eta}(d) = \boldsymbol{y}$. Then let $\boldsymbol{\gamma}_1 : [a, b+d-c] \to U$ be defined as

$$\boldsymbol{\gamma}_1(t) \equiv \begin{cases} \boldsymbol{\gamma}(t) \text{ if } t \in [a,b] \\ \boldsymbol{\eta}(t+c-b) \text{ if } t \in [b, b+d-c] \end{cases}$$

Then it is clear that $\boldsymbol{\gamma}_1$ is a continuous function mapping $\boldsymbol{p}$ to $\boldsymbol{y}$ and showing that $B(\boldsymbol{x},r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(\boldsymbol{p},\delta)$ containing $\boldsymbol{p}$ which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $\boldsymbol{z} \in B$ such that every open set containing $\boldsymbol{z}$ is not contained in B. Therefore, letting $B(\boldsymbol{z},\delta)$ be such that $\boldsymbol{z} \in B(\boldsymbol{z},\delta) \subseteq U$, there exist points of A contained in $B(\boldsymbol{z},\delta)$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected. Note that, since B is open, it contains no limit points of A and since A is open, it contains no limit points of B.

It remains to verify the connected components are open. Let $\boldsymbol{z} \in C_{\boldsymbol{p}}$ where $C_{\boldsymbol{p}}$ is the connected component determined by $\boldsymbol{p}$. Then picking $B(\boldsymbol{z},\delta) \subseteq U$, $C_{\boldsymbol{p}} \cup B(\boldsymbol{z},\delta)$ is connected and contained in U and so it must also be contained in $C_{\boldsymbol{p}}$. Thus $\boldsymbol{z}$ is an interior point of $C_{\boldsymbol{p}}$. ■

As an application, consider the following corollary.

Corollary 3.11.13 *Let $f : \Omega \to \mathbb{Z}$ be continuous where Ω is a connected nonempty open set of a metric space. Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω. To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\cup_{m \neq l}\left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left(\left(l - \frac{1}{6}, l + \frac{1}{6}\right)\right)$ also an open set. ■

3.12 Partitions of Unity in Metric Space

Lemma 3.12.1 *Let X be a metric space and let S be a nonempty subset of X.*

$$\operatorname{dist}(\boldsymbol{x},S) \equiv \inf\{d(\boldsymbol{x},\boldsymbol{z}) : \boldsymbol{z} \in S\}$$

Then

$$|\operatorname{dist}(\boldsymbol{x},S) - \operatorname{dist}(\boldsymbol{y},S)| \leq d(\boldsymbol{x},\boldsymbol{y}).$$

Proof: Say $\operatorname{dist}(\boldsymbol{x},S) \geq \operatorname{dist}(\boldsymbol{y},S)$. Then letting $\varepsilon > 0$ be given, there exists $\boldsymbol{z} \in S$ such that $d(\boldsymbol{y},\boldsymbol{z}) < \operatorname{dist}(\boldsymbol{y},S) + \varepsilon$ Then

$$|\operatorname{dist}(\boldsymbol{x},S) - \operatorname{dist}(\boldsymbol{y},S)| = \operatorname{dist}(\boldsymbol{x},S) - \operatorname{dist}(\boldsymbol{y},S) \leq \operatorname{dist}(\boldsymbol{x},S) - (d(\boldsymbol{y},\boldsymbol{z}) - \varepsilon)$$

$$\leq d(\boldsymbol{x},\boldsymbol{z})-(d(\boldsymbol{y},\boldsymbol{z})-\varepsilon)\leq d(\boldsymbol{x},\boldsymbol{y})+d(\boldsymbol{y},\boldsymbol{z})-d(\boldsymbol{y},\boldsymbol{z})+\varepsilon=d(\boldsymbol{x},\boldsymbol{y})+\varepsilon$$

Since ε is arbitrary, $|\text{dist}(\boldsymbol{x},S)-\text{dist}(\boldsymbol{y},S)|\leq d(\boldsymbol{x},\boldsymbol{y})$. The situation is completely similar if $\text{dist}(\boldsymbol{x},S)<\text{dist}(\boldsymbol{y},S)$. ■

Then this shows that $\boldsymbol{x}\to\text{dist}(\boldsymbol{x},S)$ is a continuous real valued function.

This is about partitions of unity in metric space. Assume here that closed balls are compact. For example, you might be considering $\mathbb{R}^p$ with $d(\boldsymbol{x},\boldsymbol{y})\equiv|\boldsymbol{x}-\boldsymbol{y}|$.

Definition 3.12.2 *Define* $\text{spt}(f)$ *(support of f) to be the closure of the set $\{x: f(x)\neq 0\}$. If V is an open set, $C_c(V)$ will be the set of continuous functions f, defined on Ω having* $\text{spt}(f)\subseteq V$.

Definition 3.12.3 *If K is a compact subset of an open set, V, then $K\prec\phi\prec V$ if*

$$\phi\in C_c(V),\ \phi(K)=\{1\},\ \phi(\Omega)\subseteq[0,1],$$

where Ω denotes the whole metric space. Also for $\phi\in C_c(\Omega)$, $K\prec\phi$ if

$$\phi(\Omega)\subseteq[0,1] \text{ and } \phi(K)=1.$$

and $\phi\prec V$ if

$$\phi(\Omega)\subseteq[0,1] \text{ and } \text{spt}(\phi)\subseteq V.$$

Lemma 3.12.4 *Let (Ω,d) be a metric space in which closed balls are compact. Then if K is a compact subset of an open set V, then there exists ϕ such that $K\prec\phi\prec V$.*

Proof: Since K is compact, the distance between K and V^C is positive, $\delta>0$. Otherwise there would be $x_n\in K$ and $y_n\in V^C$ with $d(x_n,y_n)<1/n$. Taking a subsequence, still denoted with n, we can assume $x_n\to x$ and $y_n\to x$ but this would imply x is in both K and V^C which is not possible. Now consider $\{B(x,\delta/2)\}$ for $x\in K$. This is an open cover and the closure of each ball is contained in V. Since K is compact, finitely many of these balls cover K. Denote their union as W. Then $\overline{W}$ is compact because it is the finite union of the closed balls. Hence $K\subseteq W\subseteq\overline{W}\subseteq V$. Now consider

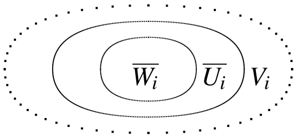
$$\phi(x)\equiv\frac{\text{dist}\left(x,W^C\right)}{\text{dist}(x,K)+\text{dist}(x,W^C)}$$

the denominator is never zero because x cannot be in both K and W^C. Thus ϕ is continuous by Lemma 3.12.1. also if $x\in K$, then $\phi(x)=1$ and if $x\notin W$, then $\phi(x)=0$. ■

Theorem 3.12.5 *(Partition of unity) Let K be a compact subset of a metric space in which closed balls are compact and suppose $K\subseteq V=\cup_{i=1}^n V_i$, V_i open. Then there exist $\psi_i\prec V_i$ with $\sum_{i=1}^n\psi_i(x)=1$ for all $x\in K$.*

Proof: Let $K_1=K\setminus\cup_{i=2}^n V_i$. Thus K_1 is compact and $K_1\subseteq V_1$. Let $K_1\subseteq W_1\subseteq\overline{W}_1\subseteq V_1$ with $\overline{W}_1$compact. To obtain W_1, use Lemma 3.12.4 to get f such that $K_1\prec f\prec V_1$ and let $W_1\equiv\{x:f(x)\neq 0\}$. Thus $W_1,V_2,\cdots V_n$ covers K and $\overline{W}_1\subseteq V_1$. Let $K_2=K\setminus(\cup_{i=3}^n V_i\cup W_1)$. Then K_2 is compact and $K_2\subseteq V_2$. Let $K_2\subseteq W_2\subseteq\overline{W}_2\subseteq V_2$ $\overline{W}_2$ compact. Continue this way finally obtaining $W_1,\cdots,W_n$, $K\subseteq W_1\cup\cdots\cup W_n$, and $\overline{W}_i\subseteq V_i$ $\overline{W}_i$ compact. Now let

$$\overline{W}_i\subseteq U_i\subseteq\overline{U}_i\subseteq V_i\ ,\overline{U}_i \text{ compact.}$$



By Lemma 3.12.4, let $\overline{U}_i \prec \phi_i \prec V_i$, $\cup_{i=1}^n \overline{W}_i \prec \gamma \prec \cup_{i=1}^n U_i$. Define

$$\psi_i(x) = \begin{cases} \gamma(x)\phi_i(x)/\sum_{j=1}^n \phi_j(x) \text{ if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 \text{ if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If x is such that $\sum_{j=1}^n \phi_j(x) = 0$, then $x \notin \cup_{i=1}^n \overline{U}_i$. Consequently $\gamma(y) = 0$ for all y near x and so $\psi_i(y) = 0$ for all y near x. Hence ψ_i is continuous at such x. If $\sum_{j=1}^n \phi_j(x) \neq 0$, this situation persists near x and so ψ_i is continuous at such points. Therefore ψ_i is continuous. If $x \in K$, then $\gamma(x) = 1$ and so $\sum_{j=1}^n \psi_j(x) = 1$. Clearly $0 \leq \psi_i(x) \leq 1$ and $\text{spt}(\psi_j) \subseteq V_j$. ■

3.13 Completion of Metric Spaces

Let (X,d) be a metric space $X \neq \emptyset$. Perhaps this is not a complete metric space. In other words, it may be that Cauchy Sequences do not converge. Of course if $x \in X$ and if $x_n = x$ for all n then $\{x_n\}$ is a Cauchy sequence and it converges to x.

Lemma 3.13.1 *Denote by $\boldsymbol{x}$ a Cauchy sequence $\boldsymbol{x}$ being short for $\{x_n\}_{n=1}^\infty$. Then if $\boldsymbol{x}, \boldsymbol{y}$ are two Cauchy sequences, $\lim_{n\to\infty} d(x_n, y_n)$ exists.*

Proof: Let $\varepsilon > 0$ be given and let N be so large that whenever $n, m \geq N$, it follows that $d(x_n, x_m), d(y_n, y_m) < \varepsilon/2$. Then for such n, m

$$\begin{aligned} |d(x_n,y_n) - d(x_m,y_m)| &\leq |d(x_n,y_n) - d(x_n,y_m)| + |d(x_n,y_m) - d(x_m,y_m)| \\ &\leq d(y_n,y_m) + d(x_n,x_m) < \varepsilon \end{aligned}$$

by Lemma 3.12.1. Therefore, $\{d(x_n,y_n)\}_n$ is a Cauchy sequence in $\mathbb{R}$ and so it converges. ■

Definition 3.13.2 *Let $\boldsymbol{x} \sim \boldsymbol{y}$ when $\lim_{n\to\infty} d(x_n, y_n) = 0$.*

Lemma 3.13.3 *$\sim$ is an equivalence relation.*

Proof: Clearly $\boldsymbol{x} \sim \boldsymbol{x}$ and if $\boldsymbol{x} \sim \boldsymbol{y}$ then $\boldsymbol{y} \sim \boldsymbol{x}$. Suppose then that $\boldsymbol{x} \sim \boldsymbol{y}$ and $\boldsymbol{y} \sim \boldsymbol{z}$. Is $\boldsymbol{x} \sim \boldsymbol{z}$?

$$d(x_n, z_n) \leq d(x_n, y_n) + d(y_n, z_n)$$

and both of those terms on the right converge to 0. ■

Definition 3.13.4 *Denote by $[\boldsymbol{x}]$ the equivalence class determined by the Cauchy sequence $\boldsymbol{x}$. Let $d([\boldsymbol{x}], [\boldsymbol{y}]) \equiv \lim_{n\to\infty} d(x_n, y_n)$.*

Theorem 3.13.5 *Denote by $\hat{X}$ the set of equivalence classes. Then d defined above is a metric, $\hat{X}$ with this is a complete metric space, and X can be considered a dense subset of $\hat{X}$.*

Proof: That d just defined is a metric is obvious from the fact that the original metric d satisfies the triangle inequality. It is also clear that $d([\boldsymbol{x}],[\boldsymbol{y}]) \geq 0$ and that if $[\boldsymbol{x}]=[\boldsymbol{y}]$ if and only if $d([\boldsymbol{x}],[\boldsymbol{y}])=0$.

It remains to show that $(\hat{X}, d)$ is complete. Let $\{[\boldsymbol{x}]_n\}_n$be a Cauchy sequence. From Theorem 3.2.2 it suffices to show the convergence of a subsequence. There is a subsequence, denoted as $\{[\boldsymbol{x}^n]\}$ where $\boldsymbol{x}^n$ is a representative of $[\boldsymbol{x}]_n$ such that $d\left([\boldsymbol{x}^n],\left[\boldsymbol{x}^{n+1}\right]\right)<4^{-n}$. Thus there is an increasing sequence $\{k_n\}$ such that $d\left(x_k^n, x_l^{n+1}\right)<2^{-n}$ if $k, l \geq k_n$ where k_n is increasing in n. Let $\boldsymbol{y}=\left\{x_{k_n}^n\right\}_{n=1}^{\infty}$. For $m \geq k_n$ and the triangle inequality,

$$\begin{aligned}
d\left(x_m^n, y_m\right) &= d\left(x_m^n, x_{k_m}^m\right) \leq d\left(x_m^n, x_{k_n}^n\right)+d\left(x_{k_n}^n, x_{k_m}^m\right) \leq 2^{-n}+\sum_{j=n}^{m-1} d\left(x_{k_j}^j, x_{k_m}^{j+1}\right) \\
&< 2^{-n}+\sum_{j=n}^{m-1} 2^{-j}<2^{-n}+2^{-(n-1)}<2^{-(n-2)}
\end{aligned}$$

Then $\boldsymbol{y}$ is a Cauchy sequence since it is a subsequence of one and also $d([\boldsymbol{x}^n],[\boldsymbol{y}]) \rightarrow 0$.

To show that X is dense in $\hat{X}$, let $[\boldsymbol{x}]$ be given. Then for m large enough, $d(x_k, x_m)<\varepsilon$ whenever $k \geq m$. It suffices to let $\boldsymbol{y}$ be the constant Cauchy sequence always equal to x_m. ■

3.14 Exercises

1. Let $d(x, y)=|x-y|$ for $x, y \in \mathbb{R}$. Show that this is a metric on $\mathbb{R}$.

2. Now consider $\mathbb{R}^n$. Let $\|\boldsymbol{x}\|_{\infty} \equiv \max\{|x_i|, i=1, \cdots, n\}$. Define $d(\boldsymbol{x}, \boldsymbol{y}) \equiv \|\boldsymbol{x}-\boldsymbol{y}\|_{\infty}$. Show that this is a metric on $\mathbb{R}^n$. In the case of $n=2$, describe the ball $B(\mathbf{0}, r)$. **Hint:** First show that $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|+\|\boldsymbol{y}\|$.

3. Let $C([0, T])$ denote the space of functions which are continuous on $[0, T]$. Define
$$\|f\| \equiv\|f\|_{\infty} \equiv \sup_{t \in[0, T]}|f(t)|=\max_{t \in[0, T]}|f(t)|$$
Verify the following. $\|f+g\| \leq\|f\|+\|g\|$. Then use to show that $d(f, g) \equiv\|f-g\|$ is a metric and that with this metric, $(C([0, T]), d)$ is a metric space.

4. Recall that $[a, b]$ is compact. Also, it is Lemma 3.5.9 above. Thus every open cover has a finite subcover of the set. Also recall that a sequence of numbers $\{x_n\}$ is a Cauchy sequence means that for every $\varepsilon>0$ there exists N such that if $m, n>N$, then $|x_n-x_m|<\varepsilon$. First show that every Cauchy sequence is bounded. Next, using the compactness of closed intervals, show that every Cauchy sequence has a convergent subsequence. By Theorem 3.2.2, the original Cauchy sequence converges. Thus $\mathbb{R}$ with the usual metric just described is complete because every Cauchy sequence converges.

5. Using the result of the above problem, show that $(\mathbb{R}^n,\|\cdot\|_{\infty})$ is a complete metric space. That is, every Cauchy sequence converges. Here $d(\boldsymbol{x}, \boldsymbol{y}) \equiv\|\boldsymbol{x}-\boldsymbol{y}\|_{\infty}$.

6. Suppose you had (X_i, d_i) is a metric space. Now consider the product space $X \equiv \prod_{i=1}^n X_i$ with $d(\boldsymbol{x}, \boldsymbol{y})=\max\{d(x_i, y_i), i=1 \cdots, n\}$. Would this be a metric space? If so, prove that this is the case.

Does triangle inequality hold? **Hint:** For each i,

$$d_i(x_i,z_i) \leq d_i(x_i,y_i) + d_i(y_i,z_i) \leq d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z})$$

Now take max of the two ends.

7. In the above example, if each (X_i,d_i) is complete, explain why (X,d) is also complete.

8. Show that $C([0,T])$ is a complete metric space. That is, show that if $\{f_n\}$ is a Cauchy sequence, then there exists $f \in C([0,T])$ such that

$$\lim_{n\to\infty} d(f,f_n) = \lim_{n\to\infty} \|f - f_n\| = 0$$

This is just a special case of theorems discussed in the chapter.

9. Let X be a nonempty set of points. Say it has infinitely many points. Define $d(x,y) = 1$ if $x \neq y$ and $d(x,y) = 0$ if $x = y$. Show that this is a metric. Show that in (X,d) every point is open and closed. In fact, show that every set is open and every set is closed. Is this a complete metric space? Explain why. Describe the open balls.

10. Show that the union of any set of open sets is an open set. Show the intersection of any set of closed sets is closed. Let A be a nonempty subset of a metric space (X,d). Then the closure of A, written as $\bar{A}$ is defined to be the intersection of all closed sets which contain A. Show that $\bar{A} = A \cup A'$. That is, to find the closure, you just take the set and include all limit points of the set. It was proved in the chapter, but go over it yourself.

11. Let A' denote the set of limit points of A, a nonempty subset of a metric space (X,d). Show that A' is closed.

12. A theorem was proved which gave three equivalent descriptions of compactness of a metric space. One of them said the following: A metric space is compact if and only if it is complete and totally bounded. Suppose (X,d) is a complete metric space and $K \subseteq X$. Then (K,d) is also clearly a metric space having the same metric as X. Show that (K,d) is compact if and only if it is **closed** and totally bounded. Note the similarity with the Heine Borel theorem on $\mathbb{R}$. Show that on $\mathbb{R}$, every bounded set is also totally bounded. Thus the earlier Heine Borel theorem for $\mathbb{R}$ is obtained.

13. Suppose (X_i,d_i) is a compact metric space. Then the Cartesian product is also a metric space. That is $(\prod_{i=1}^n X_i, d)$ is a metric space where $d(\boldsymbol{x},\boldsymbol{y}) \equiv \max\{d_i(x_i,y_i)\}$. Show that $(\prod_{i=1}^n X_i, d)$ is compact. Recall the Heine Borel theorem for $\mathbb{R}$. Explain why $\prod_{i=1}^n [a_i,b_i]$ is compact in $\mathbb{R}^n$ with the distance given by

$$d(\boldsymbol{x},\boldsymbol{y}) = \max\{|x_i - y_i|\}$$

Hint: It suffices to show that $(\prod_{i=1}^n X_i, d)$ is sequentially compact. Let $\{\boldsymbol{x}^m\}_{m=1}^\infty$ be a sequence. Then $\{x_1^m\}_{m=1}^\infty$ is a sequence in X_i. Therefore, it has a subsequence $\left\{x_1^{k_1}\right\}_{k_1=1}^\infty$ which converges to a point $x_1 \in X_1$. Now consider $\left\{x_2^{k_1}\right\}_{k_1=1}^\infty$ the second components. It has a subsequence denoted as k_2 such that $\left\{x_2^{k_2}\right\}_{k_2=1}^\infty$ converges to a point x_2 in X_2. Explain why $\lim_{k_2\to\infty} x_1^{k_2} = x_1$. Continue doing this n times. Explain why $\lim_{k_n\to\infty} x_l^{k_n} = x_l \in X_l$ for each l. Then explain why this is the same as saying $\lim_{k_n\to\infty} \boldsymbol{x}^{k_n} = \boldsymbol{x}$ in $\left(\prod_{i=1}^n X_i, d\right)$.

14. If you have a metric space (X,d) and a compact subset of (X,d) K, suppose that L is a closed subset of K. Explain why L must also be compact. **Hint:** Use the definition of compactness. Explain why every closed and bounded set in $\mathbb{R}^n$ is compact. Here the distance is given by $d(\boldsymbol{x},\boldsymbol{y}) \equiv \max_{1\leq i\leq n}\{|x_i - y_i|\}$.

15. Show that compactness is a topological property. If $(X,d),(Y,\rho)$ are both metric spaces and $f : X \to Y$ has the property that f is one to one, onto, and continuous, and also f^{-1} is one to one onto and continuous, then the two metric spaces are compact or not compact together. That is one is compact if and only if the other is.

16. Consider $\mathbb{R}$ the real numbers. Define a distance in the following way. $\rho(x,y) \equiv |\arctan(x) - \arctan(y)|$ Show this is a good enough distance and that the open sets which come from this distance are the same as the open sets which come from the usual distance $d(x,y) = |x-y|$. Explain why this yields that the identity mapping $f(x) = x$ is continuous with continuous inverse as a map from $(\mathbb{R},d)$ to $(\mathbb{R},\rho)$. To do this, you show that an open ball taken with respect to one of these is also open with respect to the other. However, $(\mathbb{R},\rho)$ is not a complete metric space while $(\mathbb{R},d)$ is. Thus, unlike compactness. Completeness is not a topological property. **Hint:** To show the lack of completeness of $(\mathbb{R},\rho)$, consider $x_n = n$. Show it is a Cauchy sequence with respect to ρ.

17. If K is a compact subset of (X,d) and $y \notin K$, show that there always exists $x \in K$ such that $d(x,y) = \operatorname{dist}(y,K)$. Give an example in $\mathbb{R}$ to show that this might not be so if K is not compact.

18. If S is a nonempty set, the diameter of S denoted as $\operatorname{diam}(S)$ is defined as follows. $\operatorname{diam}(S) \equiv \sup\{d(x,y) : x,y \in S\}$. Suppose (X,d) is a complete metric space and you have a nested sequence of closed sets whose diameters converge to 0. That is, each A_n is closed, $\cdots A_n \supseteq A_{n+1} \cdots$ and $\lim_{n\to\infty} \operatorname{diam}(A_n) = 0$. Show that there is exactly one point p contained in the intersection of all these sets A_n. Give an example which shows that if the condition on the diameters does not hold, then maybe there is no point in the intersection of these sets.

19. Two metric spaces $(X,d),(Y,\rho)$ are homeomorphic if there exists a continuous function $f : X \to Y$ which is one to one onto, and whose inverse is also continuous one to one and onto. Show that the interval $[0,1]$ is not homeomorphic to the unit circle. **Hint:** Recall that the continuous image of a connected set is connected, Theorem 3.11.3. However, if you remove a point from $[0,1]$ it is no longer connected but removing a single point from the circle results in a connected set.

20. Using the same methods in the above problem, show that the unit circle is not homeomorphic to the unit sphere $\left\{x^2 + y^2 + z^2 = 1\right\}$ and the unit circle is not homeomorphic to a figure eight.

21. The rational numbers $\mathbb{Q}$ and the natural numbers $\mathbb{N}$ have the property that there is a one to one and onto map from $\mathbb{N}$ to $\mathbb{Q}$. This is a simple consequence of the Schroeder Bernstein theorem presented earlier. Both of these are also metric spaces with respect

to the usual metric on $\mathbb{R}$. Are they homeomorphic? **Hint:** Suppose they were. Then in $\mathbb{Q}$ consider $(1,2)$, all the rationals between 1 and 2 excluding 1 and 2. This is not a closed set because 2 is a limit point of the set which is not in it. Now if you have f a homeomorphism, consider $f((1,2))$. Is this set closed?

22. If you have an open set O in $\mathbb{R}$, show that O is the countable union of disjoint open intervals. **Hint:** Consider the connected components. Go over this for yourself. It is in the chapter.

23. Addition and multiplication on $\mathbb{R}$ can be considered mappings from $\mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$ as follows. $+(x,y)\equiv x+y, \cdot(x,y)\equiv xy$. Here the metric on $\mathbb{R}\times\mathbb{R}$ can be taken as $d((x,y),(\hat{x},\hat{y}))\equiv \max(|x-\hat{x}|,|y-\hat{y}|)$. Show these operations are continuous functions.

24. Suppose K is a compact subset of a metric space (X,d) and there is an open cover $\mathscr{C}$ of K. Show that there exists a single positive $\delta>0$ such that if $x\in K, B(x,\delta)$ is contained in some set of $\mathscr{C}$. This number is called a Lebesgue number. Do this directly from the definition of compactness in terms of open covers without using the equivalence of compactness and sequential compactness.

25. Show uniform continuity of a continuous function defined on a compact set where compactness only refers to open covers. Use the above problem on existence of the Lebesgue number.

26. Let $f:D\to\mathbb{R}$ be a function. This function is said to be lower semicontinuous[1] at $x\in D$ if for any sequence $\{x_n\}\subseteq D$ which converges to x it follows $f(x)\leq \liminf_{n\to\infty} f(x_n)$. Suppose D is sequentially compact and f is lower semicontinuous at every point of D. Show that then f achieves its minimum on D. Here D is some metric space. Let $f:D\to\mathbb{R}$ be a function. This function is said to be upper semicontinuous at $x\in D$ if for any sequence $\{x_n\}\subseteq D$ which converges to x it follows $f(x)\geq \limsup_{n\to\infty} f(x_n)$. Suppose D is sequentially compact and f is upper semicontinuous at every point of D. Show that then f achieves its maximum on D.

27. Show that a real valued function defined on a metric space D is continuous if and only if it is both upper and lower semicontinuous.

28. Give an example of a lower semicontinuous function defined on $\mathbb{R}$ which is not continuous and an example of an upper semicontinuous function which is not continuous.

29. More generally, one considers functions which have values in $[-\infty,\infty]$. Then f is upper semicontinuous if, whenever $x_n\to x, f(x)\geq \limsup_{n\to\infty} f(x_n)$ and lower semicontinuous if whenever $x_n\to x,\ f(x)\leq \liminf_{n\to\infty} f(x_n)$. Suppose $\{f_\alpha:\alpha\in\Lambda\}$ is a collection of continuous real valued functions defined on a metric space. Let $F(x)\equiv \inf\{f_\alpha(x):\alpha\in\Lambda\}$. Show F is an upper semicontinuous function. Next let $G(x)\equiv \sup\{f_\alpha(x):\alpha\in\Lambda\}$. Show G is a lower semicontinuous function.

30. The result of this problem is due to Hausdorff. It says that if you have any lower semicontinuous real valued function defined on a metric space (X,d), then it is the

[1] The notion of lower semicontinuity is very important for functions which are defined on infinite dimensional sets.

limit of an increasing sequence of continuous functions. Here is an outline. You complete the details.

(a) First suppose $f(x) \geq 0$ for all x. Define $f_n(x) \equiv \inf_{z\in X}\{f(z)+nd(z,x)\}$. Then $f(x) \geq f_n(x)$ and $f_n(x)$ is increasing in n. Also each f_n is continuous because $f_n(x) \leq f(z)+nd(z,y)+nd(y,x)$. Thus $f_n(x) \leq f_n(y)+nd(y,x)$. Why? It follows that $|f_n(x)-f_n(y)| \leq nd(y,x)$. Why?

(b) Let $h(x) = \lim_{n\to\infty} f_n(x)$. Then $h(x) \leq f(x)$. Why? Now for each $\varepsilon > 0$, and fixed x, there exists z_n such that $f_n(x)+\varepsilon > f(z_n)+nd(z_n,x)$Why? Therefore, $z_n \to x$. Why?

(c) Then

$$\begin{aligned} h(x)+\varepsilon &= \lim_{n\to\infty} f_n(x)+\varepsilon \geq \liminf_{n\to\infty}(f(z_n)+nd(z_n,x)) \\ &\geq \liminf_{n\to\infty} f(z_n) \geq f(x) \end{aligned}$$

Why? Therefore, $h(x) \geq f(x)$ and so they are equal. Why?

(d) Now consider $f : X \to (-\infty,\infty)$ and is lower semicontinuous as just explained. Consider $\frac{\pi}{2}+\arctan f(x) \equiv g(x)$. Then $\arctan f(x) \in \left(-\frac{\pi}{2},\frac{\pi}{2}\right)$ because f has real values. Then $g(x)$ is also lower semicontinuous having values in $(0,\pi)$. Why? By what was just shown, there exists $g_n(x) \uparrow g(x)$ where each g_n is continuous. Consider $f_n(x) \equiv \tan\left(g_n(x)-\frac{\pi}{2}\right)$. Then f_n is continuous and increases to $f(x)$.

31. Generalize the above problem to the case where f is an upper semicontinuous real valued function. That is, $f(x) \geq \limsup_{n\to\infty} f(x_n)$ whenever $x_n \to x$. Show there are continuous functions $\{f_n(x)\}$ such that $f_n(x) \downarrow f(x)$. **Hint** To save trouble, maybe show that f is upper semicontinuous if and only if $-f$ is lower semicontinuous. Then maybe you could just use the above problem.

32. What if f is lower (upper) semicontinuous with values in $[-\infty,\infty]$? In this case, you consider $[-\infty,\infty]$ as a metric space as follows:$d(x,y) \equiv |\arctan(x)-\arctan(y)|$. Then you can generalize the above problems to show that if f is lower semicontinuous with values into $[-\infty,\infty]$ then it is the increasing limit of continuous functions with values in $[-\infty,\infty]$. Note that in this case a function identically equal to ∞ would be continuous so this is a rather odd sort of thing, a little different from what we normally like to consider. Check the details and explain why in this setting, the lower semicontinuous functions are exactly pointwise limits of increasing sequences of continuous functions and the upper semicontinuous functions are exactly pointwise limits of decreasing sequences of continuous functions.

33. This is a nice result in Taylor [42]. For a nonempty set $T, \partial T$ is the set of points p such that $B(p,r)$ contains points of T and points of T^C for each $r > 0$. Suppose you have T a proper subset of a metric space and S is a connected, nonempty set such that $S\cap T \neq \emptyset, S\cap T^C \neq \emptyset$. Show that S must contain a point of ∂T.

Chapter 4

Linear Spaces

The thing which is missing in the above material about metric spaces is any kind of algebra. In most applications, we are interested in adding things and multiplying things by scalars and so forth. This requires the notion of a vector space, also called a linear space. The simplest example is $\mathbb{R}^n$ which is described next.

In this chapter, $\mathbb{F}$ will refer to either $\mathbb{R}$ or $\mathbb{C}$. It doesn't make any difference to the arguments which it is and so $\mathbb{F}$ is written to symbolize whichever you wish to think about. When it is desired to emphasize that certain quantities are vectors, bold face will often be used. This is not necessarily done consistently. Sometimes context is considered sufficient.

4.1 Algebra in $\mathbb{F}^n$, Vector Spaces

There are exactly two algebraic operations done with elements of $\mathbb{F}^n$. One is addition and the other is multiplication by numbers, called scalars. In the case of $\mathbb{C}^n$ the scalars are complex numbers while in the case of $\mathbb{R}^n$ the only allowed scalars are real numbers. Thus, the scalars always come from $\mathbb{F}$ in either case.

Definition 4.1.1 *If $\boldsymbol{x} \in \mathbb{F}^n$ and $a \in \mathbb{F}$, also called a scalar, then $a\boldsymbol{x} \in \mathbb{F}^n$ is defined by*

$$a\boldsymbol{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{4.1}$$

This is known as scalar multiplication. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$ then $\boldsymbol{x} + \boldsymbol{y} \in \mathbb{F}^n$ and is defined by

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{4.2}$$

the points in $\mathbb{F}^n$ are also referred to as vectors.

Actually, in dealing with vectors in $\mathbb{F}^n$, it is more customary in linear algebra to write them as column vectors. To save space, I will sometimes write $(x_1, \cdots, x_n)^T$ to indicate the column vector having x_1 on the top and x_n on the bottom. With this definition, the algebraic properties satisfy the conclusions of the following theorem. These conclusions are called the vector space axioms. Any time you have a set and a field of scalars satisfying the axioms of the following theorem, it is called a vector space or linear space.

Theorem 4.1.2 *For $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{F}^n$ and α, β scalars, (real numbers), the following hold.*

$$\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}, \tag{4.3}$$

the commutative law of addition,

$$(\boldsymbol{v} + \boldsymbol{w}) + \boldsymbol{z} = \boldsymbol{v} + (\boldsymbol{w} + \boldsymbol{z}), \tag{4.4}$$

the associative law for addition,

$$\boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v}, \tag{4.5}$$

the existence of an additive identity,

$$\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}, \tag{4.6}$$

the existence of an additive inverse, Also

$$\alpha(\boldsymbol{v}+\boldsymbol{w}) = \alpha\boldsymbol{v}+\alpha\boldsymbol{w}, \tag{4.7}$$

$$(\alpha+\beta)\boldsymbol{v} = \alpha\boldsymbol{v}+\beta\boldsymbol{v}, \tag{4.8}$$

$$\alpha(\beta\boldsymbol{v}) = \alpha\beta(\boldsymbol{v}), \tag{4.9}$$

$$1\boldsymbol{v} = \boldsymbol{v}. \tag{4.10}$$

In the above $\mathbf{0} = (0,\cdots,0)$.

You should verify these properties all hold. For example, consider 4.7

$$\begin{aligned}\alpha(\boldsymbol{v}+\boldsymbol{w}) &= \alpha(v_1+w_1,\cdots,v_n+w_n)\\ &= (\alpha(v_1+w_1),\cdots,\alpha(v_n+w_n))\\ &= (\alpha v_1+\alpha w_1,\cdots,\alpha v_n+\alpha w_n)\\ &= (\alpha v_1,\cdots,\alpha v_n)+(\alpha w_1,\cdots,\alpha w_n)\\ &= \alpha\boldsymbol{v}+\alpha\boldsymbol{w}.\end{aligned}$$

As usual subtraction is defined as $\boldsymbol{x}-\boldsymbol{y} \equiv \boldsymbol{x}+(-\boldsymbol{y})$.

4.2 Subspaces Spans and Bases

As mentioned above, $\mathbb{F}^n$ is an example of a vector space. In dealing with vector spaces, the concept of linear combination is fundamental. When one considers only algebraic considerations, it makes no difference what field of scalars you are using. It could be $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$ or even a field of residue classes. However, go ahead and think $\mathbb{R}$ or $\mathbb{C}$ since the subject of interest here is analysis.

Definition 4.2.1 *Let $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p\}$ be vectors in a vector space Y having the field of scalars $\mathbb{F}$. A linear combination is any expression of the form $\sum_{i=1}^{p} c_i\boldsymbol{x}_i$ where the c_i are scalars. The set of all linear combinations of these vectors is called* $\text{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)$. *A vector $\boldsymbol{v}$ is said to be in the span of some set S of vectors if $\boldsymbol{v}$ is a linear combination of vectors of S.* ***This means: finite linear combination.*** *If $V \subseteq Y$, then V is called a subspace if it contains $\mathbf{0}$ and whenever α,β are scalars and $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of V, it follows $\alpha\boldsymbol{u}+\beta\boldsymbol{v} \in V$. That is, it is "closed under the algebraic operations of vector addition and scalar multiplication" and is therefore, a vector space. A linear combination of vectors is said to be trivial if all the scalars in the linear combination equal zero. A set of vectors is said to be linearly independent if the only linear combination of these vectors which equals the zero vector is the trivial linear combination. Thus $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ is called linearly independent if whenever $\sum_{k=1}^{n} c_k\boldsymbol{x}_k = \mathbf{0}$, it follows that all the scalars, c_k equal zero. A set of vectors, $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$, is called linearly dependent if it is not linearly independent. Thus the set of vectors is linearly dependent if there exist scalars, $c_i, i = 1,\cdots,n$, not all zero such that $\sum_{k=1}^{n} c_k\boldsymbol{x}_k = \mathbf{0}$.*

Lemma 4.2.2 *A set of vectors $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

Proof: Suppose first that $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ is linearly independent. If

$$\boldsymbol{x}_k = \sum_{j \neq k} c_j \boldsymbol{x}_j,$$

then $\mathbf{0} = 1\boldsymbol{x}_k + \sum_{j \neq k} (-c_j) \boldsymbol{x}_j$, a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.

Now suppose no vector is a linear combination of the others. Is $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ linearly independent? If it is not, there exist scalars, c_i, not all zero such that $\sum_{i=1}^n c_i \boldsymbol{x}_i = \mathbf{0}.$ Say $c_k \neq 0$. Then you can solve for $\boldsymbol{x}_k$ as $\boldsymbol{x}_k = \sum_{j \neq k} (-c_j/c_k) \boldsymbol{x}_j$ contrary to assumption. This proves the lemma. ■

The following is called the exchange theorem.

Theorem 4.2.3 *If*

$$\operatorname{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r) \subseteq \operatorname{span}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s) \equiv V$$

and $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$ are linearly independent, then $r \leq s$.

Proof: Suppose $r > s$. Let F_p denote the first p vectors in $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Let F_0 denote the empty set. Let E_p denote a finite list of vectors of $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s\}$ and let $|E_p|$ denote the number of vectors in the list. Note that, by assumption, $\operatorname{span}(F_0, E_s) = V$. For $0 \leq p \leq s$, let E_p have the property $\operatorname{span}(F_p, E_p) = V$ and $|E_p|$ is as small as possible for this to happen. If $|E_p| = 0$, then $\operatorname{span}(F_p) = V$ which would imply that, since $r > s \geq p, \boldsymbol{u}_r \in \operatorname{span}(F_s)$ contradicting the linear independence of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Assume then that $|E_p| > 0$. Then $\boldsymbol{u}_{p+1} \in \operatorname{span}(F_p, E_p)$ and so there are constants, $c_1, \cdots, c_p$ and $d_1, \cdots, d_m$ 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 $\{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_m\} \subseteq \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s\}$. Then not all the d_i can equal zero because this would violate the linear independence of the $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Therefore, you can solve for one of the $\boldsymbol{z}_k$ as a linear combination of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{p+1}\}$ and the other $\boldsymbol{z}_j$. Thus you can change F_p to F_{p+1} and include one fewer vector in E_{p+1} with $\operatorname{span}(F_{p+1}, E_{p+1}) = V$ and so $|E_{p+1}| < |E_p|$ contrary to the claim that $|E_p|$ was as small as possible. Thus $|E_p| = 0$ after all and so a contradiction results.

Alternate proof: Recall from linear algebra that if you have A an $m \times n$ matrix where $m < n$ so there are more columns than rows, then there exists a nonzero solution $\boldsymbol{x}$ to the equation $A\boldsymbol{x} = \mathbf{0}$. Recall why this was. You must have free variables. Then by assumption, you have $\boldsymbol{u}_j = \sum_{i=1}^s a_{ij} \boldsymbol{v}_i$. If $s < r$, then the matrix (a_{ij}) has more columns than rows and so there exists a nonzero vector $\boldsymbol{x} \in \mathbb{F}^r$ such that $\sum_{j=1}^r a_{ij} x_j = 0$. Then consider the following.

$$\sum_{j=1}^r x_j \boldsymbol{u}_j = \sum_{j=1}^r x_j \sum_{i=1}^s a_{ij} \boldsymbol{v}_i = \sum_i \sum_j a_{ij} x_j \boldsymbol{v}_i = \sum_i 0 \boldsymbol{v}_j = \mathbf{0}$$

and since not all $x_j = 0$, this contradicts the independence of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. ■

Definition 4.2.4 *A finite set of vectors, $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ is a basis for a vector space V if*

$$\operatorname{span}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r) = V$$

and $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ 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 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ and $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ are said to "span" V.

Corollary 4.2.5 *Let $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ and $\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$ be two bases[1] of $\mathbb{F}^n$. Then $r=s=n$. More generally, if you have two bases for a vector space V then they have the same number of vectors.*

Proof: From the exchange theorem, Theorem 4.2.3, if

$$\{\mathbf{x}_1,\cdots,\mathbf{x}_r\},\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$$

are two bases for V, then $r\le s$ and $s\le r$. Now note the vectors,

$$\mathbf{e}_i=\overbrace{(0,\cdots,0,1,0\cdots,0)^T}^{\text{1 is in the } i^{th} \text{ slot}}$$

for $i=1,2,\cdots,n$ are a basis for $\mathbb{F}^n$. ■

Lemma 4.2.6 *Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}$ be a set of vectors. Then $V\equiv\text{span}(\mathbf{v}_1,\cdots,\mathbf{v}_r)$ is a subspace.*

Proof: Suppose α,β are two scalars and let $\sum_{k=1}^r c_k\mathbf{v}_k$ and $\sum_{k=1}^r d_k\mathbf{v}_k$ are two elements of V. What about $\alpha\sum_{k=1}^r c_k\mathbf{v}_k+\beta\sum_{k=1}^r d_k\mathbf{v}_k$? Is it also in V?

$$\alpha\sum_{k=1}^r c_k\mathbf{v}_k+\beta\sum_{k=1}^r d_k\mathbf{v}_k=\sum_{k=1}^r(\alpha c_k+\beta d_k)\mathbf{v}_k\in V$$

so the answer is yes. It is clear that 0 is in $\text{span}(\mathbf{v}_1,\cdots,\mathbf{v}_r)$. This proves the lemma. ■

Definition 4.2.7 *Let V be a vector space. It is finite dimensional when it has a basis of finitely many vectors. Otherwise, it is infinite dimensional. Then* $\dim(V)$ *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.

Lemma 4.2.8 *Suppose $\mathbf{v}\notin\text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ and $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1,\cdots,\mathbf{u}_k,\mathbf{v}\}$ is also linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i\mathbf{u}_i+d\mathbf{v}=\mathbf{0}$. It is required to verify that each $c_i=0$ and that $d=0$. But if $d\neq0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$, $\mathbf{v}=-\sum_{i=1}^k\left(\frac{c_i}{d}\right)\mathbf{u}_i$ contrary to assumption. Therefore, $d=0$. But then $\sum_{i=1}^k c_i\mathbf{u}_i=0$ and the linear independence of $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ implies each $c_i=0$ also. ■

Theorem 4.2.9 *Let V be a nonzero subspace of Y a finite dimensional vector space having dimension n. Then V has a basis.*

[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.

Proof: Let $\boldsymbol{v}_1 \in V$ where $\boldsymbol{v}_1 \neq 0$. If $\operatorname{span}\{\boldsymbol{v}_1\} = V$, stop. $\{\boldsymbol{v}_1\}$ is a basis for V. Otherwise, there exists $\boldsymbol{v}_2 \in V$ which is not in $\operatorname{span}\{\boldsymbol{v}_1\}$. By Lemma 4.2.8 $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ is a linearly independent set of vectors. If $\operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\} = V$ stop, $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ is a basis for V. If $\operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\} \neq V$, then there exists $\boldsymbol{v}_3 \notin \operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem, Theorem 4.2.3, and the assumed dimension of Y. ■

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 4.2.10 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\}$ 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*

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_s\}$$

is a basis for V.

Proof: This follows immediately from the proof of Theorem 4.2.9. You do exactly the same argument except you start with $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\}$ rather than $\{\boldsymbol{v}_1\}$. ■

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 4.2.11 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and suppose $\operatorname{span}(\boldsymbol{u}_1 \cdots, \boldsymbol{u}_p) = V$ where the $\boldsymbol{u}_i$ are nonzero vectors. Then there exist vectors, $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ such that $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\} \subseteq \{\boldsymbol{u}_1 \cdots, \boldsymbol{u}_p\}$ and $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ is a basis for V.*

Proof: Let r be the smallest positive integer with the property that for some set,

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\} \subseteq \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}, \operatorname{span}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r) = V.$$

Then $r \leq p$ and it must be the case that $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ is linearly independent because if it were not so, one of the vectors, say $\boldsymbol{v}_k$ would be a linear combination of the others. But then you could delete this vector from $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ and the resulting list of $r-1$ vectors would still span V contrary to the definition of r. ■

4.3 Inner Product and Normed Linear Spaces

4.3.1 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.

Definition 4.3.1 *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$. Thus $\boldsymbol{x} = (x_1, \cdots, x_n)$ 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*

$$(\boldsymbol{x}, \boldsymbol{y}) \equiv \sum_j x_j \overline{y_j} \equiv x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.$$

Sometimes it is denoted as $\boldsymbol{x} \cdot \boldsymbol{y}$.

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 involve a conjugate. 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 $(\boldsymbol{x},\boldsymbol{x}) = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0$. If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example, $(1+i)^2 + 2^2 = 4+2i$ 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 $\boldsymbol{u} = \boldsymbol{0}$.

Note this implies $(\boldsymbol{x},\alpha\boldsymbol{y}) = \overline{\alpha}(\boldsymbol{x},\boldsymbol{y})$ because

$$(\boldsymbol{x},\alpha\boldsymbol{y}) = \overline{(\alpha\boldsymbol{y},\boldsymbol{x})} = \overline{\alpha(\boldsymbol{y},\boldsymbol{x})} = \overline{\alpha}(\boldsymbol{x},\boldsymbol{y})$$

The norm is defined as follows.

Definition 4.3.2 *For $\boldsymbol{x} \in \mathbb{F}^n$, $|\boldsymbol{x}| \equiv \left(\sum_{k=1}^n |x_k|^2\right)^{1/2} = (\boldsymbol{x},\boldsymbol{x})^{1/2}$.*

4.3.2 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.

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which holds in any inner product space. First here is a simple lemma.

Lemma 4.3.3 *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 $\theta = \dfrac{\overline{z}}{|z|}$. Recall that for $z = x+iy, \overline{z} = x-iy$ and $\overline{z}z = |z|^2$. In case z is real, there is no change in the above. ■

Theorem 4.3.4 *(Cauchy Schwarz)Let H be an inner product space. The following inequality holds for $\boldsymbol{x}$ and $\boldsymbol{y} \in H$.*

$$|(\boldsymbol{x},\boldsymbol{y})| \leq (\boldsymbol{x},\boldsymbol{x})^{1/2}(\boldsymbol{y},\boldsymbol{y})^{1/2} \tag{4.11}$$

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 $\theta(\boldsymbol{x},\boldsymbol{y}) = |(\boldsymbol{x},\boldsymbol{y})|$. Consider

$$p(t) \equiv \left(\boldsymbol{x} + \overline{\theta}t\boldsymbol{y}, \boldsymbol{x} + t\overline{\theta}\boldsymbol{y}\right)$$

where $t \in \mathbb{R}$. Then from the above list of properties of the inner product,

$$\begin{aligned} 0 &\leq p(t) = (\mathbf{x},\mathbf{x}) + t\theta(\mathbf{x},\mathbf{y}) + t\overline{\theta}(\mathbf{y},\mathbf{x}) + t^2(\mathbf{y},\mathbf{y}) \\ &= (\mathbf{x},\mathbf{x}) + t\theta(\mathbf{x},\mathbf{y}) + t\overline{\theta(\mathbf{x},\mathbf{y})} + t^2(\mathbf{y},\mathbf{y}) \\ &= (\mathbf{x},\mathbf{x}) + 2t\,\mathrm{Re}(\theta(\mathbf{x},\mathbf{y})) + t^2(\mathbf{y},\mathbf{y}) \\ &= (\mathbf{x},\mathbf{x}) + 2t\,|(\mathbf{x},\mathbf{y})| + t^2(\mathbf{y},\mathbf{y}) \end{aligned} \tag{4.12}$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(\mathbf{y},\mathbf{y}) = 0$ it must be the case that $|(\mathbf{x},\mathbf{y})| = 0$ also since otherwise the above inequality would be violated. Therefore, in this case, $|(\mathbf{x},\mathbf{y})| \leq (\mathbf{x},\mathbf{x})^{1/2}(\mathbf{y},\mathbf{y})^{1/2}$. On the other hand, if $(\mathbf{y},\mathbf{y}) \neq 0$, 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 $4|(\mathbf{x},\mathbf{y})|^2 - 4(\mathbf{x},\mathbf{x})(\mathbf{y},\mathbf{y}) \leq 0$ which is equivalent to 4.11.

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 4.11. Conversely, suppose equality does hold. Then this is equivalent to saying $4|(\mathbf{x},\mathbf{y})|^2 - 4(\mathbf{x},\mathbf{x})(\mathbf{y},\mathbf{y}) = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it t_0. Then

$$p(t_0) \equiv \left(\mathbf{x} + \overline{\theta}t_0\mathbf{y}, \mathbf{x} + t_0\overline{\theta}\mathbf{y}\right) = \left|\mathbf{x} + \overline{\theta}t\mathbf{y}\right|^2 = 0$$

and so $\mathbf{x} = -\overline{\theta}t_0\mathbf{y}$. ■

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 $(\mathbf{x},\mathbf{x}) = 0$ then $\mathbf{x} = \mathbf{0}$. That was only used to consider the case of equality.

Now the length of a vector can be defined.

Definition 4.3.5 *Let $z \in H$. Then $|z| \equiv (\mathbf{z},\mathbf{z})^{1/2}$.*

Theorem 4.3.6 *For length defined in Definition 4.3.5, the following hold.*

$$|\mathbf{z}| \geq 0 \text{ and } |\mathbf{z}| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \tag{4.13}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha\mathbf{z}| = |\alpha|\,|\mathbf{z}| \tag{4.14}$$

$$|\mathbf{z} + \mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}|. \tag{4.15}$$

Proof: The first two claims are left as exercises. To establish the third,

$$\begin{aligned} |\mathbf{z}+\mathbf{w}|^2 &\equiv (\mathbf{z}+\mathbf{w}, \mathbf{z}+\mathbf{w}) \\ &= (\mathbf{z},\mathbf{z}) + (\mathbf{w},\mathbf{w}) + (\mathbf{w},\mathbf{z}) + (\mathbf{z},\mathbf{w}) \\ &= |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\,\mathrm{Re}(\mathbf{w},\mathbf{z}) \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\,|(\mathbf{w},\mathbf{z})| \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\,|\mathbf{w}|\,|\mathbf{z}| = (|\mathbf{z}| + |\mathbf{w}|)^2. \end{aligned}$$

Note that in an inner product space, you can define $d(\mathbf{x},\mathbf{y}) \equiv |\mathbf{x} - \mathbf{y}|$ and this is a metric for this inner product space. This follows from the above since d satisfies the conditions for a metric,

$$d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}),\ d(\mathbf{x},\mathbf{y}) \geq 0 \text{ and equals 0 if and only if } \mathbf{x} = \mathbf{y}$$

$$d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) = |\mathbf{x} - \mathbf{y}| + |\mathbf{y} - \mathbf{z}| \geq |\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}| = |\mathbf{x} - \mathbf{z}| = d(\mathbf{x},\mathbf{z}).$$

It follows that all the theory of metric spaces developed earlier applies to this situation.

4.3.3 Normed Vector Spaces

The best sort of a norm is one which comes from an inner product. 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 4.13 - 4.15. That is

$$\|\boldsymbol{z}\| \geq 0 \text{ and } \|\boldsymbol{z}\| = 0 \text{ if and only if } \boldsymbol{z} = \boldsymbol{0} \tag{4.16}$$

$$\text{If } \alpha \text{ is a scalar, } \|\alpha \boldsymbol{z}\| = |\alpha| \, \|\boldsymbol{z}\| \tag{4.17}$$

$$\|\boldsymbol{z}+\boldsymbol{w}\| \leq \|\boldsymbol{z}\| + \|\boldsymbol{w}\| . \tag{4.18}$$

The last inequality above is called the triangle inequality. Another version of this is

$$|\|\boldsymbol{z}\| - \|\boldsymbol{w}\|| \leq \|\boldsymbol{z}-\boldsymbol{w}\| \tag{4.19}$$

To see that 4.19 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 $\boldsymbol{z}$ and $\boldsymbol{w}$, yields $\|\boldsymbol{w}\| - \|\boldsymbol{z}\| \leq \|\boldsymbol{z}-\boldsymbol{w}\|$ which implies 4.19.

Any normed vector space is a metric space, the distance given by $d(\boldsymbol{x},\boldsymbol{y}) \equiv \|\boldsymbol{x}-\boldsymbol{y}\|$. This satisfies all the axioms of a distance. Therefore, any normed linear space is a metric space with this metric and all the theory of metric spaces applies.

Definition 4.3.7 *When X is a normed linear space which is also complete, it is called a Banach space.*

A Banach space may or may not be finite dimensional but it is always a linear space or vector space. The field of scalars will always be $\mathbb{R}$ or $\mathbb{C}$ at least in this book. More is said about Banach spaces later.

4.3.4 The p Norms

Examples of norms are the p norms on $\mathbb{C}^n$ for $p \neq 2$. These do not come from an inner product but they are norms just the same.

Definition 4.3.8 *Let $\boldsymbol{x} \in \mathbb{C}^n$. Then define for $p \geq 1$,*

$$\|\boldsymbol{x}\|_p \equiv \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} .$$

The following inequality is called Holder's inequality.

Proposition 4.3.9 *For $\boldsymbol{x},\boldsymbol{y} \in \mathbb{C}^n$,*

$$\sum_{i=1}^n |x_i| \, |y_i| \leq \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \left(\sum_{i=1}^n |y_i|^{p'} \right)^{1/p'}$$

The proof will depend on the following lemma shown later.

Lemma 4.3.10 *If $a,b \geq 0$ and p' is defined by $\frac{1}{p}+\frac{1}{p'}=1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'} .$$

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 $A = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 4.3.10,

$$\begin{aligned}\sum_{i=1}^n \frac{|x_i|}{A}\frac{|y_i|}{B} &\leq \sum_{i=1}^n \left[\frac{1}{p}\left(\frac{|x_i|}{A}\right)^p + \frac{1}{p'}\left(\frac{|y_i|}{B}\right)^{p'}\right] \\ &= \frac{1}{p}\frac{1}{A^p}\sum_{i=1}^n |x_i|^p + \frac{1}{p'}\frac{1}{B^p}\sum_{i=1}^n |y_i|^{p'} \\ &= \frac{1}{p} + \frac{1}{p'} = 1\end{aligned}$$

and so $\sum_{i=1}^n |x_i|\,|y_i| \leq AB = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}\left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. ■

Theorem 4.3.11 *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 $\frac{p}{p'} = p-1$. Then using the Holder inequality,

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^p &= \sum_{i=1}^n |x_i+y_i|^p \leq \sum_{i=1}^n |x_i+y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i+y_i|^{p-1}|y_i| \\ &= \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|x_i| + \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|y_i| \\ &\leq \left(\sum_{i=1}^n |x_i+y_i|^p\right)^{1/p'}\left[\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}\right] \\ &= \|\boldsymbol{x}+\boldsymbol{y}\|^{p/p'}\left(\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p\right)\end{aligned}$$

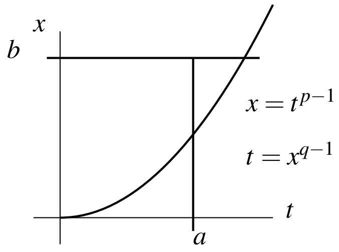
so dividing by $\|\boldsymbol{x}+\boldsymbol{y}\|^{p/p'}$, it follows

$$\|\boldsymbol{x}+\boldsymbol{y}\|^p\,\|\boldsymbol{x}+\boldsymbol{y}\|^{-p/p'} = \|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$$

$\left(p - \frac{p}{p'} = p\left(1-\frac{1}{p'}\right) = p\frac{1}{p} = 1.\right)$. ■

It only remains to prove Lemma 4.3.10.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1}dt + \int_0^b x^{q-1}dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

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

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

Then the second derivative of f is positive on $(0, \infty)$ so its graph is convex. Also $f(0) > 0$ and $\lim_{a \to \infty} f(a) = \infty$. 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

$$\|\boldsymbol{x}\|_\infty \equiv \max\{|x_k| : k = 1, 2, \cdots, n\}$$

You should verify that this satisfies all the axioms of a norm. Here is the triangle inequality.

$$\begin{aligned} \|\boldsymbol{x} + \boldsymbol{y}\|_\infty &= \max_k \{|x_k + y_k|\} \leq \max_k \{|x_k| + |y_k|\} \\ &\leq \max_k \{|x_k|\} + \max_k \{|y_k|\} = \|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty \end{aligned}$$

It turns out that in terms of analysis, it makes **absolutely no difference** which norm you use. This will be explained later. First is a short review of the notion of orthonormal bases which is not needed directly in what follows but is sufficiently important to include.

4.3.5 Orthonormal Bases

Not all bases for an inner product space H are created equal. The best bases are orthonormal.

Definition 4.3.12 *Suppose $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ is a set of vectors in an inner product space H. It is an orthonormal set if*

$$(\boldsymbol{v}_i, \boldsymbol{v}_j) = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Every orthonormal set of vectors is automatically linearly independent.

Proposition 4.3.13 *Suppose $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ is an orthonormal set of vectors. Then it is linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i \boldsymbol{v}_i = \boldsymbol{0}$. Then taking inner products with

$$\boldsymbol{v}_j, 0 = (\boldsymbol{0}, \boldsymbol{v}_j) = \sum_i c_i (\boldsymbol{v}_i, \boldsymbol{v}_j) = \sum_i c_i \delta_{ij} = c_j.$$

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. The process by which this is done is called the Gram Schmidt process.

Lemma 4.3.14 *Let X be a subspace of dimension n which is contained in an inner product space H. Let a basis for X be $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$. Then there exists an orthonormal basis for X, $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$ which has the property that for each $k \leq n$, span$(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k) =$* span$(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k)$.

Proof: Let $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ be a basis for X. Let $\boldsymbol{u}_1 \equiv \boldsymbol{x}_1 / |\boldsymbol{x}_1|$. Thus if $k = 1$, span$(\boldsymbol{u}_1) =$ span$(\boldsymbol{x}_1)$ and $\{\boldsymbol{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\boldsymbol{u}_1$, $\cdots$, $\boldsymbol{u}_k$ have been chosen such that $(\boldsymbol{u}_j, \boldsymbol{u}_l) = \delta_{jl}$ and span$(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k) =$ span$(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k)$. Then define

$$\boldsymbol{u}_{k+1} \equiv \frac{\boldsymbol{x}_{k+1} - \sum_{j=1}^{k} (\boldsymbol{x}_{k+1}, \boldsymbol{u}_j) \boldsymbol{u}_j}{\left| \boldsymbol{x}_{k+1} - \sum_{j=1}^{k} (\boldsymbol{x}_{k+1}, \boldsymbol{u}_j) \boldsymbol{u}_j \right|}, \tag{4.20}$$

where the denominator is not equal to zero because the $\boldsymbol{x}_j$ form a basis and so

$$\boldsymbol{x}_{k+1} \notin \text{span}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k) = \text{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k)$$

Thus by induction, $\boldsymbol{u}_{k+1} \in \text{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k, \boldsymbol{x}_{k+1}) = \text{span}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1})$. Also, $\boldsymbol{x}_{k+1} \in \text{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k, \boldsymbol{u}_{k+1})$ which is seen easily by solving 4.20 for $\boldsymbol{x}_{k+1}$ and it follows that span$(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}) =$ span$(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k, \boldsymbol{u}_{k+1})$. If $l \leq k$, then denoting by C the scalar $\left| \boldsymbol{x}_{k+1} - \sum_{j=1}^{k} (\boldsymbol{x}_{k+1}, \boldsymbol{u}_j) \boldsymbol{u}_j \right|^{-1}$, $(\boldsymbol{u}_{k+1}, \boldsymbol{u}_l) =$

$$\begin{aligned} C\left((\boldsymbol{x}_{k+1}, \boldsymbol{u}_l) - \sum_{j=1}^{k} (\boldsymbol{x}_{k+1}, \boldsymbol{u}_j)(\boldsymbol{u}_j, \boldsymbol{u}_l) \right) &= C\left((\boldsymbol{x}_{k+1}, \boldsymbol{u}_l) - \sum_{j=1}^{k} (\boldsymbol{x}_{k+1}, \boldsymbol{u}_j) \delta_{lj} \right) \\ &= C\left((\boldsymbol{x}_{k+1}, \boldsymbol{u}_l) - (\boldsymbol{x}_{k+1}, \boldsymbol{u}_l) \right) = 0. \end{aligned}$$

The vectors, $\{\boldsymbol{u}_j\}_{j=1}^{n}$, 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.

4.4 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. First are some simple lemmas featuring one dimensional considerations. In this case, the distance is given by $d(x, y) = |x - y|$ and so the open balls are sets of the form $(x - \delta, x + \delta)$.

Also recall the Lemma 3.5.9 which is stated next for convenience.

Lemma 4.4.1 *The closed interval $[a, b]$ is compact.*

Corollary 4.4.2 *The set $Q \equiv [a, b] + i[c, d] \subseteq \mathbb{C}$ is compact, meaning*

$$\{x + iy : x \in [a, b], y \in [c, d]\}$$

Proof: Let $\{x_n + iy_n\}$ be a sequence in Q. Then there is a subsequence such that $\lim_{k\to\infty} x_{n_k} = x \in [a, b]$. There is a further subsequence such that $\lim_{l\to\infty} y_{n_{k_l}} = y \in [c, d]$. Thus, also $\lim_{l\to\infty} x_{n_{k_l}} = x$ because subsequences of convergent sequences converge to the same point. Therefore, from the way we measure the distance in $\mathbb{C}$, it follows that $\lim_{l\to\infty} \left(x_{n_{k_l}} + y_{n_{k_l}} \right) = x + iy \in Q$. ■

The next corollary gives the definition of a closed disk and shows that, like a closed interval, a closed disk is compact.

Corollary 4.4.3 *In $\mathbb{C}$, let $D(z,r) \equiv \{w \in \mathbb{C} : |z-w| \leq r\}$. Then $D(z,r)$ is compact.*

Proof: Note that $D(z,r) \subseteq [\mathrm{Re}\, z - r, \mathrm{Re}\, z + r] + i[\mathrm{Im}\, z - r, \mathrm{Im}\, z + r]$, just shown to be compact. Also, if $w_k \to w$ where $w_k \in D(z,r)$, then by the triangle inequality,

$$|z-w| = \lim_{k\to\infty} |z-w_k| \leq r$$

and so $D(z,r)$ is a closed subset of a compact set. Hence it is compact by Proposition 3.5.2. ■

Recall that sequentially compact and compact are the same in any metric space which is the context of the assertions here.

Lemma 4.4.4 *Let K_i be a nonempty compact set in $\mathbb{F}$. Then $P \equiv \prod_{i=1}^n K_i$ is compact in $\mathbb{F}^n$.*

Proof: Let $\{\boldsymbol{x}_k\}$ be a sequence in P. Taking a succession of subsequences as in the proof of Corollary 4.4.2, there exists a subsequence, still denoted as $\{\boldsymbol{x}_k\}$ such that if x_k^i is the i^{th} component of $\boldsymbol{x}_k$, then $\lim_{k\to\infty} x_k^i = x^i \in K_i$. Thus if $\boldsymbol{x}$ is the vector of P whose i^{th} component is x^i, $\lim_{k\to\infty} |\boldsymbol{x}_k - \boldsymbol{x}| \equiv \lim_{k\to\infty} \left(\sum_{i=1}^n \left| x_k^i - x^i \right|^2 \right)^{1/2} = 0$. It follows that P is sequentially compact, hence compact. ■

A set K in $\mathbb{F}^n$ is said to be bounded if it is contained in some ball $B(\mathbf{0},r)$.

Theorem 4.4.5 *A set $K \subseteq \mathbb{F}^n$ is compact if it is closed and bounded. If $f : K \to \mathbb{R}$, then f achieves its maximum and its minimum on K.*

Proof: Say K is closed and bounded, being contained in $B(\mathbf{0},r)$. Then if $\boldsymbol{x} \in K$, $|x_i| < r$ where x_i is the i^{th} component. Hence $K \subseteq \prod_{i=1}^n D(0,r)$, a compact set by Lemma 4.4.4. By Proposition 3.5.2, since K is a closed subset of a compact set, it is compact. The last claim is just the extreme value theorem, Theorem 3.7.2. ■

Definition 4.4.6 *Let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ be a basis for V where $(V, \|\cdot\|)$ is a finite dimensional normed vector space with field of scalars equal to either $\mathbb{R}$ or $\mathbb{C}$. Define $\theta : V \to \mathbb{F}^n$ as follows. $\theta \left(\sum_{j=1}^n \alpha_j v_j \right) \equiv \boldsymbol{\alpha} \equiv (\alpha_1, \cdots, \alpha_n)^T$. Thus θ maps a vector to its coordinates taken with respect to a given basis.*

The following fundamental lemma comes from the extreme value theorem for continuous functions defined on a compact set. Let $f(\boldsymbol{\alpha}) \equiv \|\sum_i \alpha_i \boldsymbol{v}_i\| \equiv \left\| \theta^{-1} \boldsymbol{\alpha} \right\|$. Then it is clear that f is a continuous function defined on $\mathbb{F}^n$. This is because $\boldsymbol{\alpha} \to \sum_i \alpha_i \boldsymbol{v}_i$ is a continuous map into V and from the triangle inequality $\boldsymbol{x} \to \|\boldsymbol{x}\|$ is continuous as a map from V to $\mathbb{R}$.

Lemma 4.4.7 *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,

$$\delta |\boldsymbol{\alpha}| \leq \left\| \theta^{-1} \boldsymbol{\alpha} \right\| \leq \Delta |\boldsymbol{\alpha}| \tag{4.21}$$

$$\delta |\theta \boldsymbol{v}| \leq \|\boldsymbol{v}\| \leq \Delta |\theta \boldsymbol{v}| \tag{4.22}$$

Proof: These numbers exist thanks to Theorem 4.4.5. 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 $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ 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. ■

Note that these inequalities yield the fact that convergence of the coordinates with respect to a given basis is equivalent to convergence of the vectors. More precisely, to say that $\lim_{k\to\infty} \boldsymbol{v}^k = \boldsymbol{v}$ is the same as saying that $\lim_{k\to\infty} \theta \boldsymbol{v}^k = \theta \boldsymbol{v}$. Indeed, $\delta |\theta \boldsymbol{v}_n - \theta \boldsymbol{v}| \leq \|\boldsymbol{v}_n - \boldsymbol{v}\| \leq \Delta |\theta \boldsymbol{v}_n - \theta \boldsymbol{v}|$.

Now we can draw several conclusions about $(V, \|\cdot\|)$ for V finite dimensional.

Theorem 4.4.8 *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 $f : K \to \mathbb{R}$, then f achieves its maximum and minimum on K.*

Proof: First note that the inequalities 4.21 and 4.22 show that both θ^{-1} and θ are continuous. Thus these take convergent sequences to convergent sequences.

Let $\{\boldsymbol{w}_k\}_{k=1}^{\infty}$ be a Cauchy sequence. Then from 4.22, $\{\theta \boldsymbol{w}_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Thanks to Theorem 4.4.5, it converges to some $\boldsymbol{\beta} \in \mathbb{F}^n$. It follows that $\lim_{k\to\infty} \theta^{-1}\theta \boldsymbol{w}_k = \lim_{k\to\infty} \boldsymbol{w}_k = \theta^{-1}\boldsymbol{\beta} \in V$. This shows completeness.

Next let K be a closed and bounded set. Let $\{\boldsymbol{w}_k\} \subseteq K$. Then $\{\theta \boldsymbol{w}_k\} \subseteq \theta K$ which is also a closed and bounded set thanks to the inequalities 4.21 and 4.22. Thus there is a subsequence still denoted with k such that $\theta \boldsymbol{w}_k \to \boldsymbol{\beta} \in \mathbb{F}^n$. Then as just done, $\boldsymbol{w}_k \to \theta^{-1}\boldsymbol{\beta}$. Since K is closed, it follows that $\theta^{-1}\beta \in K$.

This has just shown that a closed and bounded set in V is sequentially compact hence compact.

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 $K, \{\boldsymbol{k}^m\}_{m=1}^{\infty}$ such that $\|\boldsymbol{k}^m\| \geq \left\|\boldsymbol{k}^{m-1}\right\| + 1$. It follows that it cannot have a convergent subsequence because the points are further apart from each other than 1/2. Indeed,

$$\left\|\boldsymbol{k}^m - \boldsymbol{k}^{m+1}\right\| \geq \left\|\boldsymbol{k}^{m+1}\right\| - \|\boldsymbol{k}^m\| \geq 1 > 1/2$$

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 3.1.8, there is a sequence of distinct points having no repeats and none equal to $\boldsymbol{k}$ denoted as $\{\boldsymbol{k}^m\}_{m=1}^{\infty}$ such that $\boldsymbol{k}^m \to \boldsymbol{k}$. Then this sequence $\{\boldsymbol{k}^m\}$ 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 the extreme value theorem, Theorem 3.7.2. ■

Next is the theorem which states that any two norms on a finite dimensional vector space are equivalent.

Theorem 4.4.9 *Let $\|\cdot\|, \|\cdot\|_1$ 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 $\boldsymbol{v}$,*

$$a \|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_1 \leq b \|\boldsymbol{v}\|$$

Proof: In Lemma 4.4.7, let δ,Δ go with $\|\cdot\|$ and $\hat{\delta},\hat{\Delta}$ go with $\|\cdot\|_1$. Then using the inequalities of this lemma,

$$\|\boldsymbol{v}\| \leq \Delta|\theta\boldsymbol{v}| \leq \frac{\Delta}{\hat{\delta}}\|\boldsymbol{v}\|_1 \leq \frac{\Delta\hat{\Delta}}{\hat{\delta}}|\theta\boldsymbol{v}| \leq \frac{\Delta}{\delta}\frac{\hat{\Delta}}{\hat{\delta}}\|\boldsymbol{v}\|$$

and so $\frac{\hat{\delta}}{\Delta}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_1 \leq \frac{\hat{\Delta}}{\delta}\|\boldsymbol{v}\|$. 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 4.4.10 *Consider the metric spaces $(V,\|\cdot\|_1),(V,\|\cdot\|_2)$ 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 3.6.2, the theorem about the equivalent formulations of continuity. Using this theorem, it follows from Theorem 4.4.9 that the identity map $I(\boldsymbol{x}) \equiv \boldsymbol{x}$ is continuous. The reason for this is that the inequality of this theorem implies that if $\|\boldsymbol{v}^m - \boldsymbol{v}\|_1 \to 0$ then $\|I\boldsymbol{v}^m - I\boldsymbol{v}\|_2 = \|I(\boldsymbol{v}^m - \boldsymbol{v})\|_2 \to 0$ 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 $(V,\|\cdot\|_1)$. This means it is contained in $B(\mathbf{0},r)_1$ 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 $(V,\|\cdot\|_2)$. Similarly, if S is bounded in $\|\cdot\|_2$ then it is bounded in $\|\cdot\|_1$. ■

One can show that in the case of $\mathbb{R}$ where it makes sense to consider sup and inf, convergence of Cauchy sequences can be shown to imply the other definition of completeness involving sup, and inf.

4.5 Vitali Covering Theorem

These covering theorems make sense on any finite dimensional normed linear space. There are two which are commonly used, the Vitali theorem and the Besicovitch theorem. The first adjusts the size of balls and the second does not. The Vitali theorem is the only one I will use in this book. See my larger book "Real and Abstract Analysis" for the Besicovitch theorem.

The Vitali covering theorem is a profound result about coverings of a set in $(X,\|\cdot\|)$ with balls. Usually we are interested in $\mathbb{R}^p$ with some norm. We will tacitly assume all balls have positive radius. They will not be single points. Before beginning the proof, here is a useful lemma.

Lemma 4.5.1 *In a normed linear space, $\overline{B(\boldsymbol{x},r)} = \{\boldsymbol{y} : \|\boldsymbol{y}-\boldsymbol{x}\| \leq r\}$.*

Proof: It is clear that $\overline{B(\boldsymbol{x},r)} \subseteq \{\boldsymbol{y} : \|\boldsymbol{y}-\boldsymbol{x}\| \leq r\}$ because if $\boldsymbol{y} \in \overline{B(\boldsymbol{x},r)}$, then there exists a sequence of points of $B(\boldsymbol{x},r),\{\boldsymbol{x}_n\}$ such that $\|\boldsymbol{x}_n - \boldsymbol{y}\| \to 0, \|\boldsymbol{x}_n\| < r$. However, this requires that $\|\boldsymbol{x}_n\| \to \|\boldsymbol{y}\|$ and so $\|\boldsymbol{y}\| \leq r$. Now let $\boldsymbol{y}$ be in the right side. It suffices to consider $\|\boldsymbol{y}-\boldsymbol{x}\| = 1$. Then you could consider for $t \in (0,1)$, $\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x}) = \boldsymbol{z}(t)$.

Then $\|\boldsymbol{z}(t)-\boldsymbol{x}\| = t\|\boldsymbol{y}-\boldsymbol{x}\| = tr < r$ and so $\boldsymbol{z}(t) \in B(\boldsymbol{x},r)$. But also, $\|\boldsymbol{z}(t)-\boldsymbol{y}\| = (1-t)\|\boldsymbol{y}-\boldsymbol{x}\| = (1-t)r$ so $\lim_{t\to 0}\|\boldsymbol{z}(t)-\boldsymbol{y}\| = 0$ showing that $\boldsymbol{y} \in \overline{B(\boldsymbol{x},r)}$. ■

Thus the usual way we think about the closure of a ball is completely correct in a normed linear space. Its limit points not in the ball are exactly $\boldsymbol{y}$ such that $\|\boldsymbol{y}-\boldsymbol{x}\| = r$. Recall that this lemma is not always true in the context of a metric space. Recall the discrete metric for example, in which the distance between different points is 1 and distance between a point and itself is 0. In what follows I will use the result of this lemma without comment. Balls will be either open, closed or neither. I am going to use the Hausdorff maximal theorem, Theorem 2.8.2 because it yields a very simple argument. It can be done other ways however. In the argument, the balls are not necessarily open nor closed. $\boldsymbol{y}$ is in $B(\boldsymbol{x},r)$ will mean that $\|\boldsymbol{y}-\boldsymbol{x}\| < r$ or $\|\boldsymbol{y}-\boldsymbol{x}\| = r$.

Lemma 4.5.2 *Let $\mathscr{F}$ be a nonempty collection of balls satisfying*

$$\infty > M \equiv \sup\{r : B(\boldsymbol{p},r) \in \mathscr{F}\} > 0$$

and let $k \in (0,M)$. Then there exists $\mathscr{G} \subseteq \mathscr{F}$ such that

$$\text{If } B(\boldsymbol{p},r) \in \mathscr{G}, \text{then } r > k, \tag{4.23}$$

$$\text{If } B_1, B_2 \in \mathscr{G} \text{ then } \overline{B_1} \cap \overline{B_2} = \emptyset, \tag{4.24}$$

$$\mathscr{G} \text{ is maximal with respect to 4.23 and 4.24.} \tag{4.25}$$

By this is meant that if $\mathscr{H}$ is a collection of balls satisfying 4.23 and 4.24, then $\mathscr{H}$ cannot properly contain $\mathscr{G}$.

Proof: Let $\mathfrak{S}$ denote a subset of $\mathscr{F}$ such that 4.23 and 4.24 are satisfied. Since $k < M$, 4.23 is satisfied for some ball of $\mathfrak{S}$. Thus $\mathfrak{S} \neq \emptyset$. Partially order $\mathfrak{S}$ with respect to set inclusion. Thus $\mathscr{A} \prec \mathscr{B}$ for $\mathscr{A}, \mathscr{B}$ in $\mathfrak{S}$ means that $\mathscr{A} \subseteq \mathscr{B}$. By the Hausdorff maximal theorem, there is a maximal chain in $\mathfrak{S}$ denoted by $\mathscr{C}$. Then let $\mathscr{G}$ be $\cup\mathscr{C}$. If B_1, B_2 are in $\mathscr{C}$, then since $\mathscr{C}$ is a chain, both B_1, B_2 are in some element of $\mathscr{C}$ and so $\overline{B_1} \cap \overline{B_2} = \emptyset$. The maximality of $\mathscr{C}$ is violated if there is any other element of $\mathfrak{S}$ which properly contains $\mathscr{G}$. ■

Proposition 4.5.3 *Let $\mathscr{F}$ be a collection of balls, and let*

$$A \equiv \cup\{B : B \in \mathscr{F}\}.$$

Suppose $\infty > M \equiv \sup\{r : B(\boldsymbol{p},r) \in \mathscr{F}\} > 0$. Then there exists $\mathscr{G} \subseteq \mathscr{F}$ such that $\mathscr{G}$ consists of balls whose closures are disjoint and $A \subseteq \cup\{\widehat{B} : B \in \mathscr{G}\}$ where for $B = B(\boldsymbol{x},r)$ a ball, $\widehat{B}$ denotes the open ball $B(\boldsymbol{x},5r)$.

Proof: Let $\mathscr{G}_1$ satisfy 4.23 - 4.25 for $k = \frac{2M}{3}$.

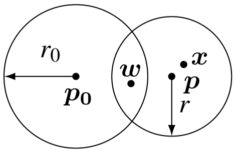
Suppose $\mathscr{G}_1, \cdots, \mathscr{G}_{m-1}$ have been chosen for $m \geq 2$. Let $\overline{\mathscr{G}_i}$ denote the collection of closures of the balls of $\mathscr{G}_i$. Then let $\mathscr{F}_m$ be those balls of $\mathscr{F}$, such that if B is one of these balls, $\overline{B}$ has empty intersection with every closed ball of $\overline{\mathscr{G}_i}$ for each $i \leq m-1$. Then using Lemma 4.5.2, let $\mathscr{G}_m$ be a maximal collection of balls from $\mathscr{F}_m$ with the property that each ball has radius larger than $\left(\frac{2}{3}\right)^m M$ and their closures are disjoint. Let $\mathscr{G} \equiv \cup_{k=1}^{\infty}\mathscr{G}_k$. Thus the closures of balls in $\mathscr{G}$ are disjoint. Let $\boldsymbol{x} \in B(\boldsymbol{p},r) \in \mathscr{F} \setminus \mathscr{G}$. Choose m such that

$$\left(\frac{2}{3}\right)^m M < r \leq \left(\frac{2}{3}\right)^{m-1} M$$

Then $\overline{B(\boldsymbol{p},r)}$ must have nonempty intersection with the closure of some ball from $\mathscr{G}_1 \cup \cdots \cup \mathscr{G}_m$ because if it didn't, then $\mathscr{G}_m$ would fail to be maximal. Denote by $B(\boldsymbol{p}_0, r_0)$ a ball in $\mathscr{G}_1 \cup \cdots \cup \mathscr{G}_m$ whose closure has nonempty intersection with $\overline{B(\boldsymbol{p},r)}$. Thus both

$$r_0, r > \left(\frac{2}{3}\right)^m M, \text{ so } r \leq \left(\frac{2}{3}\right)^{m-1} M < \frac{3}{2} r_0$$

Consider the picture, in which $\boldsymbol{w} \in \overline{B(\boldsymbol{p}_0, r_0)} \cap \overline{B(\boldsymbol{p},r)}$.



Then for $\boldsymbol{x} \in \overline{B(\boldsymbol{p},r)}$,

$$\|\boldsymbol{x}-\boldsymbol{p}_0\| \leq \|\boldsymbol{x}-\boldsymbol{p}\| + \|\boldsymbol{p}-\boldsymbol{w}\| + \overbrace{\|\boldsymbol{w}-\boldsymbol{p}_0\|}^{\leq r_0}$$

$$\leq r+r+r_0 \leq 2\overbrace{\left(\frac{2}{3}\right)^{m-1} M}^{<\frac{3}{2}r_0} + r_0 \leq 2\left(\frac{3}{2}r_0\right) + r_0 \leq 4r_0$$

Thus $B(\boldsymbol{p},r)$ is contained in $\overline{B(\boldsymbol{p}_0, 4r_0)}$. It follows that the closures of the balls of $\mathscr{G}$ are disjoint and the set $\left\{\hat{B} : B \in \mathscr{G}\right\}$ covers A. ■

Note that this theorem does not depend on the underlying space being finite dimensional. However, it is typically used in this setting.

Next is a version of the Vitali covering theorem which involves covering with disjoint closed balls. Here is the concept of a Vitali covering.

Definition 4.5.4 *Let S be a set and let $\mathscr{C}$ be a covering of S meaning that every point of S is contained in a set of $\mathscr{C}$. This covering is said to be a Vitali covering if for each $\varepsilon > 0$ and $\boldsymbol{x} \in S$, there exists a set $B \in \mathscr{C}$ containing $\boldsymbol{x}$, the diameter of B is less than ε, and there exists an upper bound to the set of diameters of sets of $\mathscr{C}$.*

The following corollary is a consequence of the above Vitali covering theorem.

Corollary 4.5.5 *Let F be a bounded set and let $\mathscr{C}$ be a Vitali covering of F consisting of closed balls. Let $r(B)$ denote the radius of one of these balls. Then assume also that $\sup\{r(B) : B \in \mathscr{C}\} = M < \infty$. Then there is a countable subset of $\mathscr{C}$ denoted by $\{B_i\}$ such that $\bar{m}_p\left(F \setminus \cup_{i=1}^N B_i\right) = 0$ for $N \leq \infty$, and $B_i \cap B_j = \emptyset$ whenever $i \neq j$.*

Proof: Let U be a bounded open set containing F such that U approximates F so well that

$$m_p(U) \leq r\bar{m}_p(F), r > 1 \text{ and very close to } 1,\ r - 5^{-p} \equiv \hat{\theta}_p < 1$$

Since this is a Vitali covering, for each $\boldsymbol{x} \in F$, there is one of these balls B containing $\boldsymbol{x}$ such that $\hat{B} \subseteq U$. Let $\widehat{\mathscr{C}}$ denote those balls of $\mathscr{C}$ such that $\hat{B} \subseteq U$ also. Thus, this is also a cover of F. By the Vitali covering theorem above, there are disjoint balls from $\mathscr{C}$, $\{B_i\}$ such that $\{\hat{B}_i\}$ covers F. Thus

$$\bar{m}_p\left(F\setminus\cup_{j=1}^{\infty}B_j\right)\leq m_p\left(U\setminus\cup_{j=1}^{\infty}B_j\right)=m_p(U)-\sum_{j=1}^{\infty}m_p(B_j)$$

$$\begin{aligned}&\leq && r\bar{m}_p(F)-5^{-p}\sum_{j=1}^{\infty}m_p\left(\widehat{B}_j\right)\leq r\bar{m}_p(F)-5^{-p}\bar{m}_p(F)\\ &\equiv && \left(r-5^{-p}\right)\bar{m}_p(F)\equiv\hat{\theta}_p\bar{m}_p(F)\end{aligned}$$

Now if n_1 is large enough and θ_p is chosen such that $1>\theta_p>\hat{\theta}_p$, then

$$\bar{m}_p\left(F\setminus\cup_{j=1}^{n_1}B_j\right)\leq m_p\left(U\setminus\cup_{j=1}^{n_1}B_j\right)\leq\theta_p\bar{m}_p(F).$$

If $\bar{m}\left(F\setminus\cup_{j=1}^{n_1}B_j\right)=0$, stop. Otherwise, do for $F\setminus\cup_{j=1}^{n_1}B_j$ exactly the same thing that was done for F. Since $\cup_{j=1}^{n_1}B_j$ is closed, you can arrange to have the approximating open set be contained in the open set $\left(\cup_{j=1}^{n_1}B_j\right)^C$. It follows there exist disjoint closed balls from $\mathscr{C}$ called $B_{n_1+1},\cdots,B_{n_2}$ such that

$$\bar{m}\left(\left(F\setminus\cup_{j=1}^{n_1}B_j\right)\setminus\cup_{j=n_{1+1}}^{n_2}B_j\right)<\theta_p\bar{m}\left(F\setminus\cup_{j=1}^{n_1}B_j\right)<\theta_p^2\bar{m}(F)$$

continuing this way and noting that $\lim_{n\to\infty}\theta_p^n=0$ while $\bar{m}(F)<\infty$, this shows the desired result. Either the process stops because $\bar{m}\left(F\setminus\cup_{j=1}^{n_k}B_j\right)=0$ or else you obtain $\bar{m}\left(F\setminus\cup_{j=1}^{\infty}B_j\right)=0$. ■

The conclusion holds for arbitrary balls, open or closed or neither. This follows from observing that the measure of the boundary of a ball is 0. Indeed, let

$$S(\boldsymbol{x},r)\equiv\{\boldsymbol{y}:|\boldsymbol{y}-\boldsymbol{x}|=r\}.$$

Then for each $\varepsilon<r$,

$$\begin{aligned}m_p(S(\boldsymbol{x},r)) &\subseteq m_p(B(\boldsymbol{x},r+\varepsilon))-m_p(B(\boldsymbol{x},r-\varepsilon))\\ &= m_p(B(\mathbf{0},r+\varepsilon))-m_p(B(\mathbf{0},r-\varepsilon))\\ &= \left(\left(\frac{r+\varepsilon}{r}\right)^p-\left(\frac{r-\varepsilon}{r}\right)^p\right)(m_p(B(\mathbf{0},r)))\end{aligned}$$

Hence $m_p(S(\boldsymbol{x},r))=0$.

Thus you can simply omit the boundaries or part of the boundary of the closed balls and there is no change in the conclusion. Just first apply the above corollary to the Vitali cover consisting of closures of the balls before omitting part or all of the boundaries. The following theorem is also obtained. You don't need to assume the set is bounded.

Theorem 4.5.6 *Let E be a bounded set and let $\mathscr{C}$ be a Vitali covering of E consisting of balls, open, closed, or neither. Let $r(B)$ denote the radius of one of these balls. Then assume also that $\sup\{r(B):B\in\mathscr{C}\}=M<\infty$. Then there is a countable subset of $\mathscr{C}$ denoted by $\{B_i\}$ such that $\bar{m}_p\left(E\setminus\cup_{i=1}^{N}B_i\right)=0,N\leq\infty$, and $B_i\cap B_j=\emptyset$ whenever $i\neq j$. Here $\bar{m}_p$ denotes the outer measure determined by m_p. The same conclusion follows if you omit the assumption that E is bounded.*

Proof: It remains to consider the last claim. Consider the balls

$$B(\mathbf{0},1), B(\mathbf{0},2), B(\mathbf{0},3), \cdots.$$

If E is some set, let E_r denote that part of E which is between $B(\mathbf{0},r-1)$ and $B(\mathbf{0},r)$ but not on the boundary of either of these balls, where $B(\mathbf{0},-1) \equiv \emptyset$. Then $\cup_{r=0}^{\infty} E_r$ differs from E by a set of measure zero and so you can apply the first part of the theorem to each E_r keeping all balls between $B(\mathbf{0},r-1)$ and $B(\mathbf{0},r)$ allowing for no intersection with any of the boundaries. Then the union of the disjoint balls associated with E_r gives the desired cover. ■

4.6 Exercises

1. Let V be a vector space with basis $\{v_1, \cdots, v_n\}$. For $v \in V$, denote its coordinate vector as $\boldsymbol{v} = (\alpha_1, \cdots, \alpha_n)$ where $v = \sum_{k=1}^{n} \alpha_k v_k$. Now define

$$\|v\| \equiv \max\{|\alpha_k| : k = 1, ..., n\}.$$

Show that this is a norm on V.

2. Let $(X, \|\cdot\|)$ be a normed linear space. You can let it be $(\mathbb{R}^n, |\cdot|)$ if you like. Recall $|\boldsymbol{x}|$ is the usual magnitude of a vector given by $|\boldsymbol{x}| = \sqrt{\sum_{k=1}^{n} |x_k|^2}$. 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 with any norm.

3. This problem is for those who have had a course in Linear algebra. A vector $\boldsymbol{v}$ is in the convex hull of S if there are finitely many vectors of $S, \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ and nonnegative scalars $\{t_1, \cdots, t_m\}$ 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 $\{t_k'\}$ summing to 1 such that $\boldsymbol{v} = \sum_{k=1}^{m-1} t_k' \boldsymbol{v}_k$.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 [40]. Convexity is more a geometric property than a topological property. **Hint:** Consider $L : \mathbb{R}^m \to V \times \mathbb{R}$ defined by $L(\boldsymbol{a}) \equiv (\sum_{k=1}^{m} a_k \boldsymbol{v}_k, \sum_{k=1}^{m} a_k)$ Explain why $\ker(L) \neq \{\mathbf{0}\}$. 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 $\boldsymbol{a} \in \ker(L) \setminus \{\mathbf{0}\}$ and $\lambda \in \mathbb{R}$, note that $\lambda\, \boldsymbol{a} \in \ker(L)$. Thus for all $\lambda \in \mathbb{R}$, $\boldsymbol{v} = \sum_{k=1}^{m} (t_k + \lambda a_k) \boldsymbol{v}_k$. Now vary λ 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. You might try to show this if you feel like it.

4. Show that the usual norm in $\mathbb{F}^n$ given by $|\boldsymbol{x}| = (\boldsymbol{x}, \boldsymbol{x})^{1/2}$ 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 $\mathbb{F}^n$.

5. Suppose K is a compact subset 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 $B(x,\delta_x)$ 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.

6. Suppose $\mathscr{C}$ is a set of compact sets 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. $K \subseteq \cup_{i=1}^m K_i^C = \left(\cap_{i=1}^m K_i\right)^C$ Therefore, you would need to have $\left\{K_1^C,\cdots,K_m^C\right\}$ is a cover of K. In other words, Now what does this say about the intersection of K with these K_i?

7. If (X,d) is a compact metric space and $f : X \to Y$ is continuous where (Y,ρ) is another metric space, show that if f is continuous on X, then it is uniformly continuous. Recall that this means that if $\varepsilon > 0$ is given, then there exists $\delta > 0$ such that if $d(x,\hat{x}) < \delta$, then $\rho(f(x),f(\hat{x})) < \varepsilon$. Compare with the definition of continuity. **Hint:** If this is not so, then there exists $\varepsilon > 0$ and $x_n,\hat{x}_n$ such that $d(x_n,\hat{x}_n) < 1/n$ but $\rho(f(x_n),f(\hat{x}_n)) \geq \varepsilon$. Now use compactness to get a contradiction.

8. Prove the above problem using another approach. Use the existence of the Lebesgue number in Problem 5 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 δ 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

$$\rho(f(x),f(\bar{x})) < \frac{\varepsilon}{2}, \rho(f(\hat{x}),f(\bar{x})) < \frac{\varepsilon}{2}.$$

Now consider the triangle inequality.

9. Let X be a vector space. A Hamel basis is a subset of X,Λ such that every vector of X can be written as a finite linear combination of vectors of Λ and the vectors of Λ are linearly independent in the sense that if $\{x_1,\cdots,x_n\} \subseteq \Lambda$ and $\sum_{k=1}^n c_k x_k = 0$ then each $c_k = 0$. Using the Hausdorff maximal theorem, show that every non-zero vector space has a Hamel basis. **Hint:** Let $x_1 \neq 0$. Let $\mathscr{F}$ denote the collection of subsets of X, Λ containing x_1 with the property that the vectors of Λ are linearly independent. Partially order $\mathscr{F}$ by set inclusion and consider the union of a maximal chain.

10. Suppose X is a nonzero real or complex normed linear space and let
$$V = \operatorname{span}(w_1, ..., w_m)$$
where $\{w_1, ..., w_m\}$ is a linearly independent set of vectors of X. Show that V is a closed subspace of X with $V \subsetneq X$. First explain why Theorem 4.2.11 implies any finite dimensional subspace of X can be written this way. **Hint:** You might want to use something like Lemma 4.4.7 to show this.

11. Suppose X is a normed linear space and its dimension is either infinite or greater than m where $V \equiv \operatorname{span}(w_1, ..., w_m)$ for $\{w_1, ..., w_m\}$ an independent set of vectors of X. Show $X \setminus V$ is a dense open subset of X which is equivalent to V containing no ball $B(v,r), \{w : \|w - v\| < r\}$. **Hint:** If $B(x,r)$ is contained in V, then show, that since V is a subspace, $B(0,r)$ is contained in V. Then show this implies $X \subseteq V$ which is not the case.

12. Show that if (X,d) is a metric space and H, K are disjoint closed sets, there are open sets U_H, U_K such that $H \subseteq U_H, K \subseteq U_K$ and $U_H \cap U_K = \emptyset$. **Hint:** Let $k \in K$. Explain why $\operatorname{dist}(k,H) \equiv \inf\{\|k - h\| : h \in H\} \equiv 2\delta_k > 0$. Now consider $U_K \equiv \cup_{k \in K} B(k, \delta_k)$. Do something similar for $h \in H$ and consider $U_H \equiv \cup_{k \in H} B(h, \delta_h)$.

13. If, in a metric space, $B(p, \delta)$ is a ball, show that
$$\overline{B(p,\delta)} \subseteq D(p,\delta) \equiv \{x : \|x - p\| \leq \delta\}$$
Now suppose (X,d) is a complete metric space and $U_n, n \in \mathbb{N}$ is a dense open set in X. Also let W be any nonempty open set. Show there exists a ball $B_1 \equiv B(p_1, r_1)$ having radius smaller than 2^{-1} such that $\overline{B_1} \subseteq U_1 \cap W_1$. Next show there exists $B_2 \equiv B(p_2, r_2)$ such that $\overline{B_2} \subseteq B_1 \cap U_2 \cap W$ with the radius of B_2 less than 2^{-2}. Continue this way. Explain why $\{p_n\}_{n=1}^{\infty}$ is a Cauchy sequence converging to some $p \in W \cap (\cup_{n=1}^{\infty} U_n)$. This is the very important Baire theorem which says that in a complete metric space, the intersection of dense open sets is dense.

14. Suppose you have a complete normed linear space, $(X, \|\cdot\|)$. Use the above problems leading to the Baire theorem in 13 to show that if $\mathscr{B}$ is a Hamel basis for for X, then $\mathscr{B}$ cannot be countable. **Hint:** If $\mathscr{B} = \{v_i\}_{i=1}^{\infty}$, consider $V_n \equiv \operatorname{span}(v_1, ..., v_n)$. Then use a problem listed above to argue that V_n^C is a dense open set. Now apply Problem 13. This shows why the idea of a Hamel basis often fails to be very useful whereas, in finite dimensional settings, it is just what is needed.

15. In any complete normed linear space which is infinite dimensional, show the unit ball is not compact. Do this by showing the existence of a sequence which cannot have a convergent subsequence. **Hint:** Pick $\|\boldsymbol{x}_1\| = 1$. Suppose $x_1, ..., x_n$ have been chosen, each $\|x_k\| = 1$. Then there is $x \notin \operatorname{span}(x_1, ..., x_n) \equiv V_n$. Now consider v such that $\|x - v\| \leq \frac{3}{2} \operatorname{dist}(x, V_n)$. Then argue that for $k \leq n$,
$$\left\| \frac{x - v}{\|x - v\|} - x_k \right\| = \left\| \frac{x - \left(\overbrace{v + \|x - v\| x_k}^{\in V_n} \right)}{\|x - v\|} \right\| \geq \frac{\operatorname{dist}(x, V_n)}{(3/2) \operatorname{dist}(x, V_n)} = \frac{2}{3}$$

16. Let X be a complete inner product space. Let $\mathscr{F}$ denote subsets $\beta \subseteq X$ such that whenever $x, y \in X, (x,y) = 0$ if $x \neq y$ and $(x,x) = 1$ if $x = y$. Thus these β are orthonormal sets. Show there exists a maximal orthonormal set. If X is separable, show that this maximal orthonormal set is countable. **Hint:** Use the Hausdorff maximal theorem. The next few problems involve linear algebra.

17. Let X be a real inner product space and let $\{v_1, ..., v_n\}$ be vectors in X. Let G be the $n \times n$ matrix $G_{ij} \equiv (v_i, v_j)$. Show that G^{-1} exists if and only if $\{v_1, ..., v_n\}$ is linearly independent. G is called the Grammian or the metric tensor.

18. ↑Let X be as above, a real inner product space, and let $V \equiv \text{span}(v_1, ..., v_n)$. Let $u \in X$ and $z \in V$. Show that $|u - z| = \inf\{|u - v| : v \in V\}$ if and only if $(u - z, v_i) = 0$ for all v_i. Note that the v_i might not be linearly independent. Also show that $|u - z|^2 = |u|^2 - (z, u)$.

19. ↑ Let G be the matrix of Problem 17 where $\{v_1, ..., v_n\}$ is linearly independent and $V \equiv \text{span}(v_1, ..., v_n) \subseteq X$, an inner product space. Let $x \equiv \sum_i x^i v_i, y \equiv \sum_i y^i v_i$ be two vectors of V. Show that $(x, y) = \sum_{i,j} x^i G_{ij} x^j$. Show that $z \equiv \sum_i z^i v_i, z$ is closest to $u \in X$ if and only if for all $i = 1, ..., n, (u, v_i) = \sum_j G_{ij} z^j$. This gives a system of linear equations which must be satisfied by the z^i in order that z just given is the best approximation to u. Next show that there exists such a solution thanks to Problem 17 which says that the matrix G is invertible, and if G^{-1} has ij^{th} component G^{ij}, one finds that $\sum_j G^{ij}(u, v_j) = z^i$.

20. ↑ In the situation of the above problems, suppose A is an $m \times n$ matrix. Use Problem 18 to show that for $\boldsymbol{y} \in \mathbb{R}^m$, there always exists a solution $\boldsymbol{x}$ to the system of equations $A^T \boldsymbol{y} = A^T A \boldsymbol{x}$. Explain how this is in a sense the best you can do to solve $\boldsymbol{y} = A\boldsymbol{x}$ even though this last system of equations might not have a solution. Here A^T is the transpose of the matrix A. The equations $A^T \boldsymbol{y} = A^T A \boldsymbol{x}$ are called the normal equations for the least squares problem. **Hint:** Verify that $\left(A^T \boldsymbol{y}, \boldsymbol{x}\right) = (\boldsymbol{y}, A\boldsymbol{x})$. Let the subspace V be $A(\mathbb{R}^n)$, the vectors spanning it being $\{A\boldsymbol{e}_1, ..., A\boldsymbol{e}_n\}$. From the above problem, there exists $A\boldsymbol{x}$ in V which is closest to $\boldsymbol{y}$. Now use the characterization of this vector $(\boldsymbol{y} - A\boldsymbol{x}, A\boldsymbol{z}) = 0$ for all $\boldsymbol{z} \in \mathbb{R}^n, A\boldsymbol{z}$ being a generic vector in $A(\mathbb{R}^n)$.

21. ↑As an example of an inner product space, consider $C([0,1])$ with the inner product $\int_0^1 f(x) g(x) dx$ where this is the ordinary integral from calculus. Abusing notation, let $\{x^{p_1}, ..., x^{p_n}\}$ with $-\frac{1}{2} < p_1 < \cdots < p_n$ be functions, (vectors) in $C([0,1])$. Verify that these vectors are linearly independent. **Hint:** You might want to use the Cauchy identity, Theorem 1.9.28.

22. ↑As above, if $\{v_1, ..., v_n\}$ is linearly independent, the Grammian is $G = G(v_1, ..., v_n)$, $G_{ij} \equiv (v_i, v_j)$, then if $u \notin \text{span}(v_1, ..., v_n) \equiv V$ you could consider $G(v_1, ..., v_n, u)$. Then if $d \equiv \min\{|u - v| : v \in \text{span}(v_1, ..., v_n)\}$, show that $d^2 = \frac{\det G(v_1, ..., v_n, u)}{\det G(v_1, ..., v_n)}$. Justify the following steps. Letting z be the closest point of V to u, from the above, $\left(u - \sum_{i=1}^n z^i v_i, v_p\right) = 0$ for each v_p and so

$$(u, v_p) = \sum_{i=1}^n (v_p, v_i) z^i \qquad (*)$$

Also, since $(u-z,v)=0$ for all $v\in V$, $|u|^2=|u-z+z|^2=|u-z|^2+|z|^2$ so

$$|u|^2=\left|u-\sum_{i=1}^{n}z^iv_i\right|^2+\left|\sum_{i=1}^{n}z^iv_i\right|^2=d^2+\left|\sum_{i=1}^{n}z^iv_i\right|^2$$

$$=d^2+\sum_j\overbrace{\sum_i(v_j,v_i)z^i}^{=(u,v_j)}z^j=d^2+\sum_j(u,v_j)z^j$$

$$=d^2+\boldsymbol{y}^T\boldsymbol{z},\ \boldsymbol{y}\equiv((u,v_1),\cdots,(u,v_n))^T,\ \boldsymbol{z}\equiv\left(z^1,\cdots,z^n\right)^T$$

From $*$, $G\boldsymbol{z}=\boldsymbol{y}$, $\begin{pmatrix}G(v_1,...,v_n) & \boldsymbol{0}\\ \boldsymbol{y}^T & 1\end{pmatrix}\begin{pmatrix}\boldsymbol{z}\\ d^2\end{pmatrix}=\begin{pmatrix}\boldsymbol{y}\\ \|u\|^2\end{pmatrix}$. Now use Cramer's rule to solve for d^2 and get

$$d^2=\frac{\det\begin{pmatrix}G(v_1,...,v_n) & \boldsymbol{y}\\ \boldsymbol{y}^T & |u|^2\end{pmatrix}}{\det(G(v_1,...,v_n))}\equiv\frac{\det G(v_1,...,v_n,u)}{\det G(v_1,...,v_n)}$$

23. In the situation of Problem 21, let $f_k(x)\equiv x^k$ and let $V\equiv\operatorname{span}(f_{p_1},...,f_{p_n})$. give an estimate for the distance d between f_m and V for m a nonnegative integer and as in the above problem $-\frac{1}{2}<p_1<\cdots<p_n$. Use Theorem 1.9.28 in the appendix and the above problem with $v_i\equiv f_{p_i}$ and $v_{n+1}\equiv f_m$. Justify the following manipulations. The numerator in the above formula for the distance is of the form $\frac{\prod_{j<i\leq n+1}(p_i-p_j)^2}{\prod_{i,j\leq n+1}(p_i+p_j+1)}$

$$=\frac{\prod_{j<i\leq n}(p_i-p_j)^2\prod_{j\leq n}(m-p_j)^2}{\prod_{i,j\leq n}(p_i+p_j+1)\prod_{i=1}^{n}(p_i+m+1)\prod_{j=1}^{n}(p_j+m+1)(2m+1)}$$

While $G(f_{p_1},...,f_{p_n})=\frac{\prod_{j<i\leq n}(p_i-p_j)^2}{\prod_{i,j\leq n}(p_i+p_j+1)}$. Thus $d=\frac{\prod_{j\leq n}|m-p_j|}{\prod_{i=1}^{n}(p_i+m+1)\left(\sqrt{2m+1}\right)}$.

24. Suppose $\sum_{k=0}^{n}a_kt^k=0$ for each $t\in(-\delta,\delta)$ where $a_k\in X$, a linear space. Show that each $a_k=0$.

25. Suppose $A\subseteq\mathbb{R}^p$ is covered by a finite collection of Balls $\mathscr{F}$. Show that then there exists a disjoint collection of these balls, $\{B_i\}_{i=1}^m$, such that $A\subseteq\cup_{i=1}^m\widehat{B_i}$ where $\widehat{B_i}$ has the same center as B_i but 3 times the radius. **Hint:** Since the collection of balls is finite, they can be arranged in order of decreasing radius. Mimic the argument for Vitali covering theorem.

Chapter 5

Functions on Normed Linear Spaces

This chapter is about the general notion of functions defined on normed linear spaces even if the linear space is not finite dimensional.

5.1 $\mathscr{L}(V,W)$ as a Vector Space

In what follows, V,W will be vector spaces.

Definition 5.1.1 *The term $\mathscr{L}(V,W)$ signifies the set of linear maps from V to W. This means that for $v,u \in V$ and α,β scalars from $\mathbb{F}$,$L(\alpha u+\beta v)=\alpha L(u)+\beta L(v)$. 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] *$(L+M)v \equiv Lv+Mv$. For α a scalar and $L \in \mathscr{L}(V,W)$, define $\alpha L \in \mathscr{L}(V,W)$ by $\alpha L(v) \equiv \alpha(Lv)$.*

Note that if you have $V=\mathbb{R}^n$ and $W=\mathbb{R}^m$, an example of something in $\mathscr{L}(V,W)$ is given by $T\boldsymbol{v} \equiv A\boldsymbol{v}$ where A is a real $m \times n$ matrix.

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. What about the dimension of $\mathscr{L}(V,W)$?

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.

Lemma 5.1.2 *Let V and W be vector spaces and suppose $\{v_1,\cdots,v_n\}$ is a basis for V. Then if $L:V \to W$ is given by $Lv_k=w_k \in W$ and $L(\sum_{k=1}^n a_k v_k) \equiv \sum_{k=1}^n a_k L v_k = \sum_{k=1}^n a_k w_k$ then L is well defined and is in $\mathscr{L}(V,W)$. Also, if L,M are two linear transformations such that $Lv_k=Mv_k$ for all k, then $M=L$.*

Proof: L is well defined on V because, since $\{v_1,\cdots,v_n\}$ 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 $V, L(\sum_{k=1}^n a_k v_k)=\sum_{k=1}^n a_k L v_k=\sum_{k=1}^n a_k M v_k = M(\sum_{k=1}^n a_k v_k)$ 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.

Theorem 5.1.3 *Let V and W be finite dimensional linear spaces of dimension n and m respectively Then* $\dim(\mathscr{L}(V,W))=mn$.

Proof: Let two sets of bases be $\{v_1,\cdots,v_n\}$ and $\{w_1,\cdots,w_m\}$ 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, $\{v_1,\cdots,v_n\}$, by $w_i v_k(v_j) \equiv w_i \delta_{jk}$ where $\delta_{ik}=1$ 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.

Then let $L \in \mathscr{L}(V,W)$. Since $\{w_1,\cdots,w_m\}$ is a basis, there exist constants, d_{jk} such that $Lv_r=\sum_{j=1}^m d_{jr} w_j$ Now consider the following sum of dyadics. $\sum_{j=1}^m \sum_{i=1}^n d_{ji} w_j v_i$. Apply this to v_r. This yields $\sum_{j=1}^m \sum_{i=1}^n d_{ji} w_j v_i(v_r)=\sum_{j=1}^m \sum_{i=1}^n d_{ji} w_j \delta_{ir}=\sum_{j=1}^m d_{jr} w_i = Lv_r$. Therefore, $L=\sum_{j=1}^m \sum_{i=1}^n d_{ji} w_j v_i$ showing the span of the dyadics is all of $\mathscr{L}(V,W)$.

[1] Note that this is the standard way of defining the sum of two functions.

Now consider whether these dyadics form a linearly independent set. Suppose that $\sum_{i,k} d_{ik} w_i v_k = \mathbf{0}$. Are all the scalars d_{ik} equal to 0? $\mathbf{0} = \sum_{i,k} d_{ik} w_i v_k (v_l) = \sum_{i=1}^m d_{il} w_i$ so, since $\{w_1, \cdots, w_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows that the dimension of $\mathscr{L}(V,W)$ is mn as claimed because there are m choices for the w_i and n choices for the v_j. ■

5.2 The Norm of a Linear Map, Operator Norm

Not surprisingly all of the above holds for a finite dimensional normed linear space. First here is an easy lemma which follows right away from Theorem 3.6.2, the theorem about equivalent formulations of continuity.

Lemma 5.2.1 *Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two normed linear spaces. Then a linear map $f : V \to W$ is continuous if and only if it takes bounded sets to bounded sets. (f is bounded) If V is finite dimensional, then f must be continuous.*

Proof: $\Longrightarrow$ Consider $f(B(0,1))$. If this is not bounded, then there exists $\|v^m\|_V \leq 1$ but $\|f(v^m)\|_W \geq m$. Then it follows that $\left\|f\left(\frac{v^m}{m}\right)\right\|_W \geq 1$ which is impossible for all m since $\left\|\frac{v^m}{m}\right\| \leq \frac{1}{m}$ and so continuity requires that $\lim_{m\to\infty} f\left(\frac{v^m}{m}\right) = 0$ (Theorem 3.6.2). Thus there exists M such that $\|f(v)\| \leq M$ whenever $v \in B(0,1)$. In general, let S be a bounded set. Then $S \subseteq B(0,r)$ for large enough r. Hence, for $v \in B$, it follows that $v/2r \in B(0,1)$. It follows that $\|f(v/2r)\|_W \leq M$ and so $\|f(v)\|_W \leq 2rM$. Thus f takes bounded sets to bounded sets.

$\Longleftarrow$ Suppose f is bounded and not continuous. Then by Theorem 3.6.2 again, there is a sequence $v_n \to v$ but $f(v_n)$ fails to converge to $f(v)$. Then there exists $\varepsilon > 0$ and a subsequence, still denoted as v_n such that $\|f(v_n) - f(v)\| = \|f(v_n - v)\| \geq \varepsilon$. Then

$$\left\|f\left(\frac{v_n - v}{\|v_n - v\|}\right)\right\| \geq \varepsilon \frac{1}{\|v_n - v\|}$$

The right side is unbounded, but the left is bounded, a contradiction.

Consider the last claim about continuity. Let $\{v_1, \cdots, v_n\}$ be a basis for V. By Lemma 4.4.7, if $y^m \to 0$, in V for $y^m = \sum_{k=1}^n y_k^m v_k$,then it follows that $\lim_{m\to\infty} y_k^m = 0$ and consequently, $f(y^m) \to f(0) = 0$. In general, if $y^m \to y$, then $(y^m - y) \to 0$ and so $f(y^m - y) = f(y^m) - f(y) \to 0$. That is, $f(y^m) \to f(y)$. ■

Definition 5.2.2 *For $f : (V, \|\cdot\|_V) \to (W, \|\cdot\|_W)$ continuous, it was just shown that there exists M such that $\|f(v)\| \leq M$, $v \in B(0,1)$. It follows that, since $\frac{v}{2\|v\|} \in B(0,1)$, then $\|f(v)\| \leq 2M\|v\|$. Therefore, letting $\|f\| \equiv \sup_{\|v\|\leq 1} \|f(v)\|$ it follows that for all $v \in V$, $\|f(v)\| \leq \|f\| \|v\|$ .Thus a linear map is bounded if and only if $\|f\| < \infty$ if and only if f is continuous. 'The number $\|f\|$ is called the operator norm. For X a real normed linear space, X' denotes the space $\mathscr{L}(X,\mathbb{R})$.*

You can show that for $\mathscr{L}(V,W)$ the space of bounded linear maps from V to W, $\mathscr{L}(V,W)$ becomes a normed linear space with this definition. This is true whether V,W are finite or infinite dimensional. You can also show that if W is complete then so is $\mathscr{L}(V,W)$. This is left as an exercise. Also, when the vector spaces are finite dimensional, Lemma 5.2.1 shows that any linear function f is automatically bounded, hence continuous, hence $\|f\|$ exists. Here is an interesting observation about the operator norm.

Lemma 5.2.3 *Let $f \in \mathscr{L}(V,W)$ and let $h \in \mathscr{L}(W,Z)$ where X,Y,Z are normed vector spaces. Then $\|h \circ f\| \leq \|h\| \|f\|$.*

Proof: This follows right away from the definition. If $\|v\| \leq 1$, then $\|f(v)\| \leq \|f\|$. This explains the first inequality in the following.

$$\sup_{\|v\| \leq 1} \|h \circ f(v)\| \leq \sup_{\|w\| \leq \|f\|} \|h(w)\| = \sup_{\|w\| \leq \|f\|} \left\| h\left(\frac{w}{\|f\|}\right)\right\| \|f\| \leq \|h\| \|f\|. \blacksquare$$

Theorem 5.2.4 *Let $(V, \|\cdot\|)$ be a normed linear space with basis $\{v_1, \cdots, v_n\}$ and field of scalars $\mathbb{F}$. Let $f : (\mathbb{F}^n, \|\cdot\|) \to (V, \|\cdot\|_V)$ be any linear map which is one to one and onto. Then both f and f^{-1} are continuous. Also the compact sets of $(V, \|\cdot\|_V)$ are exactly those which are closed and bounded.*

Proof: Define another norm $\|\cdot\|_1$ on $\mathbb{F}^n$ as follows. $\|\boldsymbol{x}\|_1 \equiv \|f(\boldsymbol{x})\|_V$. Since f is one to one and onto and linear, this is indeed a norm. The details are left as an exercise. Then from the theorem on the equivalence of norms, there are positive constants δ, Δ such that $\delta \|\boldsymbol{x}\| \leq \|f(\boldsymbol{x})\|_V \leq \Delta \|\boldsymbol{x}\|$. Since f is one to one and onto, this implies $\delta \left\|f^{-1}(\boldsymbol{v})\right\| \leq \|\boldsymbol{v}\|_V \leq \Delta \left\|f^{-1}(\boldsymbol{v})\right\|$. The first of these above inequalities implies f is continuous. The second says $\left\|f^{-1}(\boldsymbol{v})\right\| \leq \frac{1}{\delta} \|\boldsymbol{v}\|_V$ and so f^{-1} is continuous. Thus, from the above theorems, both f and f^{-1} map closed sets to closed sets, compact sets to compact sets, open sets to open sets and bounded sets to bounded sets.

Now let $K \subseteq V$ be closed and bounded. Then from the above observations, $f^{-1}(K)$ is also closed and bounded. Therefore, it is compact. Now $f\left(f^{-1}(K)\right) = K$ must be compact because the continuous image of a compact set is compact, Theorem 3.7.1. Conversely, if $K \subseteq V$ is compact, then by the theorem just mentioned, $f^{-1}(K)$ is compact and so it is closed and bounded. Hence $f\left(f^{-1}(K)\right) = K$ is also closed and bounded. ■

This is a remarkable theorem. It says that an algebraic isomorphism is also a homeomorphism which is what it means to say that the map takes open sets to open sets and the inverse does the same. In other words, there really isn't any algebraic or topological distinction between a finite dimensional normed vector space of dimension n and $\mathbb{F}^n$. Of course when one considers geometry, this is not so.

Here is another interesting theorem about coordinate maps. It follows right away from earlier theorems.

Theorem 5.2.5 *Let $f : (V, \|\cdot\|_V) \to (W, \|\cdot\|_W)$ be a continuous function where here $(V, \|\cdot\|_V)$ is a normed linear space and $(W, \|\cdot\|_W)$ is a finite dimensional normed linear space with basis $\{w_1, \cdots, w_n\}$. Thus $f(v) \equiv \sum_{k=1}^n f_k(v) w_k$. Then f is continuous if and only if each f_k is a continuous $\mathbb{F}$ valued map.*

Proof: $\Longrightarrow$ First, why is f_k linear? This follows from

$$\sum_{k=1}^n (\alpha f_k(u) + \beta f_k(v)) w_k = \alpha \sum_{k=1}^n f_k(u) w_k + \beta \sum_{k=1}^n f_k(v) w_k$$

$$= \alpha f(u) + \beta f(v) = f(\alpha u + \beta v) \equiv \sum_{k=1}^n f_k(\alpha u + \beta v) w_k$$

Why is the coordinate function f_k continuous? From Lemma 5.2.1, it suffices to verify that f_k is bounded. If this is not so, there exists $v_m, \|v_m\|_V \leq 1$ but $|f_k(v_m)|_W \geq m$. It follows that $\left|f_k\left(\frac{v_m}{m}\right)\right| \geq 1$. Since f is continuous, and $v_m/m \to 0$, it follows that $f\left(\frac{v_m}{m}\right) \to 0$ in V. However, by Lemma 4.4.7, $f_k\left(\frac{v_m}{m}\right) \to 0$, a contradiction.

$\Longleftarrow$ If each coordinate function is continuous, then

$$\|f(v)-f(\hat{v})\|_W = \left\|\sum_{k=1}^n f_k(v)\,w_k - \sum_{k=1}^n f_k(\hat{v})\,w_k\right\| \leq \sum_{k=1}^n |f_k(v)-f_k(\hat{v})|\,\|w_k\|_W$$

Since each f_k is continuous, this shows that f is also. ■

5.3 Continuous Functions in Normed Linear Space

Of course not all functions are linear. Continuous functions have already been discussed in general metric space, but now there are other considerations to consider due to the algebra available in a normed linear space. The following theorem includes these kinds of considerations for functions having values in a normed linear space.

Theorem 5.3.1 *Let f,g be continuous functions defined on D, a metric space. Also let α,β be scalars. Then the following hold.*

1. *$\alpha f+\beta g$ is continuous.*
2. *If $(W,\|\cdot\|_W)$ is an inner product space, then (f,g) defined as*

   $(f,g)(v) \equiv (f(v),g(v))$, then (f,g) is continuous.
3. *If f has values in $\mathbb{F}$ and g has values in $(W,\|\cdot\|_W)$, then fg is continuous.*

Proof: Say $v_n \to v$. Then

$$\|(\alpha f+\beta g)(v_n)-(\alpha f+\beta g)(v)\| \leq |\alpha|\,\|f(v_n)-f(v)\|+|\beta|\,\|g(v_n)-g(v)\|$$

and the right side converges to 0 as $n\to\infty$ so this shows 1.

This follows from an easy computation. From the Cauchy Schwarz inequality,

$$|(f,g)(v)-(f,g)(\hat{v})| \leq |(f(v),g(v))-(f(v),g(\hat{v}))|+|(f(v),g(\hat{v}))-(f(\hat{v}),g(\hat{v}))|$$

$$\leq \|g(v)-g(\hat{v})\|\,\|f(v)\|+\|f(v)-f(\hat{v})\|\,\|g(\hat{v})\|$$

Now since g is continuous at v and so $\|g(v)-g(\hat{v})\|<1$ provided $d(v,\hat{v})$ is small enough. Thus $\|g(\hat{v})\| \leq \|g(v)\|+1$. Hence if $d(v,\hat{v})$ is small enough,

$$|(f,g)(v)-(f,g)(\hat{v})| \leq (\|g(v)\|+1)\,\|f(v)-f(\hat{v})\|+\|f(v)\|\,\|g(v)-g(\hat{v})\|$$

Thus, by continuity of f,g at v, if $d(v,\hat{v})$ is sufficiently small, the right side is less than ε and so $f\cdot g$ is continuous at v. This shows 2. The proof of 3. is just like this. ■

Of course there are other things like cross product and determinant and so forth which are defined in terms of the component functions of f. Then these things will be continuous by an application of Theorem 5.2.5.

5.4 Polynomials

For functions of one variable, the special kind of functions known as a polynomial has a corresponding version when one considers a function of many variables. This is found in the next definition.

Definition 5.4.1 *Let α be an n dimensional multi-index. The meaning of this term is that $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ where each α_i is a positive integer or zero. Also, let $|\boldsymbol{\alpha}| \equiv \sum_{i=1}^{n} |\alpha_i|$. Then $\boldsymbol{x}^{\alpha}$means $\boldsymbol{x}^{\alpha} \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_3^{\alpha_n}$ where each $x_j \in \mathbb{F}$. An n dimensional polynomial of degree m is a function of the form $p(\boldsymbol{x}) = \sum_{|\alpha| \leq m} d_{\alpha} \boldsymbol{x}^{\alpha}$. where the d_{α} are complex or real numbers, more generally in some normed linear space X. Rational functions are defined as the quotient of two real or complex valued polynomials. Thus these functions are defined on $\mathbb{F}^n$.*

For example, $f(\boldsymbol{x}) = x_1 x_2^2 + 7x_3^4 x_1$ is a polynomial of degree 5 and $\frac{x_1 x_2^2 + 7x_3^4 x_1 + x_2^3}{4x_1^3 x_2^2 + 7x_3^2 x_1 - x_2^3}$ is a rational function.

Note that in the case of a rational function, the domain of the function might not be all of $\mathbb{F}^n$. For example, if $f(\boldsymbol{x}) = \frac{x_1 x_2^2 + 7x_3^4 x_1 + x_2^3}{x_2^2 + 3x_1^2 - 4}$,the domain of f would be all complex numbers such that $x_2^2 + 3x_1^2 \neq 4$.

By Theorem 3.6.2 all polynomials are continuous. To see this, note that the function, $\pi_k(\boldsymbol{x}) \equiv x_k$ is a continuous function because of the inequality

$$|\pi_k(\boldsymbol{x}) - \pi_k(\boldsymbol{y})| = |x_k - y_k| \leq |\boldsymbol{x} - \boldsymbol{y}|.$$

Polynomials are simple sums of scalars times products of these functions. Similarly, by this theorem, rational functions, quotients of polynomials, are continuous at points where the denominator is non zero. More generally, if V is a normed vector space, consider a V valued function of the form $\boldsymbol{f}(\boldsymbol{x}) \equiv \sum_{|\alpha| \leq m} \boldsymbol{d}_{\alpha} \boldsymbol{x}^{\alpha}$ where $\boldsymbol{d}_{\alpha} \in V$, sort of a V valued polynomial. Then such a function is continuous by application of Theorem 3.6.2 and the above observation about the continuity of the functions π_k.

Thus there are lots of examples of continuous functions. However, it is even better than the above discussion indicates. As in the case of a function of one variable, an arbitrary continuous function can typically be approximated uniformly by a polynomial. This is the n dimensional version of the Weierstrass approximation theorem.

5.5 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 5.5.1 *The following estimate holds for $x \in [0,1]$ and $m \geq 2$.*

$$\sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1 - x)^{m-k} \leq \frac{1}{4} m$$

Proof: First of all, from the binomial theorem,

$$\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)\left(e^{t(k-mx)}\right)x^{k}(1-x)^{m-k}=e^{-tmx}\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)\left(e^{tk}\right)x^{k}(1-x)^{m-k}$$

$$=e^{-tmx}\left(1-x+xe^{t}\right)^{m}=e^{-tmx}g(t)^{m},\ g(0)=1,g'(0)=g''(0)=x$$

Take a partial derivative with respect to t twice.

$$\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)(k-mx)^{2}e^{t(k-mx)}x^{k}(1-x)^{m-k}$$

$$\begin{aligned} = \quad & (mx)^{2}e^{-tmx}g(t)^{m}+2(-mx)e^{-tmx}mg(t)^{m-1}g'(t) \\ & +e^{-tmx}\left[m(m-1)g(t)^{m-2}g'(t)^{2}+mg(t)^{m-1}g''(t)\right] \end{aligned}$$

Now let $t=0$ and note that the right side is $m(x-x^2)\leq m/4$ for $x\in[0,1]$. Thus

$$\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)(k-mx)^{2}x^{k}(1-x)^{m-k}=mx-mx^{2}\leq m/4\ \blacksquare$$

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 4.3.7. Thus this is a function which is continuous with values in X as discussed earlier with metric spaces.

Theorem 5.5.2 *Let $f\in C([0,1];X)$ and let the norm on X be denoted by $\|\cdot\|$.*

$$p_{m}(x)\equiv\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}f\left(\frac{k}{m}\right)=\sum_{k=0}^{m}q_{k}(x)f\left(\frac{k}{m}\right)$$

Then these polynomials having coefficients in X converge uniformly to f on $[0,1]$. Also $q_0(0)=1,q_k(0)=0$ for $k\neq 0$, and $q_m(1)=1$ while $q_k(1)=0$ for $k\neq m$.

Proof: Let $\|f\|_{\infty}$ denote the largest value of $\|f(x)\|$. By uniform continuity of f, there exists a $\delta>0$ such that if $|x-x'|<\delta$, then $\|f(x)-f(x')\|<\varepsilon/2$. By the binomial theorem,

$$\|p_{m}(x)-f(x)\|\leq\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\|$$

$$\leq\sum_{\left|\frac{k}{m}-x\right|<\delta}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\|+2\|f\|_{\infty}\sum_{\left|\frac{k}{m}-x\right|\geq\delta}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}$$

Therefore,

$$\leq\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}\frac{\varepsilon}{2}+2\|f\|_{\infty}\sum_{(k-mx)^{2}\geq m^{2}\delta^{2}}\left(\begin{array}{c} m \\ k \end{array}\right)x^{k}(1-x)^{m-k}$$

$$\leq\frac{\varepsilon}{2}+2\|f\|_{\infty}\frac{1}{m^{2}\delta^{2}}\sum_{k=0}^{m}\left(\begin{array}{c} m \\ k \end{array}\right)(k-mx)^{2}x^{k}(1-x)^{m-k}\leq\frac{\varepsilon}{2}+2\|f\|_{\infty}\frac{1}{4}m\frac{1}{\delta^{2}m^{2}}<\varepsilon$$

provided m is large enough. Thus $\|p_m-f\|_{\infty}<\varepsilon$ when m is large enough. $\blacksquare$

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 5.5.3 *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 $[a,b]$. The m^{th} term of this sequence is $\sum_{k=0}^{m} q_k(y) f\left(l\left(\frac{k}{m}\right)\right)$ where $l : [0,1] \to [a,b]$ be one to one, linear and onto and $q_0(a) = 1$ and if $k \neq 0, q_k(a) = 0$ and $q_m(b) = 1$ and if $k \neq m$, then $q_k(b) = 0$.*

Proof: Let $l : [0,1] \to [a,b]$ be one to one, linear and onto. Then $f \circ l$ is continuous on $[0,1]$ and so if $\varepsilon > 0$ is given, if m large enough, then for all $x \in [0,1]$,

$$\left\| \sum_{k=0}^{m} \hat{q}_k(x) f\left(l\left(\frac{k}{m}\right)\right) - f \circ l(x) \right\| < \varepsilon$$

where $\hat{q}_0(0) = 1$ and $\hat{q}_k(0) = 0$ for $k \neq 0, \hat{q}_m(1) = 1, \hat{q}_k(1) = 0$ if $k \neq m$. Therefore, for all $y \in [a,b]$,

$$\left\| \sum_{k=0}^{m} \hat{q}_k\left(l^{-1}(y)\right) f\left(l\left(\frac{k}{m}\right)\right) - f(y) \right\| < \varepsilon$$

Let $q_k(y) \equiv \hat{q}_k\left(l^{-1}(y)\right)$. ■

As another corollary, here is the version which will be used in Stone's generalization later.

Corollary 5.5.4 *Let f be a continuous function defined on $[-M,M]$ with $f(0) = 0$. Then there is a sequence of polynomials $\{p_m\}$, $p_m(0) = 0$ and $\lim_{m\to\infty} \|p_m - f\|_\infty = 0$*

Proof: From Corollary 5.5.3 there exists a sequence of polynomials $\{\widehat{p_m}\}$ such that $\|\widehat{p_m} - f\|_\infty \to 0$. Simply consider $p_m = \widehat{p_m} - \widehat{p_m}(0)$. ■

5.6 Functions of Many Variables

First note that if $h : K \times H \to \mathbb{R}$ 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 $\boldsymbol{f} \in C(R_p;X)$ where $R_p = [0,1]^p$. Then let $\hat{\boldsymbol{x}}_p \equiv (x_1, ..., x_{p-1})$. By Theorem 5.5.2, if n is large enough,

$$\max_{x_p\in[0,1]} \left\| \sum_{k=0}^{n} \boldsymbol{f}\left(\cdot, \frac{k}{n}\right) \binom{n}{k} x_p^k (1-x_p)^{n-k} - \boldsymbol{f}(\cdot, x_p) \right\|_{C\left([0,1]^{p-1};X\right)} < \frac{\varepsilon}{2}$$

Now $\boldsymbol{f}\left(\cdot, \frac{k}{n}\right) \in C(R_{p-1};X)$ and so by induction, there is a polynomial $\boldsymbol{p}_k(\hat{\boldsymbol{x}}_p)$ such that

$$\max_{\hat{\boldsymbol{x}}_p\in R_{p-1}} \left\| \boldsymbol{p}_k(\hat{\boldsymbol{x}}_p) - \binom{n}{k} \boldsymbol{f}\left(\hat{\boldsymbol{x}}_p, \frac{k}{n}\right) \right\|_X < \frac{\varepsilon}{(n+1)2}$$

Thus, letting $\boldsymbol{p}(\boldsymbol{x}) \equiv \sum_{k=0}^{n} \boldsymbol{p}_k(\hat{\boldsymbol{x}}_p) x_p^k (1-x_p)^{n-k}$,

$$\|\boldsymbol{p} - \boldsymbol{f}\|_{C(R_p;X)} \leq \max_{x_p\in[0,1]} \max_{\hat{\boldsymbol{x}}_p\in R_{p-1}} \left\| \boldsymbol{p}(\hat{\boldsymbol{x}}_p, x_p) - \boldsymbol{f}(\hat{\boldsymbol{x}}_p, x_p) \right\|_X < \varepsilon$$

where $\boldsymbol{p}$ is a polynomial with coefficients in X.

In general, if $R_p \equiv \prod_{k=1}^{p}[a_k,b_k]$, note that there is a linear function $l_k : [0,1] \to [a_k,b_k]$ which is one to one and onto. Thus $\boldsymbol{l}(\boldsymbol{x}) \equiv (l_1(x_1),...,l_p(x_p))$ 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 $\|\boldsymbol{p}-\boldsymbol{f}\circ\boldsymbol{l}\|_{C([0,1]^p;X)} < \varepsilon$. Thus $\left\|\boldsymbol{p}\circ\boldsymbol{l}^{-1}-\boldsymbol{f}\right\|_{C(R_p;X)} < \varepsilon$ and $\boldsymbol{p}\circ\boldsymbol{l}^{-1}$ is a polynomial. This proves the following theorem.

Theorem 5.6.1 *Let $\boldsymbol{f}$ be a function in $C(R;X)$ for X a normed linear space where $R \equiv \prod_{k=1}^{p}[a_k,b_k]$. 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 $\lim_{n\to\infty} p_n'(x) \to f'(x)$ 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.

Lemma 5.6.2 *Let $f \in C^1([0,1])$ and let $p_m(x) \equiv \sum_{k=0}^{m}\begin{pmatrix} m \\ k\end{pmatrix} x^k(1-x)^{m-k} f\left(\frac{k}{m}\right)$ be the m^{th} Bernstein polynomial. Then in addition to $\|p_m - f\|_{[0,1]} \to 0$, it also follows that $\|p_m' - f'\|_{[0,1]} \to 0$.*

Proof: From simple computations,

$$\begin{aligned} p_m'(x) &= \sum_{k=1}^{m}\begin{pmatrix} m \\ k\end{pmatrix} kx^{k-1}(1-x)^{m-k} f\left(\frac{k}{m}\right) \\ &\quad -\sum_{k=0}^{m-1}\begin{pmatrix} m \\ k\end{pmatrix} x^k (m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right) \end{aligned}$$

$$\begin{aligned} &= \sum_{k=1}^{m}\frac{m(m-1)!}{(m-k)!(k-1)!}x^{k-1}(1-x)^{m-k} f\left(\frac{k}{m}\right) \\ &\quad -\sum_{k=0}^{m-1}\begin{pmatrix} m \\ k\end{pmatrix} x^k (m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right) \end{aligned}$$

$$\begin{aligned} &= \sum_{k=0}^{m-1}\frac{m(m-1)!}{(m-1-k)!k!}x^{k}(1-x)^{m-1-k} f\left(\frac{k+1}{m}\right) \\ &\quad -\sum_{k=0}^{m-1}\frac{m(m-1)!}{(m-1-k)!k!}x^{k}(1-x)^{m-1-k} f\left(\frac{k}{m}\right) \end{aligned}$$

$$= \sum_{k=0}^{m-1}\frac{m(m-1)!}{(m-1-k)!k!}x^{k}(1-x)^{m-1-k}\left(f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)\right)$$

$$= \sum_{k=0}^{m-1}\begin{pmatrix} m-1 \\ k\end{pmatrix} x^k(1-x)^{m-1-k}\left(\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1/m}\right)$$

By the mean value theorem, $\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1/m} = f'(x_{k,m})$, $x_{k,m} \in \left(\frac{k}{m},\frac{k+1}{m}\right)$. Now the desired result follows as before from the uniform continuity of f' on $[0,1]$. Let $\delta > 0$ be such that if $|x-y|<\delta$, then $|f'(x)-f'(y)|<\varepsilon$ 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

$$\left|f'(x)-f'\left(x_{k,m}\right)\right|=\left|f'(x)-\frac{f\left(\frac{k+1}{m}\right)-f\left(\frac{k}{m}\right)}{1/m}\right|<\varepsilon.$$

Now as before, letting $M\geq|f'(x)|$ for all x,

$$\left|p_m'(x)-f'(x)\right|\leq\sum_{k=0}^{m-1}\left(\begin{array}{c}m-1\\k\end{array}\right)x^k(1-x)^{m-1-k}\left|f'\left(x_{k,m}\right)-f'(x)\right|\leq$$

$$\sum_{\left\{x:\left|x-\frac{k}{m}\right|<\frac{\delta}{2}\right\}}\left(\begin{array}{c}m-1\\k\end{array}\right)x^k(1-x)^{m-1-k}\varepsilon+M\sum_{k=0}^{m-1}\left(\begin{array}{c}m-1\\k\end{array}\right)\frac{4(k-mx)^2}{m^2\delta^2}x^k(1-x)^{m-1-k}$$

$$\leq\varepsilon+4M\frac{1}{4}m\frac{1}{m^2\delta^2}=\varepsilon+M\frac{1}{m\delta^2}<2\varepsilon$$

whenever m is large enough. Thus this proves uniform convergence. ■

There is a more general version of the Weierstrass theorem which is easy to get. It depends on the Tietze extension theorem, a wonderful little result which is interesting for its own sake.

5.7 A Generalization

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 3.12.1.

Lemma 5.7.1 *Let H,K be two nonempty disjoint closed subsets of X. Then there exists a continuous function, $g:X\to[-1/3,1/3]$ 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 $\boldsymbol{h}_k\to\boldsymbol{x}$ 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 5.7.2 *For $f:M\subseteq X\to\mathbb{R}$, 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 5.7.3 *Suppose M is a closed set in X and suppose $f:M\to[-1,1]$ 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 5.7.1 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, $g(\boldsymbol{x}) \equiv \frac{1}{3}\frac{\text{dist}(\boldsymbol{x},H)-\text{dist}(\boldsymbol{x},K)}{\text{dist}(\boldsymbol{x},H)+\text{dist}(\boldsymbol{x},K)}$. When $\boldsymbol{x} \in H$ this gives $\frac{1}{3}\frac{-\text{dist}(\boldsymbol{x},K)}{\text{dist}(\boldsymbol{x},K)} = -\frac{1}{3}$. Then $\boldsymbol{x} \in K$, this gives $\frac{1}{3}\frac{\text{dist}(\boldsymbol{x},H)}{\text{dist}(\boldsymbol{x},H)} = \frac{1}{3}$. 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

$$\begin{aligned} g(-\boldsymbol{x}) &= \frac{1}{3}\frac{\text{dist}(-\boldsymbol{x},H)-\text{dist}(-\boldsymbol{x},K)}{\text{dist}(-\boldsymbol{x},H)+\text{dist}(-\boldsymbol{x},K)} = \frac{1}{3}\frac{\text{dist}(-\boldsymbol{x},-K)-\text{dist}(-\boldsymbol{x},-H)}{\text{dist}(-\boldsymbol{x},-K)+\text{dist}(-\boldsymbol{x},-H)} \\ &= \frac{1}{3}\frac{\text{dist}(\boldsymbol{x},K)-\text{dist}(\boldsymbol{x},H)}{\text{dist}(\boldsymbol{x},K)+\text{dist}(\boldsymbol{x},H)} = -g(\boldsymbol{x}).\ \blacksquare \end{aligned}$$

Lemma 5.7.4 *Suppose M is a closed set in X and suppose $f: M \to [-1,1]$ 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 5.7.3, let g_1 be such that $g_1(X) \subseteq [-1/3,1/3]$ and $\|f-g_1\|_M \leq \frac{2}{3}$. Suppose $g_1,\cdots,g_m$ have been chosen such that $g_j(X) \subseteq [-1/3,1/3]$ and

$$\left\| f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right\|_M < \left(\frac{2}{3}\right)^m. \tag{5.1}$$

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

$$\left\| \left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i\right) - g_{m+1} \right\|_M \leq \frac{2}{3}.$$

Thus $\left\| \left(f - \sum_{i=1}^m \left(\frac{2}{3}\right)^{i-1} g_i\right) - \left(\frac{2}{3}\right)^m g_{m+1} \right\|_M \leq \left(\frac{2}{3}\right)^{m+1}$. It follows there exists a sequence $\{g_i\}$ such that each has its values in $[-1/3,1/3]$ and for every m 5.1 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 3.9.3. The estimate 5.1 implies $f = g$ on M. The last claim follows because we can take each g_i odd. $\blacksquare$

The following is the Tietze extension theorem.

Theorem 5.7.5 *Let M be a closed nonempty subset of a metric space X and let $f : M \to [a,b]$ 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 5.7.4 and so there exists $g_1 : X \to [-1,1]$ such that g is continuous on X and equals f_1 on M. Let $g(\boldsymbol{x}) = (g_1(\boldsymbol{x}) - 1)\left(\frac{b-a}{2}\right) + b$. This works. The last claim follows from the same arguments which gave Lemma 5.7.4 or the change of variables just given. ■

Corollary 5.7.6 *Let M be a closed nonempty subset of a metric space X and let $f : M \to [a,b]$ 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 $\hat{g}(X) \subseteq [a-\varepsilon, b+\varepsilon]$. ■

With the Tietze extension theorem, here is a better version of the Weierstrass approximation theorem.

Theorem 5.7.7 *Let K be a closed and bounded subset of $\mathbb{R}^p$ and let $f : K \to \mathbb{R}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m\to\infty} (\sup\{|f(\boldsymbol{x}) - p_m(\boldsymbol{x})| : \boldsymbol{x} \in K\}) = 0.$$

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, $[a_k, b_k]$ such that $K \subseteq \prod_{k=1}^{p} [a_k, b_k] = R$. Then by the Weierstrass approximation theorem, Theorem 5.6.1 there exists a sequence of polynomials $\{p_m\}$ 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 5.7.8 *Let K be a closed and bounded subset of $\mathbb{R}^p$ and let $f : K \to \mathbb{F}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m\to\infty} (\sup\{|f(\boldsymbol{x}) - p_m(\boldsymbol{x})| : \boldsymbol{x} \in K\}) = 0.$$

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 $\mathbb{R}^p$.

5.8 An Approach to the Integral

With the Weierstrass approximation theorem, you can give a rigorous definition of the Riemann integral without wading in to Riemann sums. This shows the integral can be defined directly from very simple ideas. First is a short review of the derivative of a function of one variable.

Definition 5.8.1 *Let $f:[a,b]\to\mathbb{R}$. Then $f'(x)\equiv\lim_{x\to 0}\frac{f(x+h)-f(x)}{h}$ 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 5.8.2 *Let $f:[a,b]\to\mathbb{R}$ be continuous. Then if the maximum value of f occurs at a point $x\in(a,b)$, it follows that if $f'(x)=0$. If f achieves a minimum at $x\in(a,b)$ where $f'(x)$ exists, it also follows that $f'(x)=0$.*

Proof: By Theorem 3.7.2, f achieves a maximum at some point x. If $f'(x)$ exists, then

$$f'(x)=\lim_{h\to 0+}\frac{f(x+h)-f(x)}{h}=\lim_{h\to 0-}\frac{f(x+h)-f(x)}{h}$$

However, the first limit is non-positive while the second is non-negative and so $f'(x)=0$. 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 5.8.3 *Let f,g be continuous on $[a,b]$ and differentiable on (a,b). Then there exists $x\in(a,b)$ such that $f'(x)(g(b)-g(a))=g'(x)(f(b)-f(a))$. If $g(x)=x$, this yields $f(b)-f(a)=f'(x)(b-a)$, also $f(a)-f(b)=f'(x)(a-b)$.*

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 $h'(x)=0$. 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 $h'(x)=0$ which proves the theorem. ■

Recall that an antiderivative of a function f is just a function F such that $F'=f$. You know how to find an antiderivative for a polynomial. $\left(\frac{x^{n+1}}{n+1}\right)'=x^n$ 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 $F'(x)=0$ 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 $F'=G'$ on an open interval, F,G continuous on the closed interval, it follows that $F-G$ is a constant and so $F(b)-F(a)=G(b)-G(a)$. In words, any two antiderivatives differ by a constant.

Definition 5.8.4 *For $p(x)$ a polynomial on $[a,b]$, let $P'(x)=p(x)$. Thus, by the mean value theorem if $P',\hat{P}'$ both equal p, it follows that $P(b)-P(a)=\hat{P}(b)-\hat{P}(a)$. Then define $\int_a^b p(x)\,dx\equiv P(b)-P(a)$. If $f\in C([a,b])$, define $\int_a^b f(x)\,dx\equiv\lim_{n\to\infty}\int_a^b p_n(x)\,dx$ where $\lim_{n\to\infty}\|p_n-f\|\equiv\lim_{n\to\infty}\max_{x\in[a,b]}|f(x)-p_n(x)|=0$*

Proposition 5.8.5 *The above integral is well defined and satisfies the following properties.*

1. $\int_a^b f dx = f(\hat{x})(b-a)$ *for some* $\hat{x}$ *between* a *and* b. *Thus* $\left|\int_a^b f dx\right| \leq \|f\| |b-a|$.

2. *If* f *is continuous on an interval which contains all necessary intervals,*
$$\int_a^c f dx + \int_c^b f dx = \int_a^b f dx, \text{ so } \int_a^b f dx + \int_b^a f dx = \int_b^b f dx = 0$$

3. *If* $F(x) \equiv \int_a^x f dt$, *then* $F'(x) = f(x)$ *so any continuous function has an antiderivative, and for any* $a \neq b$, $\int_a^b f dx = G(b) - G(a)$ *whenever* $G' = f$ *on the open interval determined by* a, b *and* G *continuous on the closed interval determined by* a, b. *Also,*
$$\int_a^b (\alpha f(x) + \beta g(x)) dx = \alpha \int_a^b f(x) dx + \beta \int_a \beta g(x) dx$$
If $a < b$, *and* $f(x) \geq 0$, *then* $\int_a^b f dx \geq 0$. *Also* $\left|\int_a^b f dx\right| \leq \left|\int_a^b |f| dx\right|$.

4. $\int_a^b 1 dx = b - a$.

Proof: First, why is the integral well defined? With notation as in the above definition, the mean value theorem implies

$$\int_a^b p(x) dx \equiv P(b) - P(a) = p(\hat{x})(b-a) \tag{5.2}$$

where $\hat{x}$ is between a and b and so $\left|\int_a^b p(x) dx\right| \leq \|p\| |b-a|$. If $\|p_n - f\| \to 0$, then $\lim_{m,n\to\infty} \|p_n - p_m\| = 0$ and so

$$\left|\int_a^b p_n(x) dx - \int_a^b p_m(x) dx\right| = |(P_n(b) - P_n(a)) - (P_m(b) - P_m(a))|$$

$$= |(P_n(b) - P_m(b)) - (P_n(a) - P_m(a))| = \left|\int_a^b (p_n - p_m) dx\right| \leq \|p_n - p_m\| |b-a|$$

Thus the limit exists because $\left\{\int_a^b p_n dx\right\}_n$ is a Cauchy sequence and $\mathbb{R}$ is complete.

From 5.2, 1. holds for a polynomial $p(x)$. Let $\|p_n - f\| \to 0$. Then by definition,

$$\int_a^b f dx \equiv \lim_{n\to\infty} \int_a^b p_n dx = p_n(x_n)(b-a) \tag{5.3}$$

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 $x_n \to x \in [a,b]$. Then fixing m such that $\|f - p_n\| < \varepsilon$ whenever $n \geq m$, assume $n > m$. Then $\|p_m - p_n\| \leq \|p_m - f\| + \|f - p_n\| < 2\varepsilon$ and so

$$|f(x) - p_n(x_n)| \leq |f(x) - f(x_n)| + |f(x_n) - p_m(x_n)| + |p_m(x_n) - p_n(x_n)|$$

$$\leq |f(x) - f(x_n)| + \|f - p_m\| + \|p_m - p_n\| < |f(x) - f(x_n)| + 3\varepsilon$$

Now if n is still larger, continuity of f shows that $|f(x) - p_n(x_n)| < 4\varepsilon$. Since ε is arbitrary, $p_n(x_n) \to f(x)$ and so, passing to the limit with this subsequence in 5.3 yields 1.

Now consider 2. It holds for polynomials $p(x)$ obviously. So let $\|p_n - f\| \to 0$. Then

$$\int_a^c p_n dx + \int_c^b p_n dx = \int_a^b p_n dx$$

Pass to a limit as $n \to \infty$ and use the definition to get 2. Also note that $\int_b^b f(x)\,dx = 0$ 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, $\frac{F(x+h)-F(x)}{h} = \frac{1}{h}\int_x^{x+h} f(t)\,dt = f(x_h)$ where x_h is between x and $x+h$. Let $h \to 0$. By continuity of f, it follows that the limit of the right side exists and so

$$\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} f(x_h) = f(x)$$

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' exists on (a,b) so if G is an antiderivative,

$$\int_a^b f(t)\,dt \equiv F(b) = F(b) - F(a) = G(b) - G(a)$$

The claim that the integral is linear is obvious from this. Indeed, if $F' = f, G' = g$,

$$\begin{aligned}\int_a^b (\alpha f(t) + \beta g(t))\,dt &= \alpha F(b) + \beta G(b) - (\alpha F(a) + \beta G(a)) \\ &= \alpha(F(b) - F(a)) + \beta(G(b) - G(a)) \\ &= \alpha \int_a^b f(t)\,dt + \beta \int_a^b g(t)\,dt\end{aligned}$$

If $f \geq 0$, then the mean value theorem implies that for some

$$t \in (a,b), F(b) - F(a) = \int_a^b f dx = f(t)(b-a) \geq 0.$$

Thus $\int_a^b (|f| - f)\,dx \geq 0, \int_a^b (|f| + f)\,dx \geq 0$ and so $\int_a^b |f|\,dx \geq \int_a^b f dx$ and $\int_a^b |f|\,dx \geq -\int_a^b f dx$ so this proves $\left|\int_a^b f dx\right| \leq \int_a^b |f|\,dx$. This, along with part 2 implies the other claim that $\left|\int_a^b f dx\right| \leq \left|\int_a^b |f|\,dx\right|$.

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 $F' = f$, then $f(g(x))g'(x) = \frac{d}{dx}F(g(x))$ so that, from the above proposition, $F(g(b)) - F(g(a)) = \int_{g(a)}^{g(b)} f(y)\,dy = \int_a^b f(g(x))g'(x)\,dx$.We usually let $y = g(x)$ and $dy = g'(x)\,dx$ 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.

Definition 5.8.6 *If $f \in C_c(\mathbb{R})$, define $\int_{\mathbb{R}} f \equiv \int_{-\infty}^{\infty} f(x)\,dx$ as $\int_a^b f(x)\,dx$ where the interval $[a,b]$ is chosen such that* $\text{spt}(f) \subseteq [a,b]$.

Proposition 5.8.7 *The above definition is well defined.*

Proof: Letting $b \equiv \sup\{x : f(x) \neq 0\}$, it follows $f(b) = 0$ and $f(x) = 0$ for $x > b$. Similarly if $a \equiv \inf\{x : f(x) \neq 0\}$, it follows $f(a) = 0$ and $f(x) = 0$ for $x < a$. Thus, by the mean value theorem, $F' = f$ requires $F(x) = F(b)$ for $x > b$ and $F(x) = F(a)$ for $x < a$. It follows that the above definition is not dependent on the interval $[a,b]$ containing $\operatorname{spt}(f)$. ■

Consider the iterated integral $\int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} \alpha x_1^{\alpha_1} \cdots x_p^{\alpha_p} dx_p \cdots dx_1$. 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

$$\alpha \prod_{k=1}^{p} \left(\int_{a_k}^{b_k} x_k^{\alpha_k} dx_k \right) = \alpha \prod_{k=1}^{p} \left(\frac{b^{\alpha_k+1}}{\alpha_k+1} - \frac{a^{\alpha_k+1}}{\alpha_k+1} \right)$$

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 $(i_1, \cdots, i_p)$ is any permutation of $(1, \cdots, p)$, then for any polynomial q,

$$\int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} q(x_1, ..., x_p) \, dx_p \cdots dx_1 = \int_{a_{i_p}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} q(x_1, ..., x_p) \, dx_{i_p} \cdots dx_{i_1}$$

Now let $f : \prod_{k=1}^{p} [a_k, b_k] \to \mathbb{R}$ 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 5.8.5, $\left| \int_c^d f(x,y) \, dy - \int_c^d f(\hat{x}, y) \, dy \right| =$

$$\left| \int_c^d (f(x,y) - f(\hat{x},y)) \, dy \right| \leq \max_{y \in [c,d]} |f(x,y) - f(\hat{x},y)| < \varepsilon$$

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 $\int_a^b \int_c^d f(x,y) \, dy dx$. Then using Proposition 5.8.5 on the iterated integrals along with Theorem 5.6.1, there exists a sequence of polynomials which converges to f uniformly $\{p_n\}$. Then applying Proposition 5.8.5 repeatedly,

$$\left| \int_{a_{i_p}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} f(\boldsymbol{x}) \, dx_p \cdots dx_1 - \int_{a_{i_p}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} p_n(\boldsymbol{x}) \, dx_p \cdots dx_1 \right| \leq \|f - p_n\| \prod_{k=1}^{p} |b_k - a_k| \tag{5.4}$$

With this, it is easy to prove a rudimentary Fubini theorem valid for continuous functions.

Theorem 5.8.8 *$f : \prod_{k=1}^{p} [a_k, b_k] \to \mathbb{R}$ be continuous. Then for $(i_1, \cdots, i_p)$ any permutation of $(1, \cdots, p)$,*

$$\int_{a_{i_p}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} f(\boldsymbol{x}) \, dx_{i_p} \cdots dx_{i_1} = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(\boldsymbol{x}) \, dx_p \cdots dx_1$$

If $f \geq 0$, then the iterated integrals are nonnegative if each $a_k \leq b_k$. Also, we can define for $f \in C_c(\mathbb{R}^p)$

$$\int_{\mathbb{R}^p} f \equiv \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(\boldsymbol{x}) \, dx_p \cdots dx_1$$

where $\operatorname{spt}(f) \subseteq \prod_{k=1}^{p}[a_k, b_k]$ *and the integral does not depend on the order of the iterated integrals.*

Proof: Let $\|p_n - f\|_{\prod_{k=1}^{p}[a_k,b_k]} \to 0$ where p_n is a polynomial. Then from 5.4,

$$\int_{a_{i_1}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} f(\boldsymbol{x})\, dx_{i_p} \cdots dx_{i_1} = \lim_{n\to\infty} \int_{a_{i_p}}^{b_{i_1}} \cdots \int_{a_{i_p}}^{b_{i_p}} p_n(\boldsymbol{x})\, dx_{i_p} \cdots dx_{i_1}$$

$$= \lim_{n\to\infty} \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} p_n(\boldsymbol{x})\, dx_p \cdots dx_1 = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(\boldsymbol{x})\, dx_p \cdots dx_1$$

The fact that this integral is well defined in the last claim follows from Proposition 5.8.7. ■

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 one, consists of finitely many intervals. Thus you can integrate over all the usual sets encountered in beginning calculus.

5.9 The Stone Weierstrass Approximation Theorem

There is a profound generalization of the Weierstrass approximation theorem due to Stone. It has to be one of the most elegant things available. It holds on locally compact Hausdorff spaces but here I will show the version which is valid on compact sets.

Definition 5.9.1 *$\mathscr{A}$ is an algebra of functions if $\mathscr{A}$ is a vector space and if whenever $f, g \in \mathscr{A}$ then $fg \in \mathscr{A}$.*

To begin with assume that the field of scalars is $\mathbb{R}$. This will be generalized later. Theorem 5.5.2 implies the following corollary. See Corollary 5.5.3.

Corollary 5.9.2 *The polynomials are dense in $C([a,b])$.*

Here is another approach to proving this theorem. It is the original approach used by Weierstrass. Let $m \in \mathbb{N}$ and consider c_m such that $\int_{-1}^{1} c_m (1-x^2)^m dx = 1$. Then

$$1 = 2\int_0^1 c_m (1-x^2)^m dx \geq 2c_m \int_0^1 (1-x)^m dx = 2c_m \frac{1}{m+1}$$

so $c_m \leq m+1$. Then

$$\int_\delta^1 c_m (1-x^2)^m dx + \int_{-1}^{-\delta} c_m (1-x^2)^m dx \leq 2(m+1)\left(1-\delta^2\right)^m$$

which converges to 0. Thus

$$\lim_{m\to\infty} \sup_{x \notin [-\delta,\delta]} c_m (1-x^2)^m = 0 \tag{5.5}$$

Now let $\phi_n(t) \equiv c_m (1-t^2)^m$. Consider $f \in C([-1,1])$ and extend to let $f(x) = f(1)$ if $x > 1$ and $f(x) = f(-1)$ if $x < -1$ and define $p_m(x) \equiv \int_{-1}^{1} f(x-t)\, \phi_m(t)\, dt$. Then

$$|p_m(x) - f(x)| \leq \int_{-1}^{1} |f(x-t) - f(x)|\, \phi_m(t)\, dt \leq$$

$$\int_{-1}^{1}\mathscr{X}_{[-\delta,\delta]}(t)\left|f(x-t)-f(x)\right|\phi_m(t)\,dt+\int_{-1}^{1}\mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\left|f(x-t)-f(x)\right|\phi_m(t)\,dt$$

Choose δ so small that if $|x-y|<\delta$, then $|f(x)-f(y)|<\varepsilon$. Also let $M\geq\max_x|f(x)|$. Then

$$|p_m(x)-f(x)|\leq\varepsilon\int_{-1}^{1}\phi_m(t)\,dt+2M\int_{-1}^{1}\mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\phi_m(t)\,dt$$

$$=\varepsilon+2M\int_{-1}^{1}\mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\phi_m(t)\,dt$$

From 5.5, The second term is no larger than $2M\int_{-1}^{1}\mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\varepsilon dt\leq 4M\varepsilon$ whenever m is large enough. Hence, for large enough m, $\sup_{x\in[-1,1]}|p_m(x)-f(x)|\leq(1+4M)\,\varepsilon$. Since ε is arbitrary, this shows that the functions p_m converge uniformly to f on $[-1,1]$. However, p_m is actually a polynomial. To see this, change the variables and obtain

$$p_m(x)=\int_{x-1}^{x+1}f(t)\,\phi_m(x-t)\,dt$$

which will be a polynomial. To see this, note that a typical term is of the form

$$\int_{x-1}^{x+1}f(t)\,a\,(x-t)^k\,dt,$$

clearly a polynomial in x. This proves Corollary 5.9.2 in case $[a,b]=[-1,1]$. In the general case, there is a linear one to one onto map $l:[-1,1]\to[a,b]$.

$$l(t)=\frac{b-a}{2}(t+1)+a$$

Then if $f\in C([a,b])\,,f\circ l\in C([-1,1])$. Hence there is a polynomial p such that

$$\max_{t\in[-1,1]}|f\circ l(t)-p(t)|<\varepsilon$$

Then letting $t=l^{-1}(x)=\frac{2(x-a)}{b-a}-1$, for $x\in[a,b]\,,\max_{x\in[a,b]}\left|f(x)-p\left(l^{-1}(x)\right)\right|<\varepsilon$ but $x\to p\left(l^{-1}(x)\right)$ is a polynomial. This gives an independent proof of that corollary. ■

The next result is the key to the profound generalization of the Weierstrass theorem due to Stone in which an interval will be replaced by a compact set and polynomials will be replaced with elements of an algebra satisfying certain axioms.

Corollary 5.9.3 *On the interval $[-M,M]$, there exist polynomials p_n, $p_n(0)=0$, and $\lim_{n\to\infty}\|p_n-|\cdot|\|_\infty=0$. recall that $\|f\|_\infty\equiv\sup_{t\in[-M,M]}|f(t)|$.*

Proof: By Corollary 5.9.2 there exists a sequence of polynomials, $\{\tilde{p}_n\}$ such that $\tilde{p}_n\to|\cdot|$ uniformly. Then let $p_n(t)\equiv\tilde{p}_n(t)-\tilde{p}_n(0)$. ■

Definition 5.9.4 *An algebra of functions, $\mathscr{A}$ defined on A, annihilates no point of A if for all $x\in A$, there exists $g\in\mathscr{A}$ such that $g(x)\neq 0$. The algebra separates points if whenever $x_1\neq x_2$, then there exists $g\in\mathscr{A}$ such that $g(x_1)\neq g(x_2)$.*

The following generalization is known as the Stone Weierstrass approximation theorem.

Theorem 5.9.5 *Let A be a compact topological space and let $\mathscr{A} \subseteq C(A;\mathbb{R})$ be an algebra of functions which separates points and annihilates no point. Then $\mathscr{A}$ is dense in $C(A;\mathbb{R})$.*

Proof: First here is a lemma.

Lemma 5.9.6 *Let c_1 and c_2 be two real numbers and let $x_1 \neq x_2$ be two points of A. Then there exists a function $f_{x_1x_2}$ such that*

$$f_{x_1x_2}(x_1) = c_1,\ f_{x_1x_2}(x_2) = c_2.$$

Proof of the lemma: Let $g \in \mathscr{A}$ satisfy $g(x_1) \neq g(x_2)$. Such a g exists because the algebra separates points. Since the algebra annihilates no point, there exist functions h and k such that $h(x_1) \neq 0,\ k(x_2) \neq 0$. Then let $u \equiv gh - g(x_2)h,\ v \equiv gk - g(x_1)k$. It follows that $u(x_1) \neq 0$ and $u(x_2) = 0$ while $v(x_2) \neq 0$ and $v(x_1) = 0$. Let $f_{x_1x_2} \equiv \frac{c_1 u}{u(x_1)} + \frac{c_2 v}{v(x_2)}$. This proves the lemma. Now continue the proof of Theorem 5.9.5.

First note that $\overline{\mathscr{A}}$ satisfies the same axioms as $\mathscr{A}$ but in addition to these axioms, $\overline{\mathscr{A}}$ is closed. The closure of $\mathscr{A}$ is taken with respect to the usual norm on $C(A)$,

$$\|f\|_\infty \equiv \max\{|f(x)| : x \in A\}.$$

Suppose $f \in \overline{\mathscr{A}}$ and suppose M is large enough that $\|f\|_\infty < M$. Using Corollary 5.9.3, let p_n be a sequence of polynomials such that

$$\|p_n - |\cdot|\|_\infty \to 0,\ p_n(0) = 0.$$

It follows that $p_n \circ f \in \overline{\mathscr{A}}$ and so $|f| \in \overline{\mathscr{A}}$ whenever $f \in \overline{\mathscr{A}}$. Also note that

$$\max(f,g) = \frac{|f-g| + (f+g)}{2}$$

$$\min(f,g) = \frac{(f+g) - |f-g|}{2}.$$

Therefore, this shows that if $f,g \in \overline{\mathscr{A}}$ then $\max(f,g),\ \min(f,g) \in \overline{\mathscr{A}}$. By induction, if $f_i, i = 1,2,\cdots,m$ are in $\overline{\mathscr{A}}$ then

$$\max(f_i, i = 1,2,\cdots,m),\ \min(f_i, i = 1,2,\cdots,m) \in \overline{\mathscr{A}}.$$

Now let $h \in C(A;\mathbb{R})$ and let $x \in A$. Use Lemma 5.9.6 to obtain f_{xy}, a function of $\overline{\mathscr{A}}$ which agrees with h at x and y. Letting $\varepsilon > 0$, there exists an open set $U(y)$ containing y such that

$$f_{xy}(z) > h(z) - \varepsilon \text{ if } z \in U(y).$$

Since A is compact, let $U(y_1),\cdots,U(y_l)$ cover A. Let

$$f_x \equiv \max\left(f_{xy_1}, f_{xy_2},\cdots,f_{xy_l}\right).$$

Then $f_x \in \overline{\mathscr{A}}$ and $f_x(z) > h(z) - \varepsilon$ for all $z \in A$ and $f_x(x) = h(x)$. This implies that for each $x \in A$ there exists an open set $V(x)$ containing x such that for $z \in V(x)$, $f_x(z) < h(z) + \varepsilon$. Let $V(x_1),\cdots,V(x_m)$ cover A and let $f \equiv \min(f_{x_1},\cdots,f_{x_m})$. Therefore, $f(z) < h(z) + \varepsilon$ for all $z \in A$ and since $f_x(z) > h(z) - \varepsilon$ for all $z \in A$, it follows $f(z) > h(z) - \varepsilon$ also and so $|f(z) - h(z)| < \varepsilon$ for all z. Since ε is arbitrary, this shows $h \in \overline{\mathscr{A}}$ and proves $\overline{\mathscr{A}} = C(A;\mathbb{R})$. ■

5.10 Connectedness in Normed Linear Space

The main result is that a ball in a normed linear space is connected. This is the next lemma. From this, it follows that for an open set, it is connected if and only if it is arcwise connected.

Lemma 5.10.1 *In a normed vector space, $B(\boldsymbol{z},r)$ is arcwise connected.*

Proof: This is easy from the convexity of the set. If $\boldsymbol{x},\boldsymbol{y} \in B(\boldsymbol{z},r)$, then let $\boldsymbol{\gamma}(t) = \boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})$ for $t \in [0,1]$.

$$\|\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})-\boldsymbol{z}\| = \|(1-t)(\boldsymbol{x}-\boldsymbol{z})+t(\boldsymbol{y}-\boldsymbol{z})\|$$

$$\leq (1-t)\|\boldsymbol{x}-\boldsymbol{z}\|+t\|\boldsymbol{y}-\boldsymbol{z}\| < (1-t)r+tr = r$$

showing $\boldsymbol{\gamma}(t)$ stays in $B(\boldsymbol{z},r)$.■

Proposition 5.10.2 *If $X \neq \emptyset$ is arcwise connected, then it is connected.*

Proof: Let $p \in X$. Then by assumption, for any $x \in X$, there is an arc joining p and x. This arc is connected because it is the continuous image of an interval which is connected. Since x is arbitrary, every x is in a connected subset of X which contains p. Hence $C_p = X$ and so X is connected. ■

Theorem 5.10.3 *Let U be an open subset of a normed vector space. Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.*

Proof: By Proposition 5.10.2 it is only necessary to verify that if U is connected and open in the context of this theorem, then U is arcwise connected. Pick $\boldsymbol{p} \in U$. Say $\boldsymbol{x} \in U$ satisfies $\mathscr{P}$ if there exists a continuous function, $\boldsymbol{\gamma} : [a,b] \to U$ such that $\gamma(a) = \boldsymbol{p}$ and $\boldsymbol{\gamma}(\boldsymbol{b}) = \boldsymbol{x}$.

$$A \equiv \{\boldsymbol{x} \in U \text{ such that } \boldsymbol{x} \text{ satisfies } \mathscr{P}.\}$$

If $\boldsymbol{x} \in A$, then Lemma 5.10.1 implies $B(\boldsymbol{x},r) \subseteq U$ is arcwise connected for small enough r. Thus letting $\boldsymbol{y} \in B(\boldsymbol{x},r)$, there exist intervals, $[a,b]$ and $[c,d]$ and continuous functions having values in U, $\boldsymbol{\gamma},\boldsymbol{\eta}$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}, \boldsymbol{\gamma}(b) = \boldsymbol{x}, \boldsymbol{\eta}(c) = \boldsymbol{x}$, and $\eta(d) = \boldsymbol{y}$. Then let $\gamma_1 : [a,b+d-c] \to U$ be defined as

$$\gamma_1(t) \equiv \begin{cases} \boldsymbol{\gamma}(t) \text{ if } t \in [a,b] \\ \boldsymbol{\eta}(t+c-b) \text{ if } t \in [b,b+d-c] \end{cases}$$

Then it is clear that $\boldsymbol{\gamma}_1$ is a continuous function mapping $\boldsymbol{p}$ to $\boldsymbol{y}$ and showing that $B(\boldsymbol{x},r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(\boldsymbol{p},\boldsymbol{\delta})$ containing $\boldsymbol{p}$ which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $\boldsymbol{z} \in B$ such that every open set containing $\boldsymbol{z}$ is not contained in B. Therefore, letting $B(\boldsymbol{z},\boldsymbol{\delta})$ be such that $\boldsymbol{z} \in B(\boldsymbol{z},\boldsymbol{\delta}) \subseteq U$, there exist points of A contained in $B(\boldsymbol{z},\boldsymbol{\delta})$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected.

It remains to verify the connected components are open. Let $\boldsymbol{z} \in C_{\boldsymbol{p}}$ where $C_{\boldsymbol{p}}$ is the connected component determined by $\boldsymbol{p}$. Then picking $B(\boldsymbol{z},\delta) \subseteq U$, $C_{\boldsymbol{p}} \cup B(\boldsymbol{z},\delta)$ is connected and contained in U and so it must also be contained in $C_{\boldsymbol{p}}$. Thus $\boldsymbol{z}$ is an interior point of $C_{\boldsymbol{p}}$. ■

As an application, consider the following corollary.

Corollary 5.10.4 *Let $f : \Omega \to \mathbb{Z}$ be continuous where Ω is a connected open set in a normed vector space. Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω. To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\cup_{m \neq l}\left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left(\left(l - \frac{1}{6}, l + \frac{1}{6}\right)\right)$ also an open set. ■

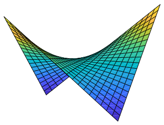
Definition 5.10.5 *An important concept in a vector space is the concept of convexity. A nonempty set K is called convex if whenever $x, y \in K$, it follows that for all $t \in [0,1], tx + (1-t)y \in K$ also. That is, the line segment joining the two points x, y is in K.*

5.11 Saddle Points*

A very useful idea in nonlinear analysis is the saddle point theorem also called the min max theorem. The proof of this theorem given here follows Brezis [6] which is where I found it. A real valued function f defined on a linear space is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

It is concave if the inequality is turned around. It can be shown that in finite dimensions, convex functions are automatically continuous, similar for concave functions. Recall the following definition of upper and lower semicontinuous functions defined on a metric space and having values in $[-\infty,\infty]$. This is about functions which look like this, convex in one direction and concave in the other.



Definition 5.11.1 *A function is upper semicontinuous if whenever $x_n \to x$, it follows that $f(x) \geq \limsup_{n\to\infty} f(x_n)$ and it is lower semicontinuous if $f(x) \leq \liminf_{n\to\infty} f(x_n)$.*

The following lemma comes directly from the definition.

Lemma 5.11.2 *If $\mathscr{F}$ is a set of functions which are upper semicontinuous, then $g(x) \equiv \inf\{f(x) : f \in \mathscr{F}\}$ is also upper semicontinuous. Similarly, if $\mathscr{F}$ is a set of functions which are lower semicontinuous, then if $g(x) \equiv \sup\{f(x) : f \in \mathscr{F}\}$ it follows that g is lower semicontinuous.*

Note that in a metric space, the above definitions of upper and lower semicontinuity in terms of sequences are equivalent to the definitions that

$$f(x) \geq \lim_{r\to 0} \sup \{f(y) : y \in B(x,r)\}, \ f(x) \leq \lim_{r\to 0} \inf \{f(y) : y \in B(x,r)\}$$

respectively.

Here is a technical lemma which will make the proof of the saddle point theorem shorter. It seems fairly interesting also.

Lemma 5.11.3 *Suppose $H : A \times B \to \mathbb{R}$ is strictly convex in the first argument and concave in the second argument where A, B are compact convex nonempty subsets of Banach spaces E, F respectively and $x \to H(x,y)$ is lower semicontinuous while $y \to H(x,y)$ is upper semicontinuous. Let $H(g(y), y) \equiv \min_{x\in A} H(x,y)$. Then $g(y)$ is uniquely defined and also for $t \in [0,1]$, $\lim_{t\to 0} g(y + t(z-y)) = g(y)$.*

Proof: First suppose both z, w yield the definition of $g(y)$. Then

$$H\left(\frac{z+w}{2}, y\right) < \frac{1}{2} H(z,y) + \frac{1}{2} H(w,y)$$

which contradicts the definition of $g(y)$. As to the existence of $g(y)$ this is nothing more than the theorem that a lower semicontinuous function defined on a compact set achieves its minimum.

Now consider the last claim about “hemicontinuity”, continuity along a line. For all $x \in A$, it follows from the definition of g that

$$H(g(y+t(z-y)), y+t(z-y)) \leq H(x, y+t(z-y))$$

By concavity of H in the second argument,

$$\begin{aligned} & (1-t) H(g(y+t(z-y)), y) + tH(g(y+t(z-y)), z) && (5.6) \\ \leq \ & H(x, y+t(z-y)) && (5.7) \end{aligned}$$

Now let $t_n \to 0$. Does $g(y+t_n(z-y)) \to g(y)$? Suppose not. By compactness, each of $g(y+t_n(z-y))$ is in a compact set and so there is a further subsequence, still denoted by t_n such that $g(y+t_n(z-y)) \to \hat{x} \in A$. Then passing to a limit in 5.7, one obtains, using the upper semicontinuity in one and lower semicontinuity in the other the following inequality.

$$\begin{aligned} H(\hat{x}, y) \quad \leq \quad & \liminf_{n\to\infty} (1-t_n) H(g(y+t_n(z-y)), y) \\ & + \liminf_{n\to\infty} t_n H(g(y+t_n(z-y)), z) \end{aligned}$$

$$\leq \liminf_{n\to\infty} \left(\begin{array}{c} (1-t_n) H(g(y+t_n(z-y)), y) \\ + t_n H(g(y+t_n(z-y)), z) \end{array} \right)$$

$$\leq \limsup_{n\to\infty} H(x, y+t_n(z-y)) \leq H(x,y)$$

This shows that $\hat{x} = g(y)$ because this holds for every x. Since $t_n \to 0$ was arbitrary, this shows that in fact $\lim_{t\to 0+} g(y+t(z-y)) = g(y)$ ■

Now with this preparation, here is the min-max theorem.

Definition 5.11.4 *A norm is called strictly convex if whenever $x \neq y$,*

$$\left\| \frac{x+y}{2} \right\| < \frac{\|x\|}{2} + \frac{\|y\|}{2}$$

Theorem 5.11.5 *Let E, F be Banach spaces with E having a strictly convex norm. Also suppose that $A \subseteq E, B \subseteq F$ are compact and convex sets and that $H : A \times B \to \mathbb{R}$ is such that*

$$x \to H(x,y) \text{ is convex}$$

$$y \to H(x,y) \text{ is concave}$$

Assume that $x \to H(x,y)$ is lower semicontinuous and $y \to H(x,y)$ is upper semicontinuous. Then $\min_{x\in A} \max_{y\in B} H(x,y) = \max_{y\in B} \min_{x\in A} H(x,y)$. This condition is equivalent to the existence of $(x_0, y_0) \in A \times B$ such that

$$H(x_0,y) \leq H(x_0,y_0) \leq H(x,y_0) \text{ for all } x,y \tag{5.8}$$

called a saddle point.

Proof: One part of the main equality is obvious.

$$\max_{y\in B} H(x,y) \geq H(x,y) \geq \min_{x\in A} H(x,y)$$

and so for each x, $\max_{y\in B} H(x,y) \geq \max_{y\in B} \min_{x\in A} H(x,y)$ and so

$$\min_{x\in A} \max_{y\in B} H(x,y) \geq \max_{y\in B} \min_{x\in A} H(x,y) \tag{5.9}$$

Next consider the other direction.

Define $H_\varepsilon(x,y) \equiv H(x,y) + \varepsilon \|x\|^2$ where $\varepsilon > 0$. Then H_ε is strictly convex in the first variable. This results from the observation that

$$\left\| \frac{x+y}{2} \right\|^2 < \left(\frac{\|x\| + \|y\|}{2} \right)^2 \leq \frac{1}{2}\left(\|x\|^2 + \|y\|^2 \right),$$

By Lemma 5.11.3 there exists a unique $x \equiv g(y)$ with $H_\varepsilon(g(y),y) \equiv \min_{x\in A} H_\varepsilon(x,y)$ and also, whenever $y,z \in A, \lim_{t\to 0+} g(y + t(z-y)) = g(y)$. Thus

$$H_\varepsilon(g(y),y) = \min_{x\in A} H_\varepsilon(x,y).$$

But also this shows that $y \to H_\varepsilon(g(y),y)$ is the minimum of functions which are upper semicontinuous and so this function is also upper semicontinuous. Hence there exists y^* such that

$$\max_{y\in B} H_\varepsilon(g(y),y) = H_\varepsilon(g(y^*),y^*) = \max_{y\in B} \min_{x\in A} H_\varepsilon(x,y) \tag{5.10}$$

Thus from concavity in the second argument and what was just defined, for $t \in (0,1)$,

$$\begin{aligned} H_\varepsilon(g(y^*),y^*) &\geq H_\varepsilon(g((1-t)y^* + ty), (1-t)y^* + ty) \\ &\geq (1-t) H_\varepsilon(g((1-t)y^* + ty), y^*) + t H_\varepsilon(g((1-t)y^* + ty), y) \\ &\geq (1-t) H_\varepsilon(g(y^*), y^*) + t H_\varepsilon(g((1-t)y^* + ty), y) \end{aligned} \tag{5.11}$$

This is because $\min_x H_\varepsilon(x,y^*) \equiv H_\varepsilon(g(y^*),y^*)$ so

$$H_\varepsilon(g((1-t)y^*+ty),y^*) \geq H_\varepsilon(g(y^*),y^*)$$

Then subtracting the first term on the right, one gets

$$tH_\varepsilon(g(y^*),y^*) \geq tH_\varepsilon(g((1-t)y^*+ty),y)$$

and cancelling the t,

$$H_\varepsilon(g(y^*),y^*) \geq H_\varepsilon(g((1-t)y^*+ty),y)$$

Now apply Lemma 5.11.3 and let $t \to 0+$. This along with lower semicontinuity yields

$$H_\varepsilon(g(y^*),y^*) \geq \lim\inf_{t\to 0+} H_\varepsilon(g((1-t)y^*+ty),y) = H_\varepsilon(g(y^*),y) \tag{5.12}$$

Hence for every x,y

$$H_\varepsilon(x,y^*) \geq H_\varepsilon(g(y^*),y^*) \geq H_\varepsilon(g(y^*),y)$$

Thus

$$\min_x H_\varepsilon(x,y^*) \geq H_\varepsilon(g(y^*),y^*) \geq \max_y H_\varepsilon(g(y^*),y)$$

and so

$$\begin{aligned}\max_{y\in B}\min_{x\in A} H_\varepsilon(x,y) &\geq \min_x H_\varepsilon(x,y^*) \geq H_\varepsilon(g(y^*),y^*)\\ &\geq \max_y H_\varepsilon(g(y^*),y) \geq \min_{x\in A}\max_{y\in B} H_\varepsilon(x,y)\end{aligned}$$

Thus, letting $C \equiv \max\{\|x\| : x \in A\}$

$$\varepsilon C^2 + \max_{y\in B}\min_{x\in A} H(x,y) \geq \min_{x\in A}\max_{y\in B} H(x,y)$$

Since ε is arbitrary, it follows that

$$\max_{y\in B}\min_{x\in A} H(x,y) \geq \min_{x\in A}\max_{y\in B} H(x,y)$$

This proves the first part because it was shown above in 5.9 that

$$\min_{x\in A}\max_{y\in B} H(x,y) \geq \max_{y\in B}\min_{x\in A} H(x,y)$$

Now consider 5.8 about the existence of a "saddle point" given the equality of min max and max min. Let

$$\alpha = \max_{y\in B}\min_{x\in A} H(x,y) = \min_{x\in A}\max_{y\in B} H(x,y)$$

Then from

$$y \to \min_{x\in A} H(x,y) \text{ and } x \to \max_{y\in B} H(x,y)$$

being upper semicontinuous and lower semicontinuous respectively, there exist y_0 and x_0 such that

$$\alpha = \min_{x\in A} H(x,y_0) = \max_{y\in B} \overset{\text{minimum of u.s.c}}{\min_{x\in A} H(x,y)} = \min_{x\in A} \overset{\text{maximum of l.s.c.}}{\max_{y\in B} H(x,y)} = \max_{y\in B} H(x_0,y)$$

Then

$$\alpha = \max_{y \in B} H(x_0, y) \geq H(x_0, y_0), \ \alpha = \min_{x \in A} H(x, y_0) \leq H(x_0, y_0)$$

so in fact $\alpha = H(x_0, y_0)$ and from the above equalities,

$$\begin{aligned} H(x_0, y_0) &= \alpha = \min_{x \in A} H(x, y_0) \leq H(x, y_0) \\ H(x_0, y_0) &= \alpha = \max_{y \in B} H(x_0, y) \geq H(x_0, y) \end{aligned}$$

and so $H(x_0, y) \leq H(x_0, y_0) \leq H(x, y_0)$. Thus if the min max condition holds, then there exists a saddle point, namely (x_0, y_0).

Finally suppose there is a saddle point (x_0, y_0) where

$$H(x_0, y) \leq H(x_0, y_0) \leq H(x, y_0)$$

Then

$$\min_{x \in A} \max_{y \in B} H(x, y) \leq \max_{y \in B} H(x_0, y) \leq H(x_0, y_0) \leq \min_{x \in A} H(x, y_0) \leq \max_{y \in B} \min_{x \in A} H(x, y)$$

However, as noted above, it is always the case that

$$\max_{y \in B} \min_{x \in A} H(x, y) \leq \min_{x \in A} \max_{y \in B} H(x, y) \ \blacksquare$$

What was really needed? You needed compactness of A, B and these sets needed to be in a linear space. Of course there needed to be a norm for which $x \to \|x\|$ is strictly convex and lower semicontinuous, so the conditions given above are sufficient but maybe not necessary.

5.12 Exercises

1. Consider the metric space $C([0,T], \mathbb{R}^n)$ 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 $\{\boldsymbol{f}_m\}$ a Cauchy sequence in $C([0,T], \mathbb{R}^n)$, then for each x, you have $\{\boldsymbol{f}_m(x)\}$ a Cauchy sequence in $\mathbb{R}^n$. Recall that this is a complete space. Thus there exists $\boldsymbol{f}(x) = \lim_{m \to \infty} \boldsymbol{f}_m(x)$. You must show that $\boldsymbol{f}$ is continuous. This was in the section on the Ascoli Arzela theorem in more generality if you need an outline of how this goes. Write down the details for this case. Note how $\boldsymbol{f}$ is in bold face. This means it is a function which has values in $\mathbb{R}^n$. $\boldsymbol{f}(t) = (f_1(t), f_2(t), \cdots, f_n(t))$.

2. For $\boldsymbol{f} \in C([0,T], \mathbb{R}^n)$, you define the Riemann integral in the usual way using Riemann sums. Alternatively, you can define it as

$$\int_0^t \boldsymbol{f}(s)\, ds = \left(\int_0^t f_1(s)\, ds, \int_0^t f_2(s)\, ds, \cdots, \int_0^t f_n(s)\, ds \right)$$

Then show that the following limit exists in $\mathbb{R}^n$ for each $t \in (0,T)$.

$$\lim_{h \to 0} \frac{\int_0^{t+h} \boldsymbol{f}(s)\, ds - \int_0^t \boldsymbol{f}(s)\, ds}{h} = \boldsymbol{f}(t).$$

You should use the fundamental theorem of calculus from one variable calculus and the definition of the norm to verify this. As a review, for $\boldsymbol{f}$ defined on an interval $[0,T]$ and $s \in [0,T]$, $\lim_{t \to s} \boldsymbol{f}(t) = \boldsymbol{l}$ means that for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $0 < |t - s| < \delta$, then $\|\boldsymbol{f}(t) - \boldsymbol{l}\|_\infty < \varepsilon$.

3. Suppose $f:\mathbb{R}\to\mathbb{R}$ and $f\geq 0$ on $[-1,1]$ with $f(-1)=f(1)=0$ and $f(x)<0$ for all $x\notin[-1,1]$. Can you use a modification of the proof of the Weierstrass approximation theorem for functions on an interval presented earlier to show that for all $\varepsilon>0$ there exists a polynomial p, such that $|p(x)-f(x)|<\varepsilon$ for $x\in[-1,1]$ and $p(x)\leq 0$ for all $x\notin[-1,1]$?

4. This and the next few problems give an alternative treatment of the Arzella Ascolli theorem of Chapter 3. collection of functions $\mathscr{F}$ of $C([0,T],\mathbb{R}^n)$ is said to be uniformly equicontinuous if for every $\varepsilon>0$ there exists $\delta>0$ such that if $\boldsymbol{f}\in\mathscr{F}$ and $|t-s|<\delta$, then $\|\boldsymbol{f}(t)-\boldsymbol{f}(s)\|_\infty<\varepsilon$. Thus the functions are uniformly continuous all at once. The single δ works for every pair t,s closer together than δ and for all functions $\boldsymbol{f}\in\mathscr{F}$. As an easy case, suppose there exists K such that for all $\boldsymbol{f}\in\mathscr{F}$, $\|\boldsymbol{f}(t)-\boldsymbol{f}(s)\|_\infty\leq K|t-s|$. Show that $\mathscr{F}$ is uniformly equicontinuous. Now suppose $\mathscr{G}$ is a collection of functions of $C([0,T],\mathbb{R}^n)$ which is bounded. That is, $\|\boldsymbol{f}\|=\max_{t\in[0,T]}\|\boldsymbol{f}(t)\|_\infty<M<\infty$ for all $\boldsymbol{f}\in\mathscr{G}$. Then let $\mathscr{F}$ denote the functions which are of the form $\boldsymbol{F}(t)\equiv\boldsymbol{y}_0+\int_0^t\boldsymbol{f}(s)\,ds$ where $\boldsymbol{f}\in\mathscr{G}$. 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 $a<b$ $\left|\int_a^b f(s)\,ds\right|\leq\int_a^b|f(s)|\,ds$. Then $\left\|\int_a^b\boldsymbol{f}(s)\,ds\right\|_\infty=\max_i\left|\int_a^b f_i(s)\,ds\right|\leq\max_i\int_a^b|f_i(s)|\,ds\leq\int_a^b\|\boldsymbol{f}(s)\|_\infty\,ds$. Reduce to the case just considered using the assumption that these $\boldsymbol{f}$ are bounded.

5. Suppose $\mathscr{F}$ is a set of functions in $C([0,T],\mathbb{R}^n)$ which is uniformly bounded and uniformly equicontinuous as described above. Show it must be totally bounded.

6. ↑If $A\subseteq(X,d)$ is totally bounded, show that $\bar{A}$ the closure of A is also totally bounded. In the above problem, explain why $\bar{\mathscr{F}}$ the closure of $\mathscr{F}$ is compact. This uses the big theorem on compactness. Try and do this on your own, but if you get stuck, it is in the section on Arzela Ascoli theorem. When you have done this problem, you have proved the important part of the Arzela Ascoli theorem in the special case where the functions are defined on an interval. You can use this to prove one of the most important results in the theory of differential equations. This theorem is a really profound result because it gives compactness in a normed linear space which is **not finite dimensional.** Thus this is a non trivial generalization of the Heine Borel theorem.

7. 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 with any norm.

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 $\{x_n\}$ be a minimizing sequence. That is $\lambda=\lim_{n\to\infty}|y-x_n|$. Explain why such a minimizing sequence exists. Next explain the following using the parallelogram identity in the above problem as follows.

$$\left|y-\frac{x_n+x_m}{2}\right|^2=\left|\frac{y}{2}-\frac{x_n}{2}+\frac{y}{2}-\frac{x_m}{2}\right|^2$$

$$= -\left|\frac{y}{2} - \frac{x_n}{2} - \left(\frac{y}{2} - \frac{x_m}{2}\right)\right|^2 + \frac{1}{2}|y - x_n|^2 + \frac{1}{2}|y - x_m|^2$$

Hence $\left|\frac{x_m - x_n}{2}\right|^2 = -\left|y - \frac{x_n + x_m}{2}\right|^2 + \frac{1}{2}|y - x_n|^2 + \frac{1}{2}|y - x_m|^2$

$$\leq -\lambda^2 + \frac{1}{2}|y - x_n|^2 + \frac{1}{2}|y - x_m|^2$$

Next explain why the right hand side converges to 0 as $m, n \to \infty$. Thus $\{x_n\}$ 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 Px. Show that Px is characterized as being the solution to the following variational inequality given by $\text{Re}\,(z - Py, y - Py) \leq 0$ for all $z \in K$. That is, show that $x = Py$ if and only if $\text{Re}\,(z - x, y - x) \leq 0$ 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

$$|x + t(z - x) - y|^2 = |x - y|^2 + t^2|z - x|^2 - t2\text{Re}\,(z - x, y - x)$$

If $x = Px$, 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 $\text{Re}\,(z - x, y - x) \leq 0$. Explain why this shows $x = Py$.

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 $|Px - Py| \leq |x - y|$.

11. Suppose, in an inner product space, you know $\text{Re}\,(x, y)$. Show that you also know $\text{Im}\,(x, y)$. That is, give a formula for $\text{Im}\,(x, y)$ in terms of $\text{Re}\,(x, y)$. **Hint:**

$$(x, iy) = -i(x, y) = -i(\text{Re}\,(x, y) + i\text{Im}\,(x, y)) = -i\text{Re}\,(x, y) + \text{Im}\,(x, y)$$

while, by definition, $(x, iy) = \text{Re}\,(x, iy) + i\text{Im}\,(x, iy)$. Now consider matching real and imaginary parts.

12. Let $h > 0$ be given and let $\boldsymbol{f}(t, \boldsymbol{x}) \in \mathbb{R}^n$ for each $\boldsymbol{x} \in \mathbb{R}^n$. Also let $(t, \boldsymbol{x}) \to \boldsymbol{f}(t, \boldsymbol{x})$ be continuous and $\sup_{t,\boldsymbol{x}} \|\boldsymbol{f}(t, \boldsymbol{x})\|_\infty < C < \infty$. Let $\boldsymbol{x}_h(t)$ be a solution to the following

$$\boldsymbol{x}_h(t) = \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}(s, \boldsymbol{x}_h(s - h))\, ds$$

where $\boldsymbol{x}_h(s - h) \equiv \boldsymbol{x}_0$ if $s - h \leq 0$. Explain why there exists a solution. **Hint:** Consider the intervals $[0, h], [h, 2h]$ and so forth. Next explain why these functions $\{\boldsymbol{x}_h\}_{h>0}$ are equicontinuous and uniformly bounded. Now use the result of Problem 6 to argue that there exists a subsequence, still denoted by $\boldsymbol{x}_h$ such that $\lim_{h\to 0} \boldsymbol{x}_h = \boldsymbol{x}$ in $C([0, T]; \mathbb{R}^n)$ as discussed in Problem 5. Use what you learned about the Riemann integral in single variable advanced calculus to explain why you can pass to a limit and conclude that $\boldsymbol{x}(t)=\boldsymbol{x}_0+\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds$ **Hint:**

$$\begin{aligned}
&\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\,ds\right\|_\infty \\
\leq\ &\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s-h))\,ds\right\|_\infty \\
&+\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s-h))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\,ds\right\|_\infty \\
\leq\ &\int_0^T \|\boldsymbol{f}(s,\boldsymbol{x}(s))-\boldsymbol{f}(s,\boldsymbol{x}(s-h))\|_\infty\,ds \\
&+\int_0^T \|\boldsymbol{f}(s,\boldsymbol{x}(s-h))-\boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\|_\infty\,ds
\end{aligned}$$

Now use Problem 2 to verify that $\boldsymbol{x}'=\boldsymbol{f}(t,\boldsymbol{x}),\ \boldsymbol{x}(0)=\boldsymbol{x}_0$. When you have done this, you will have proved the celebrated Peano existence theorem from ordinary differential equations.

13. Let $|\alpha|\equiv\sum_i \alpha_i$. Let $\mathscr{G}$ denote all finite sums of functions of the form $p(\boldsymbol{x})e^{-a|\boldsymbol{x}|^2}$ where $p(\boldsymbol{x})$ is a polynomial and $a>0$. If you consider all real valued continuous functions defined on the closed ball $\overline{B(\mathbf{0},R)}$ show that if f is such a function, then for every $\varepsilon>0$, there exists $g\in\mathscr{G}$ such that $\|f-g\|_\infty<\varepsilon$ where $\|h\|_\infty\equiv\max_{\boldsymbol{x}\in\overline{B(\mathbf{0},R)}}|h(\boldsymbol{x})|$. Thus, from multi-variable calculus, every continuous function f is uniformly close to an infinitely differentiable function on any closed ball centered at $\mathbf{0}$.

14. Suppose now that $f\in C_0(\mathbb{R}^p)$. This means that f is everywhere continuous and that $\lim_{\|\boldsymbol{x}\|\to\infty}|f(\boldsymbol{x})|=0$. Show that for every $\varepsilon>0$ there exists $g\in\mathscr{G}$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^p}|f(\boldsymbol{x})-g(\boldsymbol{x})|<\varepsilon$. Thus you can approximate such a continuous function f uniformly on all of $\mathbb{R}^p$ with a function which has infinitely many continuous partial derivatives. I assume the reader has had a beginning course in multi-variable calculus including partial derivatives. If not, a partial derivative is just a derivative with respect to one of the variables, fixing all the others.

15. In Problem 23 on Page 112, and $V\equiv\operatorname{span}(f_{p_1},...,f_{p_n}), f_r(x)\equiv x^r, x\in[0,1]$ and $-\frac{1}{2}<p_1<p_2<\cdots$ with $\lim_{k\to\infty}p_k=\infty$. The distance between f_m and V is

$$\frac{1}{\sqrt{2m+1}}\prod_{j\leq n}\frac{|m-p_j|}{(p_j+m+1)}=d$$

Let $d_n=d$ so more functions are allowed to be included in V. Show that $\sum_n\frac{1}{p_n}=\infty$ if and only if $\lim_{n\to\infty}d_n=0$. Explain, using the Weierstrass approximation theorem why this shows that if g is a function continuous on $[0,1]$, then there is a function $\sum_{k=1}^N a_k f_{p_k}$ with $\left|g-\sum_{k=1}^N a_k f_{p_k}\right|<\varepsilon$. Here $|g|^2\equiv\int_0^1|g(x)|^2\,dx$. This is Müntz's first theorem. **Hint:** $d_n\to 0$, if and only if $\ln d_n\to-\infty$ so you might want to arrange things so that this happens. You might want to use the fact that for $x\in$

$[0,1/2]\,, -x \geq \ln(1-x) \geq -2x$. See [8] which is where I read this. That product is $\prod_{j\leq n}\left(1-\left(1-\frac{|m-p_j|}{(p_j+m+1)}\right)\right)$ and so ln of this expression is

$$\sum_{j=1}^{n} \ln\left(1-\left(1-\frac{|m-p_j|}{(p_j+m+1)}\right)\right)$$

which is in the interval

$$\left[-2\sum_{j=1}^{n}\left(1-\frac{|m-p_j|}{(p_j+m+1)}\right), -\sum_{j=1}^{n}\left(1-\frac{|m-p_j|}{(p_j+m+1)}\right)\right]$$

and so $d_n \to 0$ if and only if $\sum_{j=1}^{\infty}\left(1-\frac{|m-p_j|}{(p_j+m+1)}\right) = \infty$. Since $p_n \to \infty$ it suffices to consider the convergence of $\sum_j \left(1-\frac{p_j-m}{(p_j+m+1)}\right) = \sum_j \left(\frac{2m+1}{(p_j+m+1)}\right)$. Now recall theorems from calculus.

16. For $f \in C([a,b];\mathbb{R})$, real valued continuous functions, let $|f| \equiv \left(\int_a^b |f(t)|^2\right)^{1/2} \equiv (f,f)^{1/2}$ where $(f,g) \equiv \int_a^b f(x)g(x)\,dx$. Recall the Cauchy Schwarz inequality $|(f,g)| \leq |f|\,|g|$. Now suppose $\frac{1}{2} < p_1 < p_2 \cdots$ where $\lim_{k\to\infty} p_k = \infty$. Let $V_n = \text{span}(1, f_{p_1}, f_{p_2}, ..., f_{p_n})$. For $\|\cdot\|$ the uniform approximation norm, show that for every $g \in C([0,1])$, there exists there exists a sequence of functions, $f_n \in V_n$ such that $\|g - f_n\| \to 0$. This is the second Müntz theorem. **Hint:** Show that you can approximate $x \to x^m$ uniformly. To do this, use the above Müntz to approximate mx^{m-1} with $\sum_k c_k x^{p_k-1}$ in the inner product norm. $\int_0^1 \left|mx^{m-1} - \sum_{k=1}^n c_k x^{p_k-1}\right|^2 dx \leq \varepsilon^2$. Then $x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k} = \int_0^x \left(mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1}\right) dt$. Then

$$\left|x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k}\right| \leq \int_0^x \left|mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1}\right| dt \leq \int_0^1 1 \left|mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1}\right| dt$$

Now use the Cauchy Schwarz inequality on that last integral to obtain

$$\max_{x\in[0,1]} \left|x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k}\right| \leq \varepsilon.$$

In case $m = 0$, there is nothing to show because 1 is in V_n. Explain why the result follows from this and the Weierstrass approximation theorem.

17. Suppose $f : [a,b] \to [0,1]$ is piecewise linear, equal to 1 on $[a+h, b-h]$ and 0 at a, b. Show that $\int_a^b f(x)\,dx = h + (b-a-2h) = b-a-h$.

Chapter 6

The Derivative

6.1 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 6.1.1 *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 6.1.2 *Let $\boldsymbol{f} : D(\boldsymbol{f}) \subseteq V \to W$ be a function and let $\boldsymbol{x}$ be a **limit point** of $D(\boldsymbol{f})$. Then*

$$\lim_{\boldsymbol{y} \to \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < \|\boldsymbol{y} - \boldsymbol{x}\| < \delta, \text{ and } \boldsymbol{y} \in D(\boldsymbol{f})$$

then,

$$\|\boldsymbol{L} - \boldsymbol{f}(\boldsymbol{y})\| < \varepsilon.$$

Theorem 6.1.3 *If $\lim_{\boldsymbol{y} \to \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}$ and $\lim_{y \to x} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}_1$, 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 $\boldsymbol{y} \in D(\boldsymbol{f})$, then $\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{L}\| < \varepsilon$, $\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{L}_1\| < \varepsilon$. Pick such a $\boldsymbol{y}$. There exists one because $\boldsymbol{x}$ is a limit point of $D(\boldsymbol{f})$. Then $\|\boldsymbol{L} - \boldsymbol{L}_1\| \leq \|\boldsymbol{L} - \boldsymbol{f}(\boldsymbol{y})\| + \|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{L}_1\| < \varepsilon + \varepsilon = 2\varepsilon$. Since $\varepsilon > 0$ was arbitrary, this shows $\boldsymbol{L} = \boldsymbol{L}_1$. ■

One can define what it means for $\lim_{\boldsymbol{y} \to \boldsymbol{x}} f(\boldsymbol{x}) = \pm\infty$. as in the case of real valued functions.

Definition 6.1.4 *If $f(\boldsymbol{x}) \in \mathbb{R}$, $\lim_{\boldsymbol{y} \to \boldsymbol{x}} f(\boldsymbol{x}) = \infty$ if for every number l, there exists $\delta > 0$ such that whenever $\|\boldsymbol{y} - \boldsymbol{x}\| < \delta$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $f(\boldsymbol{x}) > l$. Also the assertion that $\lim_{\boldsymbol{y} \to \boldsymbol{x}} f(\boldsymbol{x}) = -\infty$ means that for every number l, there exists $\delta > 0$ such that whenever $\|\boldsymbol{y} - \boldsymbol{x}\| < \delta$ 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 6.1.5 *Suppose $\boldsymbol{f} : D(\boldsymbol{f}) \subseteq V \to \mathbb{F}^m$. Then for $\boldsymbol{x}$ a limit point of $D(\boldsymbol{f})$,*

$$\lim_{y \to x} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L} \tag{6.1}$$

if and only if

$$\lim_{y \to x} f_k(\boldsymbol{y}) = L_k \tag{6.2}$$

where $\boldsymbol{f}(\boldsymbol{y}) \equiv (f_1(\boldsymbol{y}), \cdots, f_p(\boldsymbol{y}))$ and $\boldsymbol{L} \equiv (L_1, \cdots, L_p)$.

Suppose here that f has values in W, a normed linear space and

$$\lim_{y \to \boldsymbol{x}} f(y) = L, \ \lim_{y \to \boldsymbol{x}} g(y) = K$$

where K,$L \in W$. Then if a, $b \in \mathbb{F}$,

$$\lim_{y\to x}\left(af\left(y\right)+bg\left(y\right)\right)=aL+bK, \tag{6.3}$$

If W is an inner product space,

$$\lim_{y\to x}\left(f,g\right)\left(y\right)=\left(L,K\right) \tag{6.4}$$

If g is scalar valued with $\lim_{y\to x}g\left(y\right)=K$,

$$\lim_{y\to x}f\left(y\right)g\left(y\right)=LK. \tag{6.5}$$

Also, if h is a continuous function defined near L, then

$$\lim_{y\to x}h\circ f\left(y\right)=h\left(L\right). \tag{6.6}$$

Suppose $\lim_{y\to \boldsymbol{x}}f\left(y\right)=L$. *If* $\left\Vert f\left(y\right)-b\right\Vert \leq r$ *for all y sufficiently close to* $\boldsymbol{x}$, *then* $\left\vert L-b\right\vert \leq r$ *also.*

Proof: Suppose 6.1. Then letting $\varepsilon>0$ be given there exists $\delta>0$ such that if $0<\left\Vert y-x\right\Vert <\delta$, it follows

$$\left\vert f_k\left(y\right)-L_k\right\vert \leq\left\Vert \boldsymbol{f}\left(y\right)-\boldsymbol{L}\right\Vert <\varepsilon$$

which verifies 6.2.

Now suppose 6.2 holds. Then letting $\varepsilon>0$ be given, there exists δ_k such that if $0<\left\Vert y-x\right\Vert <\delta_k$, then $\left\vert f_k\left(y\right)-L_k\right\vert <\varepsilon$. Let $0<\delta<\min\left(\delta_1,\cdots,\delta_p\right)$. Then if $0<\left\Vert y-x\right\Vert <\delta$, it follows $\left\Vert \boldsymbol{f}\left(y\right)-\boldsymbol{L}\right\Vert_\infty<\varepsilon$. 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 6.3 is left for you. Now 6.4 is to be verified. Let $\varepsilon>0$ be given. Then by the triangle inequality,

$$\begin{aligned}\left\vert \left(f,g\right)\left(y\right)-\left(L,K\right)\right\vert &\leq\left\vert \left(f,g\right)\left(y\right)-\left(f\left(y\right),K\right)\right\vert +\left\vert \left(f\left(y\right),K\right)-\left(L,K\right)\right\vert \\ &\leq\left\Vert f\left(y\right)\right\Vert \left\Vert g\left(y\right)-K\right\Vert +\left\Vert K\right\Vert \left\Vert f\left(y\right)-L\right\Vert .\end{aligned}$$

There exists δ_1 such that if $0<\left\Vert y-x\right\Vert <\delta_1$ and $y\in D\left(f\right)$, then $\left\Vert f\left(y\right)-L\right\Vert <1$,and so for such y, the triangle inequality implies, $\left\Vert f\left(y\right)\right\Vert <1+\left\Vert L\right\Vert$. Therefore, for $0<\left\Vert y-x\right\Vert <\delta_1$,

$$\left\vert \left(f,g\right)\left(y\right)-\left(L,K\right)\right\vert \leq\left(1+\left\Vert K\right\Vert +\left\Vert L\right\Vert \right)\left[\left\Vert g\left(y\right)-K\right\Vert +\left\Vert f\left(y\right)-L\right\Vert \right]. \tag{6.7}$$

Now let $0<\delta_2$ be such that if $y\in D\left(f\right)$ and $0<\left\Vert x-y\right\Vert <\delta_2$,

$$\left\Vert f\left(y\right)-L\right\Vert <\frac{\varepsilon}{2\left(1+\left\Vert K\right\Vert +\left\Vert L\right\Vert \right)},\ \left\Vert g\left(y\right)-K\right\Vert <\frac{\varepsilon}{2\left(1+\left\Vert K\right\Vert +\left\Vert L\right\Vert \right)}.$$

Then letting $0<\delta\leq\min\left(\delta_1,\delta_2\right)$, it follows from 6.7 that $\left\vert \left(f,g\right)\left(y\right)-\left(L,K\right)\right\vert <\varepsilon$ and this proves 6.4.

The proof of 6.5 is left to you.

Consider 6.6. 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 $\lim_{y \to x} f(y) = L$, 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

$$r < \|L - b\| - \|f(y) - L\| \leq |\|b - L\| - \|f(y) - L\|| \leq \|b - f(y)\|,$$

a contradiction to the assumption that $\|b - f(y)\| \leq r$. ■

The relation between continuity and limits is as follows.

Theorem 6.1.6 *For $f : D(f) \to W$ and $x \in D(f)$ a limit point of $D(f)$, f is continuous at x if and only if* $\lim_{y \to x} f(y) = f(x)$.

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 $f(x) = \lim_{y \to x} f(y)$.

Next suppose x is a limit point of $D(f)$ and $\lim_{y \to x} f(y) = f(x)$. 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 6.1.7 *Find* $\lim_{(x,y) \to (3,1)} \left(\frac{x^2 - 9}{x - 3}, y \right)$.

It is clear that $\lim_{(x,y) \to (3,1)} \frac{x^2 - 9}{x - 3} = 6$ and $\lim_{(x,y) \to (3,1)} y = 1$. Therefore, this limit equals $(6, 1)$.

Example 6.1.8 *Find* $\lim_{(x,y) \to (0,0)} \frac{xy}{x^2 + y^2}$.

First of all, observe the domain of the function is $\mathbb{R}^2 \setminus \{(0,0)\}$, 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.

6.2 Basic Definitions

The concept of derivative generalizes right away to functions of many variables. However, no attempt will be made to consider derivatives from one side or another. This is because when you consider functions of many variables, there isn't a well defined side. However, it is certainly the case that there are more general notions which include such things. I will present a fairly general notion of the derivative of a function which is defined on a normed vector space which has values in a normed vector space. The case of most interest is that of a function which maps $\mathbb{F}^n$ to $\mathbb{F}^m$ but it is no more trouble to consider the extra generality and it is sometimes useful to have this extra generality because sometimes you want to consider functions defined, for example on subspaces of $\mathbb{F}^n$and it is nice to not have to trouble with ad hoc considerations. Also, you might want to consider $\mathbb{F}^n$ with some norm other than the usual one.

In what follows, X,Y will denote normed vector spaces. Thanks to Theorem 5.2.4 all the definitions and theorems given below work the same for any norm given on the vector spaces.

Let U be an open set in X, and let $\boldsymbol{f} : U \rightarrow Y$ be a function.

Definition 6.2.1 *A function $\boldsymbol{g}$ is $\boldsymbol{o}(\boldsymbol{v})$ if*

$$\lim_{\|\boldsymbol{v}\|\to 0} \frac{\boldsymbol{g}(\boldsymbol{v})}{\|\boldsymbol{v}\|} = \boldsymbol{0} \tag{6.8}$$

A function $\boldsymbol{f} : U \rightarrow Y$ is differentiable at $\boldsymbol{x} \in U$ if there exists a linear transformation $L \in \mathscr{L}(X,Y)$ such that

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})$$

This linear transformation L is the definition of $D\boldsymbol{f}(\boldsymbol{x})$. This derivative is often called the Frechet derivative.

Note that from Theorem 5.2.4 the question whether a given function is differentiable is independent of the norm used on the finite dimensional vector space. That is, a function is differentiable with one norm if and only if it is differentiable with another norm.

The definition 6.8 means the error $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) - L\boldsymbol{v}$ converges to $\boldsymbol{0}$ faster than $\|\boldsymbol{v}\|$. Thus the above definition is equivalent to saying

$$\lim_{\|\boldsymbol{v}\|\to 0} \frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) - L\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = 0 \tag{6.9}$$

or equivalently,

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}} \frac{\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|}{\|\boldsymbol{y}-\boldsymbol{x}\|} = 0. \tag{6.10}$$

The symbol, $\boldsymbol{o}(\boldsymbol{v})$ should be thought of as an adjective. Thus, if t and k are constants,

$$\boldsymbol{o}(\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v}) + \boldsymbol{o}(\boldsymbol{v}),\ \boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v}),\ k\boldsymbol{o}(\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v})$$

and other similar observations hold.

Theorem 6.2.2 *The derivative is well defined.*

Proof: First note that for a fixed nonzero vector $\boldsymbol{v}$, $\boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(t)$. This is because

$$\lim_{t\to 0} \frac{\boldsymbol{o}(t\boldsymbol{v})}{|t|} = \lim_{t\to 0} \|\boldsymbol{v}\| \frac{\boldsymbol{o}(t\boldsymbol{v})}{\|t\boldsymbol{v}\|} = \boldsymbol{0}$$

Now suppose both L_1 and L_2 work in the above definition. Then let $\boldsymbol{v}$ be any vector and let t be a real scalar which is chosen small enough that $t\boldsymbol{v}+\boldsymbol{x} \in U$. Then

$$\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L_1 t\boldsymbol{v} + \boldsymbol{o}(t\boldsymbol{v}),\ \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L_2 t\boldsymbol{v} + \boldsymbol{o}(t\boldsymbol{v}).$$

Therefore, subtracting these two yields $(L_2 - L_1)(t\boldsymbol{v}) = \boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(t)$. Therefore, dividing by t yields $(L_2 - L_1)(\boldsymbol{v}) = \frac{\boldsymbol{o}(t)}{t}$. Now let $t \to 0$ to conclude that $(L_2 - L_1)(\boldsymbol{v}) = 0$. Since this is true for all $\boldsymbol{v}$, it follows $L_2 = L_1$. This proves the theorem. ■

In the following lemma, $\|D\boldsymbol{f}(\boldsymbol{x})\|$ is the operator norm of the linear transformation, $D\boldsymbol{f}(\boldsymbol{x})$.

Lemma 6.2.3 *Let $\boldsymbol{f}$ be differentiable at $\boldsymbol{x}$. Then $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$ and in fact, there exists $K > 0$ such that whenever $\|\boldsymbol{v}\|$ is small enough,*

$$\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \leq K\|\boldsymbol{v}\|$$

Also if $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, then

$$\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|) = \boldsymbol{o}(\boldsymbol{v})$$

Proof: From the definition of the derivative,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}).$$

Let $\|\boldsymbol{v}\|$ be small enough that $\frac{\boldsymbol{o}(\|\boldsymbol{v}\|)}{\|\boldsymbol{v}\|} < 1$ so that $\|\boldsymbol{o}(\boldsymbol{v})\| \leq \|\boldsymbol{v}\|$. Then for such $\boldsymbol{v}$,

$$\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \leq \|D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}\| + \|\boldsymbol{v}\| \leq (\|D\boldsymbol{f}(\boldsymbol{x})\| + 1)\|\boldsymbol{v}\|$$

This proves the lemma with $K = \|D\boldsymbol{f}(\boldsymbol{x})\| + 1$. Recall the operator norm discussed in Definition 5.2.2.

The last assertion is implied by the first as follows. Define

$$\boldsymbol{h}(\boldsymbol{v}) \equiv \begin{cases} \frac{\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\|)}{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\|} \text{ if } \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \neq 0 \\ \boldsymbol{0} \text{ if } \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| = 0 \end{cases}$$

Then $\lim_{\|\boldsymbol{v}\|\to 0} \boldsymbol{h}(\boldsymbol{v}) = \boldsymbol{0}$ from continuity of $\boldsymbol{f}$ at $\boldsymbol{x}$ which is implied by the first part. Also from the above estimate, if $\|\boldsymbol{v}\|$ is sufficiently small,

$$\left\| \frac{\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|)}{\|\boldsymbol{v}\|} \right\| = \|\boldsymbol{h}(\boldsymbol{v})\| \frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|}{\|\boldsymbol{v}\|} \leq \|\boldsymbol{h}(\boldsymbol{v})\| (\|D\boldsymbol{f}(\boldsymbol{x})\| + 1)$$

and $\lim_{\|\boldsymbol{v}\|\to 0} \|\boldsymbol{h}(\boldsymbol{v})\| = 0$. This establishes the second claim. ■

6.3 The Chain Rule

With the above lemma, it is easy to prove the chain rule.

Theorem 6.3.1 *(The chain rule) Let U and V be open sets $U \subseteq X$ and $V \subseteq Y$. Suppose $\boldsymbol{f} : U \to V$ is differentiable at $\boldsymbol{x} \in U$ and suppose $\boldsymbol{g} : V \to \mathbb{F}^q$ is differentiable at $\boldsymbol{f}(\boldsymbol{x}) \in V$. Then $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ and*

$$D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))D\boldsymbol{f}(\boldsymbol{x}).$$

Proof: This follows from a computation. Let $B(\boldsymbol{x},r) \subseteq U$ and let r also be small enough that for $\|\boldsymbol{v}\| \leq r$, it follows that $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) \in V$. Such an r exists because $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$. For $\|\boldsymbol{v}\| < r$, the definition of differentiability of $\boldsymbol{g}$ and $\boldsymbol{f}$ implies

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) =$$

$$\begin{aligned}
& D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \\
= \; & D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))[D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})] + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \\
= \; & D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}) + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \qquad (6.11) \\
= \; & D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})
\end{aligned}$$

By Lemma 6.2.3. From the definition of the derivative $D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x})$ exists and equals $D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))$. ■

6.4 The Matrix of the Derivative

The case of most interest here is the only one I will discuss. It is the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, the function being defined on an open subset of $\mathbb{R}^n$. Of course this all generalizes to arbitrary vector spaces and one considers the matrix taken with respect to various bases. However, I am going to restrict to the case just mentioned here. As above, $\boldsymbol{f}$ will be defined and differentiable on an open set $U \subseteq \mathbb{R}^n$.

As discussed in the review material on linear maps, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is the matrix having the i^{th} column equal to $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{e}_i$ and so it is only necessary to compute this. Let t be a small real number such that

$$\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x})(t\boldsymbol{e}_i)}{t} = \frac{\boldsymbol{o}(t)}{t}$$

Therefore,

$$\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x})}{t} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{e}_i) + \frac{\boldsymbol{o}(t)}{t}$$

The limit exists on the right and so it exists on the left also. Thus

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} \equiv \lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x})}{t} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{e}_i)$$

and so the matrix of the derivative is just the matrix which has the i^{th} column equal to the i^{th} partial derivative of $\boldsymbol{f}$. Note that this shows that whenever $\boldsymbol{f}$ is differentiable, it follows that the partial derivatives all exist. It does not go the other way however as discussed later.

Theorem 6.4.1 *Let $\boldsymbol{f} : U \subseteq \mathbb{F}^n \to \mathbb{F}^m$ and suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$. Then all the partial derivatives $\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}$ exist and if $J\boldsymbol{f}(\boldsymbol{x})$ is the matrix of the linear transformation, $D\boldsymbol{f}(\boldsymbol{x})$ with respect to the standard basis vectors, then the ij^{th} entry is given by $\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$ also denoted as $f_{i,j}$ or f_{i,x_j}. It is the matrix whose i^{th} column is*

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} \equiv \lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x})}{t}.$$

Of course there is a generalization of this idea called the directional derivative.

Definition 6.4.2 *In general, the symbol $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is defined by*

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})}{t}$$

where $t \in \mathbb{F}$. In case $|\boldsymbol{v}| = 1, \mathbb{F} = \mathbb{R}$, and the norm is the standard Euclidean norm, this is called the directional derivative. More generally, with no restriction on the size of $\boldsymbol{v}$ and in any linear space, it is called the Gateaux derivative. $\boldsymbol{f}$ is said to be Gateaux differentiable at $\boldsymbol{x}$ if there exists $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ such that

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})}{t} = D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$$

where $\boldsymbol{v} \to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is linear. Thus we say it is Gateaux differentiable if the Gateaux derivative exists for each $\boldsymbol{v}$ and $\boldsymbol{v} \to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is linear. Note that $\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} = D_{\boldsymbol{e}_i}\boldsymbol{f}(\boldsymbol{x})$. [1]

Here is an interesting application which is used a lot in introductory courses on multivariable calculus.

Theorem 6.4.3 *Suppose U is an open set in a normed linear space X and $f : U \to \mathbb{R}$ has a Gateaux derivative $D_{\boldsymbol{v}}f$ at $\boldsymbol{x} \in U$ and that for all $\hat{\boldsymbol{x}}$ sufficiently close to $\boldsymbol{x}$, on the line through $\boldsymbol{x}$ having direction vector $\boldsymbol{v}$ it follows that $f(\boldsymbol{x}) \geq f(\hat{\boldsymbol{x}})\ (f(\boldsymbol{x}) \leq f(\hat{\boldsymbol{x}}))$. In other words, f has a local maximum/minimum at $\boldsymbol{x}$ when restricted to the line $t \to \boldsymbol{x}+t\boldsymbol{v}$, then $D_{\boldsymbol{v}}f(\boldsymbol{x}) = 0$. If $Df(\boldsymbol{x})$ exists and f has a local max/min at $\boldsymbol{x}$ for all $\boldsymbol{v}$, then $Df(\boldsymbol{x}) = 0$.*

Proof: Consider $h(t) = f(\boldsymbol{x}+t\boldsymbol{v})$. Then from single variable calculus,

$$h'(0) = D_{\boldsymbol{v}}f(\boldsymbol{x}) = 0.$$

In case f is differentiable, then for every $\boldsymbol{v}$,

$$0 = D_{\boldsymbol{v}}f(\boldsymbol{x}) = \lim_{t\to 0} \frac{f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x})}{t} = \lim_{t\to 0} \frac{Df(\boldsymbol{x})(t\boldsymbol{v}) + o(t)}{t} = Df(\boldsymbol{x})\boldsymbol{v}$$

Since this holds for every $\boldsymbol{v}$ it follows that $Df(\boldsymbol{x}) = 0$. ■

What if all the partial derivatives of $\boldsymbol{f}$ exist? Does it follow that $\boldsymbol{f}$ is differentiable? Consider the following function, $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}.$$

[1] René Gateaux was one of the many young French men killed in world war I. This derivative is named after him, but it developed naturally from ideas used in the calculus of variations which were due to Euler and Lagrange back in the 1700's.

Then from the definition of partial derivatives,

$$\lim_{h\to 0}\frac{f(h,0)-f(0,0)}{h}=\lim_{h\to 0}\frac{0-0}{h}=0$$

and

$$\lim_{h\to 0}\frac{f(0,h)-f(0,0)}{h}=\lim_{h\to 0}\frac{0-0}{h}=0$$

However f is not even continuous at $(0,0)$ which may be seen by considering the behavior of the function along the line $y=x$ and along the line $x=0$. By Lemma 6.2.3 this implies f is not differentiable. Therefore, it is necessary to consider the correct definition of the derivative given above if you want to get a notion which generalizes the concept of the derivative of a function of one variable in such a way as to preserve continuity whenever the function is differentiable.

What if the one dimensional derivative in the definition of the Gateaux derivative exists for all nonzero $\boldsymbol{v}$? Is the function differentiable then? Maybe not. See Problem 11 in the exercises for example.

6.5 A Mean Value Inequality

The following theorem will be very useful in much of what follows. It is a version of the mean value theorem as is the next lemma. The mean value theorem depends on the function having values in $\mathbb{R}$ and in the lemma and theorem, it has values in a normed vector space.

Lemma 6.5.1 *Let Y be a normed vector space and suppose $\boldsymbol{h}:[0,1]\to Y$ is continuous and differentiable from the right and satisfies $\left\|\boldsymbol{h}'(t)\right\|\leq M$, $M\geq 0$. Then $\|\boldsymbol{h}(1)-\boldsymbol{h}(0)\|\leq M$.*

Proof: Let $\varepsilon>0$ be given and let

$$S\equiv\{t\in[0,1]:\text{ for all } s\in[0,t],\|\boldsymbol{h}(s)-\boldsymbol{h}(0)\|\leq(M+\varepsilon)s\}$$

Then $0\in S$. Let $t=\sup S$. Then by continuity of $\boldsymbol{h}$ it follows

$$\|\boldsymbol{h}(t)-\boldsymbol{h}(0)\|=(M+\varepsilon)t \tag{6.12}$$

Suppose $t<1$. Then there exist positive numbers, h_k decreasing to 0 such that

$$\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(0)\|>(M+\varepsilon)(t+h_k)$$

and now it follows from 6.12 and the triangle inequality that

$$\begin{aligned}&\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|+\|\boldsymbol{h}(t)-\boldsymbol{h}(0)\|\\=&\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|+(M+\varepsilon)t>(M+\varepsilon)(t+h_k)\end{aligned}$$

Thus

$$\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|>(M+\varepsilon)h_k$$

Now dividing by h_k and letting $k\to\infty$, $\left\|\boldsymbol{h}'(t)\right\|\geq M+\varepsilon$,a contradiction. Thus $t=1$. Since ε is arbitrary, the conclusion of the lemma follows. ■

Theorem 6.5.2 *Suppose U is an open subset of X and $\boldsymbol{f} : U \to Y$ has the property that $D\boldsymbol{f}\ (\boldsymbol{x})$ exists for all $\boldsymbol{x}$ in U and that, $\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}) \in U$ for all $t \in [0,1]$. (The line segment joining the two points lies in U.) Suppose also that for all points on this line segment, $\|D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))\| \leq M$. Then $\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{f}(\boldsymbol{x})\| \leq M|\boldsymbol{y}-\boldsymbol{x}|$. More generally if $\|D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{y})\| \leq M$ for all $\boldsymbol{y}$ on the segment joining $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{v}$, then $\|\boldsymbol{f}(\boldsymbol{x}+a\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\| \leq Ma$. Also $D_{a\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x}) = aD\boldsymbol{f}(\boldsymbol{x})$ if $a \neq 0$.*

Proof: Let $\boldsymbol{h}(t) \equiv \boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))$.Then by the chain rule applied to $\boldsymbol{h}(t)$, $\boldsymbol{h}'(t) = D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})$ and so

$$\left\|\boldsymbol{h}'(t)\right\| = \|D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})\| \leq M\|\boldsymbol{y}-\boldsymbol{x}\|$$

by Lemma 6.5.1, $\|\boldsymbol{h}(1)-\boldsymbol{h}(0)\| = ||\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{f}(\boldsymbol{x})|| \leq M||\boldsymbol{y}-\boldsymbol{x}||$. For the second part, let $\boldsymbol{h}(t) \equiv \boldsymbol{f}(\boldsymbol{x}+ta\boldsymbol{v})$. Then

$$\begin{aligned} \boldsymbol{h}'(t) &= \lim_{h\to 0}\frac{\boldsymbol{h}(t+h)-\boldsymbol{h}(t)}{h} \equiv \lim_{h\to 0}\frac{a}{ha}(\boldsymbol{f}(\boldsymbol{x}+ta\boldsymbol{v}+ha\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x}+ta\boldsymbol{v})) \\ &= D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x}+ta\boldsymbol{v})a. \end{aligned}$$

This shows that $D_{a\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x}) = aD_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$. Now for the inequality, there is nothing to show if $a = 0$ so assume $a \neq 0$. Then by assumption and Lemma 6.5.1, $\|\boldsymbol{h}(1)-\boldsymbol{h}(0)\| = \|\boldsymbol{f}(\boldsymbol{x}+a\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\| \leq Ma$. ■

6.6 Existence of the Derivative, C^1 Functions

There is a way to get the differentiability of a function from the existence and continuity of one dimensional directional derivatives. The following theorem is the main result. It gives easy to verify one dimensional conditions for the existence of the derivative. The meaning of $\|\cdot\|$ will be determined by context in what follows. This theorem says that if the Gateaux derivatives exist for each vector in a basis and they are also continuous, then the function is differentiable.

Theorem 6.6.1 *Let X be a normed vector space having basis $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$ and let Y be another normed vector space. Let U be an open set in X and let $\boldsymbol{f} : U \to Y$ have the property that the one dimensional limits*

$$D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x}) \equiv \lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}_k)-\boldsymbol{f}(\boldsymbol{x})}{t}$$

exist and $\boldsymbol{x} \to D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ are continuous functions of $\boldsymbol{x} \in U$ as functions with values in Y. Then $D\boldsymbol{f}(\boldsymbol{x})$ exists and

$$D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} = \sum_{k=1}^{n} D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k$$

where $\boldsymbol{v} = \sum_{k=1}^{n} a_k\boldsymbol{v}_k$. Furthermore, $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous; that is

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}}\|D\boldsymbol{f}(\boldsymbol{y})-D\boldsymbol{f}(\boldsymbol{x})\| = 0.$$

Proof: Let $\boldsymbol{v} = \sum_{k=1}^{n} a_k\boldsymbol{v}_k$ where all a_k are small enough that for all $k \geq 0$,

$$\boldsymbol{x}+\sum_{j=1}^{k} a_j\boldsymbol{v}_j \in \overline{B(\boldsymbol{x},r)} \subseteq U, \sum_{k=1}^{0} a_k\boldsymbol{v}_k \equiv 0.$$

The mapping $\boldsymbol{v} \to (a_1,...,a_n)$ is an isomorphism of V and $\mathbb{F}^n$ and we can define a norm as $\sum_k |a_k|$ which is equivalent to the norm on V thanks to Theorem 5.2.4. Let $\boldsymbol{h}_k(\boldsymbol{x}) \equiv \boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j \boldsymbol{v}_j\right) - \boldsymbol{f}(\boldsymbol{x})$. Then collecting the terms,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = \sum_{k=1}^{n} (\boldsymbol{h}_k(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{h}_k(\boldsymbol{x})) + \sum_{k=1}^{n} (\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})) \tag{6.13}$$

Using Theorem 6.5.2,

$$\begin{aligned} \left\| D_{a_k\boldsymbol{v}_k}\boldsymbol{h}_k(\boldsymbol{x}+ta_k\boldsymbol{v}_k) \right\| &= \left\| a_k D_{\boldsymbol{v}_k}\boldsymbol{h}_k(\boldsymbol{x}+ta_k\boldsymbol{v}_k) \right\| \\ &= \left\| a_k \left(D_{\boldsymbol{v}_k}\boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j\boldsymbol{v}_j + ta_k\boldsymbol{v}_k\right) - D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x}+ta_k\boldsymbol{v}_k)\right)\right\| \\ &\leq C\|\boldsymbol{v}\|\varepsilon \end{aligned}$$

provided $\|\boldsymbol{v}\|$ is sufficiently small, thanks to the assumption that the $D_{\boldsymbol{v}_k}\boldsymbol{f}$ are continuous. It follows, since ε is arbitrary that the first sum on the right in 6.13 is $\boldsymbol{o}(\boldsymbol{v})$. Now

$$(\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})) - D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k =$$

$$\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \left(\boldsymbol{f}(\boldsymbol{x}) + D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k\right) = a_k\left(\frac{\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})}{a_k} - D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\right) = \boldsymbol{o}(\boldsymbol{v})$$

because

$$\begin{aligned} &\left\| a_k\left(\frac{\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})}{a_k} - D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\right)\right\| \\ \leq\ &\|\boldsymbol{v}\| \left\| \left(\frac{\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})}{a_k} - D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\right)\right\|. \end{aligned}$$

Collecting terms in 6.13,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{o}(\boldsymbol{v}) + \sum_{k=1}^{n} (\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{o}(\boldsymbol{v}) + \sum_{k=1}^{n} D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k$$

which shows that $D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{v}) = \sum_{k=1}^{n} D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k$ where $\boldsymbol{v} = \sum_{k=1}^{n} a_k\boldsymbol{v}_k$. This formula also shows that $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous because of the continuity of these $D_{\boldsymbol{v}_k}\boldsymbol{f}$. ■

Note how if $X = \mathbb{R}^p$ and the basis vectors are the $\boldsymbol{e}_k$, then the $\boldsymbol{a}$ are just the components of the vector $\boldsymbol{v}$ taken with respect to the usual basis vectors. Thus this gives the above result about the matrix of $D\boldsymbol{f}(\boldsymbol{x})$.

This motivates the following definition of what it means for a function to be C^1.

Definition 6.6.2 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f} : U \to Y$ another finite dimensional normed vector space. Then $\boldsymbol{f}$ is said to be C^1 if there exists a basis for $X, \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ such that the Gateaux derivatives,$D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ exist on U and are continuous functions of $\boldsymbol{x}$.*

Note that as a special case where $X = \mathbb{R}^n$, you could let the $\boldsymbol{v}_k = \boldsymbol{e}_k$ and the condition would reduce to nothing more than a statement that the partial derivatives $\frac{\partial \boldsymbol{f}}{\partial x_i}$ are all continuous. If $X = \mathbb{R}$, this is not a very interesting condition. It would say the derivative exists if the derivative exists and is continuous.

Here is another definition of what it means for a function to be C^1.

Definition 6.6.3 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f}: U \to Y$ another finite dimensional normed vector space. Then $\boldsymbol{f}$ is said to be C^1 if $\boldsymbol{f}$ is differentiable and $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous as a map from U to $\mathscr{L}(X,Y)$.*

Now the following major theorem states these two definitions are equivalent. This is obviously so in the special case where $X = \mathbb{R}^n$ and the special basis is the usual one because, as observed earlier, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is just the one which has for its columns the partial derivatives which are given to be continuous.

Theorem 6.6.4 *Let U be an open subset of a normed finite dimensional vector space X and let $\boldsymbol{f}: U \to Y$ another finite dimensional normed vector space. Then the two definitions above are equivalent.*

Proof: It was shown in Theorem 6.6.1, the one about the continuity of the Gateaux derivatives yielding differentiability, that Definition 6.6.2 implies 6.6.3. Suppose then that Definition 6.6.3 holds. Then if $\boldsymbol{v}$ is any vector,

$$\lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})}{t}=\lim_{t\to 0}\frac{D\boldsymbol{f}(\boldsymbol{x})t\boldsymbol{v}+\boldsymbol{o}(t\boldsymbol{v})}{t}=D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}+\lim_{t\to 0}\frac{\boldsymbol{o}(t\boldsymbol{v})}{t}=D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}$$

Thus $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ exists and equals $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}$. By continuity of $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$, this establishes continuity of $\boldsymbol{x} \to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ and proves the theorem. ■

Note that the proof of the theorem also implies the following corollary.

Corollary 6.6.5 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f}: U \to Y$ another finite dimensional normed vector space. Then if there is a basis of X, $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ such that the Gateaux derivatives, $D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ exist and are continuous, then all Gateaux derivatives, $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ exist and are continuous for all $\boldsymbol{v} \in X$. Also $D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{v}) = D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$.*

From now on, whichever definition is more convenient will be used.

6.7 Higher Order Derivatives

If $f: U \subseteq X \to Y$ for U an open set, then $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is a mapping from U to $\mathscr{L}(X,Y)$, a normed vector space. Therefore, it makes perfect sense to ask whether this function is also differentiable.

Definition 6.7.1 *The following is the definition of the second derivative. $D^2\boldsymbol{f}(\boldsymbol{x}) \equiv D(D\boldsymbol{f}(\boldsymbol{x}))$.*

Thus, $D\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - D\boldsymbol{f}(\boldsymbol{x}) = D^2\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})$.This implies

$$D^2\boldsymbol{f}(\boldsymbol{x}) \in \mathscr{L}(X, \mathscr{L}(X,Y)),\ D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v}) \in Y,$$

and the map $(\boldsymbol{u},\boldsymbol{v}) \to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$ is a bilinear map having values in Y. In other words, the two functions,

$$\boldsymbol{u} \to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v}),\ \boldsymbol{v} \to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$$

are both linear.

The same pattern applies to taking higher order derivatives. For example, $D^3 \boldsymbol{f}(\boldsymbol{x}) \equiv D\left(D^2 \boldsymbol{f}(\boldsymbol{x})\right)$ and $D^3 \boldsymbol{f}(\boldsymbol{x})$ may be considered as a trilinear map having values in Y. In general $D^k \boldsymbol{f}(\boldsymbol{x})$ may be considered a k linear map. This means

$$(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k) \to D^k \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}_1) \cdots (\boldsymbol{u}_k)$$

has the property $\boldsymbol{u}_j \to D^k \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}_1) \cdots (\boldsymbol{u}_j) \cdots (\boldsymbol{u}_k)$ is linear.

Also, instead of writing $D^2 \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$, or $D^3 \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})(\boldsymbol{w})$ the following notation is often used.

$$D^2 \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}) \text{ or } D^3 \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$$

with similar conventions for higher derivatives than 3. Another convention which is often used is the notation $D^k \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{v}^k$ instead of $D^k \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{v}, \cdots, \boldsymbol{v})$.

Note that for every k, $D^k \boldsymbol{f}$ maps U to a normed vector space. As mentioned above, $D\boldsymbol{f}(\boldsymbol{x})$ has values in $\mathscr{L}(X,Y)$, $D^2 \boldsymbol{f}(\boldsymbol{x})$ has values in $\mathscr{L}(X, \mathscr{L}(X,Y))$, etc. Thus it makes sense to consider whether $D^k \boldsymbol{f}$ is continuous. This is described in the following definition.

Definition 6.7.2 *Let U be an open subset of X, a normed vector space, and let $\boldsymbol{f} : U \to Y$. Then $\boldsymbol{f}$ is $C^k(U)$ if $\boldsymbol{f}$ and its first k derivatives are all continuous. Also, $D^k \boldsymbol{f}(\boldsymbol{x})$ when it exists can be considered a Y valued multi-linear function. Sometimes these are called tensors in case $\boldsymbol{f}$ has scalar values.*

6.8 Some Standard Notation

In the case where $X = \mathbb{R}^n$ there is a special notation which is often used to describe higher order mixed partial derivatives. It is called multi-index notation.

Definition 6.8.1 *$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ for $\alpha_1 \cdots \alpha_n$ positive integers is called a multi-index, as before with polynomials. For $\boldsymbol{\alpha}$ a multi-index, $|\boldsymbol{\alpha}| \equiv \alpha_1 + \cdots + \alpha_n$, and if $\boldsymbol{x} \in X$,*

$$\boldsymbol{x} = (x_1, \cdots, x_n),$$

and $\boldsymbol{f}$ a function, define

$$\boldsymbol{x}^{\boldsymbol{\alpha}} \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n},\ D^{\boldsymbol{\alpha}} \boldsymbol{f}(\boldsymbol{x}) \equiv \frac{\partial^{|\boldsymbol{\alpha}|} \boldsymbol{f}(\boldsymbol{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}.$$

Then in this special case, the following is another description of what is meant by a C^k function.

Definition 6.8.2 *Let U be an open subset of $\mathbb{R}^n$ and let $\boldsymbol{f} : U \to Y$. Then for k a nonnegative integer, a differentiable function $\boldsymbol{f}$ is C^k if for every $|\boldsymbol{\alpha}| \leq k$, $D^{\boldsymbol{\alpha}} \boldsymbol{f}$ exists and is continuous.*

Theorem 6.8.3 *Let U be an open subset of $\mathbb{R}^n$ and let $\boldsymbol{f} : U \to Y$. Then if $D^r \boldsymbol{f}(\boldsymbol{x})$ exists for $r \leq k$, then $D^r \boldsymbol{f}$ is continuous at $\boldsymbol{x}$ for $r \leq k$ if and only if $D^{\boldsymbol{\alpha}} \boldsymbol{f}$ is continuous at x for each $|\alpha| \leq k$.*

Proof: First consider the case of a single derivative. Then as shown above, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is just

$$J(\boldsymbol{x}) \equiv \begin{pmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}$$

and to say that $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous is the same as saying that each of these partial derivatives is continuous. Written out in more detail,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}+\boldsymbol{o}(\boldsymbol{v}) = \sum_{k=1}^{n} \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}) v_k + \boldsymbol{o}(\boldsymbol{v})$$

Thus $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} = \sum_{k=1}^{n} \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}) v_k$. Now consider the second derivative.

$$D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{w})(\boldsymbol{v}) =$$

$$D\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{w})\boldsymbol{v} - D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v}) = \sum_{k=1}^{n}\left(\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}+\boldsymbol{w}) - \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x})\right) v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v})$$

$$= \sum_{k=1}^{n}\left(\sum_{j=1}^{n} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j + \boldsymbol{o}(\boldsymbol{w})\right) v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v}) = \sum_{j,k} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v})$$

and so $D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{w})(\boldsymbol{v}) = \sum_{j,k} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j v_k$. Hence $D^2\boldsymbol{f}$ is continuous if and only if each of these coefficients $\boldsymbol{x} \to \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k}$ is continuous. Obviously you can continue doing this and conclude that $D^k\boldsymbol{f}$ is continuous if and only if all of the partial derivatives of order up to k are continuous. ■

In practice, this is usually what people are thinking when they say that $\boldsymbol{f}$ is C^k. But as just argued, this is the same as saying that the r linear form $\boldsymbol{x} \to D^r\boldsymbol{f}(\boldsymbol{x})$ is continuous into the appropriate space of linear transformations for each $r \leq k$.

Of course the above is based on the assumption that the first k derivatives exist and gives two equivalent formulations which state that these derivatives are continuous. Can anything be said about the existence of the derivatives based on the existence and continuity of the partial derivatives? As pointed out, if the partial derivatives exist and are continuous, then the function is differentiable and has continuous derivative. However, I want to emphasize the idea of the Cartesian product.

6.9 The Derivative and the Cartesian Product

There are theorems which can be used to get differentiability of a function based on existence and continuity of the partial derivatives. A generalization of this was given above. Here a function defined on a product space is considered. It is very much like what was presented above and could be obtained as a special case but to reinforce the ideas, I will do it from scratch because certain aspects of it are important in the statement of the implicit function theorem.

The following is an important abstract generalization of the concept of partial derivative presented above. Insead of taking the derivative with respect to one variable, it is taken with respect to several but not with respect to others. This vague notion is made precise in the following definition. First here is a lemma.

Lemma 6.9.1 *Suppose U is an open set in $X \times Y$. Then the set, $U_{\boldsymbol{y}}$ defined by*

$$U_{\boldsymbol{y}} \equiv \{\boldsymbol{x} \in X : (\boldsymbol{x},\boldsymbol{y}) \in U\}$$

is an open set in X. Here $X\times Y$ is a finite dimensional vector space in which the vector space operations are defined componentwise. Thus for $a,b\in\mathbb{F}$,

$$a(\boldsymbol{x}_1,\boldsymbol{y}_1)+b(\boldsymbol{x}_2,\boldsymbol{y}_2)=(a\boldsymbol{x}_1+b\boldsymbol{x}_2,a\boldsymbol{y}_1+b\boldsymbol{y}_2)$$

and the norm can be taken to be

$$\|(\boldsymbol{x},\boldsymbol{y})\|\equiv\max(\|\boldsymbol{x}\|,\|\boldsymbol{y}\|)$$

Proof: Recall by Theorem 5.2.4 it does not matter how this norm is defined and the definition above is convenient. It obviously satisfies most axioms of a norm. The only one which is not obvious is the triangle inequality. I will show this now.

$$\begin{aligned}\|(\boldsymbol{x},\boldsymbol{y})+(\boldsymbol{x}_1,\boldsymbol{y}_1)\| &\equiv \|(\boldsymbol{x}+\boldsymbol{x}_1,\boldsymbol{y}+\boldsymbol{y}_1)\|\equiv\max(\|\boldsymbol{x}+\boldsymbol{x}_1\|,\|\boldsymbol{y}+\boldsymbol{y}_1\|)\\ &\leq \max(\|\boldsymbol{x}\|+\|\boldsymbol{x}_1\|,\|\boldsymbol{y}\|+\|\boldsymbol{y}_1\|)\\ &\leq \max(\|\boldsymbol{x}\|,\|\boldsymbol{y}\|)+\max(\|\boldsymbol{x}_1\|,\|\boldsymbol{y}_1\|)\\ &\equiv \|(\boldsymbol{x},\boldsymbol{y})\|+\|(\boldsymbol{x}_1,\boldsymbol{y}_1)\|\end{aligned}$$

Let $\boldsymbol{x}\in U_{\boldsymbol{y}}$. Then $(\boldsymbol{x},\boldsymbol{y})\in U$ and so there exists $r>0$ such that $B((\boldsymbol{x},\boldsymbol{y}),r)\in U$. This says that if $(\boldsymbol{u},\boldsymbol{v})\in X\times Y$ such that $\|(\boldsymbol{u},\boldsymbol{v})-(\boldsymbol{x},\boldsymbol{y})\|<r$, then $(\boldsymbol{u},\boldsymbol{v})\in U$. Thus if

$$\|(\boldsymbol{u},\boldsymbol{y})-(\boldsymbol{x},\boldsymbol{y})\|=\|\boldsymbol{u}-\boldsymbol{x}\|_X<r,$$

then $(\boldsymbol{u},\boldsymbol{y})\in U$. This has just said that $B(\boldsymbol{x},r)_X$, the ball taken in X is contained in $U_{\boldsymbol{y}}$. This proves the lemma. ■

Or course one could also consider $U_{\boldsymbol{x}}\equiv\{\boldsymbol{y}:(\boldsymbol{x},\boldsymbol{y})\in U\}$ in the same way and conclude this set is open in Y. Also, the generalization to many factors yields the same conclusion. In this case, for $\boldsymbol{x}\in\prod_{i=1}^n X_i$, let

$$\|\boldsymbol{x}\|\equiv\max\left(\|\boldsymbol{x}_i\|_{X_i}:\boldsymbol{x}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\right)$$

Then a similar argument to the above shows this is a norm on $\prod_{i=1}^n X_i$. Consider the triangle inequality.

$$\|(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)+(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)\|=\max_i\left(\|\boldsymbol{x}_i+\boldsymbol{y}_i\|_{X_i}\right)\leq\max_i\left(\|\boldsymbol{x}_i\|_{X_i}+\|\boldsymbol{y}_i\|_{X_i}\right)$$

$$\leq\max_i\left(\|\boldsymbol{x}_i\|_{X_i}\right)+\max_i\left(\|\boldsymbol{y}_i\|_{X_i}\right)=\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$$

Corollary 6.9.2 *Let $U\subseteq\prod_{i=1}^n X_i$ be an open set and let*

$$U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}\equiv\left\{\boldsymbol{x}\in\mathbb{F}^{r_i}:(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)\in U\right\}.$$

Then $U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$ is an open set in $\mathbb{F}^{r_i}$.

Proof: Let $\boldsymbol{z}\in U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$. Then $(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)\equiv\boldsymbol{x}\in U$ by definition. Therefore, since U is open, there exists $r>0$ such that $B(\boldsymbol{x},r)\subseteq U$. It follows that for $B(\boldsymbol{z},r)_{X_i}$ denoting the ball in X_i, it follows that $B(\boldsymbol{z},r)_{X_i}\subseteq U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$ because to say that $\|\boldsymbol{z}-\boldsymbol{w}\|_{X_i}<r$ is to say that

$$\|(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)-(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{w},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)\|<r$$

and so $\boldsymbol{w}\in U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$. ■

Next is a generalization of the partial derivative.

Definition 6.9.3 *Let $\boldsymbol{g} : U \subseteq \prod_{i=1}^{n} X_i \rightarrow Y$, where U is an open set. Then the map*

$$\boldsymbol{z} \rightarrow \boldsymbol{g}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1}, \boldsymbol{z}, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_n)$$

is a function from the open set in X_i,

$$\{\boldsymbol{z} : \boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1}, \boldsymbol{z}, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_n) \in U\}$$

to Y. When this map is differentiable, its derivative is denoted by $D_i\boldsymbol{g}(\boldsymbol{x})$. To aid in the notation, for $\boldsymbol{v} \in X_i$, let $\theta_i \boldsymbol{v} \in \prod_{i=1}^{n} X_i$ be the vector $(\boldsymbol{0}, \cdots, \boldsymbol{v}, \cdots, \boldsymbol{0})$ where the $\boldsymbol{v}$ is in the i^{th} slot and for $\boldsymbol{v} \in \prod_{i=1}^{n} X_i$, let $\boldsymbol{v}_i$ denote the entry in the i^{th} slot of $\boldsymbol{v}$. Thus, by saying

$$\boldsymbol{z} \rightarrow \boldsymbol{g}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1}, \boldsymbol{z}, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_n)$$

is differentiable is meant that for $\boldsymbol{v} \in X_i$ sufficiently small,

$$\boldsymbol{g}(\boldsymbol{x} + \theta_i \boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{x}) = D_i\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}).$$

Note $D_i\boldsymbol{g}(\boldsymbol{x}) \in \mathscr{L}(X_i, Y)$.

As discussed above, we have the following definition of $C^1(U)$.

Definition 6.9.4 *Let $U \subseteq X$ be an open set. Then $\boldsymbol{f} : U \rightarrow Y$ is $C^1(U)$ if $\boldsymbol{f}$ is differentiable and the mapping $\boldsymbol{x} \rightarrow D\boldsymbol{f}(\boldsymbol{x})$, is continuous as a function from U to $\mathscr{L}(X, Y)$.*

With this definition of partial derivatives, here is the major theorem. Note the resemblance with the matrix of the derivative of a function having values in $\mathbb{R}^m$ in terms of the partial derivatives.

Theorem 6.9.5 *Let $\boldsymbol{g}, U, \prod_{i=1}^{n} X_i$, be given as in Definition 6.9.3. Then $\boldsymbol{g}$ is $C^1(U)$ if and only if $D_i\boldsymbol{g}$ exists and is continuous on U for each i. In this case, $\boldsymbol{g}$ is differentiable and*

$$D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{v}) = \sum_k D_k\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v}_k \tag{6.14}$$

where $\boldsymbol{v} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)$.

Proof: Suppose then that $D_i\boldsymbol{g}$ exists and is continuous for each i. Note $\sum_{j=1}^{k} \theta_j \boldsymbol{v}_j = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k, \boldsymbol{0}, \cdots, \boldsymbol{0})$. Thus $\sum_{j=1}^{n} \theta_j \boldsymbol{v}_j = \boldsymbol{v}$ and define $\sum_{j=1}^{0} \theta_j \boldsymbol{v}_j \equiv \boldsymbol{0}$. Therefore,

$$\boldsymbol{g}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{x}) = \sum_{k=1}^{n} \left[\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k} \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) \right] \tag{6.15}$$

$$= \sum_{k=1}^{n} \left[\left(\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k} \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}(\boldsymbol{x} + \theta_k \boldsymbol{v}_k) \right) - \left(\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}(\boldsymbol{x}) \right) \right]$$
$$+ \sum_{k=1}^{n} (\boldsymbol{g}(\boldsymbol{x} + \theta_k \boldsymbol{v}_k) - \boldsymbol{g}(\boldsymbol{x}))$$

If $\boldsymbol{h}_k(\boldsymbol{x}) \equiv \boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}(\boldsymbol{x})$ then the top sum is $\sum_{k=1}^{n} \boldsymbol{h}_k(\boldsymbol{x} + \theta_k \boldsymbol{v}_k) - \boldsymbol{h}_k(\boldsymbol{x})$ and from the definition of $\boldsymbol{h}_k$, $\|D\boldsymbol{h}_k(\boldsymbol{x})\| < \varepsilon$ a sufficiently small ball containing $\boldsymbol{x}$. Thus this top sum is dominated by $\varepsilon\left\|\boldsymbol{v}\right\|$ whenever $\left\|\boldsymbol{v}\right\|$ is small enough. Since ε is arbitrary, this term is $\boldsymbol{o}(\boldsymbol{v})$. The last term is $\sum_{k=1}^{n}D_k\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v}_k+\boldsymbol{o}(\boldsymbol{v}_k)$ and so, collecting terms obtains

$$\boldsymbol{g}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{g}(\boldsymbol{x})=\sum_{k=1}^{n}D_k\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v}_k+\boldsymbol{o}(\boldsymbol{v})$$

which shows $D\boldsymbol{g}(\boldsymbol{x})$ exists and equals the formula given in 6.14. Also $\boldsymbol{x}\to D\boldsymbol{g}(\boldsymbol{x})$ is continuous since each of the $D_k\boldsymbol{g}(\boldsymbol{x})$ are.

Next suppose $\boldsymbol{g}$ is C^1. I need to verify that $D_k\boldsymbol{g}(\boldsymbol{x})$ exists and is continuous. Let $\boldsymbol{v}\in X_k$ sufficiently small. Then

$$\boldsymbol{g}(\boldsymbol{x}+\theta_k\boldsymbol{v})-\boldsymbol{g}(\boldsymbol{x})=D\boldsymbol{g}(\boldsymbol{x})\theta_k\boldsymbol{v}+\boldsymbol{o}(\theta_k\boldsymbol{v})=D\boldsymbol{g}(\boldsymbol{x})\theta_k\boldsymbol{v}+\boldsymbol{o}(\boldsymbol{v})$$

since $\left\|\theta_k\boldsymbol{v}\right\|=\left\|\boldsymbol{v}\right\|$. Then $D_k\boldsymbol{g}(\boldsymbol{x})$ exists and equals $D\boldsymbol{g}(\boldsymbol{x})\circ\theta_k$. Now $\boldsymbol{x}\to D\boldsymbol{g}(\boldsymbol{x})$ is continuous. Since θ_k is linear, it follows from Lemma 5.2.1 that $\theta_k:X_k\to\prod_{i=1}^{n}X_i$ is also continuous. ■

Note that the above argument also works at a single point $\boldsymbol{x}$. That is, continuity at $\boldsymbol{x}$ of the partials implies $D\boldsymbol{g}(\boldsymbol{x})$ exists and is continuous at $\boldsymbol{x}$.

The way this is usually used is in the following corollary which has already been obtained. Remember the matrix of $D\boldsymbol{f}(\boldsymbol{x})$. Recall that if a function is C^1 in the sense that $\boldsymbol{x}\to D\boldsymbol{f}(\boldsymbol{x})$ is continuous then all the partial derivatives exist and are continuous. The next corollary says that if the partial derivatives do exist and are continuous, then the function is differentiable and has continuous derivative.

Corollary 6.9.6 *Let U be an open subset of $\mathbb{F}^n$ and let $\boldsymbol{f}:U\to\mathbb{F}^m$ be C^1 in the sense that all the partial derivatives of $\boldsymbol{f}$ exist and are continuous. Then $\boldsymbol{f}$ is differentiable and*

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})=\boldsymbol{f}(\boldsymbol{x})+\sum_{k=1}^{n}\frac{\partial\boldsymbol{f}}{\partial x_k}(\boldsymbol{x})v_k+\boldsymbol{o}(\boldsymbol{v}).$$

Similarly, if the partial derivatives up to order k exist and are continuous, then the function is C^k in the sense that the first k derivatives exist and are continuous.

6.10 Mixed Partial Derivatives

Continuing with the special case where f is defined on an open set in $\mathbb{F}^n$, I will next consider an interesting result which was known to Euler in around 1734 about mixed partial derivatives. It was proved by Clairaut some time later. It turns out that the mixed partial derivatives, if continuous will end up being equal. Recall the notation $f_x=\frac{\partial f}{\partial x}=D_{\boldsymbol{e}_1}f$ and $f_{xy}=\frac{\partial^2 f}{\partial y\partial x}=D_{\boldsymbol{e}_1\boldsymbol{e}_2}f$.

Theorem 6.10.1 *Suppose $f:U\subseteq\mathbb{F}^2\to\mathbb{R}$ where U is an open set on which f_x,f_y, f_{xy} and f_{yx} exist. Then if f_{xy} and f_{yx} are continuous at the point $(x,y)\in U$, it follows*

$$f_{xy}(x,y)=f_{yx}(x,y).$$

Proof: Since U is open, there exists $r>0$ such that $B((x,y),r)\subseteq U$. Now let $|t|,|s|<r/2,t,s$ real numbers and consider

$$\Delta(s,t)\equiv\frac{1}{st}\{\overbrace{f(x+t,y+s)-f(x+t,y)}^{h(t)}-\overbrace{(f(x,y+s)-f(x,y))}^{h(0)}\}. \tag{6.16}$$

Note that $(x+t,y+s)\in U$ because

$$\begin{aligned}|(x+t,y+s)-(x,y)| &= |(t,s)| = \left(t^2+s^2\right)^{1/2} \\ &\leq \left(\frac{r^2}{4}+\frac{r^2}{4}\right)^{1/2} = \frac{r}{\sqrt{2}} < r.\end{aligned}$$

As implied above, $h(t)\equiv f(x+t,y+s)-f(x+t,y)$. Therefore, by the mean value theorem from one variable calculus and the (one variable) chain rule,

$$\begin{aligned}\Delta(s,t) &= \frac{1}{st}(h(t)-h(0)) = \frac{1}{st}h'(\alpha t)t \\ &= \frac{1}{s}\left(f_x(x+\alpha t,y+s)-f_x(x+\alpha t,y)\right)\end{aligned}$$

for some $\alpha\in(0,1)$. Applying the mean value theorem again,

$$\Delta(s,t) = f_{xy}(x+\alpha t,y+\beta s)$$

where $\alpha,\beta\in(0,1)$.

If the terms $f(x+t,y)$ and $f(x,y+s)$ are interchanged in 6.16, $\Delta(s,t)$ is unchanged and the above argument shows there exist $\gamma,\delta\in(0,1)$ such that

$$\Delta(s,t) = f_{yx}(x+\gamma t,y+\delta s).$$

Letting $(s,t)\to(0,0)$ and using the continuity of f_{xy} and f_{yx} at (x,y),

$$\lim_{(s,t)\to(0,0)}\Delta(s,t) = f_{xy}(x,y) = f_{yx}(x,y).\blacksquare$$

The following is obtained from the above by simply fixing all the variables except for the two of interest.

Corollary 6.10.2 *Suppose U is an open subset of X and $f:U\to\mathbb{R}$ has the property that for two indices, k,l, f_{x_k}, f_{x_l}, $f_{x_lx_k}$, and $f_{x_kx_l}$ exist on U and $f_{x_kx_l}$ and $f_{x_lx_k}$ are both continuous at $\boldsymbol{x}\in U$. Then $f_{x_kx_l}(\boldsymbol{x}) = f_{x_lx_k}(\boldsymbol{x})$.*

By considering the real and imaginary parts of f in the case where f has values in $\mathbb{C}$ you obtain the following corollary.

Corollary 6.10.3 *Suppose U is an open subset of $\mathbb{F}^n$ and $f:U\to\mathbb{F}$ has the property that for two indices, k,l, f_{x_k}, f_{x_l}, $f_{x_lx_k}$, and $f_{x_kx_l}$ exist on U and $f_{x_kx_l}$ and $f_{x_lx_k}$ are both continuous at $\boldsymbol{x}\in U$. Then $f_{x_kx_l}(\boldsymbol{x}) = f_{x_lx_k}(\boldsymbol{x})$.*

Finally, by considering the components of $\boldsymbol{f}$ you get the following generalization.

Corollary 6.10.4 *Suppose U is an open subset of $\mathbb{F}^n$ and $\boldsymbol{f}:U\to\mathbb{F}^m$ has the property that for two indices, k,l, $\boldsymbol{f}_{x_k}$, $\boldsymbol{f}_{x_l}$, $\boldsymbol{f}_{x_lx_k}$, and $\boldsymbol{f}_{x_kx_l}$ exist on U and $\boldsymbol{f}_{x_kx_l}$ and $\boldsymbol{f}_{x_lx_k}$ are both continuous at $\boldsymbol{x}\in U$. Then $\boldsymbol{f}_{x_kx_l}(\boldsymbol{x}) = \boldsymbol{f}_{x_lx_k}(\boldsymbol{x})$.*

This can be generalized to functions which have values in a normed linear space, but I plan to stop with what is given above. One way to proceed would be to reduce to a consideration of the coordinate maps and then apply the above. It would even hold in infinite dimensions through the use of the Hahn Banach theorem. The idea is to reduce to the scalar valued case as above.

In addition, it is obvious that for a function of many variables you could pick any pair and say these are equal if they are both continuous.

It is necessary to assume the mixed partial derivatives are continuous in order to assert they are equal. The following is a well known example [2].

Example 6.10.5 *Let*

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

From the definition of partial derivatives it follows that $f_x(0,0) = f_y(0,0) = 0$. Using the standard rules of differentiation, for $(x,y) \neq (0,0)$,

$$f_x = y\frac{x^4 - y^4 + 4x^2y^2}{(x^2+y^2)^2}, \ f_y = x\frac{x^4 - y^4 - 4x^2y^2}{(x^2+y^2)^2}$$

Now

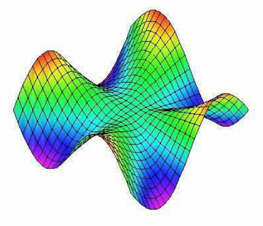
$$f_{xy}(0,0) \equiv \lim_{y\to 0} \frac{f_x(0,y) - f_x(0,0)}{y} = \lim_{y\to 0} \frac{-y^4}{(y^2)^2} = -1$$

while

$$f_{yx}(0,0) \equiv \lim_{x\to 0} \frac{f_y(x,0) - f_y(0,0)}{x} = \lim_{x\to 0} \frac{x^4}{(x^2)^2} = 1$$

showing that although the mixed partial derivatives do exist at $(0,0)$, they are not equal there.

Incidentally, the graph of this function appears very innocent. Its fundamental sickness is not apparent. It is like one of those whited sepulchers mentioned in the Bible.



6.11 A Cofactor Identity

Lemma 6.11.1 *Suppose* $\det(A) = 0$. *Then for all sufficiently small nonzero* ε, *it follows that* $\det(A + \varepsilon I) \neq 0$.

Proof: Let $\det(\lambda I - A) = \lambda^p + a_1\lambda^{p-1} + \cdots + a_{p-1}\lambda + a_p$. First suppose A is a $p \times p$ matrix. If $\det(A) \neq 0$, this will still be true for all ε small enough. Now suppose also that

$\det(A) = 0$. Thus, the constant term of $\det(\lambda I - A)$ is 0. Consider $\varepsilon I + A \equiv A_\varepsilon$ for small real ε. The characteristic polynomial of A_ε is

$$\det(\lambda I - A_\varepsilon) = \det((\lambda - \varepsilon) I - A)$$

This is of the form

$$(\lambda - \varepsilon)^p + a_1 (\lambda - \varepsilon)^{p-1} + \cdots + (\lambda - \varepsilon)^m a_m$$

where the a_j are the coefficients in the characteristic polynomial for A and $a_k = 0$ for $k > m, a_m \neq 0$. The constant term of this polynomial in λ must be nonzero for all ε small enough because it is of the form

$$(-1)^m \varepsilon^m a_m + (\text{higher order terms in } \varepsilon) = \varepsilon^m [a_m (-1)^m + \varepsilon C(\varepsilon)]$$

which is nonzero for all positive but very small ε. Thus $\varepsilon I + A$ is invertible for all ε small enough but nonzero. ■

Recall that for A an $p \times p$ matrix, $\operatorname{cof}(A)_{ij}$ is the determinant of the matrix which results from deleting the i^{th} row and the j^{th} column and multiplying by $(-1)^{i+j}$. In the proof and in what follows, I am using $D\boldsymbol{g}$ to equal the matrix of the linear transformation $D\boldsymbol{g}$ taken with respect to the usual basis on $\mathbb{R}^p$. Thus $(D\boldsymbol{g})_{ij} = \partial g_i / \partial x_j$ where $\boldsymbol{g} = \sum_i g_i \boldsymbol{e}_i$ for the $\boldsymbol{e}_i$ the standard basis vectors.

Lemma 6.11.2 *Let $\boldsymbol{g} : U \to \mathbb{R}^p$ be C^2 where U is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^{p} \operatorname{cof}(D\boldsymbol{g})_{ij,j} = 0,$$

where here $(D\boldsymbol{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij} = \frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}}$.

Proof: From the cofactor expansion theorem,

$$\delta_{kj} \det(D\boldsymbol{g}) = \sum_{i=1}^{p} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij} \tag{6.17}$$

This is because if $k \neq j$, that on the right is the cofactor expansion of a determinant with two equal columns while if $k = j$, it is just the cofactor expansion of the determinant. In particular,

$$\frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}} = \operatorname{cof}(D\boldsymbol{g})_{ij} \tag{6.18}$$

which shows the last claim of the lemma. Assume that $D\boldsymbol{g}(\boldsymbol{x})$ is invertible to begin with. Differentiate 6.17 with respect to x_j and sum on j. This yields

$$\sum_{r,s,j} \delta_{kj} \frac{\partial (\det D\boldsymbol{g})}{\partial g_{r,s}} g_{r,sj} = \sum_{ij} g_{i,kj} (\operatorname{cof}(D\boldsymbol{g}))_{ij} + \sum_{ij} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij,j}.$$

Hence, using $\delta_{kj} = 0$ if $j \neq k$ and 6.18,

$$\sum_{rs} (\operatorname{cof}(D\boldsymbol{g}))_{rs} g_{r,sk} = \sum_{rs} g_{r,ks} (\operatorname{cof}(D\boldsymbol{g}))_{rs} + \sum_{ij} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij,j}.$$

Subtracting the first sum on the right from both sides and using the equality of mixed partials,

$$\sum_i g_{i,k}\left(\sum_j (\operatorname{cof}(D\boldsymbol{g}))_{ij,j}\right) = 0.$$

Since it is assumed $D\boldsymbol{g}$ is invertible, this shows $\sum_j (\operatorname{cof}(D\boldsymbol{g}))_{ij,j} = 0$. If $\det(D\boldsymbol{g}) = 0$, use Lemma 6.11.1 to let $\boldsymbol{g}_k(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) + \varepsilon_k \boldsymbol{x}$ where $\varepsilon_k \to 0$ and $\det(D\boldsymbol{g} + \varepsilon_k I) \equiv \det(D\boldsymbol{g}_k) \neq 0$. Then

$$\sum_j (\operatorname{cof}(D\boldsymbol{g}))_{ij,j} = \lim_{k\to\infty}\sum_j (\operatorname{cof}(D\boldsymbol{g}_k))_{ij,j} = 0 \blacksquare$$

6.12 Exercises

1. Here are some scalar valued functions of several variables. Determine which of these functions are $o(\boldsymbol{v})$. Here $\boldsymbol{v}$ is a vector in $\mathbb{R}^n$, $\boldsymbol{v} = (v_1, \cdots, v_n)$.

(a) $v_1 v_2$

(b) $v_2 \sin(v_1)$

(c) $v_1^2 + v_2$

(d) $v_2 \sin(v_1 + v_2)$

(e) $v_1(v_1 + v_2 + xv_3)$

(f) $(e^{v_1} - 1 - v_1)$

(g) $(\boldsymbol{x}\cdot\boldsymbol{v})|\boldsymbol{v}|$

2. Here is a function of two variables. $f(x,y) = x^2y + x^2$. Find $Df(x,y)$ directly from the definition. Recall this should be a linear transformation which results from multiplication by a 1 × 2 matrix. Find this matrix.

3. Let $\boldsymbol{f}(x,y) = \begin{pmatrix} x^2+y \\ y^2 \end{pmatrix}$. Compute the derivative directly from the definition. This should be the linear transformation which results from multiplying by a 2 × 2 matrix. Find this matrix.

4. You have $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ Here $\boldsymbol{x} \in \mathbb{R}^n$**,** $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ and $\boldsymbol{g}(\boldsymbol{y}) \in \mathbb{R}^p$. where $\boldsymbol{f},\boldsymbol{g}$ are appropriately differentiable. Thus $D\boldsymbol{h}(\boldsymbol{x})$ results from multiplication by a matrix. Using the chain rule, give a formula for the ij^{th} entry of this matrix. How does this relate to multiplication of matrices? In other words, you have two matrices which correspond to $D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ and $D\boldsymbol{f}(\boldsymbol{x})$ Call $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{y}), \boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. Then

$$D\boldsymbol{g}(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial \boldsymbol{z}}{\partial y_1} & \cdots & \frac{\partial \boldsymbol{z}}{\partial y_m} \end{pmatrix}, D\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial \boldsymbol{y}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_n} \end{pmatrix}$$

Explain the manner in which the ij^{th} entry of $D\boldsymbol{h}(\boldsymbol{x})$ is $\sum_k \frac{\partial z_i}{\partial y_k}\frac{\partial y_y}{\partial x_j}$. This is a review of the way we multiply matrices. what is the i^{th} row of $D\boldsymbol{g}(\boldsymbol{y})$ and the j^{th} column of $D\boldsymbol{f}(\boldsymbol{x})$?

5. Find $f_x, f_y, f_z, f_{xy}, f_{yx}, f_{zy}$ for the following. Verify the mixed partial derivatives are equal.

(a) $x^2y^3z^4 + \sin(xyz)$

(b) $\sin(xyz) + x^2yz$

6. As an important application of Theorem 6.4.3 consider the following. Experiments are done at n times, $t_1, t_2, \cdots, t_n$ and at each time there results a collection of numerical outcomes. Denote by $\{(t_i, x_i)\}_{i=1}^{p}$ the set of all such pairs and try to find numbers a and b such that the line $x = at + b$ approximates these ordered pairs as well as possible in the sense that out of all choices of a and b, $\sum_{i=1}^{p} (at_i + b - x_i)^2$ is as small as possible. In other words, you want to minimize the function of two variables $f(a,b) \equiv \sum_{i=1}^{p} (at_i + b - x_i)^2$. Find a formula for a and b in terms of the given ordered pairs. You will be finding the formula for the least squares regression line.

7. Let f be a function which has continuous derivatives. Show that $u(t,x) = f(x - ct)$ solves the wave equation $u_{tt} - c^2 \Delta u = 0$. What about $u(x,t) = f(x + ct)$? Here $\Delta u = u_{xx}$.

8. Show that if $\Delta u = \lambda u$ where u is a function of only x, then $e^{\lambda t} u$ solves the heat equation $u_t - \Delta u = 0$. Here $\Delta u = u_{xx}$.

9. Show that if $f(x) = o(x)$, then $f'(0) = 0$.

10. Let $f(x,y)$ be defined on $\mathbb{R}^2$ as follows. $f\left(x, x^2\right) = 1$ if $x \neq 0$. Define $f(0,0) = 0$, and $f(x,y) = 0$ if $y \neq x^2$. Show that f is not continuous at $(0,0)$ but that
$$\lim_{h \to 0} \frac{f(ha, hb) - f(0,0)}{h} = 0$$
for (a,b) an arbitrary vector. Thus the Gateaux derivative exists at $(0,0)$ in every direction but f is not even continuous there.

11. Let
$$f(x,y) \equiv \begin{cases} \frac{xy^4}{x^2+y^8} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$
Show that this function is not continuous at $(0,0)$ but that the Gateaux derivative $\lim_{h \to 0} \frac{f(ha,hb) - f(0,0)}{h}$ exists and equals 0 for every vector (a,b).

12. Let U be an open subset of $\mathbb{R}^n$ and suppose that $f : [a,b] \times U \to \mathbb{R}$ satisfies
$$(x,y) \to \frac{\partial f}{\partial y_i}(x,y), (x,y) \to f(x,y)$$
are all continuous. Show that $\int_a^b f(x,y)\,dx$, $\int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$ all make sense and that in fact
$$\frac{\partial}{\partial y_i}\left(\int_a^b f(x,y)\,dx\right) = \int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$$
Also explain why $y \to \int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$ is continuous. **Hint:** You will need to use the theorems from one variable calculus about the existence of the integral for a continuous function. You may also want to use theorems about uniform continuity of continuous functions defined on compact sets.

13. I found this problem in Apostol's book [1]. This is a very important result and is obtained very simply. Read it and fill in any missing details. Let $g(x) \equiv \int_0^1 \frac{e^{-x^2\left(1+t^2\right)}}{1+t^2} dt$ and $f(x) \equiv \left(\int_0^x e^{-t^2} dt\right)^2$. Note $\frac{\partial}{\partial x}\left(\frac{e^{-x^2\left(1+t^2\right)}}{1+t^2}\right) = -2xe^{-x^2\left(1+t^2\right)}$. Explain why this is so. Also show the conditions of Problem 12 are satisfied so that

$$g'(x) = \int_0^1 \left(-2xe^{-x^2\left(1+t^2\right)}\right) dt.$$

Now use the chain rule and the fundamental theorem of calculus to find $f'(x)$. Then change the variable in the formula for $f'(x)$ to make it an integral from 0 to 1 and show $f'(x) + g'(x) = 0$. Now this shows $f(x) + g(x)$ is a constant. Show the constant is $\pi/4$ by letting $x \to 0$. Next take a limit as $x \to \infty$ to obtain the following formula for the improper integral, $\int_0^\infty e^{-t^2} dt$, $\left(\int_0^\infty e^{-t^2} dt\right)^2 = \pi/4$. In passing to the limit in the integral for g as $x \to \infty$ you need to justify why that integral converges to 0. To do this, argue the integrand converges uniformly to 0 for $t \in [0,1]$ and then explain why this gives convergence of the integral. Thus $\int_0^\infty e^{-t^2} dt = \sqrt{\pi}/2$.

14. Recall the treatment of integrals of continuous functions in Proposition 5.8.5 or what you used in beginning calculus. The gamma function is defined for $x > 0$ as $\Gamma(x) \equiv \int_0^\infty e^{-t} t^{x-1} dt \equiv \lim_{R\to\infty} \int_0^R e^{-t} t^{x-1} dt$. Show this limit exists. Note you might have to give a meaning to $\int_0^R e^{-t} t^{x-1} dt$ if $x < 1$. Also show that $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$. How does $\Gamma(n)$ for n an integer compare with $(n-1)!$?

15. Show the mean value theorem for integrals. Suppose $f \in C([a,b])$. Then there exists $x \in (a,b)$, not just in $[a,b]$ such that $f(x)(b-a) = \int_a^b f(t) dt$. **Hint:** Let $F(x) \equiv \int_a^x f(t) dt$ and use the mean value theorem, Theorem 5.8.3 along with $F'(x) = f(x)$.

16. Show, using the Weierstrass approximation theorem that linear combinations of the form $\sum_{i,j} a_{ij} g_i(s) h_j(t)$ where g_i, h_j are continuous functions on $[0,b]$ are dense in $C([0,b] \times [0,b])$, the continuous functions defined on $[0,b] \times [0,b]$ with norm given by

$$\|f\| \equiv \max\{|f(x,y)| : (x,y) \in [0,b] \times [0,b]\}$$

Show that for h, g continuous, $\int_0^b \int_0^s g(s) h(t) dt ds - \int_0^b \int_t^b g(s) h(t) ds dt = 0$. Now explain why if f is in $C([0,b] \times [0,b])$,

$$\int_0^b \int_0^s f(s,t) dt ds - \int_0^b \int_t^b f(s,t) ds dt = 0.$$

17. Let $f(x) \equiv \left(\int_0^x e^{-t^2} dt\right)^2$. Use Proposition 5.8.5 which includes the fundamental theorem of calculus and elementary change of variables, show that

$$f'(x) = 2e^{-x^2}\left(\int_0^x e^{-t^2} dt\right) = 2e^{-x^2}\left(\int_0^1 e^{-(xs)^2} x ds\right) = \int_0^1 2xe^{-x^2\left(1+s^2\right)} ds.$$

Now show

$$f(x) = \int_0^1 \int_0^x 2te^{-t^2\left(1+s^2\right)} dt ds.$$

Show $\lim_{x\to\infty} \int_0^x e^{-t^2} dt = \frac{1}{2}\sqrt{\pi}$

Chapter 7

Implicit Function Theorem

The implicit function theorem is one of the greatest theorems in mathematics. There are many versions of this theorem which are of far greater generality than the one given here. The proof given here is like one found in one of Caratheodory's books on the calculus of variations. It is not as elegant as some of the others which are based on a contraction mapping principle but it may be shorter and is based on more elementary ideas. For a more elegant proof which generalizes better see my book Real and Abstract Analysis. The proof given here is based on a mean value theorem in the following lemma.

Lemma 7.0.1 *Let U be an open set in $\mathbb{R}^p$ which contains the line segment $t \to \boldsymbol{y} + t(\boldsymbol{z}-\boldsymbol{y})$ for $t \in [0,1]$ and let $f : U \to \mathbb{R}$ be differentiable at $\boldsymbol{y}+t(\boldsymbol{z}-\boldsymbol{y})$ for $t \in (0,1)$ and continuous for $t \in [0,1]$. Then there exists $\boldsymbol{x}$ on this line segment such that $f(\boldsymbol{z}) - f(\boldsymbol{y}) = Df(\boldsymbol{x})(\boldsymbol{z}-\boldsymbol{y})$.*

Proof: Let $h(t) \equiv f(\boldsymbol{y}+t(\boldsymbol{z}-\boldsymbol{y}))$ for $t \in [0,1]$. Then h is continuous on $[0,1]$ and has a derivative, $h'(t) = Df(\boldsymbol{y}+t(\boldsymbol{z}-\boldsymbol{y}))(\boldsymbol{z}-\boldsymbol{y})$, this by the chain rule. Then by the mean value theorem of one variable calculus, there exists $t \in (0,1)$ such that

$$f(\boldsymbol{z}) - f(\boldsymbol{y}) = h(1) - h(0) = h'(t) = Df(\boldsymbol{y}+t(\boldsymbol{z}-\boldsymbol{y}))(\boldsymbol{z}-\boldsymbol{y})$$

and we let $\boldsymbol{x} = \boldsymbol{y}+t(\boldsymbol{z}-\boldsymbol{y})$ for this t. ■

Also of use is the following lemma.

Lemma 7.0.2 *Let A be an $m \times n$ matrix and suppose that for all i,j, $|A_{ij}| \leq C$. Then the operator norm satisfies $\|A\| \leq Cmn$.*

Proof: Note that if $\boldsymbol{z}$ is a vector, $|\boldsymbol{z}| = \sup_{|\boldsymbol{y}|\leq 1}(\boldsymbol{z},\boldsymbol{y})$. Indeed, for $|\boldsymbol{y}| \leq 1$, the right side is no more than $|\boldsymbol{z}|$ thanks to the Cauchy Schwarz inequality and this can be achieved by letting $\boldsymbol{y} = \boldsymbol{z}/|\boldsymbol{z}|$.

$$\begin{aligned} \|A\| &\equiv \sup_{|\boldsymbol{x}|\leq 1} |A\boldsymbol{x}| = \sup_{|\boldsymbol{x}|\leq 1}\sup_{|\boldsymbol{y}|\leq 1} |(A\boldsymbol{x},\boldsymbol{y})| = \sup_{|\boldsymbol{x}|\leq 1}\sup_{|\boldsymbol{y}|\leq 1}\left|\sum_i\sum_j A_{ij}x_j y_i\right| \\ &\leq \sup_{|\boldsymbol{x}|\leq 1}\sup_{|\boldsymbol{y}|\leq 1}\sum_i\sum_j C|x_j||y_i| \leq C\sum_i\sum_j |\boldsymbol{x}||\boldsymbol{y}| = Cmn. \blacksquare \end{aligned}$$

Definition 7.0.3 *Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$ and $(\boldsymbol{x},\boldsymbol{y})$ will denote a typical point of $\mathbb{R}^n \times \mathbb{R}^m$ with $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$. Let $\boldsymbol{f} : U \to \mathbb{R}^p$ be in $C^1(U)$ meaning that all partial derivatives exist and are continuous. Then define*

$$\begin{aligned} D_1\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) &\equiv \begin{pmatrix} f_{1,x_1}(\boldsymbol{x},\boldsymbol{y}) & \cdots & f_{1,x_n}(\boldsymbol{x},\boldsymbol{y}) \\ \vdots & & \vdots \\ f_{p,x_1}(\boldsymbol{x},\boldsymbol{y}) & \cdots & f_{p,x_n}(\boldsymbol{x},\boldsymbol{y}) \end{pmatrix}, \\ D_2\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) &\equiv \begin{pmatrix} f_{1,y_1}(\boldsymbol{x},\boldsymbol{y}) & \cdots & f_{1,y_m}(\boldsymbol{x},\boldsymbol{y}) \\ \vdots & & \vdots \\ f_{p,y_1}(\boldsymbol{x},\boldsymbol{y}) & \cdots & f_{p,y_m}(\boldsymbol{x},\boldsymbol{y}) \end{pmatrix}. \end{aligned}$$

Definition 7.0.4 *Let $\delta,\eta>0$ satisfy: $\overline{B(\boldsymbol{x}_0,\delta)}\times\overline{B(\boldsymbol{y}_0,\eta)}\subseteq U$ where $\boldsymbol{f}:U\subseteq\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^p$ is given as*

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\begin{pmatrix} f_1(\boldsymbol{x},\boldsymbol{y})\\ f_2(\boldsymbol{x},\boldsymbol{y})\\ \vdots\\ f_p(\boldsymbol{x},\boldsymbol{y})\end{pmatrix}$$

and for $\begin{pmatrix}\boldsymbol{x}^1 & \cdots & \boldsymbol{x}^n\end{pmatrix}\in\overline{B(\boldsymbol{x}_0,\delta)}^p$ and $\boldsymbol{y}\in B(\boldsymbol{y}_0,\hat{\eta})$ define

$$J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^p,\boldsymbol{y}\right)\equiv\begin{pmatrix} f_{1,x_1}\left(\boldsymbol{x}^1,\boldsymbol{y}\right) & \cdots & f_{1,x_n}\left(\boldsymbol{x}^1,\boldsymbol{y}\right)\\ \vdots & & \vdots\\ f_{p,x_1}\left(\boldsymbol{x}^p,\boldsymbol{y}\right) & \cdots & f_{p,x_n}\left(\boldsymbol{x}^p,\boldsymbol{y}\right)\end{pmatrix}. \qquad (*)$$

Thus, its i^{th} row is $D_1f_i\left(\boldsymbol{x}^i,\boldsymbol{y}\right)$. Let K,r be constants.

By Theorem 6.9.5 $\boldsymbol{f}$ is differentiable and its derivative is the $p\times(n+m)$ matrix,

$$\begin{pmatrix} D_1\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) & D_2\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\end{pmatrix}.$$

Also, by Lemma 7.0.1, and $(\boldsymbol{x},\boldsymbol{y})\in B(\boldsymbol{x}_0,\delta)\times B(\boldsymbol{y}_0,\eta)\subseteq U$, and $\boldsymbol{h},\boldsymbol{k}$ sufficiently small, there are $\boldsymbol{x}^i$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{h}$ such that

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{y}+\boldsymbol{k})-\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) &= \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{y}+\boldsymbol{k})-\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}+\boldsymbol{k})+\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}\ +\boldsymbol{k})-\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\\ &= J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^p,\boldsymbol{y}+\boldsymbol{k}\right)\boldsymbol{h}+D_2\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{k}+\boldsymbol{o}(\boldsymbol{k}) \qquad (7.1)\end{aligned}$$

Proposition 7.0.5 *Suppose $g:\overline{B(\boldsymbol{x}_0,\delta)}\times\overline{B(\boldsymbol{y}_0,\eta_0)}\to[0,\infty)$ is continuous and also that $g(\boldsymbol{x}_0,\boldsymbol{y}_0)=0$ and if $\boldsymbol{x}\neq\boldsymbol{x}_0, g(\boldsymbol{x},\boldsymbol{y}_0)>0$. Then there exists $\eta<\eta_0$ such that if $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta)$, then the function $\boldsymbol{x}\to g(\boldsymbol{x},\boldsymbol{y})$ achieves its minimum on the open set $B(\boldsymbol{x}_0,\delta)$.*

Proof: If not, then there is a sequence $\boldsymbol{y}_k\to\boldsymbol{y}_0$ but the minimum of $\boldsymbol{x}\to g(\boldsymbol{x},\boldsymbol{y}_k)$ for $\boldsymbol{x}\in\overline{B(\boldsymbol{x}_0,\delta)}$ happens on $\partial B(\boldsymbol{x}_0,\delta)\equiv\partial B\equiv\{\boldsymbol{x}:|\boldsymbol{x}-\boldsymbol{x}_0|=\delta\}$ at $\boldsymbol{x}_k$. Now ∂B is closed and bounded and so compact. Hence there is a subsequence, still denoted with subscript k such that $\boldsymbol{x}_k\to\boldsymbol{x}\in\partial B$ and $\boldsymbol{y}_k\to\boldsymbol{y}_0$. Let $0<2\varepsilon<\min\{g(\hat{\boldsymbol{x}},\boldsymbol{y}_0):\hat{\boldsymbol{x}}\in\partial B\}$.

Then for k large, $|g(\boldsymbol{x}_k,\boldsymbol{y}_k)-g(\boldsymbol{x},\boldsymbol{y}_0)|<\varepsilon$, and $|g(\boldsymbol{x}_k,\boldsymbol{y}_k)-g(\boldsymbol{x}_k,\boldsymbol{y}_0)|<\varepsilon$, the second inequality from uniform continuity. Then from these inequalities, for k large,

$$\begin{aligned} g(\boldsymbol{x}_0,\boldsymbol{y}_k) &\geq g(\boldsymbol{x}_k,\boldsymbol{y}_k)>g(\boldsymbol{x}_k,\boldsymbol{y}_0)-\varepsilon\\ &> \min\{g(\hat{\boldsymbol{x}},\boldsymbol{y}_0):\hat{\boldsymbol{x}}\in\partial B\}-\varepsilon>2\varepsilon-\varepsilon=\varepsilon\end{aligned}$$

Now let $k\to\infty$ to conclude that $g(\boldsymbol{x}_0,\boldsymbol{y}_0)\geq\varepsilon$, a contradiction. ■

Here is the implicit function theorem. It is based on the mean value theorem from one variable calculus, the extreme value theorem from calculus, and the formula for the inverse of a matrix in terms of the transpose of the cofactor matrix divided by the determinant.

Theorem 7.0.6 *(implicit function theorem) Suppose U is an open set in $\mathbb{R}^n\times\mathbb{R}^m$. Let $\boldsymbol{f}:U\to\mathbb{R}^n$ be in $C^1(U)$ and suppose*

$$\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)=\boldsymbol{0},\ D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}\ \textit{exists.} \qquad (7.2)$$

Then there exist positive constants δ,η, such that for every $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta)$ there exists a unique $\boldsymbol{x}(\boldsymbol{y})\in B(\boldsymbol{x}_0,\delta)$ such that

$$\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})=\mathbf{0}. \tag{7.3}$$

Furthermore, the mapping, $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ is in $C^1(B(\boldsymbol{y}_0,\eta))$.

Proof: Let $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\begin{pmatrix} f_1(\boldsymbol{x},\boldsymbol{y}) & f_2(\boldsymbol{x},\boldsymbol{y}) & \cdots & f_n(\boldsymbol{x},\boldsymbol{y})\end{pmatrix}^T$. Also define the expression $J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right)$ to be given above in $*$. Then by the assumption of continuity of all the partial derivatives, there exists $r>0$ and $\delta_0,\eta_0>0$ such that if $\delta\leq\delta_0$ and $\eta\leq\eta_0$, it follows that for all $\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n\right)\in\overline{B(\boldsymbol{x}_0,\delta)}^n\equiv\overline{B(\boldsymbol{x}_0,\delta)}\times\overline{B(\boldsymbol{x}_0,\delta)}\times\cdots\times\overline{B(\boldsymbol{x}_0,\delta)}$, and $\boldsymbol{y}\in\overline{B(\boldsymbol{y}_0,\eta)}$,

$$\det J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right)\notin(-r,r). \tag{7.4}$$

and $\overline{B(\boldsymbol{x}_0,\delta_0)}\times\overline{B(\boldsymbol{y}_0,\eta_0)}\subseteq U$. Therefore, from the formula for the inverse of a matrix and continuity of all entries of the various matrices, there exists a constant K such that all entries of $J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right), J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right)^{-1}$, and $D_2\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ have absolute value smaller than K on the convex set $\overline{B(\boldsymbol{x}_0,\delta)}^n\times\overline{B(\boldsymbol{y}_0,\eta)}$ whenever δ,η are sufficiently small. It is always tacitly assumed that these radii are this small.

Next it is shown that for a given $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta),\eta\leq\eta_0$, there is at most one $\boldsymbol{x}\in B(\boldsymbol{x}_0,\delta_0)$ such that $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\mathbf{0}$.

Pick $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta)$ and suppose there exist $\boldsymbol{x},\boldsymbol{z}\in\overline{B(\boldsymbol{x}_0,\delta)}$ such that

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{f}(\boldsymbol{z},\boldsymbol{y})=\mathbf{0}.$$

Then applying Lemma 7.0.1 on the components, there are $\boldsymbol{x}^i$ such that

$$J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right)(\boldsymbol{z}-\boldsymbol{x})=\mathbf{0}$$

and so from 7.4 $\boldsymbol{z}-\boldsymbol{x}=\mathbf{0}$. (The matrix $J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}\right)$ is invertible since its determinant is nonzero.) Now it will be shown that if η is chosen sufficiently small, then for all $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta)$, there exists a unique $\boldsymbol{x}(\boldsymbol{y})\in B(\boldsymbol{x}_0,\delta)$ such that $\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})=\mathbf{0}$.

Claim: If η is small enough, then the function, $\boldsymbol{x}\to h_{\boldsymbol{y}}(\boldsymbol{x})\equiv|\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})|^2$ achieves its minimum value on $\overline{B(\boldsymbol{x}_0,\delta)}$ at a point of $B(\boldsymbol{x}_0,\delta)$. This is Proposition 7.0.5.

Choose $\eta<\eta_0$ and also small enough that the above claim holds and let $\boldsymbol{x}(\boldsymbol{y})$ denote a point of $B(\boldsymbol{x}_0,\delta)$ at which the minimum of $h_{\boldsymbol{y}}$ on $\overline{B(\boldsymbol{x}_0,\delta)}$ is achieved. Since $\boldsymbol{x}(\boldsymbol{y})$ is an interior point, it follows that you can consider $h_{\boldsymbol{y}}(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v})$ for $|t|$ small and conclude this function of t has a zero derivative at $t=0$. Now

$$h_{\boldsymbol{y}}(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v})=\sum_{i=1}^n f_i^2(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v},\boldsymbol{y})$$

and so from the chain rule,

$$\frac{d}{dt}h_{\boldsymbol{y}}(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v})=\sum_{i=1}^n\sum_{j=1}^n 2f_i(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v},\boldsymbol{y})\frac{\partial f_i(\boldsymbol{x}(\boldsymbol{y})+t\boldsymbol{v},\boldsymbol{y})}{\partial x_j}v_j.$$

Therefore, letting $t=0$, it is required that for every $\boldsymbol{v}$,

$$\sum_{i=1}^n\sum_{j=1}^n 2f_i(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\frac{\partial f_i(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})}{\partial x_j}v_j=0.$$

In terms of matrices this reduces to $0 = 2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^T D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{v}$ for every vector $\boldsymbol{v}$. Therefore, $\boldsymbol{0} = \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^T D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})$. From 7.4, it follows $\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) = \boldsymbol{0}$. (Multiply by $D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1}$ on the right.) This proves the existence of the function $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ such that $\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) = \boldsymbol{0}$ for all $\boldsymbol{y}\in B(\boldsymbol{y}_0,\eta)$.

It remains to verify this function is a C^1 function. To do this, let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be points of $B(\boldsymbol{y}_0,\eta)$. Then as before, consider the i^{th} component of $\boldsymbol{f}$ and consider the same argument using the mean value theorem to write

$$0 = f_i(\boldsymbol{x}(\boldsymbol{y}_1),\boldsymbol{y}_1) - f_i(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}_2)$$

$$= f_i(\boldsymbol{x}(\boldsymbol{y}_1),\boldsymbol{y}_1) - f_i(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}_1) + f_i(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}_1) - f_i(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}_2)$$
$$= D_1 f_i\left(\boldsymbol{x}^i,\boldsymbol{y}_1\right)(\boldsymbol{x}(\boldsymbol{y}_1) - \boldsymbol{x}(\boldsymbol{y}_2)) + D_2 f_i\left(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}^i\right)(\boldsymbol{y}_1 - \boldsymbol{y}_2) \tag{7.5}$$

where $\boldsymbol{y}^i$ is a point on the line segment joining $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ and $\boldsymbol{x}^i$ is a point on the line segment joining $\boldsymbol{x}(\boldsymbol{y}_1)$ and $\boldsymbol{x}(\boldsymbol{y}_2)$. Thus

$$(\boldsymbol{x}(\boldsymbol{y}_1) - \boldsymbol{x}(\boldsymbol{y}_2)) = -J\left(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^n,\boldsymbol{y}_1\right)^{-1} M(\boldsymbol{y}_1 - \boldsymbol{y}_2)$$

where M denotes the matrix having the i^{th} row equal to $D_2 f_i\left(\boldsymbol{x}(\boldsymbol{y}_2),\boldsymbol{y}^i\right)$ all entries being bounded by K. It follows that

$$|\boldsymbol{x}(\boldsymbol{y}_1) - \boldsymbol{x}(\boldsymbol{y}_2)| \leq Kn|M(\boldsymbol{y}_1 - \boldsymbol{y}_2)| \leq K^2 nm|\boldsymbol{y}_1 - \boldsymbol{y}_2|$$

Thus $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ is continuous near $\boldsymbol{y}_0$.

Now let $\boldsymbol{y}_2 = \boldsymbol{y}, \boldsymbol{y}_1 = \boldsymbol{y} + h\boldsymbol{e}_k$ for small h. Then M described above depends on h and $\lim_{h\to 0} M(h) = D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})$ thanks to the continuity of $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ just shown. Also,

$$\frac{\boldsymbol{x}(\boldsymbol{y}+h\boldsymbol{e}_k) - \boldsymbol{x}(\boldsymbol{y})}{h} = -J\left(\boldsymbol{x}^1(h),\cdots,\boldsymbol{x}^n(h),\boldsymbol{y}+h\boldsymbol{e}_k\right)^{-1} M(h)\boldsymbol{e}_k$$

Passing to a limit and using the formula for the inverse of a matrix in terms of the cofactor matrix, and the continuity of $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ shown above, this yields

$$\frac{\partial\boldsymbol{x}}{\partial y_k} = -D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1} D_2 f_i(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{e}_k$$

Then continuity of $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ and the assumed continuity of the partial derivatives of $\boldsymbol{f}$ shows that each partial derivative of $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ exists and is continuous. ■

This theorem implies the inverse function theorem stated next.

Theorem 7.0.7 *(inverse function theorem) Let $\boldsymbol{x}_0 \in U$, an open set in $\mathbb{R}^n$, and let $\boldsymbol{f} : U \to \mathbb{R}^n$. Suppose*

$$\boldsymbol{f} \text{ is } C^1(U) \text{ , and } D\boldsymbol{f}(\boldsymbol{x}_0)^{-1} \text{ exists.} \tag{7.6}$$

Then there exist open sets W, and V such that $\boldsymbol{x}_0 \in W \subseteq U$, $\boldsymbol{f} : W \to V$ is one to one and onto, $\boldsymbol{f}^{-1}$ is C^1.

Proof: Apply the implicit function theorem to the function $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) \equiv \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}$ where $\boldsymbol{y}_0 \equiv \boldsymbol{f}(\boldsymbol{x}_0)$. Thus the function $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ defined in that theorem is $\boldsymbol{f}^{-1}$ and there is $B(\boldsymbol{y}_0,\eta)$ where this function is defined. Now let $W \equiv \boldsymbol{f}^{-1}(B(\boldsymbol{y}_0,\eta))$ and $V \equiv B(\boldsymbol{y}_0,\eta)$. ■

7.1 More Continuous Partial Derivatives

The implicit function theorem will now be improved slightly. If $\boldsymbol{f}$ is C^k, it follows that the function which is implicitly defined is also C^k, not just C^1, meaning all mixed partial derivatives of $\boldsymbol{f}$ up to order k are continuous. Since the inverse function theorem comes as a case of the implicit function theorem, this shows that the inverse function also inherits the property of being C^k. First some notation is convenient. Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ where each α_i is a nonnegative integer. Then letting $|\alpha| = \sum_i \alpha_i$,

$$D^\alpha \boldsymbol{f}(\boldsymbol{x}) \equiv \frac{\partial^{|\alpha|} \boldsymbol{f}}{\partial^{\alpha_1} \partial^{\alpha_2} \dots \partial^{\alpha_n}} (\boldsymbol{x}),\ D^0 \boldsymbol{f}(\boldsymbol{x}) \equiv \boldsymbol{f}(\boldsymbol{x})$$

The symbol on the right means to take the α_n partial derivative with respect to x_n, then the α_{n-1} partial derivative with respect to x_{n-1} of what you just got and so on till you take the α_1 partial derivative with respect to x_1. The idea is to show that all mixed partial derivatives such that $|\alpha| \leq k$ exist and are continuous.

Theorem 7.1.1 *(implicit function theorem) Suppose U is an open set in $\mathbb{F}^n \times \mathbb{F}^m$. Let $\boldsymbol{f} : U \to \mathbb{F}^n$ be in $C^k(U)$ and suppose*

$$\boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{0},\ D_1 \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{y}_0)^{-1} \in \mathscr{L}(\mathbb{F}^n, \mathbb{F}^n). \tag{7.7}$$

Then there exist positive constants δ, η, such that for every $\boldsymbol{y} \in B(\boldsymbol{y}_0, \eta)$ there exists a unique $\boldsymbol{x}(\boldsymbol{y}) \in B(\boldsymbol{x}_0, \delta)$ such that

$$\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y}) = \boldsymbol{0}. \tag{7.8}$$

Furthermore, the mapping $\boldsymbol{y} \to \boldsymbol{x}(\boldsymbol{y})$ is in $C^k(B(\boldsymbol{y}_0, \eta))$.

Proof: From the implicit function theorem $\boldsymbol{y} \to \boldsymbol{x}(\boldsymbol{y})$ is C^1. It remains to show that it is C^k for $k > 1$ assuming that $\boldsymbol{f}$ is C^k. From 7.8

$$\frac{\partial \boldsymbol{x}}{\partial y^l} = -D_1 \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})^{-1} \frac{\partial \boldsymbol{f}}{\partial y^l}.$$

By the formula for the inverse in terms of cofactors, if $\boldsymbol{f}$ is C^2, one can use the chain rule to take another continuous derivative. Thus, the following formula holds for $q = 1$ and $|\alpha| = q$.

$$D^\alpha \boldsymbol{x}(\boldsymbol{y}) = \sum_{|\beta| \leq q} M_\beta(\boldsymbol{x}, \boldsymbol{y}) D^\beta \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{7.9}$$

where M_β is a matrix whose entries are differentiable functions of $D^\gamma \boldsymbol{x}$ for $|\gamma| < q$ and $D^\tau \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ for $|\tau| \leq q$. This follows easily from the description of $D_1 \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})^{-1}$ in terms of the cofactor matrix and the determinant of $D_1 \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$. Suppose 7.9 holds for $|\alpha| = q < k$. Then by induction, this yields $\boldsymbol{x}$ is C^q. Then

$$\frac{\partial D^\alpha \boldsymbol{x}(\boldsymbol{y})}{\partial y^p} = \sum_{|\beta| \leq |\alpha|} \frac{\partial M_\beta(\boldsymbol{x}, \boldsymbol{y})}{\partial y^p} D^\beta \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) + M_\beta(\boldsymbol{x}, \boldsymbol{y}) \frac{\partial D^\beta \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})}{\partial y^p}.$$

By the chain rule $\frac{\partial M_\beta(\boldsymbol{x}, \boldsymbol{y})}{\partial y^p}$ is a matrix whose entries are differentiable functions of the matrix $D^\tau \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ for $|\tau| \leq q+1$ and $D^\gamma \boldsymbol{x}$ for $|\gamma| < q+1$. It follows, since y^p was arbitrary, that for any $|\alpha| = q+1$, a formula like 7.9 holds with q being replaced by $q+1$. Continuing this way, $\boldsymbol{x}$ is C^k. ■

As a simple corollary, this yields the inverse function theorem. You just let $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})$ and apply the implicit function theorem.

Theorem 7.1.2 *(inverse function theorem) Let $\boldsymbol{x}_0 \in U \subseteq \mathbb{F}^n$ and let $\boldsymbol{f} : U \to \mathbb{F}^n$. Suppose for k a positive integer, $\boldsymbol{f}$ is $C^k(U)$, and $D\boldsymbol{f}(\boldsymbol{x}_0)^{-1} \in \mathscr{L}(\mathbb{F}^n,\mathbb{F}^n)$. Then there exist open sets W, and V such that $\boldsymbol{x}_0 \in W \subseteq U$, $\boldsymbol{f} : W \to V$ is one to one and onto, $\boldsymbol{f}^{-1}$ is C^k.*

7.2 Normed Linear Space

The implicit function theorem and inverse function theorem continue to hold if $\mathbb{R}^n$ and $\mathbb{R}^m$ are replaced by finite dimensional normed linear spaces X,Y respectively of dimension n and m.

Theorem 7.2.1 *(implicit function theorem) Suppose U is an open set in $X \times Y$ where X,Y are normed linear space of dimension n,m and suppose $\boldsymbol{f} : U \to Z$ be in $C^k(U)$ where Z is an n dimensional normed linear space. Suppose also*

$$f(x_0,y_0) = 0,\ D_1 f(x_0,y_0)^{-1} \text{ exists.} \tag{7.10}$$

Then there exist positive constants δ,η, such that for every $y \in B(y_0,\eta)$ there exists a unique $x(y) \in B(x_0,\delta)$ such that

$$f(x(y),y) = 0. \tag{7.11}$$

Furthermore, the mapping, $y \to x(y)$ is in $C^k(B(y_0,\eta))$.

Proof: Denote the coordinate maps for X,Y,Z in terms of bases for these spaces by $\theta_X,\theta_Y,\theta_Z$. These are all linear maps and so, since we are in finite dimensions, they are each C^k for every positive integer k with respect to any norm on $\mathbb{R}^n,\mathbb{R}^m$ thanks to Theorem 4.4.9 on equivalence of norms and the same is true of their inverses. Denote by $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$ the coordinate vectors for $x,y,z \in X,Y,Z$ respectively. Let $\boldsymbol{f} = \theta_Z f$ and note that the conditions for the implicit function theorem, Theorem 7.1.1 for $\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0) = \boldsymbol{0}$ all hold and so this proves the theorem. Since we are in finite dimensions, $D_1 f(x_0,y_0)^{-1}$ exists if $D_1 f(x_0,y_0)$ is one to one which implies $D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}$ exists. ■

Of course the inverse function theorem follows from this in the case of normed linear spaces. This also illustrates how you can always reduce to $\mathbb{R}^p$ by doing everything in terms of coordinates.

7.3 The Method of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers from calculus. Recall the problem is to maximize or minimize a function subject to equality constraints. Let $f : U \to \mathbb{R}$ be a C^1 function where $U \subseteq \mathbb{R}^n$ and let

$$g_i(\boldsymbol{x}) = 0,\ i = 1,\cdots,m \tag{7.12}$$

be a collection of equality constraints with $m < n$. Now consider the system of nonlinear equations

$$f(\boldsymbol{x}) = a,\ \ g_i(\boldsymbol{x}) = 0,\ i = 1,\cdots,m.$$

$\boldsymbol{x}_0$ is a local maximum if $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x}$ near $\boldsymbol{x}_0$ which also satisfies the constraints 7.12. A local minimum is defined similarly. Let $\boldsymbol{F} : U \times \mathbb{R} \to \mathbb{R}^{m+1}$ be defined by

$$\boldsymbol{F}(\boldsymbol{x},a) \equiv \begin{pmatrix} f(\boldsymbol{x}) - a \\ g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{pmatrix}. \tag{7.13}$$

Now consider the $m+1 \times n$ Jacobian matrix, the matrix of the linear transformation, $D_1\boldsymbol{F}(\boldsymbol{x},a)$ with respect to the usual basis for $\mathbb{R}^n$ and $\mathbb{R}^{m+1}$.

$$\begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) & \cdots & f_{x_n}(\boldsymbol{x}_0) \\ g_{1x_1}(\boldsymbol{x}_0) & \cdots & g_{1x_n}(\boldsymbol{x}_0) \\ \vdots & & \vdots \\ g_{mx_1}(\boldsymbol{x}_0) & \cdots & g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix}.$$

If this matrix has rank $m+1$ then some $m+1 \times m+1$ submatrix has nonzero determinant. It follows from the implicit function theorem that there exist $m+1$ variables, $x_{i_1}, \cdots, x_{i_{m+1}}$ such that the system

$$\boldsymbol{F}(\boldsymbol{x},a) = \boldsymbol{0} \tag{7.14}$$

specifies these $m+1$ variables as a function of the remaining $n-(m+1)$ variables and a in an open set of $\mathbb{R}^{n-m}$. Thus there is a solution $(\boldsymbol{x},a)$ to 7.14 for some $\boldsymbol{x}$ close to $\boldsymbol{x}_0$ whenever a is in some open interval. Therefore, $\boldsymbol{x}_0$ cannot be either a local minimum or a local maximum. It follows that if $\boldsymbol{x}_0$ is either a local maximum or a local minimum, then the above matrix must have rank less than $m+1$ which requires the rows to be linearly dependent. Thus, there exist m scalars $\lambda_1, \cdots, \lambda_m$, and a scalar μ, not all zero such that

$$\mu \begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) \\ \vdots \\ f_{x_n}(\boldsymbol{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix}. \tag{7.15}$$

If the column vectors

$$\begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix}, \cdots \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix} \tag{7.16}$$

are linearly independent, then, $\mu \neq 0$ and dividing by μ yields an expression of the form

$$\begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) \\ \vdots \\ f_{x_n}(\boldsymbol{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix} \tag{7.17}$$

at every point $\boldsymbol{x}_0$ which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 7.3.1 *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a C^1 function. Then if $\boldsymbol{x}_0 \in U$ is either a local maximum or local minimum of f subject to the constraints 7.12, then 7.15 must hold for some scalars $\mu, \lambda_1, \cdots, \lambda_m$ not all equal to zero. If the vectors in 7.16 are linearly independent, it follows that an equation of the form 7.17 holds.*

7.4 Taylor Approximations

First recall the following one variable calculus theorem. It is in my on line book "Calculus of One and Many Variables" or in any elementary Calculus book. See Problem 3 below on Page 176.

Theorem 7.4.1 *Let $h : (-\delta, 1+\delta) \to \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in (0,1)$ such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Now suppose U is an open set in $\mathbb{R}^p$ and $f : U \to \mathbb{R}$ is C^{m+1} with $\boldsymbol{x}_0 \in U$. For $\boldsymbol{x} \in B(\boldsymbol{x}_0, r) \subseteq U$, let $h(t) = f(\boldsymbol{x}_0 + t(\boldsymbol{x} - \boldsymbol{x}_0)), t \in (0,1)$. Then

$$h'(t) = \sum_i \frac{\partial f(\boldsymbol{x}_0 + t(\boldsymbol{x} - \boldsymbol{x}_0))}{\partial x_i}(x_i - x_{0i}),\ h''(t) = \sum_{i_1, i_2} \frac{\partial^2 f}{\partial x_{i_1} \partial x_{i_2}}(x_{i_1} - x_{0i_1})(x_{i_2} - x_{0i_2})$$

and continuing this way,

$$h^{(k)}(t) = \sum_{i_1, \cdots, i_k} \frac{\partial^k f}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_k}} \prod_{j=1}^{k} (x_{i_j} - x_{0i_j}) \tag{7.18}$$

Then the Taylor approximation is of the form $h(1) = f(\boldsymbol{x}) =$

$$\begin{aligned} &f(\boldsymbol{x}_0) + \sum_{k=1}^{m} \frac{1}{k!} \sum_{i_1, \cdots, i_k} \frac{\partial^k f(\boldsymbol{x}_0)}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_k}} \prod_{j=1}^{k} (x_{i_j} - x_{0i_j}) \\ &+ \frac{1}{(m+1)!} \sum_{i_1, \cdots, i_{m+1}} \frac{\partial^{m+1} f(\boldsymbol{x}_0 + t(\boldsymbol{x} - \boldsymbol{x}_0))}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_{m+1}}} \prod_{j=1}^{m+1} (x_{i_j} - x_{0i_j}) \end{aligned} \tag{7.19}$$

The last term being the remainder with $t \in (0,1)$. Thus, if the $(m+1)^{st}$ partial derivatives are all bounded, this shows that if $\|\boldsymbol{x} - \boldsymbol{x}_0\|$ is sufficiently small, then the difference between $f(\boldsymbol{x})$ and that series on the right in 7.19 other than the remainder term will also be very small.

7.5 Second Derivative Test

Now consider the case where $U \subseteq \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is $C^2(U)$. Then from Taylor's theorem, if $\boldsymbol{v}$ is small enough, there exists $t \in (0,1)$ such that

$$f(\boldsymbol{x} + \boldsymbol{v}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})\boldsymbol{v} + \frac{D^2 f(\boldsymbol{x} + t\boldsymbol{v})\boldsymbol{v}^2}{2}. \tag{7.20}$$

Consider

$$\begin{aligned} D^2 f(\boldsymbol{x} + t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) &\equiv D(D(f(\boldsymbol{x} + t\boldsymbol{v}))\boldsymbol{e}_i)\boldsymbol{e}_j \\ &= D\left(\frac{\partial f(\boldsymbol{x} + t\boldsymbol{v})}{\partial x_i}\right)\boldsymbol{e}_j = \frac{\partial^2 f(\boldsymbol{x} + t\boldsymbol{v})}{\partial x_j \partial x_i} \end{aligned}$$

where $\boldsymbol{e}_i$ are the usual basis vectors. Lettin $\boldsymbol{v} = \sum_{i=1}^{n} v_i \boldsymbol{e}_i$, the second derivative term in 7.20 reduces to

$$\frac{1}{2}\sum_{i,j} D^2 f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) v_i v_j = \frac{1}{2}\sum_{i,j} H_{ij}(\boldsymbol{x}+t\boldsymbol{v}) v_i v_j$$

where

$$H_{ij}(\boldsymbol{x}+t\boldsymbol{v}) = D^2 f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) = \frac{\partial^2 f(\boldsymbol{x}+t\boldsymbol{v})}{\partial x_j \partial x_i}.$$

Definition 7.5.1 *The matrix whose ij^{th} entry is $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i}$ is called the Hessian matrix, denoted as $\boldsymbol{H}(\boldsymbol{x})$.*

From Theorem 6.10.1, this is a symmetric real matrix, thus self adjoint. By the continuity of the second partial derivative,

$$f(\boldsymbol{x}+\boldsymbol{v}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T H(\boldsymbol{x})\boldsymbol{v} +$$
$$\frac{1}{2}\left(\boldsymbol{v}^T (H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x}))\boldsymbol{v}\right). \tag{7.21}$$

where the last two terms involve ordinary matrix multiplication and

$$\boldsymbol{v}^T = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$$

for v_i the components of $\boldsymbol{v}$ relative to the standard basis.

Definition 7.5.2 *Let $f : D \to \mathbb{R}$ where D is a subset of some normed vector space. Then f has a local minimum at $\boldsymbol{x} \in D$ if there exists $\delta > 0$ such that for all $\boldsymbol{y} \in B(\boldsymbol{x}, \delta)$, $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$. Also f has a local maximum at $\boldsymbol{x} \in D$ if there exists $\delta > 0$ such that for all $\boldsymbol{y} \in B(\boldsymbol{x}, \delta)$, $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$.*

Theorem 7.5.3 *If $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^n$ and f is C^2, suppose $Df(\boldsymbol{x}) = 0$. Then if $H(\boldsymbol{x})$ has all positive eigenvalues, $\boldsymbol{x}$ is a local minimum. If the Hessian matrix $H(\boldsymbol{x})$ has all negative eigenvalues, then $\boldsymbol{x}$ is a local maximum. If $H(\boldsymbol{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\boldsymbol{x}$, while if $H(\boldsymbol{x})$ has a negative eigenvalue, there exists a direction in which $H(\boldsymbol{x})$ has a local maximum at $\boldsymbol{x}$.*

Proof: Since $Df(\boldsymbol{x}) = 0$, formula 7.21 holds and by continuity of the second derivative, $H(\boldsymbol{x})$ is a symmetric matrix. Thus $H(\boldsymbol{x})$ has all real eigenvalues. Suppose first that $H(\boldsymbol{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$. Then by Theorem 1.4.1, $H(\boldsymbol{x})$ has an orthonormal basis of eigenvectors, $\{\boldsymbol{v}_i\}_{i=1}^n$ and if $\boldsymbol{u}$ is an arbitrary vector, such that $\boldsymbol{u} = \sum_{j=1}^n u_j \boldsymbol{v}_j$ where $u_j = \boldsymbol{u} \cdot \boldsymbol{v}_j$, then

$$\boldsymbol{u}^T H(\boldsymbol{x})\boldsymbol{u} = \sum_{j=1}^n u_j \boldsymbol{v}_j^T H(\boldsymbol{x}) \sum_{j=1}^n u_j \boldsymbol{v}_j = \sum_{j=1}^n u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^n u_j^2 = \delta^2 |\boldsymbol{u}|^2.$$

From 7.21 and the continuity of H, if $\boldsymbol{v}$ is small enough,

$$f(\boldsymbol{x}+\boldsymbol{v}) \geq f(\boldsymbol{x}) + \frac{1}{2}\delta^2 |\boldsymbol{v}|^2 - \frac{1}{4}\delta^2 |\boldsymbol{v}|^2 = f(\boldsymbol{x}) + \frac{\delta^2}{4}|\boldsymbol{v}|^2.$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H(\boldsymbol{x})$ has a positive eigenvalue λ^2. Then let $\boldsymbol{v}$ be an eigenvector for this eigenvalue. Then from 7.21,

$$f(\boldsymbol{x}+t\boldsymbol{v}) = f(\boldsymbol{x}) + \frac{1}{2}t^2\boldsymbol{v}^T H(\boldsymbol{x})\boldsymbol{v} + \frac{1}{2}t^2\left(\boldsymbol{v}^T\left(H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x})\right)\boldsymbol{v}\right)$$

which implies

$$\begin{aligned} f(\boldsymbol{x}+t\boldsymbol{v}) &= f(\boldsymbol{x}) + \frac{1}{2}t^2\lambda^2|v|^2 + \frac{1}{2}t^2\left(\boldsymbol{v}^T\left(H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x})\right)\boldsymbol{v}\right) \\ &\geq f(\boldsymbol{x}) + \frac{1}{4}t^2\lambda^2|\boldsymbol{v}|^2 \end{aligned}$$

whenever t is small enough. Thus in the direction $\boldsymbol{v}$ the function has a local minimum at $\boldsymbol{x}$. The assertion about the local maximum in some direction follows similarly. This proves the theorem. ■

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1(x,y) = x^4 + y^2, \ f_2(x,y) = -x^4 + y^2.$$

Then $Df_i(0,0) = \mathbf{0}$ and for both functions, the Hessian matrix evaluated at $(0,0)$ equals

$$\begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.

7.6 The Rank Theorem

This is a very interesting result. The proof follows Marsden and Hoffman. First here is some linear algebra.

Theorem 7.6.1 *Let $L : \mathbb{R}^n \to \mathbb{R}^N$ have rank m. Then there exists a basis*

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$$

such that a basis for $\ker(L)$ *is* $\{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$.

Proof: Since L has rank m, there is a basis for $L(\mathbb{R}^n)$ which is of the form

$$\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$$

Then if $\sum_i c_i\boldsymbol{u}_i = 0$ you can do L to both sides and conclude that each $c_i = 0$. Hence $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ is linearly independent. Let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ be a basis for $\ker(L)$. Let $\boldsymbol{x} \in \mathbb{R}^n$. Then $L\boldsymbol{x} = \sum_{i=1}^m c_i L\boldsymbol{u}_i$ for some choice of scalars c_i. Hence $L(\boldsymbol{x} - \sum_{i=1}^m c_i\boldsymbol{u}_i) = \mathbf{0}$ which shows that there exist d_j such that $\boldsymbol{x} = \sum_{i=1}^m c_i\boldsymbol{u}_i + \sum_{j=1}^k d_j\boldsymbol{v}_j$ It follows that

$$\operatorname{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_k) = \mathbb{R}^n$$

Is this set of vectors linearly independent? Suppose $\sum_{i=1}^{m} c_i \boldsymbol{u}_i + \sum_{j=1}^{k} d_j \boldsymbol{v}_j = \mathbf{0}$ Do L to both sides to get $\sum_{i=1}^{m} c_i L\boldsymbol{u}_i = \mathbf{0}$ Thus each $c_i = 0$. Hence $\sum_{j=1}^{k} d_j \boldsymbol{v}_j = \mathbf{0}$ and so each $d_j = 0$ also. It follows that $k = n - m$ and we can let

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\} = \{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}. \blacksquare$$

Another useful linear algebra result is the following lemma.

Lemma 7.6.2 *Let $V \subseteq \mathbb{R}^n$ be a subspace and suppose $A(\boldsymbol{x}) \in \mathscr{L}(V, \mathbb{R}^N)$ for $\boldsymbol{x}$ in some open set U. Also suppose $\boldsymbol{x} \to A(\boldsymbol{x})$ is continuous for $\boldsymbol{x} \in U$. Then if $A(\boldsymbol{x}_0)$ is one to one on V for some $\boldsymbol{x}_0 \in U$, then it follows that for all $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, $A(\boldsymbol{x})$ is also one to one on V.*

Proof: Consider V as an inner product space with the inner product from $\mathbb{R}^n$ and $A(\boldsymbol{x})^* A(\boldsymbol{x})$. Then $A(\boldsymbol{x})^* A(\boldsymbol{x}) \in \mathscr{L}(V, V)$ and $\boldsymbol{x} \to A(\boldsymbol{x})^* A(\boldsymbol{x})$ is also continuous. Also for $\boldsymbol{v} \in V$,

$$\left(A(\boldsymbol{x})^* A(\boldsymbol{x}) \boldsymbol{v}, \boldsymbol{v}\right)_V = (A(\boldsymbol{x}) \boldsymbol{v}, A(\boldsymbol{x}) \boldsymbol{v})_{\mathbb{R}^N}$$

If $A(\boldsymbol{x}_0)^* A(\boldsymbol{x}_0) \boldsymbol{v} = \mathbf{0},$ then from the above, it follows that $A(\boldsymbol{x}_0) \boldsymbol{v} = \mathbf{0}$ also. Therefore, $\boldsymbol{v} = \mathbf{0}$ and so $A(\boldsymbol{x}_0)^* A(\boldsymbol{x}_0)$ is one to one on V. For all $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, it follows from continuity that $A(\boldsymbol{x})^* A(\boldsymbol{x})$ is also one to one. Thus, for such $\boldsymbol{x},$ if $A(\boldsymbol{x}) \boldsymbol{v} = \mathbf{0},$ Then $A(\boldsymbol{x})^* A(\boldsymbol{x}) \boldsymbol{v} = \mathbf{0}$ and so $\boldsymbol{v} = \mathbf{0}$. Thus, for $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, it follows that $A(\boldsymbol{x})$ is also one to one on V. $\blacksquare$

Theorem 7.6.3 *Let $\boldsymbol{f} : A \subseteq \mathbb{R}^n \to \mathbb{R}^N$ where A is open in $\mathbb{R}^n$. Let $\boldsymbol{f}$ be a C^r function and suppose that $D\boldsymbol{f}(\boldsymbol{x})$ has rank m for all $\boldsymbol{x} \in A$. Let $\boldsymbol{x}_0 \in A$. Then there are open sets $U, V \subseteq \mathbb{R}^n$ with $\boldsymbol{x}_0 \in V$, and a C^r function $\boldsymbol{h} : U \to V$ with inverse $\boldsymbol{h}^{-1} : V \to U$ also C^r such that $\boldsymbol{f} \circ \boldsymbol{h}$ depends only on $(x_1, \cdots, x_m)$.*

Proof: Let $L = D\boldsymbol{f}(x_0)$, and $N_0 = \ker L$. Using the above linear algebra theorem, there exists

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$$

such that $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for $L\mathbb{R}^n$. Extend to form a basis for $\mathbb{R}^n$,

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$$

such that a basis for $N_0 = \ker L$ is $\{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$. Let

$$M \equiv \text{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m).$$

Let the coordinate maps be ψ_k so that if $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{x} = \psi_1(\boldsymbol{x}) \boldsymbol{u}_1 + \cdots + \psi_n(\boldsymbol{x}) \boldsymbol{u}_n$$

Since these coordinate maps are linear, they are infinitely differentiable.

Next I will define coordinate maps for $\boldsymbol{x} \in \mathbb{R}^N$. Then by the above construction, $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for $L(\mathbb{R}^n)$. Let a basis for $\mathbb{R}^N$ be

$$\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m, \boldsymbol{v}_{m+1}, \cdots, \boldsymbol{v}_N\}$$

(Note that, since the rank of $\boldsymbol{Df}(\boldsymbol{x}) = m$ you must have $N \geq m$.) The coordinate maps ϕ_i will be defined as follows for $\boldsymbol{x} \in \mathbb{R}^N$.

$$\boldsymbol{x} = \phi_1(\boldsymbol{x}) L\boldsymbol{u}_1 + \cdots \phi_m(\boldsymbol{x}) L\boldsymbol{u}_m + \phi_{m+1}(\boldsymbol{x}) \boldsymbol{v}_{m+1} + \cdots + \phi_N(\boldsymbol{x}) \boldsymbol{v}_N$$

Now define two infinitely differentiable maps $G : \mathbb{R}^n \to \mathbb{R}^n$ and $H : \mathbb{R}^N \to \mathbb{R}^n$,

$$G(\boldsymbol{x}) \equiv \left(0, \cdots, 0, \psi_{m+1}(\boldsymbol{x}), \cdots, \psi_n(\boldsymbol{x})\right)$$

$$H(\boldsymbol{y}) \equiv (\phi_1(\boldsymbol{y}), \cdots, \phi_m(\boldsymbol{y}), 0, \cdots, 0)$$

For $\boldsymbol{x} \in A \subseteq \mathbb{R}^n$, let

$$\boldsymbol{g}(\boldsymbol{x}) \equiv H(\boldsymbol{f}(\boldsymbol{x})) + G(\boldsymbol{x}) \in \mathbb{R}^n$$

Thus the first term picks out the first m entries of $\boldsymbol{f}(\boldsymbol{x})$ and the second term the last $n - m$ entries of $\boldsymbol{x}$. It is of the form

$$\left(\phi_1(\boldsymbol{f}(\boldsymbol{x})), \cdots, \phi_m(\boldsymbol{f}(\boldsymbol{x})), \psi_{m+1}(\boldsymbol{x}), \cdots, \psi_n(\boldsymbol{x})\right)$$

Then

$$D\boldsymbol{g}(\boldsymbol{x}_0)(\boldsymbol{v}) = HL(\boldsymbol{v}) + G\,\boldsymbol{v} = HL\boldsymbol{v} + G\boldsymbol{v} \tag{7.22}$$

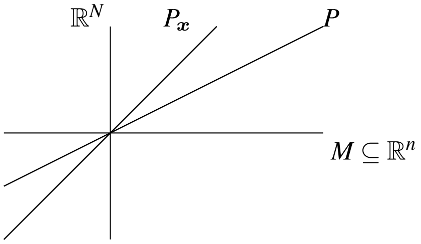
which is of the form

$$D\boldsymbol{g}(\boldsymbol{x}_0)(\boldsymbol{v}) = \left(\phi_1(L\boldsymbol{v}), \cdots, \phi_m(L\boldsymbol{v}), \psi_{m+1}(\boldsymbol{v}), \cdots, \psi_n(\boldsymbol{v})\right)$$

If this equals $\mathbf{0}$**,** then all the components of $\boldsymbol{v}$, $\psi_{m+1}(\boldsymbol{v}), \cdots, \psi_n(\boldsymbol{v})$ are equal to 0. Hence $\boldsymbol{v} = \sum_{i=1}^m c_i \boldsymbol{u}_i$. But also the coordinates of $L\boldsymbol{v}$**,**$\phi_1(L\boldsymbol{v}), \cdots, \phi_m(L\boldsymbol{v})$ are all zero so $L\boldsymbol{v} = \mathbf{0}$ and so $\mathbf{0} = \sum_{i=1}^m c_i L\boldsymbol{u}_i$ so by independence of the $L\boldsymbol{u}_i$, each $c_i = 0$ and consequently $\boldsymbol{v} = \mathbf{0}$**.**

This proves the conditions for the inverse function theorem are valid for $\boldsymbol{g}$. Therefore, there is an open ball U and an open set V, $\boldsymbol{x}_0 \in V$, such that $\boldsymbol{g} : V \to U$ is a C^r map and its inverse $\boldsymbol{g}^{-1} : U \to V$ is also. We can assume by continuity and Lemma 7.6.2 that V and U are small enough that for each $\boldsymbol{x} \in V, D\boldsymbol{g}(\boldsymbol{x})$ is one to one. This follows from the fact that $\boldsymbol{x} \to D\boldsymbol{g}(\boldsymbol{x})$ is continuous.

Since it is assumed that $D\boldsymbol{f}(\boldsymbol{x})$ is of rank $m, D\boldsymbol{f}(\boldsymbol{x})(\mathbb{R}^n)$ is a subspace which is m dimensional, denoted as $P_{\boldsymbol{x}}$. Also denote $L(\mathbb{R}^n) = L(M)$ as P.



Thus $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for P. Using Lemma 7.6.2 again, by making V, U smaller if necessary, one can also assume that for each $\boldsymbol{x} \in V$, $D\boldsymbol{f}(\boldsymbol{x})$ is one to one on M (although not on $\mathbb{R}^n$) and $HD\boldsymbol{f}(\boldsymbol{x})$ is one to one on M. This follows from continuity and the fact that $L = D\boldsymbol{f}(\boldsymbol{x}_0)$ is one to one on M. Therefore, it is also the case that $D\boldsymbol{f}(\boldsymbol{x})$ maps the m dimensional space M **onto** the m dimensional space $P_{\boldsymbol{x}}$ and H is one to one on $P_{\boldsymbol{x}}$. The reason for this last claim is as follows: If $H\boldsymbol{z} = \mathbf{0}$ where $\boldsymbol{z} \in P_{\boldsymbol{x}}$, then $HD\boldsymbol{f}(\boldsymbol{x})\boldsymbol{w} = \mathbf{0}$

where $\boldsymbol{w} \in M$ and $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{w} = \boldsymbol{z}$. Hence $\boldsymbol{w} = \mathbf{0}$ because $HD\boldsymbol{f}(\boldsymbol{x})$ is one to one, and so $\boldsymbol{z} = \mathbf{0}$ which shows that indeed H is one to one on $P_{\boldsymbol{x}}$.

Denote as $L_{\boldsymbol{x}}$ the inverse of H which is defined on $\mathbb{R}^m \times \mathbf{0}$, $L_{\boldsymbol{x}} : \mathbb{R}^m \times \mathbf{0} \to P_{\boldsymbol{x}}$. That $\mathbf{0}$ refers to the $N - m$ string of zeros in the definition given above for H.

Define $\boldsymbol{h} \equiv \boldsymbol{g}^{-1}$ and consider $\boldsymbol{f}_1 \equiv \boldsymbol{f} \circ \boldsymbol{h}$. It is desired to show that $\boldsymbol{f}_1$ depends only on $x_1, \cdots, x_m$. Let D_1 refer to $(x_1, \cdots, x_m)$ and let D_2 refer to $(x_{m+1}, \cdots, x_n)$. Then $\boldsymbol{f} = \boldsymbol{f}_1 \circ \boldsymbol{g}$ and so by the chain rule

$$D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) = D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \tag{7.23}$$

Now as in 7.22, for $\boldsymbol{y} \in \mathbb{R}^n$,

$$D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) = HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) + G\boldsymbol{y}$$

$$= \big(\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\big)$$

Recall that from the above definitions of H and G,

$$G(\boldsymbol{y}) \equiv \big(0, \cdots, 0, \psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\big)$$

$$H(D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})) = (\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), 0, \cdots, 0)$$

Let $\pi_1 : \mathbb{R}^n \to \mathbb{R}^m$ denote the projection onto the first m positions and π_2 the projection onto the last $n - m$. Thus

$$\begin{aligned} \pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) &= (\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y})) \\ \pi_2 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) &= \big(\psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\big) \end{aligned}$$

Now in general, for $\boldsymbol{z} \in \mathbb{R}^n$,

$$D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\boldsymbol{z} = D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1\boldsymbol{z} + D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2\boldsymbol{z}$$

Therefore, it follows that $D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y})$ is given by

$$\begin{aligned} D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) &= D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \\ &= D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) + D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \end{aligned}$$

$$\begin{aligned} D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) &= D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) = D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\overbrace{\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})}^{=\pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y})} \\ &\quad + D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} \end{aligned}$$

We need to verify the last term equals 0. Solving for this term,

$$D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})$$

As just explained, $L_{\boldsymbol{x}} \circ H$ is the identity on $P_{\boldsymbol{x}}$, the image of $D\boldsymbol{f}(\boldsymbol{x})$. Then

$$\begin{aligned} D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} &= L_{\boldsymbol{x}} \circ HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) \\ &= \Big(L_{\boldsymbol{x}} \circ \underline{HD\boldsymbol{f}(\boldsymbol{x})} - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 \underline{HD\boldsymbol{f}(\boldsymbol{x})}\Big)(\boldsymbol{y}) \end{aligned}$$

Factoring out that underlined term,

$$D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} = [L_{\boldsymbol{x}} - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1] HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})$$

Now $D\boldsymbol{f}(\boldsymbol{x}) : M \to P_{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{x})(\mathbb{R}^n)$ is onto. (This is based on the assumption that $D\boldsymbol{f}(\boldsymbol{x})$ has rank m.) Thus it suffices to consider only $\boldsymbol{y} \in M$ in the right side of the above. However, for such $\boldsymbol{y}, \pi_2 G\boldsymbol{y} = \boldsymbol{0}$ because to be in M, $\psi_k(\boldsymbol{y}) = 0$ if $k \geq m+1$, and so the left side of the above equals **0.** Thus it appears this term on the left is **0 for any $\boldsymbol{y}$ chosen.** How can this be so? It can only take place if $D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{0}$ for every $\boldsymbol{x} \in V$. Thus, since $\boldsymbol{g}$ is onto, it can only take place if $D_2\boldsymbol{f}_1(\boldsymbol{x}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in U$. Therefore on U it must be the case that $\boldsymbol{f}_1$ depends only on $x_1, \cdots, x_m$ as desired. ■

7.7 Exercises

1. Consider the question about level surfaces. Let $S = \left\{\boldsymbol{x} \in \mathbb{R}^{n+1} : f(\boldsymbol{x}) = c\right\}$. We usually refer to this as a level surface in $\mathbb{R}^{n+1}$ and we give examples of things like ellipsoids and spheres. Then everyone is deceived into thinking they know what is going on because of the examples. After this deception, and this is indeed what it is, we give specious arguments to justify the method of Lagrange multipliers (I have spent my career giving such specious arguments.) by showing that the gradient of the objective function is perpendicular to the direction vector of every smooth curve lying in S at a point where the maximum or minimum exists using the chain rule. One thing which is missing in this kind of stupidity is a consideration whether there even exist such smooth curves. Use the implicit function theorem to give conditions which imply the existence of such smooth curves near a point on S.

2. State and give a short proof of the inverse function theorem for normed linear spaces using Theorem 7.2.1.

3. Prove Theorem 7.4.1. **Hint:** Let K be such that
$$h(1) = h(0) + \sum_{k=1}^{m} \frac{1}{k!} h^{(k)}(0) + K.$$
Now define $g(u) \equiv h(1) - \left(h(u) + \sum_{k=1}^{m} \frac{1}{k!} h^{(k)}(u)(1-u)^k + K(1-u)^{m+1}\right)$. Then $g(0) = 0$ and $g(1) = 0$ so by the mean value theorem, there is $t \in (0,1)$ where $g'(t) = 0$. Compute $g'(u)$ and simplify then choose the t just mentioned and solve for K.

4. Let $\boldsymbol{f} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$
$$\boldsymbol{f}(x,y,\lambda) = \begin{pmatrix} x + xy + y^2 + \sin(\lambda) \\ x + y^2 - x^2 + \lambda \end{pmatrix}$$
Then $\boldsymbol{f}(0,0,\lambda) = \boldsymbol{0}$,
$$D_1\boldsymbol{f}(x,y,\lambda) = \begin{pmatrix} 1+y & x+2y \\ 1-2x & 2y \end{pmatrix} \text{ so } D_1\boldsymbol{f}((0,0),0) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Thus you can't say $\boldsymbol{f}(x,y,\lambda)=\mathbf{0}$ defines (x,y) as a function of λ near $(0,0,0)$. However, let

$$\begin{aligned} Q\begin{pmatrix}\alpha\\ \beta\end{pmatrix} &\equiv \begin{pmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}=\begin{pmatrix}\frac{\alpha+\beta}{2}\\ \frac{\alpha+\beta}{2}\end{pmatrix} \\ (I-Q)\begin{pmatrix}\alpha\\ \beta\end{pmatrix} &= \begin{pmatrix}\alpha\\ \beta\end{pmatrix}-\begin{pmatrix}\frac{\alpha+\beta}{2}\\ \frac{\alpha+\beta}{2}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\alpha-\frac{1}{2}\beta\\ \frac{1}{2}\beta-\frac{1}{2}\alpha\end{pmatrix}\end{aligned}$$

The equation $\boldsymbol{f}(x,y,\lambda)=\mathbf{0}$ can be written in the form

$$Q\boldsymbol{f}(x,y,\lambda)=\begin{pmatrix}-\frac{1}{2}x^2+\frac{1}{2}xy+x+y^2+\frac{1}{2}\lambda+\frac{1}{2}\sin\lambda\\ -\frac{1}{2}x^2+\frac{1}{2}xy+x+y^2+\frac{1}{2}\lambda+\frac{1}{2}\sin\lambda\end{pmatrix}=\mathbf{0} \tag{7.24}$$

$$(I-Q)\boldsymbol{f}(x,y,\lambda)=\begin{pmatrix}\frac{1}{2}x^2+\frac{1}{2}yx-\frac{1}{2}\lambda+\frac{1}{2}\sin\lambda\\ -\frac{1}{2}x^2-\frac{1}{2}yx+\frac{1}{2}\lambda-\frac{1}{2}\sin\lambda\end{pmatrix}=\mathbf{0}$$

$D_xQ\boldsymbol{f}(0,0,0)=\begin{pmatrix}1\\ 1\end{pmatrix}$ which is one to one on $\mathbb{R}$. Indeed, if $\begin{pmatrix}1\\ 1\end{pmatrix}u=\begin{pmatrix}0\\ 0\end{pmatrix}$, then $u=0$. By Theorem 7.2.1, the first equation in 7.24 defines $x=x(y,\lambda)$ for small y,λ. Also, you know it is a C^k function for every k so you can use Taylor approximation for functions of many variables to approximate $x(y,\lambda)$. In the top equation, $x_y=0$. Also $x_\lambda=-1$ so $x(y,\lambda)\approx-\lambda$ other than higher order terms for small y,λ. Now plug in to the bottom equation

$$\begin{aligned} & \frac{1}{2}x^2(y,\lambda)+\frac{1}{2}yx(y,\lambda)-\frac{1}{2}\lambda+\frac{1}{2}\sin\lambda \\ = & \frac{1}{2}(-\lambda)^2+\frac{1}{2}y(-\lambda)-\frac{1}{2}\lambda+\frac{1}{2}\sin\lambda=0\end{aligned}$$

Solve this for y to find $y(\lambda)=-1+\frac{\sin(\lambda)}{\lambda}+\lambda$ at least approximately. This kind of procedure is called the Lyapunov Schmidt procedure. It deals with the case where the partial derivative used in the statement of the implicit function theorem is not invertible. Note how it was possible to solve for a solution $\boldsymbol{f}(x,y,\lambda)=\mathbf{0}$ in this example.

5. Let $\boldsymbol{f}((x,y),\lambda)=\begin{pmatrix}x+xy+y^2+x\sin(\lambda)\\ x+y^2-x^2+x\lambda\end{pmatrix}$. One solution to $\boldsymbol{f}((x,y),\lambda)=\mathbf{0}$ is $x(\lambda)=y(\lambda)=0$. Use the above procedure to show there is a nonzero solution to this non-linear system of equations for small λ.

6. Let X,Y be finite dimensional vector spaces and let $L\in\mathscr{L}(X,Y)$. Let $\{Lx_1,...,Lx_m\}$ be a basis for $L(X)$. Show that if $\{z_1,...,z_r\}$ is a basis for $\ker(L)$, then a basis for X is $\{x_1,...,x_m,z_1,...,z_r\}$ is a basis for X. Show that L is one to one on $X_1\equiv\operatorname{span}(x_1,...,x_m)$.

7. Go through the details of the following argument. Let $\boldsymbol{f}:U\subseteq\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$ where U is open in $\mathbb{R}^n\times\mathbb{R}^m$, $(\mathbf{0},\mathbf{0})\in U$. Let $\boldsymbol{f}$ be C^k for $k\geq1$. Also suppose $\boldsymbol{f}(\mathbf{0},\mathbf{0})=\mathbf{0}$. If $L=D_1\boldsymbol{f}(0,0)$ and if L^{-1} exists, then by the implicit function theorem, the

equation $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}$ defines $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\lambda})$ for small $\boldsymbol{\lambda}$ and $\boldsymbol{x}$ is C^k. Let $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m\}$ be a basis for $L(\mathbb{R}^n)$ and enlarge to get $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m,\boldsymbol{w}_{m+1},\ldots,\boldsymbol{w}_n\}$ as a basis for $\mathbb{R}^n$. Letting $L\boldsymbol{x}_k = \boldsymbol{y}_k$ use the above problem to have a basis for X which is of the form $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{z}_{m+1},\ldots,\boldsymbol{z}_n\}$ with $\{\boldsymbol{z}_{m+1},\ldots,\boldsymbol{z}_n\}$ a basis for $\ker(L)$. Thus, from the above problem L is one to one on $X_1 \equiv \operatorname{span}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)$. For $\hat{\boldsymbol{x}} \in X_1$, show $D_{\hat{\boldsymbol{x}}}\boldsymbol{f}(\mathbf{0},\mathbf{0})$ is the restriction of L to X_1 and so $D_{\hat{\boldsymbol{x}}}\boldsymbol{f}(\mathbf{0},\mathbf{0})$ is one to one on X_1. Now define the linear map $Q:\mathbb{R}^n \to \mathbb{R}^n$ by $Q\left(\sum_{k=1}^m a_k\boldsymbol{y}_k + \sum_{k=m+1}^n b_k\boldsymbol{w}_k\right) \equiv \sum_{k=1}^m a_k\boldsymbol{y}_k$. Thus $Q^2 = Q$. We can write the original equations $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}$ as

$$\begin{gathered} Q\boldsymbol{f}(\hat{\boldsymbol{x}},\tilde{\boldsymbol{x}},\boldsymbol{\lambda}) = Q\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}, \tilde{\boldsymbol{x}} \in \ker(L) \\ (I-Q)\boldsymbol{f}(\hat{\boldsymbol{x}},\tilde{\boldsymbol{x}},\boldsymbol{\lambda}) = \mathbf{0} \end{gathered}$$

Thus $Q\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\lambda}) \in \operatorname{span}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m) \equiv Y_1$. Now show that for $\hat{\boldsymbol{x}}$ the variable in X_1, and if $\boldsymbol{v} \in X_1$, and $D_{\hat{\boldsymbol{x}}}Q\boldsymbol{f}(\mathbf{0},\mathbf{0},\mathbf{0})\boldsymbol{v} = \mathbf{0}$, then $\boldsymbol{v} = \mathbf{0}$ and so we can apply the implicit function theorem to obtain $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\tilde{\boldsymbol{x}},\boldsymbol{\lambda})$ as the solution to $Q\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}$ for $\tilde{\boldsymbol{x}},\boldsymbol{\lambda}$ small where here $\tilde{\boldsymbol{x}}$ is in $\ker(L)$. Since everything in sight is C^k, one can use Taylor series for functions of many variables to approximate the solution in these two equations. See the Taylor formula 7.19. This is the general idea in the above two problems.

Chapter 8

Measures and Measurable Functions

The Lebesgue integral is much better than the Rieman integral. This has been known for over 100 years. It is **much easier** to generalize to many dimensions and it is much easier to use in applications. It is also this integral which is most important in probability. However, this integral is more abstract. This chapter will develop the abstract machinery for this integral.

The next definition describes what is meant by a σ algebra. This is the fundamental object which is studied in probability theory. The events come from a σ algebra of sets. Recall that $\mathscr{P}(\Omega)$ is the set of all subsets of the given set Ω. It may also be denoted by 2^{Ω} but I won't refer to it this way.

Definition 8.0.1 *$\mathscr{F} \subseteq \mathscr{P}(\Omega)$, the set of all subsets of Ω, is called a σ algebra if it contains $\emptyset, \Omega$, and is closed with respect to countable unions and complements. That is, if $\{A_n\}_{n=1}^{\infty}$ is countable and each $A_n \in \mathscr{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathscr{F}$ also and if $A \in \mathscr{F}$, then $\Omega \setminus A \in \mathscr{F}$. It is clear that any intersection of σ algebras is a σ algebra. If $\mathscr{K} \subseteq \mathscr{P}(\Omega)$, $\sigma(\mathscr{K})$ is the smallest σ algebra which contains $\mathscr{K}$. In fact, the intersection of all σ algebras containing $\mathscr{K}$ is obviously a σ algebra so this intersection is $\sigma(\mathscr{K})$.*

If $\mathscr{F}$ is a σ algebra, then it is also closed with respect to countable intersections. Here is why. Let $\{F_k\}_{k=1}^{\infty} \subseteq \mathscr{F}$. Then $(\cap_k F_k)^C = \cup_k F_k^C \in \mathscr{F}$ and so $\cap_k F_k = \left((\cap_k F_k)^C\right)^C = \left(\cup_k F_k^C\right)^C \in \mathscr{F}$.

Example 8.0.2 *You could consider $\mathbb{N}$ and for your σ algebra, you could have $\mathscr{P}(\mathbb{N})$. This satisfies all the necessary requirements. Note that in fact, $\mathscr{P}(S)$ works for any S. However, useful examples are not typically the set of all subsets.*

8.1 Simple Functions and Measurable Functions

A σ algebra is a collection of subsets of a set Ω which includes $\emptyset, \Omega$, and is closed with respect to countable unions and complements.

Definition 8.1.1 *A measurable space, denoted as $(\Omega, \mathscr{F})$, is one for which $\mathscr{F}$ is a σ algebra contained in $\mathscr{P}(\Omega)$. Let $f : \Omega \to X$ where X is a metric space. Then f is said to be measurable means $f^{-1}(U) \in \mathscr{F}$ whenever U is open.*

It is important to have a theorem about pointwise limits of measurable functions. The following is a fairly general such theorem which holds in the situations to be considered in this book. First recall $\operatorname{dist}(x,S)$ in Lemma 3.12.1 which impliles that $x \to \operatorname{dist}(x,S)$ is continuous.

Theorem 8.1.2 *Let $\{f_n\}$ be a sequence of measurable functions mapping Ω to the metric space (X,d) where $(\Omega, \mathscr{F})$ is a measureable space. Suppose the pointwise limit $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ for all ω. Then f is also a measurable function.*

Proof: It is required to show $f^{-1}(U)$ is measurable for all U open. Let

$$V_m \equiv \left\{ x \in U : \operatorname{dist}\left(x, U^C\right) > \frac{1}{m} \right\}.$$

Thus, since dist is continuous, (Lemma 3.12.1), $V_m \subseteq \left\{x \in U : \text{dist}\left(x, U^C\right) \geq \frac{1}{m}\right\}$, $V_m \subseteq \overline{V_m} \subseteq V_{m+1}$, and $\cup_m V_m = U$. Then since V_m is open, $f^{-1}(V_m) = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m)$ and so

$$\begin{aligned} f^{-1}(U) &= \cup_{m=1}^{\infty} f^{-1}(V_m) = \cup_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m) \\ &\subseteq \cup_{m=1}^{\infty} f^{-1}\left(\overline{V_m}\right) = f^{-1}(U) \end{aligned}$$

which shows $f^{-1}(U)$ is measurable. ■

Important examples of a metric spaces are $\mathbb{R}, \mathbb{C}, \mathbb{F}^n$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. However, it is also very convenient to consider the metric space $(-\infty, \infty]$, the real line with ∞ tacked on at the end. This can be considered as a metric space in a very simple way.

$$\rho(x, y) = |\arctan(x) - \arctan(y)|$$

with the understanding that $\arctan(\infty) \equiv \pi/2$. It is easy to show that this metric restricted to $\mathbb{R}$ gives the same open sets on $\mathbb{R}$ as the usual metric given by $d(x, y) = |x - y|$ but in addition, allows the inclusion of that ideal point out at the end of the real line denoted as ∞. This is considered mainly because it makes the development of the theory easier. The open sets in $(-\infty, \infty]$ are described in the following lemma.

Lemma 8.1.3 *The open balls in $(-\infty, \infty]$ consist of sets of the form (a, b) for a, b real numbers and $(a, \infty]$. This is a separable metric space.*

Proof: If the center of the ball is a real number, then the ball will result in an interval (a, b) where a, b are real numbers. If the center of the ball is ∞, then the ball results in something of the form $(a, \infty]$. It is obvious that this is a separable metric space with the countable dense set being $\mathbb{Q}$ since every ball contains a rational number. ■

If you kept both $-\infty$ and ∞ with the obvious generalization that $\arctan(-\infty) \equiv -\frac{\pi}{2}$, then the resulting metric space would be a complete separable metric space. However, it is not convenient to include $-\infty$, so this won't be done. The reason is that it will be desired to make sense of things like $f + g$.

Then for functions which have values in $(-\infty, \infty]$ we have the following extremely useful description of what it means for a function to be measurable.

Lemma 8.1.4 *Let $f : \Omega \to (-\infty, \infty]$ where $\mathscr{F}$ is a σ algebra of subsets of Ω. Here $(-\infty, \infty]$ is the metric space just described with the metric given by*

$$\rho(x, y) = |\arctan(x) - \arctan(y)|.$$

Then the following are equivalent.

$$f^{-1}((d, \infty]) \in \mathscr{F}, \textit{for all finite } d,$$

$$f^{-1}((-\infty, d)) \in \mathscr{F}, \textit{for all finite } d,$$

$$f^{-1}([d, \infty]) \in \mathscr{F}, \textit{for all finite } d,$$

$$f^{-1}((-\infty, d]) \in \mathscr{F}, \textit{for all finite } d,$$

$$f^{-1}((a, b)) \in \mathscr{F} \textit{ for all } a < b, -\infty < a < b < \infty.$$

Any of these equivalent conditions is equivalent to the function being measurable.

Proof: First note that the first and the third are equivalent. To see this, observe $f^{-1}([d,\infty]) = \cap_{n=1}^{\infty} f^{-1}((d-1/n,\infty])$, and so if the first condition holds, then so does the third. $f^{-1}((d,\infty]) = \cup_{n=1}^{\infty} f^{-1}([d+1/n,\infty])$, and so if the third condition holds, so does the first.

Similarly, the second and fourth conditions are equivalent. Now from the definition of inverse image, $f^{-1}((-\infty,d]) = (f^{-1}((d,\infty]))^C$ so the first and fourth conditions are equivalent. Thus the first four conditions are equivalent and if any of them hold, then for $-\infty < a < b < \infty$, $f^{-1}((a,b)) = f^{-1}((-\infty,b)) \cap f^{-1}((a,\infty]) \in \mathscr{F}$. Finally, if the last condition holds, $f^{-1}([d,\infty]) = \left(\cup_{k=1}^{\infty} f^{-1}((-k+d,d))\right)^C \in \mathscr{F}$ and so the third condition holds. Therefore, all five conditions are equivalent.

Since $(-\infty,\infty]$ is a separable metric space, it follows from Theorem 3.4.2 that every open set U is a countable union of open intervals $U = \cup_k I_k$ where I_k is of the form (a,b) or $(a,\infty]$ and, as just shown if any of the equivalent conditions holds, then $f^{-1}(U) = \cup_k f^{-1}(I_k) \in \mathscr{F}$. Conversely, if $f^{-1}(U) \in \mathscr{F}$ for any open set $U \in (-\infty,\infty]$, then in particular, $f^{-1}((a,b)) \in \mathscr{F}$ which is one of the equivalent conditions and so all the equivalent conditions hold. ■

Note that if f is continuous and g is measurable, then $f \circ g$ is always measurable. This is because, for U open, $(f \circ g)^{-1}(U) = g^{-1}\left(f^{-1}(U)\right) = g^{-1}(\text{open})$ which is measurable.

There is a fundamental theorem about the relationship of simple functions to measurable functions given in the next theorem.

Definition 8.1.5 *Let $E \in \mathscr{F}$ for $\mathscr{F}$ a σ algebra. Then*

$$\mathscr{X}_E(\omega) \equiv \begin{cases} 1 \textit{ if } \omega \in E \\ 0 \textit{ if } \omega \notin E \end{cases}$$

This is called the indicator function of the set E. Let $s : (\Omega,\mathscr{F}) \to \mathbb{R}$. Then s is a simple function if it is of the form $s(\omega) = \sum_{i=1}^{n} c_i \mathscr{X}_{E_i}(\omega)$ where $E_i \in \mathscr{F}$ and $c_i \in \mathbb{R}$, the E_i being disjoint. Thus simple functions are those which have finitely many values and are measurable. In the next theorem, it will also be assumed that each $c_i \geq 0$.

Each simple function is measurable. This is easily seen as follows. First of all, you can assume the c_i are distinct because if not, you could just replace those E_i which correspond to a single value with their union. Then if you have any open interval (a,b), $s^{-1}((a,b)) = \cup\{E_i : c_i \in (a,b)\}$ and this is measurable because it is the finite union of measurable sets.

Theorem 8.1.6 *Let $f \geq 0$ be measurable. Then there exists a sequence of nonnegative simple functions $\{s_n\}$ satisfying*

$$0 \leq s_n(\omega) \tag{8.1}$$

$$\cdots s_n(\omega) \leq s_{n+1}(\omega) \cdots$$

$$f(\omega) = \lim_{n\to\infty} s_n(\omega) \textit{ for all } \omega \in \Omega. \tag{8.2}$$

If f is bounded, the convergence is actually uniform. Conversely, if f is nonnegative and is the pointwise limit of such simple functions, then f is measurable.

Proof: Letting $I \equiv \{\omega : f(\omega) = \infty\}$, define

$$t_n(\omega) = \sum_{k=0}^{2^n} \frac{k}{n} \mathscr{X}_{f^{-1}\left(\left[\frac{k}{n},\frac{k+1}{n}\right)\right)}(\omega) + 2^n \mathscr{X}_I(\omega).$$

Then $t_n(\omega) \leq f(\omega)$ for all ω and $\lim_{n\to\infty} t_n(\omega) = f(\omega)$ for all ω. This is because $t_n(\omega) = 2^n$ for $\omega \in I$ and if $f(\omega) \in [0, \frac{2^n+1}{n})$, then

$$0 \leq f(\omega) - t_n(\omega) \leq \frac{1}{n}. \tag{8.3}$$

Thus whenever $\omega \notin I$, the above inequality will hold for all n large enough. Let

$$s_1 = t_1,\ s_2 = \max(t_1, t_2),\ s_3 = \max(t_1, t_2, t_3), \cdots.$$

Then the sequence $\{s_n\}$ satisfies 8.1-8.2. Also each s_n has finitely many values and is measurable. To see this, note that $s_n^{-1}((a,\infty]) = \cup_{k=1}^n t_k^{-1}((a,\infty]) \in \mathscr{F}$

To verify the last claim, note that in this case the term $2^n \mathscr{X}_I(\omega)$ is not present and for n large enough, $2^n/n$ is larger than all values of f. Therefore, for all n large enough, 8.3 holds for all ω. Thus the convergence is uniform.

The last claim follows right away from Theorem 8.1.2. ■

Another useful observation is that the set where a sequence of measurable functions converges is also a measurable set.

Proposition 8.1.7 *Let $\{f_n\}$ be measurable with values in $(-\infty,\infty)$. Let*

$$A \equiv \{\omega : \{f_n(\omega)\} \text{ converges}\}.$$

Then A is measurable.

Proof: The set A is the same as the set on which $\{f_n(\omega)\}$ is a Cauchy sequence. This set is

$$\cap_{n=1}^{\infty} \cup_{m=1}^{\infty} \cap_{p,q>m} \left[\left\| f_p(\omega) - f_q(\omega) \right\| < \frac{1}{n} \right]$$

which is a measurable set thanks to the measurability of each f_n. ■

8.2 Measures and Their Properties

What is meant by a **measure?**

Definition 8.2.1 *Let $(\Omega, \mathscr{F})$ be a measurable space. Here $\mathscr{F}$ is a σ algebra of sets of Ω. Then $\mu : \mathscr{F} \to [0,\infty]$ is called a measure if whenever $\{F_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets of $\mathscr{F}$, it follows that*

$$\mu(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mu(E_i)$$

Note that the series could equal ∞. If $\mu(\Omega) < \infty$, then μ is called a finite measure. An important case is when $\mu(\Omega) = 1$ when it is called a probability measure.

Note that $\mu(\emptyset) = \mu(\emptyset \cup \emptyset) = \mu(\emptyset) + \mu(\emptyset)$ and so $\mu(\emptyset) = 0$.

Example 8.2.2 *You could have $\mathscr{P}(\mathbb{N}) = \mathscr{F}$ and you could define $\mu(S)$ to be the number of elements of S. This is called counting measure. It is left as an exercise to show that this is a measure.*

Example 8.2.3 *Here is a pathological example. Let Ω be uncountable and $\mathscr{F}$ will be those sets which have the property that either the set is countable or its complement is countable. Let $\mu(E) = 0$ if E is countable and $\mu(E) = 1$ if E is uncountable. It is left as an exercise to show that this is a measure.*

Of course the most important measure in this book will be Lebesgue measure which gives the "volume" of a subset of $\mathbb{R}^n$.

Lemma 8.2.4 *If μ is a measure and $F_i \in \mathscr{F}$, then $\mu(\cup_{i=1}^\infty F_i) \leq \sum_{i=1}^\infty \mu(F_i)$. Also if $F_n \in \mathscr{F}$ and $F_n \subseteq F_{n+1}$ for all n, then if $F = \cup_n F_n$,*

$$\mu(F) = \lim_{n\to\infty} \mu(F_n)$$

If $F_n \supseteq F_{n+1}$ for all n, then if $\mu(F_1) < \infty$ and $F = \cap_n F_n$, then

$$\mu(F) = \lim_{n\to\infty} \mu(F_n)$$

Proof: Let $G_1 = F_1$ and if $G_1, \cdots, G_n$ have been chosen disjoint, let $G_{n+1} \equiv F_{n+1} \setminus \cup_{i=1}^n G_i$. Thus the G_i are disjoint. In addition, these are all measurable sets. Now

$$\mu(G_{n+1}) + \mu(F_{n+1} \cap (\cup_{i=1}^n G_i)) = \mu(F_{n+1})$$

and so $\mu(G_n) \leq \mu(F_n)$. Therefore,

$$\mu(\cup_{i=1}^\infty G_i) = \sum_i \mu(G_i) \leq \sum_i \mu(F_i).$$

Now consider the increasing sequence of $F_n \in \mathscr{F}$. If $F \subseteq G$ and these are sets of $\mathscr{F}$, then $\mu(G) = \mu(F) + \mu(G \setminus F)$ so $\mu(G) \geq \mu(F)$. Also $F = \cup_{i=1}^\infty (F_{i+1} \setminus F_i) + F_1$. Then $\mu(F) = \sum_{i=1}^\infty \mu(F_{i+1} \setminus F_i) + \mu(F_1)$. Now $\mu(F_{i+1} \setminus F_i) + \mu(F_i) = \mu(F_{i+1})$. If any $\mu(F_i) = \infty$, there is nothing to prove. Assume then that these are all finite. Then $\mu(F_{i+1} \setminus F_i) = \mu(F_{i+1}) - \mu(F_i)$ and so

$$\begin{aligned}
\mu(F) &= \sum_{i=1}^\infty \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) \\
&= \lim_{n\to\infty} \sum_{i=1}^n \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) = \lim_{n\to\infty} \mu(F_{n+1})
\end{aligned}$$

Next suppose $\mu(F_1) < \infty$ and $\{F_n\}$ is a decreasing sequence. Then $F_1 \setminus F_n$ is increasing to $F_1 \setminus F$ and so by the first part,

$$\mu(F_1) - \mu(F) = \mu(F_1 \setminus F) = \lim_{n\to\infty} \mu(F_1 \setminus F_n) = \lim_{n\to\infty} (\mu(F_1) - \mu(F_n))$$

This is justified because $\mu(F_1 \setminus F_n) + \mu(F_n) = \mu(F_1)$ and all numbers are finite by assumption. Hence $\mu(F) = \lim_{n\to\infty} \mu(F_n)$. ■

I like to remember this as $E_n \uparrow E \Rightarrow \mu(E_n) \uparrow \mu(E)$ and $E_n \downarrow E \Rightarrow \mu(E_n) \downarrow \mu(E)$ if $\mu(E_1) < \infty$.

There is a monumentally important theorem called the Borel Cantelli lemma. This is next.

Lemma 8.2.5 *If $(\Omega, \mathscr{F}, \mu)$ is a measure space and if $\{E_i\} \subseteq \mathscr{F}$ and $\sum_{i=1}^{\infty} \mu(E_i) < \infty$, then there exists a set N of measure 0 ($\mu(N) = 0$) such that if $\omega \notin N$, then ω is in only finitely many of the E_i.*

Proof: The set of ω in infinitely many E_i is $N \equiv \cap_{n=1}^{\infty} \cup_{k \geq n} E_k$ because this consists of those ω which are in some E_k for $k \geq n$ for any choice of n. Now $\mu(N) \leq \sum_{k=n}^{\infty} \mu(E_k)$ which is just the tail of a convergent series. Thus, it converges to 0 as $n \to \infty$. Hence it is less than ε for n large enough. Thus $\mu(N)$ is no more than ε for any $\varepsilon > 0$. ■

8.3 Dynkin's Lemma

Dynkin's lemma is a very useful result. It is like something used in other books called monotone classes containing something called an algebra of sets, but it is easier to use.

Definition 8.3.1 *Let Ω be a set and let $\mathscr{K}$ be a collection of subsets of Ω. Then $\mathscr{K}$ is called a π system if $\emptyset, \Omega \in \mathscr{K}$ and whenever $A, B \in \mathscr{K}$, it follows $A \cap B \in \mathscr{K}$.*

The following is the fundamental lemma which shows these π systems are useful. This is due to Dynkin.

Lemma 8.3.2 *Let $\mathscr{K}$ be a π system of subsets of Ω, a set. Also let $\mathscr{G}$ be a collection of subsets of Ω which satisfies the following three properties.*

1. $\mathscr{K} \subseteq \mathscr{G}$
2. *If $A \in \mathscr{G}$, then $A^C \in \mathscr{G}$*
3. *If $\{A_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets from $\mathscr{G}$ then $\cup_{i=1}^{\infty} A_i \in \mathscr{G}$.*

Then $\mathscr{G} \supseteq \sigma(\mathscr{K})$, where $\sigma(\mathscr{K})$ is the smallest σ algebra which contains $\mathscr{K}$.

Proof: First note that if

$$\mathscr{H} \equiv \{\mathscr{G} : 1 \text{ - } 3 \text{ all hold}\}$$

then $\cap \mathscr{H}$ yields a collection of sets which also satisfies 1 - 3. Therefore, I will assume in the argument that $\mathscr{G}$ is the smallest collection satisfying 1 - 3. Let $A \in \mathscr{K}$ and define

$$\mathscr{G}_A \equiv \{B \in \mathscr{G} : A \cap B \in \mathscr{G}\}.$$

I want to show $\mathscr{G}_A$ satisfies 1 - 3 because then it must equal $\mathscr{G}$ since $\mathscr{G}$ is the smallest collection of subsets of Ω which satisfies 1 - 3. This will give the conclusion that for $A \in \mathscr{K}$ and $B \in \mathscr{G}$, $A \cap B \in \mathscr{G}$. This information will then be used to show that if $A, B \in \mathscr{G}$ then $A \cap B \in \mathscr{G}$. From this it will follow very easily that $\mathscr{G}$ is a σ algebra which will imply it contains $\sigma(\mathscr{K})$. Now here are the details of the argument.

Since $\mathscr{K}$ is given to be a π system contained in $\mathscr{G}$, $\mathscr{K} \subseteq \mathscr{G}_A$. Indeed, if $C \in \mathscr{K}$ then $A \cap C \in \mathscr{K} \subseteq \mathscr{G}$ so $C \in \mathscr{G}_A$. Property 3 is obvious because if $\{B_i\}$ is a sequence of disjoint sets in $\mathscr{G}_A$, then

$$A \cap \cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A \cap B_i \in \mathscr{G}$$

because $A \cap B_i \in \mathscr{G}$ and the property 3 of $\mathscr{G}$.

It remains to verify Property 2 so let $B \in \mathscr{G}_A$. I need to verify that $B^C \in \mathscr{G}_A$. In other words, I need to show that $A \cap B^C \in \mathscr{G}$. However,

$$A \cap B^C = \left(A^C \cup (A \cap B)\right)^C \in \mathscr{G}$$

Here is why. Since $B \in \mathscr{G}_A$, $A \cap B \in \mathscr{G}$ and since $A \in \mathscr{K} \subseteq \mathscr{G}$ it follows $A^C \in \mathscr{G}$ by assumption 2. It follows from assumption 3 the union of the disjoint sets, A^C and $(A \cap B)$ is in $\mathscr{G}$ and then from 2 the complement of their union is in $\mathscr{G}$. Thus $\mathscr{G}_A$ satisfies 1 - 3 and this implies, since $\mathscr{G}$ is the smallest such, that $\mathscr{G}_A \supseteq \mathscr{G}$. However, $\mathscr{G}_A$ is constructed as a subset of $\mathscr{G}$. This proves that for every $B \in \mathscr{G}$ and $A \in \mathscr{K}$, $A \cap B \in \mathscr{G}$. Now pick $B \in \mathscr{G}$ and consider

$$\mathscr{G}_B \equiv \{A \in \mathscr{G} : A \cap B \in \mathscr{G}\}.$$

I just proved $\mathscr{K} \subseteq \mathscr{G}_B$. The other arguments are identical to show $\mathscr{G}_B$ satisfies 1 - 3 and is therefore equal to $\mathscr{G}$. This shows that whenever $A, B \in \mathscr{G}$ it follows $A \cap B \in \mathscr{G}$.

This implies $\mathscr{G}$ is a σ algebra. To show this, all that is left is to verify $\mathscr{G}$ is closed under countable unions because then it follows $\mathscr{G}$ is a σ algebra. Let $\{A_i\} \subseteq \mathscr{G}$. Then let $A_1' = A_1$ and

$$A_{n+1}' \equiv A_{n+1} \setminus (\cup_{i=1}^n A_i) = A_{n+1} \cap \left(\cap_{i=1}^n A_i^C\right) = \cap_{i=1}^n \left(A_{n+1} \cap A_i^C\right) \in \mathscr{G}$$

because finite intersections of sets of $\mathscr{G}$ are in $\mathscr{G}$. Since the A_i' are disjoint, it follows $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty A_i' \in \mathscr{G}$. Therefore, $\mathscr{G} \supseteq \sigma(\mathscr{K})$. ■

Corollary 8.3.3 *Given 2, closed with respect to complements, the condition that $\mathscr{G}$ is closed with respect to countable disjoint unions is equivalent to $\mathscr{G}$ the condition that $\mathscr{G}$ is closed with respect to countable intersections.*

Proof: ⇒ Consider $\cap_k E_k$ where $E_k \in \mathscr{G}$. Then $\cap_k E_k = \left(\cup_k E_k^C\right)^C$. Now the E_k^C are not necessarily disjoint, but each is in $\mathscr{G}$ and so one can use the scheme of the last part of the proof of Lemma 8.3.2 to reduce to this case and conclude $\cup_k E_k^C \in \mathscr{G}$. Then the countable intersection is just the complement of this last set.

⇐ Suppose the countable intersection of sets of $\mathscr{G}$ is in $\mathscr{G}$ and consider a countable union $\cup_k E_k$ of sets of $\mathscr{G}$. Then $\cup_k E_k = \left(\cap_k E_k^C\right)^C \in \mathscr{G}$. ■

8.4 Outer Measures

There is also something called an outer measure which is defined on the set of all subsets.

Definition 8.4.1 *Let Ω be a nonempty set and let $\lambda : \mathscr{P}(\Omega) \to [0, \infty)$ satisfy the following:*

1. $\lambda(\emptyset) = 0$
2. *If* $A \subseteq B$, *then* $\lambda(A) \leq \lambda(B)$
3. $\lambda(\cup_{i=1}^\infty E_i) \leq \sum_{i=1}^\infty \lambda(E_i)$

Then λ is called an outer measure.

Every measure determines an outer measure. For example, suppose that μ is a measure on $\mathscr{F}$ a σ algebra of subsets of Ω. Then define

$$\bar{\mu}(S) \equiv \inf\{\mu(E) : E \supseteq S,\ E \in \mathscr{F}\}.$$

This is easily seen to be an outer measure. Also, we have the following Proposition.

Proposition 8.4.2 *Let μ be a measure defined on a σ algebra of subsets $\mathscr{F}$ of Ω as just described. Then $\bar{\mu}$ as defined above, is an outer measure and also, if $E \in \mathscr{F}$, then $\bar{\mu}(E) = \mu(E)$.*

Proof: The first two properties of an outer measure are obvious. What of the third? If any $\bar{\mu}(E_i) = \infty$, then there is nothing to show so suppose each of these is finite. Let $F_i \supseteq E_i$ such that $F_i \in \mathscr{F}$ and $\bar{\mu}(E_i) + \frac{\varepsilon}{2^i} > \mu(F_i)$. Then

$$\bar{\mu}(\cup_{i=1}^{\infty} E_i) \leq \mu(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mu(F_i) < \sum_{i=1}^{\infty} \left(\bar{\mu}(E_i) + \frac{\varepsilon}{2^i}\right) = \sum_{i=1}^{\infty} \bar{\mu}(E_i) + \varepsilon$$

Since ε is arbitrary, this establishes the third condition. Finally, if $E \in \mathscr{F}$, then by definition, $\bar{\mu}(E) \leq \mu(E)$ because $E \supseteq E$. Also, $\mu(E) \leq \mu(F)$ for all $F \in \mathscr{F}$ such that $F \supseteq E$. It follows that $\mu(E)$ is a lower bound of all such $\mu(F)$ and so $\bar{\mu}(E) \geq \mu(E)$.■

8.5 Measures From Outer Measures

There is a general procedure for constructing a σ algebra and a measure from an outer measure which is due to Caratheodory about 1918.

Thus, when you have a measure on $(\Omega, \mathscr{F})$, you can obtain an outer measure on $(\Omega, \mathscr{P}(\Omega))$ from this measure as in Proposition 8.4.2, and if you have an outer measure on $(\Omega, \mathscr{P}(\Omega))$, this will define a σ algebra $\mathscr{F}$ and a measure on $(\Omega, \mathscr{F})$. This last assertion is the topic of this section.

Definition 8.5.1 *Let Ω be a nonempty set and let $\mu : \mathscr{P}(\Omega) \to [0,\infty]$ be an outer measure. For $E \subseteq \Omega$, E is μ measurable if for all $S \subseteq \Omega$,*

$$\mu(S) = \mu(S \setminus E) + \mu(S \cap E). \tag{8.4}$$

To help in remembering 8.4, think of a measurable set E, as a process which divides a given set into two pieces, the part in E and the part not in E as in 8.4. In the Bible, there are several incidents recorded in which a process of division resulted in more stuff than was originally present.[1] Measurable sets are exactly those which are incapable of such a miracle. With an outer measure, it is always the case that $\mu(S) \leq \mu(S \setminus E) + \mu(S \cap E)$. The set is measurable, when equality is always obtained for any choice of $S \in \mathscr{P}(\Omega)$. You might think of the measurable sets as the non-miraculous sets. The idea is to show that these sets form a σ algebra on which the outer measure μ is a measure.

First here is a definition and a lemma.

[1] 1 Kings 17, 2 Kings 4, Mathew 14, and Mathew 15 all contain such descriptions. The stuff involved was either oil, bread, flour or fish. In mathematics such things have also been done with sets. In the book by Bruckner Bruckner and Thompson there is an interesting discussion of the Banach Tarski paradox which says it is possible to divide a ball in $\mathbb{R}^3$ into five disjoint pieces and assemble the pieces to form two disjoint balls of the same size as the first. The details can be found in: The Banach Tarski Paradox by Wagon, Cambridge University press. 1985. It is known that all such examples must involve the axiom of choice.

Definition 8.5.2 *$(\mu\lfloor S)(A) \equiv \mu(S\cap A)$ for all $A\subseteq\Omega$. Thus $\mu\lfloor S$ is the name of a new outer measure, called μ restricted to S.*

The next lemma indicates that the property of measurability is not lost by considering this restricted measure.

Lemma 8.5.3 *If A is μ measurable, then for any S, A is $\mu\lfloor S$ measurable.*

Proof: Suppose A is μ measurable. It is desired to to show that for all $T\subseteq\Omega$,

$$(\mu\lfloor S)(T) = (\mu\lfloor S)(T\cap A) + (\mu\lfloor S)(T\setminus A).$$

Thus it is desired to show

$$\mu(S\cap T) = \mu(T\cap A\cap S) + \mu(T\cap S\cap A^C). \tag{8.5}$$

But 8.5 holds because A is μ measurable. Apply Definition 8.5.1 to $S\cap T$ instead of S. ■

If A is $\mu\lfloor S$ measurable, it does not follow that A is μ measurable. Indeed, if you believe in the existence of non measurable sets which is discussed later, you could let $A=S$ for such a μ non measurable set and verify that S is $\mu\lfloor S$ measurable.

The next theorem is the main result on outer measures which shows that starting with an outer measure you can obtain a measure.

Theorem 8.5.4 *Let Ω be a set and let μ be an outer measure on $\mathscr{P}(\Omega)$. The collection of μ measurable sets $\mathscr{S}$, forms a σ algebra and*

$$\text{If } F_i\in\mathscr{S},\ F_i\cap F_j=\emptyset,\ \text{then } \mu(\cup_{i=1}^{\infty}F_i) = \sum_{i=1}^{\infty}\mu(F_i). \tag{8.6}$$

If $\cdots F_n\subseteq F_{n+1}\subseteq\cdots$, then if $F=\cup_{n=1}^{\infty}F_n$ and $F_n\in\mathscr{S}$, it follows that

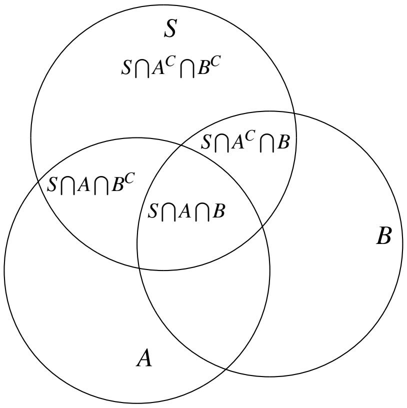
$$\mu(F) = \lim_{n\to\infty}\mu(F_n). \tag{8.7}$$

If $\cdots F_n\supseteq F_{n+1}\supseteq\cdots$, and if $F=\cap_{n=1}^{\infty}F_n$ for $F_n\in\mathscr{S}$ then if $\mu(F_1)<\infty$,

$$\mu(F) = \lim_{n\to\infty}\mu(F_n). \tag{8.8}$$

This measure space is also complete which means that if $\mu(F)=0$ for some $F\in\mathscr{S}$ then if $G\subseteq F$, it follows $G\in\mathscr{S}$ also.

Proof: First note that $\emptyset$ and Ω are obviously in $\mathscr{S}$. Now suppose $A,B\in\mathscr{S}$. I will show $A\setminus B\equiv A\cap B^C$ is in $\mathscr{S}$. To do so, consider the following picture.



It is required to show that $\mu(S) = \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))$. First consider $S \setminus (A \setminus B)$. From the picture, it equals

$$\left(S \cap A^C \cap B^C\right) \cup (S \cap A \cap B) \cup \left(S \cap A^C \cap B\right)$$

Therefore, $\mu(S) \leq \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))$

$$\begin{aligned}
&\leq \mu\left(S \cap A^C \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) + \mu(S \cap (A \setminus B)) \\
&= \mu\left(S \cap A^C \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) + \mu\left(S \cap A \cap B^C\right) \\
&= \mu\left(S \cap A^C \cap B^C\right) + \mu\left(S \cap A \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) \\
&= \mu\left(S \cap B^C\right) + \mu(S \cap B) = \mu(S)
\end{aligned}$$

and so this shows that $A \setminus B \in \mathscr{S}$ whenever $A, B \in \mathscr{S}$.

Since $\Omega \in \mathscr{S}$, this shows that $A \in \mathscr{S}$ if and only if $A^C \in \mathscr{S}$. Now if $A, B \in \mathscr{S}$, $A \cup B = (A^C \cap B^C)^C = (A^C \setminus B)^C \in \mathscr{S}$. By induction, if $A_1, \cdots, A_n \in \mathscr{S}$, then so is $\cup_{i=1}^n A_i$. If $A, B \in \mathscr{S}$, with $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu((A \cup B) \cap A) + \mu((A \cup B) \setminus A) = \mu(A) + \mu(B).$$

By induction, if $A_i \cap A_j = \emptyset$ and $A_i \in \mathscr{S}$,

$$\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i). \tag{8.9}$$

Now let $A = \cup_{i=1}^\infty A_i$ where $A_i \cap A_j = \emptyset$ for $i \neq j$. $\sum_{i=1}^\infty \mu(A_i) \geq \mu(A) \geq \mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$. Since this holds for all n, you can take the limit as $n \to \infty$ and conclude, $\sum_{i=1}^\infty \mu(A_i) = \mu(A)$ which establishes 8.6.

Consider part 8.7. Without loss of generality $\mu(F_k) < \infty$ for all k since otherwise there is nothing to show. Suppose $\{F_k\}$ is an increasing sequence of sets of $\mathscr{S}$. Then letting $F_0 \equiv \emptyset$, $\{F_{k+1} \setminus F_k\}_{k=0}^\infty$ is a sequence of disjoint sets of $\mathscr{S}$ since it was shown above that the difference of two sets of $\mathscr{S}$ is in $\mathscr{S}$. Also note that from 8.9

$$\mu(F_{k+1} \setminus F_k) + \mu(F_k) = \mu(F_{k+1})$$

and so if $\mu(F_k) < \infty$, then

$$\mu(F_{k+1} \setminus F_k) = \mu(F_{k+1}) - \mu(F_k).$$

Therefore, letting $F \equiv \cup_{k=1}^\infty F_k$ which also equals $\cup_{k=1}^\infty (F_{k+1} \setminus F_k)$, it follows from part 8.6 just shown that

$$\begin{aligned}
\mu(F) &= \sum_{k=0}^\infty \mu(F_{k+1} \setminus F_k) = \lim_{n\to\infty} \sum_{k=0}^n \mu(F_{k+1} \setminus F_k) \\
&= \lim_{n\to\infty} \sum_{k=0}^n \mu(F_{k+1}) - \mu(F_k) = \lim_{n\to\infty} \mu(F_{n+1}).
\end{aligned}$$

In order to establish 8.8, let the F_n be as given there. Then, since $(F_1 \setminus F_n)$ increases to $(F_1 \setminus F)$, 8.7 implies

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \lim_{n\to\infty} \mu(F_1 \setminus F_n) = \mu(F_1 \setminus F).$$

The problem is, I don't know $F \in \mathscr{S}$ and so it is not clear that $\mu(F_1 \setminus F) = \mu(F_1) - \mu(F)$. However, $\mu(F_1 \setminus F) + \mu(F) \geq \mu(F_1)$ and so $\mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$. Hence

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$$

which implies $\lim_{n\to\infty} \mu(F_n) \leq \mu(F)$. But since $F \subseteq F_n$, $\mu(F) \leq \lim_{n\to\infty} \mu(F_n)$ and this establishes 8.8. Note that it was assumed $\mu(F_1) < \infty$ because $\mu(F_1)$ was subtracted from both sides.

It remains to show $\mathscr{S}$ is closed under countable unions. Recall that if $A \in \mathscr{S}$, then $A^C \in \mathscr{S}$ and $\mathscr{S}$ is closed under finite unions. Let $A_i \in \mathscr{S}$, $A = \cup_{i=1}^{\infty} A_i$, $B_n = \cup_{i=1}^{n} A_i$. Then

$$\begin{aligned} \mu(S) &= \mu(S \cap B_n) + \mu(S \setminus B_n) \qquad (8.10) \\ &= (\mu\lfloor S)(B_n) + (\mu\lfloor S)(B_n^C). \end{aligned}$$

By Lemma 8.5.3 B_n is $(\mu\lfloor S)$ measurable and so is B_n^C. I want to show $\mu(S) \geq \mu(S \setminus A) + \mu(S \cap A)$. If $\mu(S) = \infty$, there is nothing to prove. Assume $\mu(S) < \infty$. Then apply Parts 8.8 and 8.7 to the outer measure $\mu\lfloor S$ in 8.10 and let $n \to \infty$. Thus $B_n \uparrow A$, $B_n^C \downarrow A^C$ and this yields $\mu(S) = (\mu\lfloor S)(A) + (\mu\lfloor S)(A^C) = \mu(S \cap A) + \mu(S \setminus A)$.

Therefore $A \in \mathscr{S}$ and this proves Parts 8.6, 8.7, and 8.8.

It only remains to verify the assertion about completeness. Letting G and F be as described above, let $S \subseteq \Omega$. I need to verify $\mu(S) \geq \mu(S \cap G) + \mu(S \setminus G)$. However,

$$\begin{aligned} \mu(S \cap G) + \mu(S \setminus G) &\leq \mu(S \cap F) + \mu(S \setminus F) + \mu(F \setminus G) \\ &= \mu(S \cap F) + \mu(S \setminus F) = \mu(S) \end{aligned}$$

because by assumption, $\mu(F \setminus G) \leq \mu(F) = 0$. ■

Corollary 8.5.5 *Completeness is the same as saying that if $(E \setminus E') \cup (E' \setminus E) \subseteq N \in \mathscr{F}$ and $\mu(N) = 0$, then if $E \in \mathscr{F}$, it follows that $E' \in \mathscr{F}$ also.*

Proof: If the new condition holds, then suppose $G \subseteq F$ where $\mu(F) = 0, F \in \mathscr{F}$. Then $\overbrace{(G \setminus F)}^{=\emptyset} \cup (F \setminus G) \subseteq F$ and $\mu(F)$ is given to equal 0. Therefore, $G \in \mathscr{F}$.

Now suppose the earlier version of completeness and let

$$\left(E \setminus E'\right) \cup \left(E' \setminus E\right) \subseteq N \in \mathscr{F}$$

where $\mu(N) = 0$ and $E \in \mathscr{F}$. Then we know $(E \setminus E'), (E' \setminus E) \in \mathscr{F}$ and all have measure zero. It follows $E \setminus (E \setminus E') = E \cap E' \in \mathscr{F}$. Hence

$$E' = \left(E \cap E'\right) \cup \left(E' \setminus E\right) \in \mathscr{F} \blacksquare$$

Given a measure space $(\Omega, \mathscr{F}, \mu)$ we can always complete the measure space by considering the outer measure described above in Proposition 8.4.2. Denoting this outer measure by $\bar{\mu}$, the completion will be $(\Omega, \mathscr{S}, \bar{\mu})$ where $\mathscr{S}$ will be the sets measurable in the sense of Caratheodory just described as in Proposition 8.4.2, the new measure $\bar{\mu}$ will coincide with μ on $\mathscr{F}$ but will be a complete measure on the larger σ algebra $\mathscr{S}$.

Proposition 8.5.6 *Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space, $\mu(\Omega) < \infty$. Then if $E \in (\Omega, \mathscr{S}, \bar{\mu})$, the complete measure space obtained as the above using Caratheodory's approach and $\bar{\mu}$ is the outer measure defined as in Proposition 8.4.2, then there exists $F \in \mathscr{F}$*

such that $F \supseteq E$ and $\bar{\mu}(F) = \mu(F) = \bar{\mu}(E)$. The same conclusion holds if $(\Omega, \mathscr{F}, \mu)$ is a σ finite measure space meaning that $\Omega = \cup_{k=1}^{\infty} \Omega_k$ where the $\Omega_k \in \mathscr{F}$ and are disjoint with $\mu(\Omega_k) < \infty$.

Proof: $\bar{\mu}(E) \equiv \inf\{\mu(F) : F \supseteq E, F \in \mathscr{F}\}$. Let $F_n \in \mathscr{F}$, $F_n \supseteq E$ and $\bar{\mu}(E) + \frac{1}{n} > \mu(F_n)$. By taking intersections, we can also assume that $F_n \supseteq F_{n+1}$. Let $F \equiv \cap_n F_n$. Then using Proposition 8.4.2 we have $\mu(F) = \bar{\mu}(F) = \bar{\mu}(E)$. In σ finite case, it was just shown that there exists $F_k \subseteq \Omega_k, F_k \in \mathscr{F}$ such that $\bar{\mu}(F_k) = \mu(F_k) = \bar{\mu}(E \cap \Omega_k)$. Then let $F \equiv \cup_k F_k$. $\bar{\mu}(E) = \sum_{k=1}^{\infty} \bar{\mu}(E \cap \Omega_k) = \sum_{k=1}^{\infty} \mu(F_k) = \mu(F) = \bar{\mu}(F)$ ■

As a corollary, we can say something about functions.

Corollary 8.5.7 *Let $(\Omega, \mathscr{F}, \mu)$ be σ finite and let $(\Omega, \mathscr{S}, \bar{\mu})$ be the completion just discussed. Then if $f \geq 0$ and $\mathscr{S}$ measurable, there exists $h \geq f$ such that $h = f$ for $\bar{\mu}$ a.e. and h is $\mathscr{F}$ measurable.*

Proof: From Theorem 8.1.6 there is a sequence of $\mathscr{S}$ measurable simple nonnegative functions $s_n(\omega) = \sum_{k=1}^{m_n} c_k^n \mathscr{X}_{E_k^n}(\omega)$ which converges pointwise to f. From Proposition 8.5.6, $s_n(\omega) = \hat{s}_n(\omega)$ where $\hat{s}_n(\omega) \equiv \sum_{k=1}^{m_n} c_k^n \mathscr{X}_{\hat{E}_k^n}(\omega)$ with $\hat{E}_k^n \supseteq E_k^n$, $\bar{\mu}\left(\hat{E}_k^n \setminus E_k^n\right) = 0$. Then letting $h(\omega) \equiv \limsup_{n\to\infty} \hat{s}_n(\omega)$, it follows that $h(\omega)$ is $\mathscr{F}$ measurable, $h(\omega) = f(\omega)$ $\bar{\mu}$ a.e., and $h(\omega) \geq f(\omega)$. ■

8.6 Measurable Sets Include Borel Sets?

If you have an outer measure, it determines a measure. This section gives a very convenient criterion which allows you to conclude right away that the measure is a Borel measure.

Theorem 8.6.1 *Let μ be an outer measure on the subsets of (X, d), a metric space. If $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $\operatorname{dist}(A, B) > 0$, then the σ algebra of measurable sets $\mathscr{S}$ contains the Borel sets.*

Proof: It suffices to show that closed sets are in $\mathscr{S}$, the σ-algebra of measurable sets, because then the open sets are also in $\mathscr{S}$ and consequently $\mathscr{S}$ contains the Borel sets. Let K be closed and let S be a subset of Ω. Is $\mu(S) \geq \mu(S \cap K) + \mu(S \setminus K)$? It suffices to assume $\mu(S) < \infty$. Let $K_n \equiv \left\{x : \operatorname{dist}(x, K) \leq \frac{1}{n}\right\}$. By Lemma 3.12.1 on Page 83, $x \to \operatorname{dist}(x, K)$ is continuous and so K_n is a closed set having K as a subset. That in K_n^C is at a positive distance from K. By the assumption of the theorem,

$$\mu(S) \geq \mu((S \cap K) \cup (S \setminus K_n)) = \mu(S \cap K) + \mu(S \setminus K_n) \tag{8.11}$$

Now

$$\mu(S \setminus K_n) \leq \mu(S \setminus K) \leq \mu(S \setminus K_n) + \mu((K_n \setminus K) \cap S). \tag{8.12}$$

If $\lim_{n\to\infty} \mu((K_n \setminus K) \cap S) = 0$ then the theorem will be proved because this limit along with 8.12 implies $\lim_{n\to\infty} \mu(S \setminus K_n) = \mu(S \setminus K)$ and then taking a limit in 8.11, $\mu(S) \geq \mu(S \cap K) + \mu(S \setminus K)$ as desired. Therefore, it suffices to establish this limit.

Since K is closed, a point, $x \notin K$ must be at a positive distance from K and so

$$K_n \setminus K = \cup_{k=n}^{\infty} K_k \setminus K_{k+1}.$$

Therefore

$$\mu(S \cap (K_n \setminus K)) \leq \sum_{k=n}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})). \tag{8.13}$$

If

$$\sum_{k=1}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})) < \infty, \tag{8.14}$$

then $\mu(S \cap (K_n \setminus K)) \to 0$ because it is dominated by the tail of a convergent series so it suffices to show 8.14.

$$\sum_{k=1}^{M} \mu(S \cap (K_k \setminus K_{k+1})) =$$

$$\sum_{k \text{ } even, \, k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ } odd, \, k \leq M} \mu(S \cap (K_k \setminus K_{k+1})). \tag{8.15}$$

By the construction, the distance between any pair of sets, $S \cap (K_k \setminus K_{k+1})$ for different even values of k is positive and the distance between any pair of sets, $S \cap (K_k \setminus K_{k+1})$ for different odd values of k is positive. Therefore,

$$\sum_{k \text{ } even, \, k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ } odd, \, k \leq M} \mu(S \cap (K_k \setminus K_{k+1})) \leq$$

$$\mu\left(\bigcup_{k \text{ } even, \, k \leq M} (S \cap (K_k \setminus K_{k+1}))\right) + \mu\left(\bigcup_{k \text{ } odd, \, k \leq M} (S \cap (K_k \setminus K_{k+1}))\right)$$

$$\leq \mu(S) + \mu(S) = 2\mu(S)$$

and so for all M, $\sum_{k=1}^{M} \mu(S \cap (K_k \setminus K_{k+1})) \leq 2\mu(S)$ showing 8.14. ■

8.7 Regular Measures

In using measures defined on a σ algebra of subsets of a metric space, the idea of regularity is fundamental.

Definition 8.7.1 *A measure μ defined on a σ algebra $\mathscr{F}$ of sets in a metric space X which includes the Borel sets $\mathscr{B}(X)$ will be called inner regular on $\mathscr{F}$ if for all $F \in \mathscr{F}$,*

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is closed}\} \tag{8.16}$$

A measure, μ defined on $\mathscr{F}$ will be called outer regular on $\mathscr{F}$ if for all $F \in \mathscr{F}$,

$$\mu(F) = \inf\{\mu(V) : V \supseteq F \text{ and } V \text{ is open}\} \tag{8.17}$$

When a measure is both inner and outer regular, it is called regular. Actually, it is more useful and likely more standard to refer to μ being inner regular as

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\} \tag{8.18}$$

Thus the word "closed" is replaced with "compact". A complete measure defined on a σ algebra $\mathscr{F}$ which includes the Borel sets which is finite on compact sets and also satisfies 8.17 and 8.18 for each $F \in \mathscr{F}$ is called a Radon measure. A G_δ set, pronounced as G delta is the countable intersection of open sets. An F_σ set, pronounced F sigma is the countable union of closed sets.

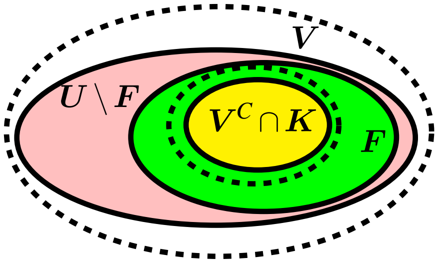
In every case which has been of interest to me, the measure has been σ finite.

Definition 8.7.2 *If $(X,\mathscr{F},\mu)$ is a measure space, it is called σ finite if there are $X_n \in \mathscr{F}$ with $\cup_n X_n = X$ and $\mu(X_n) < \infty$. Note that by considering $Y_n = X_n \setminus X_{n-1}, X_0 \equiv \emptyset$ we could assume $X = \cup_n Y_n$ where the Y_n are disjoint.*

Then there is a useful general result.

Theorem 8.7.3 *Let (X,d) be a metric space and suppose μ is σ finite and outer regular. Then μ is inner regular. If every closed set is the countable union of compact sets, then in the definition of inner regular, one can replace "closed" with "compact".*

Proof: Whenever $\mu(F), \mu(G) < \infty$ for $G \supseteq F, \mu(G \setminus F) = \mu(G) - \mu(F)$. I will use this simple observation without comment in the following. To show the measure space is regular, the following picture might help or it might not. V is between the dotted lines.



Let F be a bounded measurable set and let $\mu(U \setminus F) < \varepsilon$ where U is open and let $K \subseteq U$, K closed and $\mu(U \setminus K) < \varepsilon$. I can get such a K because every open set is the countable union of closed sets

$$U = \cup_{k=1}^{\infty} \left\{ x : \text{dist}\left(x, U^C\right) \geq \frac{1}{k} \right\} \equiv \cup_{k=1}^{\infty} K_k, \ ...K_k \subseteq K_{k+1}...$$

thus $\mu(U) < \mu(K_k)$ for all k large enough since $\mu(U) = \lim_{k\to\infty} \mu(K_k)$ by Lemma 8.2.4. Then let V be open and $\mu(V \setminus (U \setminus F)) < \varepsilon$ where $V \supseteq U \setminus F = U \cap F^C$ so $V^C \subseteq U^C \cup F$. This is possible because all sets are in $\mathscr{F}$. Then $V^C \cap K \subseteq \left(U^C \cup F\right) \cap K = F \cap K \subseteq F$. Now $V^C \cap K$ is compact and

$$\begin{aligned} \mu\left(F \setminus \left(K \cap V^C\right)\right) &= \mu\left(F \cap \left(K^C \cup V\right)\right) = \mu(F \cap V) + \mu\left(F \cap K^C\right) \\ &\leq \mu(F \cap V) + \mu(U \setminus K) < \mu(F \cap V) + \varepsilon \end{aligned}$$

However, $\varepsilon > \mu(V \setminus (U \setminus F)) = \mu\left(V \cap \left(U \cap F^C\right)^C\right) = \mu\left(V \cap \left(U^C \cup F\right)\right) \geq \mu(V \cap F)$ and so $\mu\left(F \setminus \left(K \cap V^C\right)\right) \leq 2\varepsilon$. That $\mu(F) = \sup\{\mu(K) : K \subseteq F\}$ follows from observing that $\mu(F) = \lim_{n\to\infty} \mu(F \cap B(0,n))$ and then applying what was just shown to a suitable $F \cap B(0,n)$. As to the last claim, it follows from observing that for K a closed set, there is an increasing sequence of compact sets $\{K_n\}$ whose union is K and then using Lemma 8.2.4. ■

The following is a nice formulation of the above and also gives a useful claim about uniqueness.

Theorem 8.7.4 *Suppose $(X,\mathscr{F},\mu), \mathscr{F} \supseteq \mathscr{B}(X)$ is a measure space for X a metric space and μ is σ finite, $X = \cup_n X_n$ with $\mu(X_n) < \infty$ and the X_n disjoint Borel sets. Suppose also that μ is outer regular. Then for each $E \in \mathscr{F}$, there exists F, G an F_σ and G_δ set* *respectively such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$. In particular, μ is inner and outer regular on $\mathscr{F}$. If $(X, \hat{\mathscr{F}}, \hat{\mu})$ has the same properties, outer regular, and σ finite, and $\mu = \hat{\mu}$ on open sets, then if both $\mu, \hat{\mu}$ are complete measures, it follows that $\mu = \hat{\mu}$ and $\mathscr{F} = \hat{\mathscr{F}}$.*

Proof: Since μ is outer regular and $\mu(X_n) < \infty$, there exists an open set $V_n \supseteq E \cap X_n$ such that

$$\mu(V_n \setminus (E \cap X_n)) = \mu(V_n) - \mu(E \cap X_n) < \frac{\varepsilon}{2^n}.$$

Then let $V \equiv \cup_n V_n$ so that $V \supseteq E$. Then $E = \cup_n E \cap X_n$ and so

$$\mu(V \setminus E) \leq \mu(\cup_n (V_n \setminus (E \cap X_n))) \leq \sum_n \mu(V_n \setminus (E \cap X_n)) < \sum_n \frac{\varepsilon}{2^n} = \varepsilon$$

Similarly, there exists U_n open such that $\mu\left(U_n \setminus \left(E^C \cap X_n\right)\right) < \frac{\varepsilon}{2^n}, U_n \supseteq E^C \cap X_n$ so if $U \equiv \cup_n U_n, \mu\left(U \setminus E^C\right) = \mu\left(E \setminus U^C\right) < \varepsilon$. Now U^C is closed and contained in E because $U \supseteq E^C$. Hence, letting $\varepsilon = \frac{1}{2^n}$, there exist closed sets C_n, and open sets V_n such that $C_n \subseteq E \subseteq V_n$ and $\mu(V_n \setminus C_n) < \frac{1}{2^{n-1}}$. Letting $G \equiv \cap_n V_n, F \equiv \cup_n C_n, F \subseteq E \subseteq G$ and $\mu(G \setminus F) \leq \mu(V_n \setminus C_n) < \frac{1}{2^{n-1}}$. Since n is arbitrary, $\mu(G \setminus F) = 0$.

Let the disjoint sets X_n work for $\hat{\mu}$ as well as for μ. One can simply take an enumeration of $X_n \cap \hat{X}_m$ where the $\hat{X}_m$ work for $\hat{\mu}$. Let $\mathscr{K}$ consist of the open sets.This is clearly a π system because finite intersections remain in $\mathscr{K}$. Also $\mu = \hat{\mu}$ on $\mathscr{K}$ by assumption. Let $\mathscr{G}$ be those Borel sets F such that $\mu(F \cap X_n) = \hat{\mu}(F \cap X_n)$. Then $\mathscr{G}$ is clearly closed with respect to complements and countable disjoint intersections so $\mathscr{G} = \mathscr{B}(X)$. Taking unions, it follows that $\hat{\mu} = \mu$ on the Borel sets. Now by the first part, there is G a G_δ set and F an F_σ such that $\mu(G \setminus F) = \hat{\mu}(G \setminus F) = 0$ and $G \supseteq E \supseteq F$ for $E \in \mathscr{F}$. Then by completeness of $\hat{\mu}$, it follows that $E \in \hat{\mathscr{F}}$. Thus $\mathscr{F} \subseteq \hat{\mathscr{F}}$. Similarly $\hat{\mathscr{F}} \subseteq \mathscr{F}$. Also, $\mu(E) = \mu(G) = \hat{\mu}(G) = \hat{\mu}(E)$ so $\mu = \hat{\mu}$. ■

8.8 Constructing Measures From Functionals

Here is a theorem which is the main result on measures and functionals defined on a space of continuous functions. The typical situation is of a metric space in which closed balls are compact like $\mathbb{R}^p$.

Definition 8.8.1 *$C_c(X)$ will denote the complex values functions which have compact support in some metric space X. This is clearly a linear space. Then a linear function $L : C_c(X) \to \mathbb{C}$ is called "postitive" if whenever $f \geq 0$, then $Lf \geq 0$.*

The following theorem is called the Riesz representation theorem for positive linear functionals. I will make the way in which it represents something more clear later on. For now it will just produce lots of measures. Recall that $K \prec f \prec V$ means that f is 1 on the compact set K, has compact support in the open set V and takes values in $[0,1]$. Also $f \prec V$ means f has values in $[0,1]$ and has compact support in the open set V.

Theorem 8.8.2 *Let $L : C_c(X) \to \mathbb{C}$ be a positive linear functional where X is a metric space and X is a countable union of compact sets. Then there exists a complete measure μ defined on a σ algebra $\mathscr{F}$ which contains the Borel sets $\mathscr{B}(X)$ which is finite on compact sets and has the following properties. μ is regular. If E is measurable, there are F_σ and G_δ sets F, G such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$ so $\mu)F = \mu(E) = \mu(G)$. Then μ and $\mathscr{F}$ are uniquely determined.*

Proof: See the notation and lemmas near Definition 3.12.3 having to do with partitions of unity on a metric space for what is needed in this proof. For V open, let $\bar{\mu}(V) \equiv \sup\{Lf : f \prec V\}$. Then for an arbitrary set F, let $\bar{\mu}(F) \equiv \inf\{\bar{\mu}(V) : V \supseteq F\}, \bar{\mu}(\emptyset) \equiv 0$. In what follows, V will be an open set and K a compact set.

Claim 1: $\bar{\mu}$ is well defined.

Proof of Claim 1: Note there are two descriptions of $\bar{\mu}(V)$ for V open. They need to be the same. Let $\bar{\mu}_1$ be the definition involving supremums of Lf and let $\bar{\mu}$ be the general definition. Let $V \subseteq U$ where V,U open. Then by definition, $\bar{\mu}(V) \leq \bar{\mu}_1(U)$ and so $\bar{\mu}(V) \equiv \inf\{\bar{\mu}_1(U) : U \supseteq V\} \geq \bar{\mu}_1(V)$. However, $V \subseteq V$ and so $\bar{\mu}(V) \leq \bar{\mu}_1(V)$. ■

Claim 2: $\bar{\mu}$ is finite on compact sets. Also, if $K \prec f$, it follows that $\bar{\mu}(K) \leq L(f) < \infty$.

Proof of Claim 2: Let $K \prec f \prec X$. Let $V_\varepsilon \equiv \{x : f(x) > 1-\varepsilon\}$, an open set since f is continuous. Then let $g \prec V_\varepsilon$ so it follows that $\frac{f}{1-\varepsilon} \geq g$. Then $L(g) \leq \frac{1}{1-\varepsilon}L(f) < \infty$. Then taking the sup over all such g, it follows that $\bar{\mu}(K) \leq \bar{\mu}(V_\varepsilon) \leq \frac{1}{1-\varepsilon}Lf$. Now let $\varepsilon \to 0$ and conclude that $\bar{\mu}(K) \leq L(f)$. ■

Claim 3: $\bar{\mu}$ is subadditive: $\bar{\mu}(\cup_i E_i) \leq \sum_i \bar{\mu}(E_i)$.

Proof of Claim 3: First consider the case of open sets. Let $V = \cup_i V_i$. Let $l < \bar{\mu}(V)$. Then there exists $f \prec V$ with $Lf > l$. Then $\sup(f)$ is contained in $\cup_{i=1}^n V_i$. Now let $\sup \psi_i \subseteq V_i$ and $\sum_{i=1}^n \psi_i = 1$ on $\sup(f)$. This is from Theorem 3.12.5. Then

$$l < Lf = \sum_{i=1}^n L(\psi_i f) \leq \sum_{i=1}^n \bar{\mu}(V_i) \leq \sum_i \bar{\mu}(V_i).$$

Since l is arbitrary, it follows that $\bar{\mu}(V) \leq \sum_i \bar{\mu}(V_i)$. Now consider the general case. Let $E = \cup_i E_i$. If $\sum_i \bar{\mu}(E_i) = \infty$, there is nothing to show. Assume then that this sum is finite and let $V_i \supseteq E_i, \bar{\mu}(E_i) + \frac{\varepsilon}{2^i} > \bar{\mu}(V)$. Then

$$\bar{\mu}(E) \leq \bar{\mu}(\cup_i V_i) \leq \sum_i \bar{\mu}(V_i) \leq \sum_i \left(\bar{\mu}(E_i) + \frac{\varepsilon}{2^i}\right) = \sum_i \bar{\mu}(E_i) + \varepsilon$$

Since ε is arbitrary, this shows $\bar{\mu}$ is subadditive. ■

Claim 4: If $\operatorname{dist}(A,B) = \delta > 0$, then $\bar{\mu}(A \cup B) = \bar{\mu}(A) + \bar{\mu}(B)$.

Proof of Claim 4: If the right side is infinite, there is nothing to show so we can assume that $\bar{\mu}(A), \bar{\mu}(B)$ are both finite. First suppose U,V are open and disjoint having finite outer measure. Let $\bar{\mu}(U) \leq Lf_1 + \varepsilon$ where $f_1 \prec U$ and let $f_2 \prec V$ with $\bar{\mu}(V) \leq L(f_2) + \varepsilon$. Then

$$\bar{\mu}(U \cup V) \leq \bar{\mu}(U) + \bar{\mu}(V) \leq Lf_1 + Lf_2 + 2\varepsilon \leq L(f_1 + f_2) + 2\varepsilon \leq \bar{\mu}(U \cup V) + 2\varepsilon$$

Since ε is arbitrary, this shows that $\bar{\mu}(U \cup V) = \bar{\mu}(U) + \bar{\mu}(V)$. Now in case A,B are as assumed, let $U \equiv \cup_{x \in U} B(x, \delta/3), V \equiv \cup_{x \in V} B(x, \delta/3)$. Then these are disjoint open sets containing A and B respectively. Then there is O open, $O \supseteq A \cup B$ such that $\bar{\mu}(A \cup B) + \varepsilon > \bar{\mu}(O)$. Replacing U with $U \subseteq O$ and V with $V \cap O$, we can assume $\bar{\mu}(A \cup B) + \varepsilon > \bar{\mu}(U \cup V)$. Then

$$\begin{aligned} \bar{\mu}(A) + \bar{\mu}(B) &\leq \bar{\mu}(U) + \bar{\mu}(V) = \bar{\mu}(U \cup V) \\ &< \varepsilon + \bar{\mu}(A \cup B) \leq \varepsilon + \bar{\mu}(A) + \bar{\mu}(B) \end{aligned}$$

Since ε is arbitrary, this shows that $\bar{\mu}(A) + \bar{\mu}(B) = \bar{\mu}(A \cup B)$.

From Theorem 8.5.4 there is a complete measure μ defined on a σ algebra $\mathscr{F}$ which equals $\bar{\mu}$ on $\mathscr{F}$. From Claim 4 and Theorem 8.6.1, $\mathscr{F}$ contains the Borel sets $\mathscr{B}(X)$. From the definition, μ is outer regular and so it follows from Theorem 8.7.4 that μ is regular because it is finite on compact sets and X is the union of countably many compact sets so μ is σ finite. Thus $\mu(F) \leq \mu(E) \leq \mu(G) = \mu(F) + \mu(G \setminus F) = \mu(F)$.

The measure μ is uniquely determined. If μ, ν are two, then if K is compact, there is a sequence of open sets V_n decreasing to K such that $\mu(K) = \lim_{n\to\infty} \mu(V_n), \nu(K) = \lim_{n\to\infty} \nu(V_n)$. Now let $K \prec f_n \prec V_n$. By Claim 2, $Lf_n \to \mu(K)$ and $Lf_n \to \nu(K)$ so $\mu = \nu$ on all compact sets. Therefore, $\mu = \nu$ on all F_σ sets. Now every open set V is the countable union of a sequence of increasing compact sets from the assumptions on X so $\mu = \nu$ on every open set. It follows that the two σ algebras are the same and the measures are equal. To see this, $\mu(\hat{F}) \leq \mu(E) \leq \mu(\hat{G}), \nu(\tilde{F}) \leq \nu(E) \leq \nu(\tilde{G})$ where these G and F are respectively G_δ and F_σ with $\mu(\hat{G} \setminus \hat{F}) = 0$ similar with the other pair. So let $G = \tilde{G} \cap \hat{G}, F = \tilde{F} \cup \hat{F}$ and then $\mu(G \setminus F) = 0, \nu(G \setminus F) = 0$. Now if the two σ algebras are $\mathscr{F}_\mu, \mathscr{F}_\nu$ then if $E \in \mathscr{F}_\mu$, then E differs from F by a subset of a set of ν measure zero. Therefore, by completeness of ν it follows that $E \in \mathscr{F}_\nu$. Also $\mu(E) = \mu(F) = \nu(F) = \nu(E)$. The same argument shows that $\mathscr{F}_\nu \subseteq \mathscr{F}_\mu$. ■

Definition 8.8.3 *Let Lf be given by Theorem 5.8.8. That is*

$$\int_{\mathbb{R}^p} f dx = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(x_1, x_2, ..., x_p)\, dx_p \cdots dx_1$$

whenever f vanishes outside of $\prod_{i=1}^p (a_i, b_i)$. The resulting measure defined in Theorem 8.8.2, denoted as m_p is Lebesgue measure.

From the above theorem m_p is a Borel measure meaning that the Borel sets are measurable. Also it has the regularity properties. What does it do to boxes?

Theorem 8.8.4 *Lebesgue mesure is translation invariant. This terminology means that $m_p(E) = m_p(E + \boldsymbol{z})$. Also $m_p\left(\prod_{i=1}^p (a_i, b_i)\right) = m_p\left(\prod_{i=1}^p [a_i, b_i]\right) = \prod_{i=1}^p (b_i - a_i)$.*

Proof: What is $m_p(R)$ where $R = \prod_{i=1}^p (a_i, b_i)$? Let $R_n = \prod_{i=1}^p \left(a_i + \frac{1}{n}, b_i - \frac{1}{n}\right)$ and let $f_n = 1$ on R_n while vanishing off of R_{2n} and piecewise linear in each variable. Then from the definition, there is $g \in C_c(R)$ such that $m_p(R) < Lg + \varepsilon$ where here $g \prec R$. However, since the distance from the support of g to the boundary of R is positive, it follows that for all n large enough, $g \leq f_n$ and so $m_p(R) < Lf_n + \varepsilon$. Now letting $n \to \infty$ and computing $\int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f_n(x_1, x_2, ..., x_p)\, dx_p \cdots dx_1$, (You could use Problem 17 on Page 140.) it follows that $m_p(R) < \prod_{i=1}^p (b_i - a_i) + \varepsilon$. Since ε is arbitrary, it follows that $m_p(R) \leq \prod_{i=1}^p (b_i - a_i)$. In fact these will be equal because for each $n, Lf_n \leq m_p(R)$ and as just observed, $Lf_n \to \prod_{i=1}^p (b_i - a_i)$. In the case of a closed box, $\prod_{i=1}^p (a_i + \delta, b_i - \delta) \subseteq \prod_{i=1}^p [a_i, b_i] \subseteq \prod_{i=1}^p (a_i - \delta, b_i + \delta)$ and so for every $\delta > 0$ and sufficiently small,

$$m_p\left(\prod_{i=1}^p [a_i, b_i]\right) \in \left[\prod_{i=1}^p (b_i - a_i - 2\delta), \prod_{i=1}^p (b_i - a_i + 2\delta)\right]$$

and so $m_p\left(\prod_{i=1}^p [a_i, b_i]\right) = m_p\left(\prod_{i=1}^p (a_i, b_i)\right) = \prod_{i=1}^p (b_i - a_i)$.

Let $\mathscr{K}$ be all sets of the form $\prod_{i=1}^p (a_i, b_i)$ where $-\infty \leq a_i < b_i \leq \infty$. Clearly $\mathscr{F}$ is closed with respect to finite intersections. Let $\mathscr{G}$ be the Borel sets F such that

$$m_p(\boldsymbol{z} + F \cap (-m, m)^p) = m_p(F \cap (-m, m)^p).$$

Then $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. From what was just shown about rectangles, $\mathscr{G} \supseteq \mathscr{K}$ and so by Dynkin's lemma, it follows that $\mathscr{G} = \mathscr{B}(\mathbb{R}^p)$. Since m is arbitrary, this proves the theorem in case E is Borel.

From regularity, if E is only Lebesgue measurable, there is F an F_σ set and G a G_δ set such that $F \subseteq E \subseteq G$ and $m_p(G \setminus F) = 0$. Then from what was just shown, it follows that

$$\begin{aligned} m_p(E) &= m_p(F) = m_p(\boldsymbol{z}+F) \leq m_p(\boldsymbol{z}+E) \\ &\leq m_p(\boldsymbol{z}+G) = m_p(G) = m_p(F) \leq m_p(E) \end{aligned}$$

By completeness of m_p it follows that $\boldsymbol{z}+E$ is measurable because it lies between the F_σ set $\boldsymbol{z}+F$ and the G_δ set $\boldsymbol{z}+G$ and

$$m_p(\boldsymbol{z}+G \setminus (\boldsymbol{z}+F)) = m_p(\boldsymbol{z}+G \setminus F) = m_p(G \setminus F) = 0 \blacksquare$$

Example 8.8.5 *On $\mathbb{R}$ you could take an increasing function F and for the functional consider $L(f) \equiv \int f dF$ where this is the Riemann Stieltjes integral. This would give a measure μ_F with all the properties of the above Theorem 8.8.2.*

8.9 Exercises

1. Show carefully that if $\mathfrak{S}$ is a set whose elements are σ algebras which are subsets of $\mathscr{P}(\Omega)$, then $\cap\mathfrak{S}$ is also a σ algebra. Now let $\mathscr{G} \subseteq \mathscr{P}(\Omega)$ satisfy property P if $\mathscr{G}$ is closed with respect to complements and countable disjoint unions as in Dynkin's lemma, and contains $\emptyset$ and Ω. If $\mathfrak{H} \subseteq \mathscr{G}$ is any set whose elements are subsets of $\mathscr{P}(\Omega)$ which satisfies property P, then $\cap\mathfrak{H}$ also satisfies property P. Thus there is a smallest subset of $\mathscr{G}$ satisfying P. In other words, verify the details of the proof of Dynkin's lemma.

2. The Borel sets of a metric space (X,d) are the sets in the smallest σ algebra which contains the open sets. These sets are denoted as $\mathscr{B}(X)$. Thus $\mathscr{B}(X) = \sigma$ (open sets) where $\sigma(\mathscr{F})$ simply means the smallest σ algebra which contains $\mathscr{F}$. Show that in $\mathbb{R}^n$, $\mathscr{B}(\mathbb{R}^n) = \sigma(\mathscr{P})$ where $\mathscr{P}$ consists of the half open rectangles which are of the form $\prod_{i=1}^n [a_i, b_i)$.

3. Recall that $f : (\Omega, \mathscr{F}) \to X$ where X is a metric space is measurable means that inverse images of open sets are in $\mathscr{F}$. Show that if E is any set in $\mathscr{B}(X)$, then $f^{-1}(E) \in \mathscr{F}$. Thus, inverse images of Borel sets are measurable. Next consider $f : (\Omega, \mathscr{F}) \to X$ being measurable and $g : X \to Y$ is Borel measurable, meaning that $g^{-1}(open) \in \mathscr{B}(X)$. Explain why $g \circ f$ is measurable. **Hint:** You know that $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$. For your information, it does not work the other way around. That is, measurable composed with Borel measurable is not necessarily measurable. In fact examples exist which show that if g is measurable and f is continuous, then $g \circ f$ may fail to be measurable. An example is given later.

4. If you have X_i is a metric space, let $X = \prod_{i=1}^n X_i$ with the metric

$$d(\boldsymbol{x}, \boldsymbol{y}) \equiv \max\{d_i(x_i, y_i), i = 1, 2, \cdots, n\}$$

Show that any set of the form $\prod_{i=1}^n E_i,\ E_i \in \mathscr{B}(X_i)$ is a Borel set. That is, the product of Borel sets is Borel. **Hint:** You might consider the continuous functions

$\pi_i : \prod_{j=1}^{n} X_j \rightarrow X_i$ which are the projection maps. Thus $\pi_i(\boldsymbol{x}) \equiv x_i$. Then $\pi_i^{-1}(E_i)$ would have to be Borel measurable whenever $E_i \in \mathscr{B}(X_i)$. Explain why. You know π_i is continuous. Why would $\pi_i^{-1}(Borel)$ be a Borel set? Then you might argue that $\prod_{i=1}^{n} E_i = \cap_{i=1}^{n} \pi_i^{-1}(E_i)$.

5. You have two finite measures defined on $\mathscr{B}(X)$ μ, ν. Suppose these are equal on every open set. Show that these must be equal on every Borel set. **Hint:** You should use Dynkin's lemma to show this very easily.

6. Show that $(\mathbb{N}, \mathscr{P}(\mathbb{N}), \mu)$ is a measure space where $\mu(S)$ equals the number of elements of S. You need to verify that if the sets E_i are disjoint, then $\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$.

7. Let Ω be an uncountable set and let $\mathscr{F}$ denote those subsets of Ω, F such that either F or F^C is countable. Show that this is a σ algebra. Next define the following measure. $\mu(A) = 1$ if A is uncountable and $\mu(A) = 0$ if A is countable. Show that μ is a measure. This is a perverted example.

8. Let $\mu(E) = 1$ if $0 \in E$ and $\mu(E) = 0$ if $0 \notin E$. Show this is a measure on $\mathscr{P}(\mathbb{R})$.

9. Give an example of a measure μ and a measure space and a decreasing sequence of measurable sets $\{E_i\}$ such that $\lim_{n\to\infty} \mu(E_n) \neq \mu(\cap_{i=1}^{\infty} E_i)$.

10. Let $K \subseteq V$ where K is closed and V is open. Consider the following function.

$$f(x) = \frac{\operatorname{dist}(x, V^C)}{\operatorname{dist}(x, K) + \operatorname{dist}(x, V^C)}$$

Explain why this function is continuous, equals 0 off V and equals 1 on K. It is in the book earlier, but go through the details.

11. Let $(\Omega, \mathscr{F})$ be a measurable space and let $f : \Omega \rightarrow X$ be a measurable function. Then $\sigma(f)$ denotes the smallest σ algebra such that f is measurable with respect to this σ algebra. Show that $\sigma(f) = \left\{f^{-1}(E) : E \in \mathscr{B}(X)\right\}$.

12. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. A sequence of functions $\{f_n\}$ is said to converge in measure to a measurable function f if and only if for each

$$\varepsilon > 0, \lim_{n\to\infty} \mu(\omega : |f_n(\omega) - f(\omega)| > \varepsilon) = 0.$$

Show that if this happens, then there exists a subsequence $\left\{f_{n_k}\right\}$ and a set of measure N such that if $\omega \notin N$, then $\lim_{k\to\infty} f_{n_k}(\omega) = f(\omega)$. Also show that if $\lim_{n\to\infty} f_n(\omega) = f(\omega)$, and $\mu(\Omega) < \infty$, then f_n converges in measure to f. **Hint:**For the subsequence, let $\mu\left(\omega : \left|f_{n_k}(\omega) - f(\omega)\right| > \varepsilon\right) < 2^{-k}$ and use Borel Cantelli lemma.

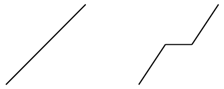
13. Let X, Y be separable metric spaces. Then $X \times Y$ can also be considered as a metric space with the metric $\rho((x,y),(\hat{x},\hat{y})) \equiv \max(d_X(x,\hat{x}), d_Y(y,\hat{y}))$. Verify this. Then show that if $\mathscr{K}$ consists of sets $A \times B$ where A, B are Borel sets in X and Y respectively, then it follows that $\sigma(\mathscr{K}) = \mathscr{B}(X \times Y)$, the Borel sets from $X \times Y$. Extend to the Cartesian product $\prod_i X_i$ of finitely many separable metric spaces.

14. Suppose you have $(X, \mathscr{F}, \mu)$ where $\mathscr{F} \supseteq \mathscr{B}(X)$ and also $\mu(B(x_0, r)) < \infty$ for all $r > 0$. Let $S(x_0, r) \equiv \{x \in X : d(x, x_0) = r\}$. Show that $\{r > 0 : \mu(S(x_0, r)) > 0\}$ cannot be uncountable. Explain why there exists a strictly increasing sequence $r_n \to \infty$ such that $\mu(x : d(x, x_0) = r_n) = 0$. In other words, the skin of the ball has measure zero except for possibly countably many values of the radius r.

15. Lebesgue measure was discussed. Recall that $m((a, b)) = b - a$ and it is defined on a σ algebra which contains the Borel sets, more generally on $\mathscr{P}(\mathbb{R})$. Also recall that m is translation invariant. Let $x \sim y$ if and only if $x - y \in \mathbb{Q}$. Show this is an equivalence relation. Now let W be a set of positive measure which is contained in $(0, 1)$. For $x \in W$, let $[x]$ denote those $y \in W$ such that $x \sim y$. Thus the equivalence classes partition W. Use axiom of choice to obtain a set $S \subseteq W$ such that S consists of exactly one element from each equivalence class. Let $\mathbb{T}$ denote the rational numbers in $[-1, 1]$. Consider $\mathbb{T} + S \subseteq [-1, 2]$. Explain why $\mathbb{T} + S \supseteq W$. For $\mathbb{T} \equiv \{r_j\}$, explain why the sets $\{r_j + S\}_j$ are disjoint. Now suppose S is measurable. Then show that you have a contradiction if $m(S) = 0$ since $m(W) > 0$ and you also have a contradiction if $m(S) > 0$ because $\mathbb{T} + S$ consists of countably many disjoint sets. Explain why S cannot be measurable. Thus there exists $T \subseteq \mathbb{R}$ such that $m(T) < m(T \cap S) + m(T \cap S^C)$. Is there an open interval (a, b) such that if $T = (a, b)$, then the above inequality holds?

16. Consider the following nested sequence of compact sets, $\{P_n\}$.Let $P_1 = [0, 1]$, $P_2 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$, etc. To go from P_n to P_{n+1}, delete the open interval which is the middle third of each closed interval in P_n. Let $P = \cap_{n=1}^{\infty} P_n$. By the finite intersection property of compact sets, $P \neq \emptyset$. Show $m(P) = 0$. If you feel ambitious also show there is a one to one onto mapping of $[0, 1]$ to P. The set P is called the Cantor set. Thus, although P has measure zero, it has the same number of points in it as $[0, 1]$ in the sense that there is a one to one and onto mapping from one to the other. **Hint:** There are various ways of doing this last part but the most enlightenment is obtained by exploiting the topological properties of the Cantor set rather than some silly representation in terms of sums of powers of two and three. All you need to do is use the Schroder Bernstein theorem and show there is an onto map from the Cantor set to $[0, 1]$. If you do this right and remember the theorems about characterizations of compact metric spaces, Proposition 3.5.8 on Page 70, you may get a pretty good idea why every compact metric space is the continuous image of the Cantor set.

17. Consider the sequence of functions defined in the following way. Let $f_1(x) = x$ on $[0, 1]$. To get from f_n to f_{n+1}, let $f_{n+1} = f_n$ on all intervals where f_n is constant. If f_n is nonconstant on $[a, b]$, let $f_{n+1}(a) = f_n(a)$, $f_{n+1}(b) = f_n(b)$, f_{n+1} is piecewise linear and equal to $\frac{1}{2}(f_n(a) + f_n(b))$ on the middle third of $[a, b]$. Sketch a few of these and you will see the pattern. The process of modifying a nonconstant section of the graph of this function is illustrated in the following picture.



Show $\{f_n\}$ converges uniformly on $[0, 1]$. If $f(x) = \lim_{n \to \infty} f_n(x)$, show that $f(0) = 0$, $f(1) = 1$, f is continuous, and $f'(x) = 0$ for all $x \notin P$ where P is the Cantor set of Problem 16. This function is called the Cantor function.It is a very important example to remember. Note it has derivative equal to zero a.e. and yet it succeeds in climbing from 0 to 1. Explain why this interesting function cannot be recovered by integrating its derivative. (It is not absolutely continuous, explained later.) **Hint:** This isn't too hard if you focus on getting a careful estimate on the difference between two successive functions in the list considering only a typical small interval in which the change takes place. The above picture should be helpful.

18. ↑ This problem gives a very interesting example found in the book by McShane [33]. Let $g(x) = x + f(x)$ where f is the strange function of Problem 17. Let P be the Cantor set of Problem 16. Let $[0,1] \setminus P = \cup_{j=1}^{\infty} I_j$ where I_j is open and $I_j \cap I_k = \emptyset$ if $j \neq k$. These intervals are the connected components of the complement of the Cantor set. Show $m(g(I_j)) = m(I_j)$ so $m(g(\cup_{j=1}^{\infty} I_j)) = \sum_{j=1}^{\infty} m(g(I_j)) = \sum_{j=1}^{\infty} m(I_j) = 1$. Thus $m(g(P)) = 1$ because $g([0,1]) = [0,2]$. By Problem 15 there exists a set, $A \subseteq g(P)$ which is non measurable. Define $\phi(x) = \mathscr{X}_A(g(x))$. Thus $\phi(x) = 0$ unless $x \in P$. Tell why ϕ is measurable. (Recall $m(P) = 0$ and Lebesgue measure is complete.) Now show that $\mathscr{X}_A(y) = \phi(g^{-1}(y))$ for $y \in [0,2]$. Tell why g is strictly increasing and g^{-1} is continuous but $\phi \circ g^{-1}$ is not measurable. (This is an example of measurable $\circ$ continuous $\neq$ measurable.) Show there exist Lebesgue measurable sets which are not Borel measurable. **Hint**: The function, ϕ is Lebesgue measurable. Now recall that Borel $\circ$ measurable = measurable.

19. Show that every countable set of real numbers is of Lebesgue measure zero.

20. The Cantor set is obtained by starting with $[0,1]$, delete the middle third, the open set $(1/3, 2/3)$. Now do the same for the two remaining closed intervals. This results in a nested sequence of compact sets. The intersection of all of these is the Cantor set. Show that you can take out open intervals in the middle which are not necessarily middle thirds, and end up with a set C which has Lebesgue measure equal to $1 - \varepsilon$. Also show if you can that there exists a continuous and one to one map $f : C \to J$ where J is the usual Cantor set which also has measure 0.

21. Suppose you have a π system $\mathscr{K}$ of sets of Ω and suppose $\mathscr{G} \supseteq \mathscr{K}$ and that $\mathscr{G}$ is closed with respect to complements and that whenever $\{F_k\}$ is a decreasing sequence of sets of $\mathscr{G}$ it follows that $\cap_k F_k \in \mathscr{G}$. Show that then $\mathscr{G}$ contains $\sigma(\mathscr{K})$. This is an alternative formulation of Dynkin's lemma. It was shown after the Dynkin lemma that closure with respect to countable intersections is equivalent.

22. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $s(\omega) = \sum_{i=0}^{n} c_i \mathscr{X}_{E_i}(\omega)$ where the E_i are distinct measurable sets but the c_i might not be. Thus the c_i are the finitely many values of s. Say each $c_i \geq 0$ and $c_0 = 0$. Define $\int s d\mu$ as $\sum_i c_i \mu(E_i)$. Show that this is well defined and that if you have $s(\omega) = \sum_{i=1}^{n} c_i \mathscr{X}_{E_i}(\omega), t(\omega) = \sum_{j=1}^{m} d_j \mathscr{X}_{F_j}(\omega)$, then for a, b nonnegative numbers, $as(\omega) + bt(\omega)$ can be written also in this form and that $\int (as + bt) d\mu = a \int s d\mu + b \int t d\mu$. **Hint:** $s(\omega) = \sum_i \sum_j c_i \mathscr{X}_{E_i \cap F_j}(\omega) = \sum_j \sum_i c_i \mathscr{X}_{E_i \cap F_j}(\omega)$ and $(as + bt)(\omega) = \sum_j \sum_i (ac_i + bd_j) \mathscr{X}_{E_i \cap F_j}(\omega)$.

23. ↑Having defined the integral of nonnegative simple functions in the above problem, letting f be nonnegative and measurable. Define

$$\int f d\mu \equiv \sup \left\{ \int s d\mu : 0 \leq s \leq f, s \text{ simple} \right\}.$$

Show that if f_n is nonnegative and measurable and $n \to f_n(\omega)$ is increasing, show that for $f(\omega) = \lim_{n\to\infty} f_n(\omega)$, it follows that $\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$. **Hint:** Show $\int f_n d\mu$ is increasing to something $\alpha \leq \infty$. Explain why $\int f d\mu \geq \alpha$. Now pick a nonnegative simple function $s \leq f$. For $r \in (0,1)$, $[f_n > rs] \equiv E_n$ is increasing in n and $\cup_n E_n = \Omega$. Tell why $\int f_n d\mu \geq \int \mathscr{X}_{E_n} f_n d\mu \geq r \int s d\mu$. Let $n \to \infty$ and show that $\alpha \geq r \int s d\mu$. Now explain why $\alpha \geq r \int f d\mu$. Since r is arbitrary, $\alpha \geq \int f d\mu \geq \alpha$.

24. ↑Show that if f, g are nonnegative and measurable and $a, b \geq 0$, then

$$\int (af + bg)\, d\mu = a \int f d\mu + b \int g d\mu$$

25. Let F be increasing on $\mathbb{R}$. Consider the measure μ_F from Theorem 8.8.2 in which the functional is the Riemann Stieltjes integral $\int f dF$. Show that

$$\mu_F((a_i, b_i)) = F(b_i-) - F(a_i+),$$
$$\mu_F([a,b)) = F(b-) - F(a-),$$
$$\mu_F((a,b]) = F(b+) - F(a+),$$
$$\mu_F([a,b]) = F(b+) - F(a-).$$

Here $F(b+) = \lim_{x\to b+} F(x)$ that is, it is the limit from the right. Other notation is similar. This will give the Lebesgue Stieltjes measures. These measures will NOT be translation invariant. Why? However, they still have the regularity properties.

Chapter 9

The Lebesgue Integral

The presentation in terms of simple functions of the Lebesgue integral is presented in Problems starting with 22 on Page 199. I will present it a different way here. The general Lebesgue integral requires a measure space, $(\Omega, \mathscr{F}, \mu)$ and, to begin with, a nonnegative measurable function. I will use Lemma 2.5.3 about interchanging two supremums frequently. Also, I will use the observation that if $\{a_n\}$ is an increasing sequence of points of $[0,\infty]$, then $\sup_n a_n = \lim_{n\to\infty} a_n$ which is obvious from the definition of sup.

9.1 Nonnegative Measurable Functions

9.1.1 Riemann Integrals for Decreasing Functions

First of all, the notation $[g < f]$ means $\{\omega \in \Omega : g(\omega) < f(\omega)\}$ with other variants of this notation being similar. Also, the convention, $0 \cdot \infty = 0$ will be used to simplify the presentation whenever it is convenient to do so. The notation $a \wedge b$ means the minimum of a and b.

Definition 9.1.1 *Let $f : [a,b] \to [0,\infty]$ be decreasing. Note that ∞ is a possible value. Define*

$$\int_a^b f(\lambda)\, d\lambda \equiv \lim_{M\to\infty} \int_a^b M \wedge f(\lambda)\, d\lambda = \sup_M \int_a^b M \wedge f(\lambda)\, d\lambda$$

where $a \wedge b$ means the minimum of a and b. Note that for f bounded,

$$\sup_M \int_a^b M \wedge f(\lambda)\, d\lambda = \int_a^b f(\lambda)\, d\lambda$$

where the integral on the right is the usual Riemann integral because eventually $M > f$. For f a nonnegative decreasing function defined on $[0,\infty)$,

$$\int_0^\infty f d\lambda \equiv \lim_{R\to\infty} \int_0^R f d\lambda = \sup_{R>1} \int_0^R f d\lambda = \sup_R \sup_{M>0} \int_0^R f \wedge M d\lambda$$

Since decreasing bounded functions are Riemann integrable, the above definition is well defined. For a discussion of this, see Calculus of One and Many Variables on the web site or the single variable advanced calculus book. Now here is an obvious property.

Lemma 9.1.2 *Let f be a decreasing nonnegative function defined on an interval $[a,b]$. Then if $[a,b] = \cup_{k=1}^m I_k$ where $I_k \equiv [a_k, b_k]$ and the intervals I_k are non overlapping, it follows*

$$\int_a^b f d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda.$$

Proof: This follows from the computation,

$$\int_a^b f d\lambda \equiv \lim_{M\to\infty} \int_a^b f \wedge M d\lambda = \lim_{M\to\infty} \sum_{k=1}^m \int_{a_k}^{b_k} f \wedge M d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda$$

Note both sides could equal $+\infty$. ■

In all considerations below, assume h is fairly small, certainly much smaller than R. Thus $R - h > 0$.

Lemma 9.1.3 *Let g be a decreasing nonnegative function defined on an interval $[0,R]$. Then*

$$\int_0^R g\wedge M d\lambda = \sup_{h>0}\sum_{i=1}^{m(R,h)}\left(g\left(ih\right)\wedge M\right)h$$

where $m(h,R)\in\mathbb{N}$ satisfies $R-h<hm(h,R)\leq R$.

Proof: Since $g\wedge M$ is a decreasing bounded function the lower sums converge to the integral as $h\to 0$. Thus

$$\int_0^R g\wedge M d\lambda = \lim_{h\to 0}\left(\sum_{i=1}^{m(R,h)}\left(g\left(ih\right)\wedge M\right)h+\left(g\left(R\right)\wedge M\right)\left(R-hm\left(h,R\right)\right)\right)$$

Now the last term in the above is no more than Mh and so the above is

$$\lim_{h\to 0}\left(\sum_{i=1}^{m(R,h)}\left(g\left(ih\right)\wedge M\right)h\right)=\sup_{h>0}\left(\sum_{i=1}^{m(R,h)}\left(g\left(ih\right)\wedge M\right)h\right). \blacksquare$$

9.1.2 The Lebesgue Integral for Nonnegative Functions

Here is the definition of the Lebesgue integral of a function which is measurable and has values in $[0,\infty]$.

Definition 9.1.4 *Let $(\Omega,\mathscr{F},\mu)$ be a measure space and suppose $f:\Omega\to[0,\infty]$ is measurable. Then define $\int fd\mu\equiv\int_0^\infty\mu\left([f>\lambda]\right)d\lambda$ which makes sense because $\lambda\to\mu\left([f>\lambda]\right)$ is nonnegative and decreasing.*

Note that if $f\leq g$, then $\int fd\mu\leq\int gd\mu$ because $\mu\left([f>\lambda]\right)\leq\mu\left([g>\lambda]\right)$.
For convenience $\sum_{i=1}^{0}a_i\equiv 0$.

Lemma 9.1.5 *In the above definition, $\int fd\mu=\sup_{h>0}\sum_{i=1}^{\infty}\mu\left([f>hi]\right)h$*

Proof: Let $m(h,R)\in\mathbb{N}$ satisfy $R-h<hm(h,R)\leq R$. Then $\lim_{R\to\infty}m(h,R)=\infty$ and so from Lemma 9.1.3,

$$\begin{aligned}\int fd\mu &\equiv \int_0^\infty\mu\left([f>\lambda]\right)d\lambda=\sup_M\sup_R\int_0^R\mu\left([f>\lambda]\right)\wedge Md\lambda\\ &= \sup_M\sup_{R>0}\sup_{h>0}\sum_{k=1}^{m(h,R)}\left(\mu\left([f>kh]\right)\wedge M\right)h\end{aligned}$$

Hence, switching the order of the sups, this equals

$$\sup_{R>0}\sup_{h>0}\sup_M\sum_{k=1}^{m(h,R)}\left(\mu\left([f>kh]\right)\wedge M\right)h=\sup_{R>0}\sup_{h>0}\lim_{M\to\infty}\sum_{k=1}^{m(h,R)}\left(\mu\left([f>kh]\right)\wedge M\right)h$$

$$=\sup_{h>0}\sup_R\sum_{k=1}^{m(R,h)}\left(\mu\left([f>kh]\right)\right)h=\sup_{h>0}\sum_{k=1}^{\infty}\left(\mu\left([f>kh]\right)\right)h. \blacksquare$$

9.2 Nonnegative Simple Functions

To begin with, here is a useful lemma.

Lemma 9.2.1 *If $f(\lambda)=0$ for all $\lambda>a$, where f is a decreasing nonnegative function, then $\int_0^\infty f(\lambda)\,d\lambda=\int_0^a f(\lambda)\,d\lambda$.*

Proof: From the definition,

$$\begin{aligned}\int_0^\infty f(\lambda)\,d\lambda &= \lim_{R\to\infty}\int_0^R f(\lambda)\,d\lambda=\sup_{R>1}\int_0^R f(\lambda)\,d\lambda=\sup_{R>1}\sup_M\int_0^R f(\lambda)\wedge M d\lambda\\ &= \sup_M\sup_{R>1}\int_0^R f(\lambda)\wedge M d\lambda=\sup_M\sup_{R>1}\int_0^a f(\lambda)\wedge M d\lambda\\ &= \sup_M\int_0^a f(\lambda)\wedge M d\lambda\equiv\int_0^a f(\lambda)\,d\lambda.\ \blacksquare\end{aligned}$$

Now the Lebesgue integral for a nonnegative function has been defined, what does it do to a nonnegative simple function? Recall a nonnegative simple function is one which has finitely many nonnegative real values which it assumes on measurable sets. Thus a simple function can be written in the form $s(\omega)=\sum_{i=1}^n c_i\mathscr{X}_{E_i}(\omega)$ where the c_i are each nonnegative, the distinct nonzero values of s.

Lemma 9.2.2 *Let $s(\omega)=\sum_{i=1}^p a_i\mathscr{X}_{E_i}(\omega)$ be a nonnegative simple function where the E_i are distinct but the a_i might not be. Thus the values of s are the a_i. Then*

$$\int s\,d\mu=\sum_{i=1}^p a_i\mu(E_i). \tag{9.1}$$

Proof: Without loss of generality, assume $0\equiv a_0<a_1\leq a_2\leq\cdots\leq a_p$ and that $\mu(E_i)<\infty, i>0$. Here is why. If $\mu(E_i)=\infty,$ then letting $a\in(a_{i-1},a_i)$, by Lemma 9.2.1, the left side is

$$\begin{aligned}\int_0^{a_p}\mu([s>\lambda])\,d\lambda &\geq \int_{a_0}^{a_i}\mu([s>\lambda])\,d\lambda\\ &\equiv \sup_M\int_0^{a_i}\mu([s>\lambda])\wedge M d\lambda\geq\sup_M M\mu(E_i)\,a_i=\infty\end{aligned}$$

and so both sides of 9.1 are equal to ∞. Thus it can be assumed for each $i,\mu(E_i)<\infty$. Then it follows from Lemma 9.2.1 and Lemma 9.1.2,

$$\int_0^\infty\mu([s>\lambda])\,d\lambda=\int_0^{a_p}\mu([s>\lambda])\,d\lambda=\sum_{k=1}^p\int_{a_{k-1}}^{a_k}\mu([s>\lambda])\,d\lambda$$

$$=\sum_{k=1}^p(a_k-a_{k-1})\sum_{i=k}^p\mu(E_i)=\sum_{i=1}^p\mu(E_i)\sum_{k=1}^i(a_k-a_{k-1})=\sum_{i=1}^p a_i\mu(E_i)\ \blacksquare$$

Note that this is the same result as in Problem 22 on Page 199 but here there is no question about the definition of the integral of a simple function being well defined.

Lemma 9.2.3 *If $a,b\geq 0$ and if s and t are nonnegative simple functions, then*

$$\int as+bt\,d\mu=a\int s\,d\mu+b\int t\,d\mu.$$

Proof: Let $s(\omega) = \sum_{i=1}^{n} \alpha_i \mathscr{X}_{A_i}(\omega)$, $t(\omega) = \sum_{i=1}^{m} \beta_j \mathscr{X}_{B_j}(\omega)$ where α_i are the distinct values of s and the β_j are the distinct values of t. Clearly $as + bt$ is a nonnegative simple function because it has finitely many values on measurable sets. In fact, $(as+bt)(\omega) = \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mathscr{X}_{A_i \cap B_j}(\omega)$ where the sets $A_i \cap B_j$ are disjoint and measurable. By Lemma 9.2.2,

$$\begin{aligned} \int as + bt d\mu &= \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mu(A_i \cap B_j) \\ &= \sum_{i=1}^{n} a \sum_{j=1}^{m} \alpha_i \mu(A_i \cap B_j) + b \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^{n} \alpha_i \mu(A_i) + b \sum_{j=1}^{m} \beta_j \mu(B_j) = a \int s d\mu + b \int t d\mu. \blacksquare \end{aligned}$$

9.3 The Monotone Convergence Theorem

The following is called the monotone convergence theorem. This theorem and related convergence theorems are the reason for using the Lebesgue integral. If $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ and f_n is increasing in n, then clearly f is also measurable because

$$f^{-1}((a,\infty]) = \cup_{k=1}^{\infty} f_k^{-1}((a,\infty]) \in \mathscr{F}$$

For a different approach to this, see Problem 22 on Page 199.

Theorem 9.3.1 *(Monotone Convergence theorem) Suppose that the function f has all values in $[0,\infty]$ and suppose $\{f_n\}$ is a sequence of nonnegative measurable functions having values in $[0,\infty]$ and satisfying*

$$\lim_{n\to\infty} f_n(\omega) = f(\omega) \text{ for each } \omega.$$

$$\cdots f_n(\omega) \leq f_{n+1}(\omega) \cdots$$

Then f is measurable and $\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$.

Proof: By Lemma 9.1.5 $\lim_{n\to\infty} \int f_n d\mu = \sup_n \int f_n d\mu$

$$\begin{aligned} &= \sup_n \sup_{h>0} \sum_{k=1}^{\infty} \mu([f_n > kh]) h = \sup_{h>0} \sup_N \sup_n \sum_{k=1}^{N} \mu([f_n > kh]) h \\ &= \sup_{h>0} \sup_N \sum_{k=1}^{N} \mu([f > kh]) h = \sup_{h>0} \sum_{k=1}^{\infty} \mu([f > kh]) h = \int f d\mu. \blacksquare \end{aligned}$$

Note how it was important to have $\int_0^\infty [f > \lambda] d\lambda$ in the definition of the integral and **not** $[f \geq \lambda]$. You need to have $[f_n > kh] \uparrow [f > kh]$ so $\mu([f_n > kh]) \to \mu([f > kh])$. To illustrate what goes wrong without the Lebesgue integral, consider the following example.

Example 9.3.2 *Let $\{r_n\}$ denote the rational numbers in $[0,1]$ and let*

$$f_n(t) \equiv \begin{cases} 1 \text{ if } t \notin \{r_1, \cdots, r_n\} \\ 0 \text{ otherwise} \end{cases}$$

Then $f_n(t) \uparrow f(t)$ where f is the function which is one on the rationals and zero on the irrationals. Each f_n is Riemann integrable (why?) but f is not Riemann integrable because it is everywhere discontinuous. Also, there is a gap between all upper sums and lower sums. Therefore, you can't write $\int f dx = \lim_{n\to\infty} \int f_n dx$.

An observation which is typically true related to this type of example is this. If you can choose your functions, you don't need the Lebesgue integral. The Riemann Darboux integral is just fine. It is when you can't choose your functions and they come to you as pointwise limits that you really need the superior Lebesgue integral or at least something more general than the Riemann integral. The Riemann integral is entirely adequate for evaluating the seemingly endless lists of boring problems found in calculus books. It is shown later that the two integrals coincide when the Lebesgue integral is taken with respect to Lebesgue measure and the function being integrated is continuous. It has been correctly observed that we never compute a Lebesgue integral. We compute Riemann integrals and sometimes take limits.

9.4 Other Definitions

To review and summarize the above, if $f \geq 0$ is measurable,

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda]) d\lambda \tag{9.2}$$

another way to get the same thing for $\int f d\mu$ is to take an increasing sequence of nonnegative simple functions, $\{s_n\}$ with $s_n(\omega) \to f(\omega)$ and then by monotone convergence theorem, $\int f d\mu = \lim_{n\to\infty} \int s_n$ where if $s_n(\omega) = \sum_{j=1}^m c_i \mathscr{X}_{E_i}(\omega)$, $\int s_n d\mu = \sum_{i=1}^m c_i \mu(E_i)$. Similarly this also shows that for such nonnegative measurable function,

$$\int f d\mu = \sup \left\{ \int s : 0 \leq s \leq f,\ s \text{ simple} \right\}.$$

Here is an equivalent definition of the integral of a nonnegative measurable function. The fact it is well defined has been discussed above.

Definition 9.4.1 *For s a nonnegative simple function,*

$$s(\omega) = \sum_{k=1}^n c_k \mathscr{X}_{E_k}(\omega), \int s = \sum_{k=1}^n c_k \mu(E_k).$$

For f a nonnegative measurable function,

$$\int f d\mu = \sup \left\{ \int s : 0 \leq s \leq f,\ s \textit{ simple} \right\}.$$

Proof: Let V be an open set and let $V = \cup_n K_n$ where $K_n \subseteq K_{n+1}$ for all n. Let

$$g_n(x) \equiv 1 - \frac{\text{dist}(x, K_n)}{\text{dist}(x, K_n) + \text{dist}(x, V^C)}, f_n \equiv \max\{g_k : k \leq n\}$$

Then using the monotone convergence theorem, it follows that $\mu = \nu$ on all open sets. The conclusion follows from Theorem 8.7.4. ■

9.5 Fatou's Lemma

The next theorem, known as Fatou's lemma is another important theorem which justifies the use of the Lebesgue integral.

Theorem 9.5.1 *(Fatou's lemma) Let f_n be a nonnegative measurable function. Let $g(\omega) = \liminf_{n\to\infty} f_n(\omega)$. Then g is measurable and $\int g d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$. In other words, $\int (\liminf_{n\to\infty} f_n)\, d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$.*

Proof: Let $g_n(\omega) = \inf\{f_k(\omega) : k \geq n\}$. Then

$$g_n^{-1}([a,\infty]) = \cap_{k=n}^{\infty} f_k^{-1}([a,\infty]) = \left(\cup_{k=n}^{\infty} f_k^{-1}([a,\infty])^C\right)^C \in \mathscr{F}.$$

Thus g_n is measurable by Lemma 8.1.4. Also $g(\omega) = \lim_{n\to\infty} g_n(\omega)$ so g is measurable because it is the pointwise limit of measurable functions. Now the functions g_n form an increasing sequence of nonnegative measurable functions so the monotone convergence theorem applies. This yields

$$\int g d\mu = \lim_{n\to\infty} \int g_n d\mu \leq \lim \inf_{n\to\infty} \int f_n d\mu.$$

The last inequality holding because $\int g_n d\mu \leq \int f_n d\mu$. (Note that it is not known whether $\lim_{n\to\infty} \int f_n d\mu$ exists.) ■

9.6 The Integral's Righteous Algebraic Desires

The monotone convergence theorem shows the integral wants to be linear. This is the essential content of the next theorem.

Theorem 9.6.1 *Let f, g be nonnegative measurable functions and let a, b be nonnegative numbers. Then $af + bg$ is measurable and*

$$\int (af + bg)\, d\mu = a \int f d\mu + b \int g d\mu. \tag{9.3}$$

Proof: By Theorem 8.1.6 on Page 181 there exist increasing sequences of nonnegative simple functions, $s_n \to f$ and $t_n \to g$. Then $af + bg$, being the pointwise limit of the simple functions $as_n + bt_n$, is measurable. Now by the monotone convergence theorem and Lemma 9.2.3,

$$\begin{aligned} \int (af + bg)\, d\mu &= \lim_{n\to\infty} \int as_n + bt_n d\mu = \lim_{n\to\infty} \left(a \int s_n d\mu + b \int t_n d\mu \right) \\ &= a \int f d\mu + b \int g d\mu. \ ■ \end{aligned}$$

As long as you are allowing functions to take the value $+\infty$, you cannot consider something like $f + (-g)$ and so you can't very well expect a satisfactory statement about the integral being linear until you restrict yourself to functions which have values in a vector space. To be linear, a function must be defined on a vector space. This is discussed next.

9.7 The Lebesgue Integral, L^1

The functions considered here have values in $\mathbb{C}$, which is a vector space. A function f with values in $\mathbb{C}$ is of the form $f = \operatorname{Re} f + i \operatorname{Im} f$ where $\operatorname{Re} f$ and $\operatorname{Im} f$ are real valued functions. In fact $\operatorname{Re} f = \frac{f+\overline{f}}{2}$, $\operatorname{Im} f = \frac{f-\overline{f}}{2i}$.

Definition 9.7.1 *Let $(\Omega, \mathscr{S}, \mu)$ be a measure space and suppose $f : \Omega \to \mathbb{C}$. Then f is said to be measurable if both $\operatorname{Re} f$ and $\operatorname{Im} f$ are measurable real valued functions.*

Of course there is another definition of measurability which says that inverse images of open sets are measurable. This is equivalent to this new definition.

Lemma 9.7.2 *Let $f : \Omega \to \mathbb{C}$. Then f is measurable if and only if $\operatorname{Re} f, \operatorname{Im} f$ are both real valued measurable functions. Also if f, g are complex measurable functions and a, b are complex scalars, then $af + bg$ is also measurable.*

Proof: $\Rightarrow$Suppose first that f is measurable. Recall that $\mathbb{C}$ is considered as $\mathbb{R}^2$ with (x,y) being identified with $x+iy$. Thus the open sets of $\mathbb{C}$ can be obtained with either of the two equivlanent norms $|z| \equiv \sqrt{(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2}$ or $\|z\|_\infty = \max(\operatorname{Re} z, \operatorname{Im} z)$. Therefore, if f is measurable $\operatorname{Re} f^{-1}(a,b) \cap \operatorname{Im} f^{-1}(c,d) = f^{-1}((a,b) + i(c,d)) \in \mathscr{F}$. In particular, you could let $(c,d) = \mathbb{R}$ and conclude that $\operatorname{Re} f$ is measurable because in this case, the above reduces to the statement that $\operatorname{Re} f^{-1}(a,b) \in \mathscr{F}$. Similarly $\operatorname{Im} f$ is measurable.

$\Leftarrow$ Next, if each of $\operatorname{Re} f$ and $\operatorname{Im} f$ are measurable, then

$$f^{-1}((a,b) + i(c,d)) = \operatorname{Re} f^{-1}(a,b) \cap \operatorname{Im} f^{-1}(c,d) \in \mathscr{F}$$

and so, since every open set is the countable union of sets of the form $(a,b) + i(c,d)$, it follows that f is measurable.

Now consider the last claim. Let $h : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ be given by $h(z,w) \equiv az + bw$. Then h is continuous. If f, g are complex valued measurable functions, consider the complex valued function, $h \circ (f,g) : \Omega \to \mathbb{C}$. Then

$$(h \circ (f,g))^{-1}(\text{open}) = (f,g)^{-1}\left(h^{-1}(\text{open})\right) = (f,g)^{-1}(\text{open})$$

Now letting U, V be open in $\mathbb{C}$, $(f,g)^{-1}(U \times V) = f^{-1}(U) \cap g^{-1}(V) \in \mathscr{F}$. Since every open set in $\mathbb{C} \times \mathbb{C}$ is the countable union of sets of the form $U \times V$, it follows that $(f,g)^{-1}(\text{open})$ is in $\mathscr{F}$. Thus $af + bg$ is also complex measurable. ■

As is always the case for complex numbers, $|\boldsymbol{z}|^2 = (\operatorname{Re} z)^2 + (\operatorname{Im} z)^2$. Also, for g a real valued function, one can consider its positive and negative parts defined respectively as

$$g^+(x) \equiv \frac{g(x) + |g(x)|}{2},\ g^-(x) = \frac{|g(x)| - g(x)}{2}.$$

Thus $|g| = g^+ + g^-$ and $g = g^+ - g^-$ and both g^+ and g^- are measurable nonnegative functions if g is measurable.

Then the following is the definition of what it means for a complex valued function f to be in $L^1(\Omega)$.

Definition 9.7.3 *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Then a complex valued measurable function f is in $L^1(\Omega)$ if $\int |f| d\mu < \infty$. For a function in $L^1(\Omega)$, the integral is defined as follows.*

$$\int f d\mu \equiv \int (\operatorname{Re} f)^+ d\mu - \int (\operatorname{Re} f)^- d\mu + i\left[\int (\operatorname{Im} f)^+ d\mu - \int (\operatorname{Im} f)^- d\mu\right]$$

I will show that with this definition, the integral is linear and well defined. First note that it is clearly well defined because all the above integrals are of nonnegative functions and are each equal to a nonnegative real number because for h equal to any of the functions, $|h| \leq |f|$ and $\int |f| d\mu < \infty$.

Here is a lemma which will make it possible to show the integral is linear.

Lemma 9.7.4 *Let g, h, g', h' be nonnegative measurable functions in $L^1(\Omega)$ and suppose that $g - h = g' - h'$.Then $\int g d\mu - \int h d\mu = \int g' d\mu - \int h' d\mu$.*

Proof: By assumption, $g + h' = g' + h$. Then from the Lebesgue integral's righteous algebraic desires, Theorem 9.6.1, $\int g d\mu + \int h' d\mu = \int g' d\mu + \int h d\mu$ which implies the claimed result. ■

Lemma 9.7.5 *Let $\operatorname{Re}\left(L^1(\Omega)\right)$ denote the vector space of real valued functions in $L^1(\Omega)$ where the field of scalars is the real numbers. Then $\int d\mu$ is linear on $\operatorname{Re}\left(L^1(\Omega)\right)$, the scalars being real numbers.*

Proof: First observe that from the definition of the positive and negative parts of a function, $(f+g)^+ - (f+g)^- = f^+ + g^+ - (f^- + g^-)$ because both sides equal $f+g$. Therefore from Lemma 9.7.4 and the definition, it follows from Theorem 9.6.1 that

$$\begin{aligned}\int f + g d\mu &\equiv \int (f+g)^+ - (f+g)^- d\mu = \int f^+ + g^+ d\mu - \int f^- + g^- d\mu \\ &= \int f^+ d\mu + \int g^+ d\mu - \left(\int f^- d\mu + \int g^- d\mu\right) = \int f d\mu + \int g d\mu.\end{aligned}$$

what about taking out scalars? First note that if a is real and nonnegative, then $(af)^+ = af^+$ and $(af)^- = af^-$ while if $a < 0$, then $(af)^+ = -af^-$ and $(af)^- = -af^+$. These claims follow immediately from the above definitions of positive and negative parts of a function. Thus if $a < 0$ and $f \in L^1(\Omega)$, it follows from Theorem 9.6.1 that

$$\begin{aligned}\int af d\mu &\equiv \int (af)^+ d\mu - \int (af)^- d\mu = \int (-a) f^- d\mu - \int (-a) f^+ d\mu \\ &= -a \int f^- d\mu + a \int f^+ d\mu = a\left(\int f^+ d\mu - \int f^- d\mu\right) \equiv a \int f d\mu.\end{aligned}$$

The case where $a \geq 0$ works out similarly but easier. ■

Now here is the main result.

Theorem 9.7.6 *$\int d\mu$ is linear on $L^1(\Omega)$ and $L^1(\Omega)$ is a complex vector space. If $f \in L^1(\Omega)$, then $\operatorname{Re} f, \operatorname{Im} f$, and $|f|$ are all in $L^1(\Omega)$. Furthermore, for $f \in L^1(\Omega)$,*

$$\begin{aligned}\int f d\mu &\equiv \int (\operatorname{Re} f)^+ d\mu - \int (\operatorname{Re} f)^- d\mu + i\left[\int (\operatorname{Im} f)^+ d\mu - \int (\operatorname{Im} f)^- d\mu\right] \\ &\equiv \int \operatorname{Re} f d\mu + i \int \operatorname{Im} f d\mu\end{aligned}$$

and the triangle inequality holds,

$$\left|\int f d\mu\right| \leq \int |f| \, d\mu. \tag{9.4}$$

Also, for every $f \in L^1(\Omega)$ it follows that for every $\varepsilon > 0$ there exists a simple function s such that $|s| \leq |f|$ and $\int |f - s| \, d\mu < \varepsilon$.

Proof: First consider the claim that the integral is linear. It was shown above that the integral is linear on $\operatorname{Re}\left(L^1(\Omega)\right)$. Then letting $a+ib, c+id$ be scalars and f, g functions in $L^1(\Omega)$,

$$(a+ib)f + (c+id)g = (a+ib)(\operatorname{Re} f + i \operatorname{Im} f) + (c+id)(\operatorname{Re} g + i \operatorname{Im} g)$$

$$= c\operatorname{Re}(g) - b\operatorname{Im}(f) - d\operatorname{Im}(g) + a\operatorname{Re}(f) + i(b\operatorname{Re}(f) + c\operatorname{Im}(g) + a\operatorname{Im}(f) + d\operatorname{Re}(g))$$

It follows from the definition that

$$\int (a+ib)f + (c+id)g d\mu = \int (c\operatorname{Re}(g) - b\operatorname{Im}(f) - d\operatorname{Im}(g) + a\operatorname{Re}(f)) \, d\mu$$

$$+ i\int (b\operatorname{Re}(f) + c\operatorname{Im}(g) + a\operatorname{Im}(f) + d\operatorname{Re}(g)) \tag{9.5}$$

Also, from the definition,

$$\begin{aligned}(a+ib)\int f d\mu + (c+id)\int g d\mu &= (a+ib)\left(\int \operatorname{Re} f d\mu + i\int \operatorname{Im} f d\mu\right) \\ &\quad + (c+id)\left(\int \operatorname{Re} g d\mu + i\int \operatorname{Im} g d\mu\right)\end{aligned}$$

which equals

$$\begin{aligned} = \; & a\int \operatorname{Re} f d\mu - b\int \operatorname{Im} f d\mu + ib\int \operatorname{Re} f d\mu + ia\int \operatorname{Im} f d\mu \\ & + c\int \operatorname{Re} g d\mu - d\int \operatorname{Im} g d\mu + id\int \operatorname{Re} g d\mu - d\int \operatorname{Im} g d\mu.\end{aligned}$$

Using Lemma 9.7.5 and collecting terms, it follows that this reduces to 9.5. Thus the integral is linear as claimed.

Consider the claim about approximation with a simple function. Letting h equal any of

$$(\operatorname{Re} f)^+, (\operatorname{Re} f)^-, (\operatorname{Im} f)^+, (\operatorname{Im} f)^-, \tag{9.6}$$

It follows from the monotone convergence theorem and Theorem 8.1.6 on Page 181 there exists a nonnegative simple function $s \leq h$ such that $\int |h - s| \, d\mu < \frac{\varepsilon}{4}$. Therefore, letting s_1, s_2, s_3, s_4 be such simple functions, approximating respectively the functions listed in 9.6, and $s \equiv s_1 - s_2 + i(s_3 - s_4)$,

$$\int |f - s| \, d\mu \leq \int \left|(\operatorname{Re} f)^+ - s_1\right| d\mu + \int \left|(\operatorname{Re} f)^- - s_2\right| d\mu$$

$$+ \int \left|(\operatorname{Im} f)^+ - s_3\right| d\mu + \int \left|(\operatorname{Im} f)^- - s_4\right| d\mu < \varepsilon$$

It is clear from the construction that $|s| \leq |f|$.

What about 9.4? Let $\theta \in \mathbb{C}$ be such that $|\theta| = 1$ and $\theta \int f d\mu = |\int f d\mu|$. Then from what was shown above about the integral being linear,

$$\left|\int f d\mu\right| = \theta \int f d\mu = \int \theta f d\mu = \int \operatorname{Re}(\theta f) d\mu \leq \int |f| d\mu.$$

If $f, g \in L^1(\Omega)$, then it is known that for a, b scalars, it follows that $af + bg$ is measurable. See Lemma 9.7.2. Also $\int |af + bg| d\mu \leq \int |a||f| + |b||g| d\mu < \infty$. ■

The following corollary follows from this. The conditions of this corollary are sometimes taken as a definition of what it means for a function f to be in $L^1(\Omega)$.

Corollary 9.7.7 *$f \in L^1(\Omega)$ if and only if there exists a sequence of complex simple functions, $\{s_n\}$ such that*

$$\begin{array}{l} s_n(\omega) \to f(\omega) \text{ for all } \omega \in \Omega \\ \lim_{m,n\to\infty} \int (|s_n - s_m|) d\mu = 0 \end{array} \tag{9.7}$$

When $f \in L^1(\Omega)$,

$$\int f d\mu \equiv \lim_{n\to\infty} \int s_n. \tag{9.8}$$

Proof: From the above theorem, if $f \in L^1$ there exists a sequence of simple functions $\{s_n\}$ such that

$$\int |f - s_n| d\mu < 1/n,\ s_n(\omega) \to f(\omega) \text{ for all } \omega$$

Then $\int |s_n - s_m| d\mu \leq \int |s_n - f| d\mu + \int |f - s_m| d\mu \leq \frac{1}{n} + \frac{1}{m}$.

Next suppose the existence of the approximating sequence of simple functions. Then f is measurable because its real and imaginary parts are the limit of measurable functions. By Fatou's lemma, $\int |f| d\mu \leq \liminf_{n\to\infty} \int |s_n| d\mu < \infty$ because $|\int |s_n| d\mu - \int |s_m| d\mu| \leq \int |s_n - s_m| d\mu$ which is given to converge to 0. Thus $\{\int |s_n| d\mu\}$ is a Cauchy sequence and is therefore, bounded.

In case $f \in L^1(\Omega)$, letting $\{s_n\}$ be the approximating sequence, Fatou's lemma implies

$$\left|\int f d\mu - \int s_n d\mu\right| \leq \int |f - s_n| d\mu \leq \lim \inf_{m\to\infty} \int |s_m - s_n| d\mu < \varepsilon$$

provided n is large enough. Hence 9.8 follows. ■

This is a good time to observe the following fundamental observation which follows from a repeat of the above arguments.

Theorem 9.7.8 *Suppose $\Lambda(f) \in [0, \infty]$ for all nonnegative measurable functions and suppose that for $a, b \geq 0$ and f, g nonnegative measurable functions,*

$$\Lambda(af + bg) = a\Lambda(f) + b\Lambda(g).$$

In other words, Λ wants to be linear. Then Λ has a unique linear extension to the set of measurable functions $\{f \text{ measurable } : \Lambda(|f|) < \infty\}$, this set being a vector space.

9.8 The Dominated Convergence Theorem

One of the major theorems in this theory is the dominated convergence theorem. Before presenting it, here is a technical lemma about $\limsup$ and $\liminf$ which is really pretty obvious from the definition.

Lemma 9.8.1 *Let $\{a_n\}$ be a sequence in $[-\infty,\infty]$. Then $\lim_{n\to\infty} a_n$ exists if and only if $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$ and in this case, the limit equals the common value of these two numbers.*

Proof: Suppose first $\lim_{n\to\infty} a_n = a \in \mathbb{R}$. Letting $\varepsilon > 0$ be given, $a_n \in (a-\varepsilon, a+\varepsilon)$ for all n large enough, say $n \geq N$. Therefore, both $\inf\{a_k : k \geq n\}$ and $\sup\{a_k : k \geq n\}$ are contained in $[a-\varepsilon, a+\varepsilon]$ whenever $n \geq N$. It follows $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$ are both in $[a-\varepsilon, a+\varepsilon]$, showing $|\liminf_{n\to\infty} a_n - \limsup_{n\to\infty} a_n| < 2\varepsilon$. Since ε is arbitrary, the two must be equal and they both must equal a. Next suppose $\lim_{n\to\infty} a_n = \infty$. Then if $l \in \mathbb{R}$, there exists N such that for $n \geq N, l \leq a_n$ and therefore, for such $n, l \leq \inf\{a_k : k \geq n\} \leq \sup\{a_k : k \geq n\}$ and this shows, since l is arbitrary that $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \infty$. The case for $-\infty$ is similar.

Conversely, suppose $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = a$. Suppose first that $a \in \mathbb{R}$. Then, letting $\varepsilon > 0$ be given, there exists N such that if $n \geq N, \sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$. Therefore, if $k, m > N$, and $a_k > a_m$,

$$|a_k - a_m| = a_k - a_m \leq \sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$$

showing that $\{a_n\}$ is a Cauchy sequence. Therefore, it converges to $a \in \mathbb{R}$, and as in the first part, the $\liminf$ and $\limsup$ both equal a. If $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \infty$, then given $l \in \mathbb{R}$, there exists N such that for $n \geq N$, $\inf_{n>N} a_n > l$.Therefore, $\lim_{n\to\infty} a_n = \infty$. The case for $-\infty$ is similar. ■

Here is the dominated convergence theorem.

Theorem 9.8.2 *(Dominated Convergence theorem) Let $f_n \in L^1(\Omega)$ and suppose that $f(\omega) = \lim_{n\to\infty} f_n(\omega)$, and there exists a measurable function g, with values in $[0,\infty]$,[1] such that $|f_n(\omega)| \leq g(\omega)$ and $\int g(\omega)d\mu < \infty$.Then $f \in L^1(\Omega)$ and $0 = \lim_{n\to\infty} \int |f_n - f| d\mu = \lim_{n\to\infty} |\int f d\mu - \int f_n d\mu|$.*

Proof: f is measurable by Theorem 8.1.2. Since $|f| \leq g$, it follows that

$$f \in L^1(\Omega) \text{ and } |f - f_n| \leq 2g.$$

By Fatou's lemma (Theorem 9.5.1),

$$\int 2g d\mu \leq \liminf_{n\to\infty} \int 2g - |f - f_n| d\mu = \int 2g d\mu - \limsup_{n\to\infty} \int |f - f_n| d\mu.$$

Subtracting $\int 2g d\mu$, $0 \leq -\limsup_{n\to\infty} \int |f - f_n| d\mu$. Hence

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty} \left(\int |f - f_n| d\mu \right) \\ &\geq \liminf_{n\to\infty} \left(\int |f - f_n| d\mu \right) \geq \liminf_{n\to\infty} \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \end{aligned}$$

This proves the theorem by Lemma 9.8.1 because the $\limsup$ and $\liminf$ are equal. ■

[1] Note that, since g is allowed to have the value ∞, it is not known that $g \in L^1(\Omega)$.

Corollary 9.8.3 *Suppose $f_n \in L^1(\Omega)$ and $f(\omega) = \lim_{n\to\infty} f_n(\omega)$. Suppose also there exist measurable functions, g_n, g with values in $[0,\infty]$ such that $\lim_{n\to\infty}\int g_n d\mu = \int g d\mu$, $g_n(\omega) \to g(\omega)$ μ a.e. and both $\int g_n d\mu$ and $\int g d\mu$ are finite. Also suppose $|f_n(\omega)| \leq g_n(\omega)$. Then $\lim_{n\to\infty}\int |f - f_n| d\mu = 0$.*

Proof: It is just like the above. This time $g + g_n - |f - f_n| \geq 0$ and so by Fatou's lemma,

$$\int 2g d\mu - \limsup_{n\to\infty}\int |f - f_n| d\mu = \lim_{n\to\infty}\int (g_n + g) d\mu - \limsup_{n\to\infty}\int |f - f_n| d\mu$$

$$\begin{aligned} &= \liminf_{n\to\infty}\int (g_n + g) d\mu - \limsup_{n\to\infty}\int |f - f_n| d\mu \\ &= \liminf_{n\to\infty}\int ((g_n + g) - |f - f_n|) d\mu \geq \int 2g d\mu \end{aligned}$$

and so $-\limsup_{n\to\infty}\int |f - f_n| d\mu \geq 0$. Thus

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty}\left(\int |f - f_n| d\mu\right) \\ &\geq \liminf_{n\to\infty}\left(\int |f - f_n| d\mu\right) \geq \left|\int f d\mu - \int f_n d\mu\right| \geq 0. \blacksquare \end{aligned}$$

Definition 9.8.4 *Let E be a measurable subset of Ω. $\int_E f d\mu \equiv \int f \mathscr{X}_E d\mu$.*

If $L^1(E)$ is written, the σ algebra is defined as $\{E \cap A : A \in \mathscr{F}\}$ and the measure is μ restricted to this smaller σ algebra. Clearly, if $f \in L^1(\Omega)$, then $f\mathscr{X}_E \in L^1(E)$ and if $f \in L^1(E)$, then letting $\tilde{f}$ be the 0 extension of f off of E, it follows $\tilde{f} \in L^1(\Omega)$.

Another very important observation applies to the case where Ω is also a metric space. In this lemma, $\operatorname{spt}(f)$ denotes the closure of the set on which f is nonzero.

Definition 9.8.5 *Let K be a set and let V be an open set containing K. Then the notation $K \prec f \prec V$ means that $f(x) = 1$ for all $x \in K$ and $\operatorname{spt}(f)$ is a compact subset of V. $\operatorname{spt}(f)$ is defined as the closure of the set where f is not zero. It is called the "support" of f. A function $f \in C_c(\Omega)$ for Ω a metric space if f is continuous on Ω and $\operatorname{spt}(f)$ is compact. This $C_c(\Omega)$ is called the continuous functions with compact support.*

Now that the Lebesgue integral has been presented, it is time to show the way that the measure of Theorem 8.8.2 represents the functional.

Proposition 9.8.6 *Let L be a positive linear functional on $C_c(X)$ for X a metric space and let μ be the measure described by Theorem 8.8.2. Then for all $f \in C_c(X)$, $L(f) = \int_X f d\mu$ where this is the Lebesgue integral just described.*

Proof: Let $f \in C_c(X)$, f real-valued, and suppose $f(X) \subseteq [a,b]$. Choose $t_0 < a$ and let $t_0 < t_1 < \cdots < t_n = b$, $t_i - t_{i-1} < \varepsilon$. Let $E_i = f^{-1}((t_{i-1}, t_i]) \cap \operatorname{spt}(f)$. Note that $\cup_{i=1}^n E_i$ is a closed set equal to $\operatorname{spt}(f)$. $\cup_{i=1}^n E_i = \operatorname{spt}(f)$. Since $X = \cup_{i=1}^n f^{-1}((t_{i-1}, t_i])$. Let $V_i \supseteq E_i, V_i$ is open and let V_i satisfy

$$f(x) < t_i + \varepsilon \text{ for all } x \in V_i,\ \mu(V_i \setminus E_i) < \varepsilon/n. \tag{9.9}$$

By Theorem 3.12.5, there exists $h_i \in C_c(X)$ such that

$$h_i \prec V_i,\ \sum_{i=1}^{n} h_i(x) = 1 \text{ on } \operatorname{spt}(f).$$

Now note that for each i, $f(x)h_i(x) \leq h_i(x)(t_i+\varepsilon)$. Therefore,

$$\begin{aligned} Lf &= L(\sum_{i=1}^{n} fh_i) \leq L(\sum_{i=1}^{n} h_i(t_i+\varepsilon)) = \sum_{i=1}^{n}(t_i+\varepsilon)L(h_i) \\ &= \sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)L(h_i) - |t_0|L\left(\sum_{i=1}^{n} h_i\right). \end{aligned}$$

Now note that $|t_0|+t_i+\varepsilon \geq 0$ and so from the definition of μ and **Claim 2** of the proof of Theorem 8.8.2, this is no larger than

$$\sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)\mu(V_i) - |t_0|\mu(\operatorname{spt}(f)) \leq \sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)(\mu(E_i)+\varepsilon/n) - |t_0|\mu(\operatorname{spt}(f))$$

$$\leq |t_0|\overbrace{\sum_{i=1}^{n}\mu(E_i)}^{\mu(\operatorname{spt}(f))} + \frac{\varepsilon}{n}n|t_0| + \sum_i t_i\mu(E_i) + \sum_i t_i\frac{\varepsilon}{n} + \sum_i \varepsilon\mu(E_i) + \frac{\varepsilon^2}{n} - |t_0|\mu(\operatorname{spt}(f))$$

$$\begin{aligned} &\leq \varepsilon|t_0| + \varepsilon(|t_0|+|b|) + \varepsilon\mu(\operatorname{spt}(f)) + \varepsilon^2 + \sum_i t_i\mu(E_i) \\ &\leq \varepsilon|t_0| + \varepsilon(|t_0|+|b|) + 2\varepsilon\mu(\operatorname{spt}(f)) + \varepsilon^2 + \sum_{i=1}^{n} t_{i-1}\mu(E_i) \end{aligned}$$

$$\leq \varepsilon(2|t_0|+|b|+2\mu(\operatorname{spt}(f))+\varepsilon) + \int f d\mu$$

Since $\varepsilon > 0$ is arbitrary, $Lf \leq \int f d\mu$ for all $f \in C_c(X)$, f real. Hence equality holds because $L(-f) \leq -\int f d\mu$ so $L(f) \geq \int f d\mu$. Thus $Lf = \int f d\mu$ for all $f \in C_c(X)$. Just apply the result for real functions to the real and imaginary parts of f. ■

9.9 Some Important General Theory

9.9.1 Eggoroff's Theorem

You might show that a sequence of measurable real or complex valued functions converges on a measurable set. This is Proposition 8.1.7 above. Eggoroff's theorem says that if the set of points where a sequence of measurable functions converges is all but a set of measure zero, then the sequence almost converges uniformly in a certain sense.

Theorem 9.9.1 *(Egoroff) Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space, $\mu(\Omega) < \infty$ and let f_n, f be complex valued functions such that $\operatorname{Re} f_n, \operatorname{Im} f_n$ are all measurable and also that $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ for all $\omega \notin E$ where $\mu(E) = 0$. Then for every $\varepsilon > 0$, there exists a set, $F \supseteq E$, $\mu(F) < \varepsilon$, such that f_n converges uniformly to f on F^C.*

Proof: First suppose $E=\emptyset$ so that convergence is pointwise everywhere. It follows then that $\operatorname{Re} f$ and $\operatorname{Im} f$ are pointwise limits of measurable functions and are therefore measurable. Let $E_{km}=\{\omega\in\Omega:|f_n(\omega)-f(\omega)|\geq 1/m \text{ for some } n>k\}$. Note that

$$|f_n(\omega)-f(\omega)|=\sqrt{(\operatorname{Re} f_n(\omega)-\operatorname{Re} f(\omega))^2+(\operatorname{Im} f_n(\omega)-\operatorname{Im} f(\omega))^2}$$

and so, $\left[|f_n-f|\geq\frac{1}{m}\right]$ is measurable. Hence E_{km} is measurable because

$$E_{km}=\cup_{n=k+1}^{\infty}\left[|f_n-f|\geq\frac{1}{m}\right].$$

For fixed $m,\cap_{k=1}^{\infty}E_{km}=\emptyset$ because f_n converges to f. Therefore, if $\omega\in\Omega$ there exists k such that if $n>k$, $|f_n(\omega)-f(\omega)|<\frac{1}{m}$ which means $\omega\notin E_{km}$. Note also that $E_{km}\supseteq E_{(k+1)m}$. Since $\mu(E_{1m})<\infty$, Theorem 8.2.4 on Page 183 implies

$$0=\mu(\cap_{k=1}^{\infty}E_{km})=\lim_{k\to\infty}\mu(E_{km}).$$

Let $k(m)$ be chosen such that $\mu(E_{k(m)m})<\varepsilon 2^{-m}$ and let $F=\cup_{m=1}^{\infty}E_{k(m)m}$. Then $\mu(F)<\varepsilon$ because $\mu(F)\leq\sum_{m=1}^{\infty}\mu\left(E_{k(m)m}\right)<\sum_{m=1}^{\infty}\varepsilon 2^{-m}=\varepsilon$.

Now let $\eta>0$ be given and pick m_0 such that $m_0^{-1}<\eta$. If $\omega\in F^C$, then $\omega\in\bigcap_{m=1}^{\infty}E_{k(m)m}^C$. Hence $\omega\in E_{k(m_0)m_0}^C$ so $|f_n(\omega)-f(\omega)|<1/m_0<\eta$ for all $n>k(m_0)$. This holds for all $\omega\in F^C$and so f_n converges uniformly to f on F^C.

Now if $E\neq\emptyset$, consider $\{\mathscr{X}_{E^C}f_n\}_{n=1}^{\infty}$. Each $\mathscr{X}_{E^C}f_n$ has real and imaginary parts measurable and the sequence converges pointwise to $\mathscr{X}_E f$ everywhere. Therefore, from the first part, there exists a set of measure less than ε,F such that on $F^C,\{\mathscr{X}_{E^C}f_n\}$ converges uniformly to $\mathscr{X}_{E^C}f$. Therefore, on $(E\cup F)^C$, $\{f_n\}$ converges uniformly to f. This proves the theorem. ■

9.9.2 The Vitali Convergence Theorem

The Vitali convergence theorem is a convergence theorem which in the case of a finite measure space is superior to the dominated convergence theorem.

Definition 9.9.2 *Let $(\Omega,\mathscr{F},\mu)$ be a measure space and let $\mathfrak{S}\subseteq L^1(\Omega)$. $\mathfrak{S}$ is uniformly integrable if for every $\varepsilon>0$ there exists $\delta>0$ such that for all $f\in\mathfrak{S}$*

$$|\int_E fd\mu|<\varepsilon \text{ whenever } \mu(E)<\delta.$$

Lemma 9.9.3 *If $\mathfrak{S}$ is uniformly integrable, then $|\mathfrak{S}|\equiv\{|f|:f\in\mathfrak{S}\}$ is uniformly integrable. Also $\mathfrak{S}$ is uniformly integrable if $\mathfrak{S}$ is finite.*

Proof: Let $\varepsilon>0$ be given and suppose $\mathfrak{S}$ is uniformly integrable. First suppose the functions are real valued. Let δ be such that if $\mu(E)<\delta$, then $|\int_E fd\mu|<\frac{\varepsilon}{2}$ for all $f\in\mathfrak{S}$. Let $\mu(E)<\delta$. Then if $f\in\mathfrak{S}$,

$$\begin{aligned}\int_E|f|d\mu &\leq \int_{E\cap[f\leq 0]}(-f)d\mu+\int_{E\cap[f>0]}fd\mu=\left|\int_{E\cap[f\leq 0]}fd\mu\right|+\left|\int_{E\cap[f>0]}fd\mu\right|\\ &< \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.\end{aligned}$$

In general, if $\mathfrak{S}$ is a uniformly integrable set of complex valued functions, the inequalities,

$$\left|\int_E \operatorname{Re} f d\mu\right| \leq \left|\int_E f d\mu\right|, \left|\int_E \operatorname{Im} f d\mu\right| \leq \left|\int_E f d\mu\right|,$$

imply $\operatorname{Re}\mathfrak{S} \equiv \{\operatorname{Re} f : f \in \mathfrak{S}\}$ and $\operatorname{Im}\mathfrak{S} \equiv \{\operatorname{Im} f : f \in \mathfrak{S}\}$ are also uniformly integrable. Therefore, applying the above result for real valued functions to these sets of functions, it follows $|\mathfrak{S}|$ is uniformly integrable also.

For the last part, is suffices to verify a single function in $L^1(\Omega)$ is uniformly integrable. To do so, note that from the dominated convergence theorem, $\lim_{R\to\infty}\int_{[|f|>R]}|f|\,d\mu = 0$. Let $\varepsilon > 0$ be given and choose R large enough that $\int_{[|f|>R]}|f|\,d\mu < \frac{\varepsilon}{2}$. Now let $\mu(E) < \frac{\varepsilon}{2R}$. Then

$$\begin{aligned}\int_E |f|\,d\mu &= \int_{E\cap[|f|\leq R]}|f|\,d\mu + \int_{E\cap[|f|>R]}|f|\,d\mu \\ &< R\mu(E) + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

This proves the lemma. ■

The following gives a nice way to identify a uniformly integrable set of functions.

Lemma 9.9.4 *Let $\mathfrak{S}$ be a subset of $L^1(\Omega,\mu)$ where $\mu(\Omega) < \infty$. Let $t \to h(t)$ be a continuous function which satisfies $\lim_{t\to\infty}\frac{h(t)}{t} = \infty$. Then $\mathfrak{S}$ is uniformly integrable and bounded in $L^1(\Omega)$ if $\sup\{\int_\Omega h(|f|)\,d\mu : f \in \mathfrak{S}\} = N < \infty$.*

Proof: First I show $\mathfrak{S}$ is bounded in $L^1(\Omega;\mu)$ which means there exists a constant M such that for all $f \in \mathfrak{S}$, $\int_\Omega |f|\,d\mu \leq M$. From the properties of h, there exists R_n such that if $t \geq R_n$, then $h(t) \geq nt$. Therefore, $\int_\Omega |f|\,d\mu = \int_{[|f|\geq R_n]}|f|\,d\mu + \int_{[|f|<R_n]}|f|\,d\mu$. Letting $n = 1$, and $f \in \mathfrak{S}$,

$$\begin{aligned}\int_\Omega |f|\,d\mu &= \int_{[|f|\geq R_1]}|f|\,d\mu + \int_{[|f|<R_1]}|f|\,d\mu \\ &\leq \int_{[|f|\geq R_1]} h(|f|)\,d\mu + R_1\mu([|f| < R_1]) \leq N + R_1\mu(\Omega) \equiv M. \quad (9.10)\end{aligned}$$

Next let E be a measurable set. Then for every $f \in \mathfrak{S}$, it follows from 9.10

$$\int_E |f|\,d\mu = \int_{[|f|\geq R_n]\cap E}|f|\,d\mu + \int_{[|f|<R_n]\cap E}|f|\,d\mu$$

$$\leq \frac{1}{n}\int_\Omega |f|\,d\mu + R_n\mu(E) \leq \frac{M}{n} + R_n\mu(E) \quad (9.11)$$

Let n be large enough that $M/n < \varepsilon/2$ and then let $\mu(E) < \varepsilon/2R_n$. Then 9.11 is less than $\varepsilon/2 + R_n(\varepsilon/2R_n) = \varepsilon$ ■

Letting $h(t) = t^2$, it follows that if all the functions in $\mathfrak{S}$ are bounded, then the collection of functions is uniformly integrable. Another way to discuss uniform integrability is the following. This other way involving equi-integrability is used a lot in probability.

Definition 9.9.5 *Let $(\Omega,\mathscr{F},\mu)$ be a measure space with $\mu(\Omega) < \infty$. A set $\mathfrak{S} \subseteq L^1(\Omega)$ is said to be equi-integrable if for every $\varepsilon > 0$ there exists $\lambda > 0$ sufficiently large, such that $\int_{[|f|>\lambda]}|f|\,d\mu < \varepsilon$ for all $f \in \mathfrak{S}$.*

Then the relation between this and uniform integrability is as follows.

Proposition 9.9.6 *In the context of the above definition, $\mathfrak{S}$ is equi-integrable if and only if it is a bounded subset of $L^1(\Omega)$ which is also uniformly integrable.*

Proof: $\Rightarrow$ I need to show $\mathfrak{S}$ is bounded and uniformly integrable. First consider bounded. Choose λ to work for $\varepsilon = 1$. Then for all $f \in \mathfrak{S}$,

$$\int |f| \, d\mu = \int_{[|f|>\lambda]} |f| \, d\mu + \int_{[|f|\leq\lambda]} |f| \leq 1 + \lambda\mu(\Omega)$$

Thus it is bounded. Now let E be a measurable subset of Ω. Let λ go with $\varepsilon/2$ in the definition of equi-integrable. Then for all $f \in \mathfrak{S}$,

$$\int_E |f| \, d\mu \leq \int_{[|f|>\lambda]} |f| \, d\mu + \int_{E\cap[|f|\leq\lambda]} |f| \, d\mu \leq \frac{\varepsilon}{2} + \lambda\mu(E)$$

Then let $\mu(E)$ be small enough that $\lambda\mu(E) < \varepsilon/2$ and this shows uniform integrability.

$\Leftarrow$ I need to verify equi-integrable from bounded and uniformly integrable. Let δ be such that if $\mu(E) < \delta$, then $\int_E |f| \, d\mu < \varepsilon$ for all $f \in \mathfrak{S}$. If not, then there exists $f_n \in \mathfrak{S}$ with $[|f_n| > n] > \delta$. Thus $\int |f_n| \, d\mu \geq \int_{[|f_n|>n]} |f_n| \, d\mu \geq n\mu([|f_n| > n]) > n\delta$ and so $\mathfrak{S}$ is not bounded after all. ■

The following theorem is Vitali's convergence theorem.

Theorem 9.9.7 *Let $\{f_n\}$ be a uniformly integrable set of complex valued functions, $\mu(\Omega) < \infty$, and $f_n(x) \to f(x)$ a.e. where f is a measurable complex valued function. Then $f \in L^1(\Omega)$ and $\lim_{n\to\infty} \int_\Omega |f_n - f| \, d\mu = 0$.*

Proof: First it will be shown that $f \in L^1(\Omega)$. By uniform integrability, there exists $\delta > 0$ such that if $\mu(E) < \delta$, then $\int_E |f_n| \, d\mu < 1$ for all n. By Egoroff's theorem, there exists a set E of measure less than δ such that on E^C, $\{f_n\}$ converges uniformly. Therefore, for p large enough, and $n > p$, $\int_{E^C} |f_p - f_n| \, d\mu < 1$ which implies $\int_{E^C} |f_n| \, d\mu < 1 + \int_\Omega |f_p| \, d\mu$.Then since there are only finitely many functions, f_n with $n \leq p$, there exists a constant, M_1 such that for all n, $\int_{E^C} |f_n| \, d\mu < M_1$. But also,

$$\int_\Omega |f_m| \, d\mu = \int_{E^C} |f_m| \, d\mu + \int_E |f_m| \leq M_1 + 1 \equiv M.$$

Therefore, by Fatou's lemma, $\int_\Omega |f| \, d\mu \leq \liminf_{n\to\infty} \int |f_n| \, d\mu \leq M$, showing that $f \in L^1$ as hoped.

Now $\mathfrak{S} \cup \{f\}$ is uniformly integrable so there exists $\delta_1 > 0$ such that if $\mu(E) < \delta_1$, then $\int_E |g| \, d\mu < \varepsilon/3$ for all $g \in \mathfrak{S} \cup \{f\}$.

By Egoroff's theorem, there exists a set, F with $\mu(F) < \delta_1$ such that f_n converges uniformly to f on F^C. Therefore, there exists m such that if $n > m$, then $\int_{F^C} |f - f_n| \, d\mu < \frac{\varepsilon}{3}$. It follows that for $n > m$,

$$\int_\Omega |f - f_n| \, d\mu \leq \int_{F^C} |f - f_n| \, d\mu + \int_F |f| \, d\mu + \int_F |f_n| \, d\mu < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,$$

which verifies the claim of the theorem. ■

9.10 The Distribution Function

For $(\Omega, \mathscr{F}, \mu)$ a measure space, the integral of a nonnegative measurable function was defined earlier as $\int f d\mu \equiv \int_0^\infty \mu([f > t]) dt$. This idea will be developed more in this section.

Definition 9.10.1 *Let $f \geq 0$ and suppose f is measurable. The distribution function is the function defined by $t \to \mu([t < f])$.*

Lemma 9.10.2 *If $\{f_n\}$ is an increasing sequence of functions converging pointwise to f then $\mu([f > t]) = \lim_{n\to\infty} \mu([f_n > t])$.*

Proof: The sets, $[f_n > t]$ are increasing and their union is $[f > t]$ because if $f(\omega) > t$, then for all n large enough, $f_n(\omega) > t$ also. Therefore, the desired conclusion follows from properties of measures, the one which says that if $E_n \uparrow E$, then $\mu(E_n) \uparrow \mu(E)$. ■

Note how it was important to have strict inequality in the definition.

Lemma 9.10.3 *Suppose $s \geq 0$ is a simple function, $s(\omega) \equiv \sum_{k=1}^n a_k \mathscr{X}_{E_k}(\omega)$ where the a_k are the distinct nonzero values of $s, 0 < a_1 < a_2 < \cdots < a_n$ on the measurable sets E_k. Suppose ϕ is a C^1 function defined on $[0, \infty)$ which has the properties that $\phi(0) = 0$, and also that $\phi'(t) > 0$ for all t. Then*

$$\int_0^\infty \phi'(t) \mu([s > t]) dm(t) = \int \phi(s) d\mu(s).$$

Proof: First note that if $\mu(E_k) = \infty$ for any k then both sides equal ∞ and so without loss of generality, assume $\mu(E_k) < \infty$ for all k. Letting $a_0 \equiv 0$, the left side equals

$$\sum_{k=1}^n \int_{a_{k-1}}^{a_k} \phi'(t) \mu([s > t]) dm(t) = \sum_{k=1}^n \int_{a_{k-1}}^{a_k} \phi'(t) \sum_{i=k}^n \mu(E_i) dm$$

$$\begin{aligned} &= \sum_{k=1}^n \sum_{i=k}^n \mu(E_i) \int_{a_{k-1}}^{a_k} \phi'(t) dm = \sum_{k=1}^n \sum_{i=k}^n \mu(E_i)(\phi(a_k) - \phi(a_{k-1})) \\ &= \sum_{i=1}^n \mu(E_i) \sum_{k=1}^i (\phi(a_k) - \phi(a_{k-1})) = \sum_{i=1}^n \mu(E_i) \phi(a_i) = \int \phi(s) d\mu. \ \blacksquare \end{aligned}$$

With this lemma the next theorem which is the main result follows easily.

Theorem 9.10.4 *Let $f \geq 0$ be measurable and let ϕ be a C^1 function defined on $[0, \infty)$ which satisfies $\phi'(t) > 0$ for all $t > 0$ and $\phi(0) = 0$. Then*

$$\int \phi(f) d\mu = \int_0^\infty \phi'(t) \mu([f > t]) dm.$$

Proof: By Theorem 8.1.6 on Page 181 there exists an increasing sequence of nonnegative simple functions, $\{s_n\}$ which converges pointwise to f. By the monotone convergence theorem and Lemma 9.10.2,

$$\begin{aligned} \int \phi(f) d\mu &= \lim_{n\to\infty} \int \phi(s_n) d\mu = \lim_{n\to\infty} \int_0^\infty \phi'(t) \mu([s_n > t]) dm \\ &= \int_0^\infty \phi'(t) \mu([f > t]) dm \ \blacksquare \end{aligned}$$

9.11 Radon Nikodym Theorem

Let μ,ν be two finite measures on the measurable space $(\Omega,\mathscr{F})$ and let $\alpha \geq 0$. Let $\lambda \equiv \nu - \alpha\mu$. Then it is clear that if $\{E_i\}_{i=1}^{\infty}$ are disjoint sets of $\mathscr{F}$, then $\lambda(\cup_i E_i) = \sum_{i=1}^{\infty} \lambda(E_i)$ and that the series converges. The next proposition is fairly obvious.

Proposition 9.11.1 *Let $(\Omega,\mathscr{F},\lambda)$ be a measure space and let $\lambda : \mathscr{F} \to [0,\infty)$ be a measure. Then λ is a finite measure.*

Proof: Since $\lambda(\Omega) < \infty$ this is a finite measure. ■

Definition 9.11.2 *Let $(\Omega,\mathscr{F})$ be a measurable space and let $\lambda : \mathscr{F} \to \mathbb{R}$ satisfy: If $\{E_i\}_{i=1}^{\infty}$ are disjoint sets of $\mathscr{F}$, then $\lambda(\cup_i E_i) = \sum_{i=1}^{\infty} \lambda(E_i)$ and the series converges. Such a real valued function is called a signed measure. In this context, a set $E \in \mathscr{F}$ is called positive if whenever F is a measurable subset of E, it follows $\lambda(F) \geq 0$. A negative set is defined similarly. Note that this requires $\lambda(\Omega) \in \mathbb{R}$.*

Lemma 9.11.3 *The countable union of disjoint positive sets is positive.*

Proof: Let E_i be positive and consider $E \equiv \cup_{i=1}^{\infty} E_i$. If $A \subseteq E$ with A measurable, then $A \cap E_i \subseteq E_i$ and so $\lambda(A \cap E_i) \geq 0$. Hence $\lambda(A) = \sum_i \lambda(A \cap E_i) \geq 0$. ■

Lemma 9.11.4 *Let λ be a signed measure on $(\Omega,\mathscr{F})$. If $E \in \mathscr{F}$ with $0 < \lambda(E)$, then E has a measurable subset which is positive.*

Proof: If every measurable subset F of E has $\lambda(F) \geq 0$, then E is positive and we are done. Otherwise there exists measurable $F \subseteq E$ with $\lambda(F) < 0$. Let the elements of $\mathfrak{F}$ consist of sets of disjoint sets of measurable subsets of E each of which has measure less than 0. Partially order $\mathfrak{F}$ by set inclusion. By the Hausdorff maximal theorem, Theorem 2.8.4, there is a maximal chain $\mathscr{C}$. Then $\cup\mathscr{C}$ is a set consisting of disjoint measurable sets $F \in \mathscr{F}$ such that $\lambda(F) < 0$. Since each set in $\cup\mathscr{C}$ has measure strictly less than 0, it follows that $\cup\mathscr{C}$ is a countable set, $\{F_i\}_{i=1}^{\infty}$. Otherwise, there would exist an infinite subset of $\cup\mathscr{C}$ with each set having measure less than $-\frac{1}{n}$ for some $n \in \mathbb{N}$ so λ would not be real valued. Letting $F = \cup_i F_i$, then $E \setminus F$ has no measurable subsets S for which $\lambda(S) < 0$ since, if it did, $\mathscr{C}$ would not have been maximal. Thus $E \setminus F$ is positive. ■

A major result is the following, called a Hahn decomposition.

Theorem 9.11.5 *Let λ be a signed measure on a measurable space $(\Omega,\mathscr{F})$. Then there are disjoint measurable sets P,N such that P is a positive set, N is a negative set, and $P \cup N = \Omega$.*

Proof: If Ω is either positive or negative, there is nothing to show, so suppose Ω is neither positive nor negative. $\mathfrak{F}$ will consist of collections of disjoint measurable sets F such that $\lambda(F) > 0$. Thus each element of $\mathfrak{F}$ is necessarily countable. Partially order $\mathfrak{F}$ by set inclusion and use the Hausdorff maximal theorem to get $\mathscr{C}$ a maximal chain. Then, as in the above lemma, $\cup\mathscr{C}$ is countable, say $\{P_i\}_{i=1}^{\infty}$ because $\lambda(F) > 0$ for each $F \in \cup\mathscr{C}$ and λ has values in $\mathbb{R}$. The sets in $\cup\mathscr{C}$ are disjoint because if A,B are two of them, then they are both in a single element of $\mathscr{C}$. Letting $P \equiv \cup_i P_i$, and $N = P^C$, it follows from Lemma 9.11.3 that P is positive. It is also the case that N must be negative because otherwise, $\mathscr{C}$ would not be maximal. ■

Clearly a Hahn decomposition is not unique. For example, you could have obtained a different Hahn decomposition if you had considered disjoint negative sets F for which $\lambda(F) < 0$ in the above argument. I will only use the case where $\nu \ll \mu$ which is to say that ν is absolutely continuous with respect to μ which is defined next.

Definition 9.11.6 *Let μ, ν be finite measures on $(\Omega, \mathscr{F})$. Then $\nu \ll \mu$ means that whenever $\mu(E) = 0$, it follows that $\nu(E) = 0$.*

Let $k \in \mathbb{N}$, $\left\{\alpha_n^k\right\}_{n=0}^{\infty}$ be equally spaced points $\alpha_n^k = 2^{-k}n$. Then $\alpha_{2n}^k = 2^{-k}(2n) = 2^{-(k-1)}n \equiv \alpha_n^{k-1}$ and $\alpha_{2n}^{k+1} \equiv 2^{-(k+1)}2n = \alpha_n^k$. Similarly $N_{2n}^{k+1} = N_n^k$ because these depend on the α_n^k. Also let $\left(P_n^k, N_n^k\right)$ be a Hahn decomposition for the signed measure $\nu - \alpha_n^k\mu$ where ν, μ are two finite measures. Now from the definition, $N_{n+1}^k \setminus N_n^k = N_{n+1}^k \cap P_n^k$. Also, $N_n \subseteq N_{n+1}$ for each n and we can take $N_0 = \emptyset$ because $\nu(N_0) \leq 0$ Then $\left\{N_{n+1}^k \setminus N_n^k\right\}_{n=0}^{\infty}$ covers all of Ω except for a set of ν measure 0.

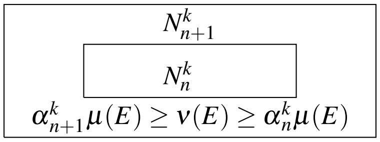
Lemma 9.11.7 *Let $S \equiv \Omega \setminus \left(\cup_n N_n^k\right) = \Omega \setminus \left(\cup_n N_n^l\right)$ for any l. Then $\mu(S) = 0$.*

Proof: $S = \cap_n P_n^k$ so for all $n, \nu(S) - \alpha_n^k\mu(S) \geq 0$. But letting $n \to \infty$, it must be that $\mu(S) = 0$. ■

By the assumption that $\nu \ll \mu$, we can neglect S because this also implies $\nu(S) = 0$. Thus, asside from a set of μ and ν measure zero, $\Omega = \cup_n N_n^k$.

As just noted, if $E \subseteq N_{n+1}^k \setminus N_n^k$, then

$$\nu(E) - \alpha_n^k\mu(E) \geq 0 \geq \nu(E) - \alpha_{n+1}^k\mu(E), \text{ so } \alpha_{n+1}^k\mu(E) \geq \nu(E) \geq \alpha_n^k\mu(E) \tag{9.12}$$



Then define $f^k(\omega) \equiv \sum_{n=0}^{\infty} \alpha_n^k \mathscr{X}_{\Delta_n^k}(\omega)$ where $\Delta_m^k \equiv N_{m+1}^k \setminus N_m^k$. Thus,

$$f^k = \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\left(N_{2n+2}^{k+1} \setminus N_{2n}^{k+1}\right)} = \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n+1}^{k+1}} + \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n}^{k+1}}$$

$$\leq \sum_{n=0}^{\infty} \alpha_{2n+1}^{k+1} \mathscr{X}_{\Delta_{2n+1}^{k+1}} + \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n}^{k+1}} = f^{k+1} \tag{9.13}$$

Thus $k \to f^k(\omega)$ is increasing. Let $f(\omega) \equiv \lim_{k\to\infty} f^k(\omega)$. Then using 9.12, if E is a measurable set,

$$\int \mathscr{X}_E f^k d\mu \leq \sum_{n=0}^{\infty} \alpha_{n+1}^k \mu\left(E \cap \Delta_n^k\right) \leq \sum_{n=0}^{\infty} \alpha_n^k \mu\left(E \cap \Delta_n^k\right) + \sum_{n=0}^{\infty} 2^{-k}\mu\left(E \cap \Delta_n^k\right)$$

$$\leq \sum_{n=0}^{\infty} \nu\left(E \cap \Delta_n^k\right) + 2^{-k}\mu(E) = \nu(E) + 2^{-k}\mu(E) \leq \int \mathscr{X}_E f^k d\mu + 2^{-k}\mu(E) \tag{9.14}$$

From the monotone convergence theorem it follows $\nu(E) = \int \mathscr{X}_E f d\mu$.

This proves the following major theorem called the Radon Nikodym theorem.

Theorem 9.11.8 *Let ν and μ be finite measures defined on a measurable space $(\Omega, \mathscr{F})$ where $\nu \ll \mu$. Then there exists a unique up to a set of measure zero nonnegative measurable function $\omega \to f(\omega)$ such that $\nu(E) = \int_E f d\mu$.*

Proof: If you had $\hat{f}$ which also works, then consider the set E_n where $\hat{f}(\omega) > f(\omega) + 1/n$. Then $0 = \int_{E_n} \left(\hat{f}(\omega) - f(\omega)\right) d\mu \geq \frac{1}{n}\mu(E_n)$. Thus $\mu(E_n) = 0$ and so also

$$\left[\hat{f} - f > 0\right] = \cup_n E_n$$

is a set of measure 0. Similarly $\left[f - \hat{f} > 0\right]$ is a set of measure zero and so $f = \hat{f}$ for a.e. ω. ■

Sometimes people write $f = \frac{d\lambda}{d\mu}$ and $\frac{d\lambda}{d\mu}$ is called the Radon Nikodym derivative.

Corollary 9.11.9 *In the above situation, let λ be a signed measure and let $\lambda \ll \mu$ meaning that if $\mu(E) = 0 \Rightarrow \lambda(E) = 0$. Here assume that μ is a finite measure. Then there exists a unique up to a set of measure zero $h \in L^1$ such that $\lambda(E) = \int_E h d\mu$.*

Proof: Let $P \cup N$ be a Hahn decomposition of λ. Let

$$\lambda_+(E) \equiv \lambda(E \cap P), \ \ \lambda_-(E) \equiv -\lambda(E \cap N).$$

Then both λ_+ and λ_- are absolutely continuous measures and so there are nonnegative h_+ and h_- with $\lambda_-(E) = \int_E h_- d\mu$ and a similar equation for λ_+. Then $0 \leq -\lambda(\Omega \cap N) \leq \lambda_-(\Omega) < \infty$, similar for λ_+ so both of these measures are necessarily finite. Hence both h_- and h_+ are in L^1 so $h \equiv h_+ - h_-$ is also in L^1 and $\lambda(E) = \lambda_+(E) - \lambda_-(E) = \int_E (h_+ - h_-) d\mu$. ■

Definition 9.11.10 *A measure space $(\Omega, \mathscr{F}, \mu)$ is σ finite if there are countably many measurable sets $\{\Omega_n\}$ such that μ is finite on measurable subsets of Ω_n.*

There is a routine corollary of the above theorem.

Corollary 9.11.11 *Suppose μ, ν are both σ finite measures defined on $(\Omega, \mathscr{F})$ with $\nu \ll \mu$. Then a similar conclusion to the above theorem can be obtained. $\nu(E) = \int_E f d\mu$ for f a nonnegative measurable function. If $\nu(\Omega) < \infty$, then $f \in L^1(\Omega)$. This f is unique up to a set of μ measure zero.*

Proof: Since both μ, ν are σ finite, there are $\left\{\tilde{\Omega}_k\right\}_{k=1}^{\infty}$ such that $\nu\left(\tilde{\Omega}_k\right), \mu\left(\tilde{\Omega}_k\right)$ are finite. Let $\Omega_0 = \emptyset$ and $\Omega_k \equiv \tilde{\Omega}_k \setminus \left(\cup_{j=0}^{k-1} \tilde{\Omega}_j\right)$ so that μ, ν are finite on Ω_k and the Ω_k are disjoint. Let $\mathscr{F}_k$ be the measurable subsets of Ω_k, equivalently the intersections with Ω_k with sets of $\mathscr{F}$. Now let $\nu_k(E) \equiv \nu(E \cap \Omega_k)$, similar for μ_k. By Theorem 9.11.8, there exists f_k as described there, unique up to sets of μ measure 0. Thus $\nu_k(E) = \int_{E \cap \Omega_k} f_k d\mu_k$.Now let $f(\omega) \equiv f_k(\omega)$ for $\omega \in \Omega_k$. Thus $\nu(E \cap \Omega_k) = \int_{E \cap \Omega_k} f d\mu$. Summing over all $k, \nu(E) = \int_E f d\mu$. ■

9.12 Iterated Integrals

This is about what can be said for the σ algebra of product measurable sets. First it is necessary to define what this means.

Definition 9.12.1 *A measure space $(\Omega,\mathscr{F},\mu)$ is called σ finite if there are measurable subsets Ω_n such that $\mu(\Omega_n)<\infty$ and $\Omega=\cup_{n=1}^{\infty}\Omega_n$.*

Next is a σ algebra which comes from two σ algebras.

Definition 9.12.2 *Let $(X,\mathscr{E}),(Y,\mathscr{F})$ be measurable spaces. That is, a set with a σ algebra of subsets. Then $\mathscr{E}\times\mathscr{F}$ will be the smallest σ algebra which contains the measurable rectangles, sets of the form $E\times F$ where $E\in\mathscr{E}$, $F\in\mathscr{F}$. The sets in this new σ algebra are called product measurable sets.*

Definition 9.12.3 *Given two finite measure spaces, $(X,\mathscr{E},\mu)$ and $(Y,\mathscr{F},\nu)$,one can define a new measure $\mu\times\nu$ defined on $\mathscr{E}\times\mathscr{F}$ by specifying what it does to measurable rectangles as follows:*

$$(\mu\times\nu)(A\times B)=\mu(A)\nu(B)$$

whenever $A\in\mathscr{E}$ and $B\in\mathscr{F}$.

We also have the following important proposition which holds in every context independent of any measure.

Proposition 9.12.4 *Let $E\subseteq\mathscr{E}\times\mathscr{F}$ be product measurable $\mathscr{E}\times\mathscr{F}$ where $\mathscr{E}$ is a σ algebra of sets of X and $\mathscr{F}$ is a σ algebra of sets of Y. then if $E_x\equiv\{y\in Y:(x,y)\in E\}$ and $E_y\equiv\{x\in X:(x,y)\in E\}$, then $E_x\in\mathscr{E}$ and $E_y\in\mathscr{F}$.*

Proof: It is obvious that if $\mathscr{K}$ is the measurable rectangles, then the conclusion of the proposition holds. If $\mathscr{G}$ consists of the sets of $\mathscr{E}\times\mathscr{F}$ for which the proposition holds, then it is clearly closed with respect to countable disjoint unions and complements. This is obvious in the case of a countable disjoint union since $\left(\cup_i E^i\right)_x=\cup_i E^i_x$, similar for y. As to complement, if $E\in\mathscr{G}$, then $E_x\in\mathscr{F}$ and so $\left(E^C\right)_x=(E_x)^C\in\mathscr{F}$. It is similar for y. By Lemma 8.3.2, Dynkin's lemma, $\mathscr{G}\supseteq\mathscr{E}\times\mathscr{F}$. However $\mathscr{G}$ was defined as a subset of $\mathscr{E}\times\mathscr{F}$ so these are equal. ■

Let $(X,\mathscr{E},\mu)$ and $(Y,\mathscr{F},\nu)$ be two finite measure spaces. Define $\mathscr{K}$ to be the set of measurable rectangles, $A\times B$, $A\in\mathscr{E}$ and $B\in\mathscr{F}$. Let

$$\mathscr{G}\equiv\left\{E\subseteq X\times Y:\int_Y\int_X\mathscr{X}_E d\mu d\nu=\int_X\int_Y\mathscr{X}_E d\nu d\mu\right\}\tag{9.15}$$

where in the above, part of the requirement is for all integrals to make sense.

Then $\mathscr{K}\subseteq\mathscr{G}$. This is obvious.

Next I want to show that if $E\in\mathscr{G}$ then $E^C\in\mathscr{G}$. Observe $\mathscr{X}_{E^C}=1-\mathscr{X}_E$ and so

$$\begin{aligned}\int_Y\int_X\mathscr{X}_{E^C}d\mu d\nu &= \int_Y\int_X(1-\mathscr{X}_E)\,d\mu d\nu=\int_X\int_Y(1-\mathscr{X}_E)\,d\nu d\mu\\ &= \int_X\int_Y\mathscr{X}_{E^C}d\nu d\mu\end{aligned}$$

which shows that if $E\in\mathscr{G}$, then $E^C\in\mathscr{G}$.

Next is to show $\mathscr{G}$ is closed under countable unions of disjoint sets of $\mathscr{G}$. Let $\{A_i\}$ be a sequence of disjoint sets from $\mathscr{G}$. Then, using the monotone convergence theorem as needed,

$$\int_Y\int_X\mathscr{X}_{\cup_{i=1}^{\infty}A_i}d\mu d\nu=\int_Y\int_X\sum_{i=1}^{\infty}\mathscr{X}_{A_i}d\mu d\nu=\int_Y\sum_{i=1}^{\infty}\int_X\mathscr{X}_{A_i}d\mu d\nu$$

$$= \sum_{i=1}^{\infty} \int_Y \int_X \mathscr{X}_{A_i} d\mu d\nu = \sum_{i=1}^{\infty} \int_X \int_Y \mathscr{X}_{A_i} d\nu d\mu$$

$$= \int_X \sum_{i=1}^{\infty} \int_Y \mathscr{X}_{A_i} d\nu d\mu = \int_X \int_Y \sum_{i=1}^{\infty} \mathscr{X}_{A_i} d\nu d\mu = \int_X \int_Y \mathscr{X}_{\cup_{i=1}^{\infty} A_i} d\nu d\mu, \tag{9.16}$$

Thus $\mathscr{G}$ is closed with respect to countable disjoint unions.

From Lemma 8.3.2, $\mathscr{G} \supseteq \sigma(\mathscr{K})$, the smallest σ algebra containing $\mathscr{K}$. Also the computation in 9.16 implies that on $\sigma(\mathscr{K})$ one can define a measure, denoted by $\mu \times \nu$ and that for every $E \in \sigma(\mathscr{K})$,

$$(\mu \times \nu)(E) = \int_Y \int_X \mathscr{X}_E d\mu d\nu = \int_X \int_Y \mathscr{X}_E d\nu d\mu. \tag{9.17}$$

with each iterated integral making sense.

Next is product measure. First is the case of finite measures. Then this will extend to σ finite measures. The following theorem is Fubini's theorem.

Theorem 9.12.5 *Let $f : X \times Y \to [0,\infty]$ be measurable with respect to the σ algebra, $\sigma(\mathscr{K}) \equiv \mathscr{E} \times \mathscr{F}$ just defined and let $\mu \times \nu$ be the product measure of 9.17 where μ and ν are finite measures on $(X,\mathscr{E})$ and $(Y,\mathscr{F})$ respectively. Then*

$$\int_{X \times Y} f d(\mu \times \nu) = \int_Y \int_X f d\mu d\nu = \int_X \int_Y f d\nu d\mu.$$

Proof: Let $\{s_n\}$ be an increasing sequence of $\sigma(\mathscr{K}) \equiv \mathscr{E} \times \mathscr{F}$ measurable simple functions which converges pointwise to f. The above equation holds for s_n in place of f from what was shown above. The final result follows from passing to the limit and using the monotone convergence theorem. ■

Of course one can generalize right away to measures which are only σ finite. This is also called Fubini's theorem.

Definition 9.12.6 *Let $(X,\mathscr{E},\mu),(Y,\mathscr{F},\nu)$ both be σ finite. Thus there exist **disjoint** measurable X_n with $\cup_{n=1}^{\infty} X_n$ and **disjoint** measurable Y_n with $\cup_{n=1}^{\infty} Y_n = Y$ such that μ,ν restricted to X_n, Y_n respectively are finite measures. Let $\mathscr{E}_n$ be intersections of sets of $\mathscr{E}$ with X_n and $\mathscr{F}_n$ similarly defined. Then letting $\mathscr{K}$ consist of all measurable rectangles $A \times B$ for $A \in \mathscr{E}, B \in \mathscr{F}$, and letting $\mathscr{E} \times \mathscr{F} \equiv \sigma(\mathscr{K})$ define the product measure of E contained in this σ algebra as $(\mu \times \nu)(E) \equiv \sum_n \sum_m (\mu_n \times \nu_m)(E \cap (X_n \times Y_m))$.*

Lemma 9.12.7 *The above definition yields a well defined measure on $\mathscr{E} \times \mathscr{F}$.*

Proof: This follows from the standard theorems about sums of nonnegative numbers. See Theorem 2.5.4. For example if you have two other disjoint sequences X_k, Y_l on which the measures are finite, then

$$\begin{aligned} (\mu \times \nu)(E) &= \sum_n \sum_m \sum_k \sum_l (\mu_n \times \nu_m)(E \cap (X_n \cap X_k \times Y_m \cap Y_l)) \\ &= \sum_k \sum_l \sum_n \sum_m (\mu_k \times \nu_l)(E \cap (X_n \cap X_k \times Y_m \cap Y_l)) \end{aligned}$$

and so the definition with respect to the two different increasing sequences gives the same thing. Thus the definition is well defined. $(\mu \times \nu)$ is a measure because if the E_i are disjoint $\mathscr{E} \times \mathscr{F}$ measurable sets and $E = \cup_i E_i$,

$$(\mu \times \nu)(E) \equiv \sum_n \sum_m (\mu_n \times \nu_m)(\cup_i E_i \cap (X_n \times Y_m)) = \sum_n \sum_m \sum_i (\mu_n \times \nu_m)(E_i \cap (X_n \times Y_m))$$

$$= \sum_i \sum_n \sum_m (\mu_n \times \nu_m)(E_i \cap (X_n \times Y_m)) \equiv \sum_i (\mu \times \nu)(E_i) \ \blacksquare$$

Theorem 9.12.8 *Let $f : X \times Y \to [0,\infty]$ be measurable with respect to the σ algebra, $\sigma(\mathscr{K})$ just defined as the smallest σ algebra containing the measurable rectangles, and let $\mu \times \nu$ be the product measure of 9.17 where μ and ν are σ finite measures on $(X, \mathscr{E})$ and $(Y, \mathscr{F})$ respectively. (9.12.1) Then*

$$\int_{X \times Y} f d(\mu \times \nu) = \int_Y \int_X f d\mu d\nu = \int_X \int_Y f d\nu d\mu. \tag{9.18}$$

Proof: Letting $E \in \mathscr{E} \times \mathscr{F}$,

$$\int_{X \times Y} \mathscr{X}_E d(\mu \times \nu) \equiv (\mu \times \nu)(E) \equiv \sum_n \sum_m (\mu_n \times \nu_m)(E \cap (X_n \times Y_m))$$

$$= \sum_n \sum_m \int_{Y_n} \int_{X_n} \mathscr{X}_E d\mu_n d\nu_n = \int_Y \int_X \mathscr{X}_E d\mu d\nu$$

the last coming from a use of the monotone convergence theorem applied to sums. It follows that 9.18 holds for simple functions and then from monotone convergence theorem and Theorem 8.1.6, it holds for nonnegative $\mathscr{E} \times \mathscr{F}$ measurable functions. $\blacksquare$

It is also useful to note that all the above holds for $\prod_{i=1}^p X_i$ in place of $X \times Y$ and μ_i a measure on $\mathscr{E}_i$ a σ algebra of sets of X_i. You would simply modify the definition of $\mathscr{G}$ in 9.15 including all permutations for the iterated integrals and for $\mathscr{K}$ you would use sets of the form $\prod_{i=1}^p A_i$ where A_i is measurable. Everything goes through exactly as above.

Thus the following is mostly obtained.

Theorem 9.12.9 *Let $\{(X_i, \mathscr{E}_i, \mu_i)\}_{i=1}^p$ be σ finite measure spaces and $\prod_{i=1}^p \mathscr{E}_i$ denotes the smallest σ algebra which contains the measurable boxes of the form $\prod_{i=1}^p A_i$ where $A_i \in \mathscr{E}_i$. Then there exists a measure λ defined on a σ algebra $\prod_{i=1}^p \mathscr{E}_i$ such that if $f : \prod_{i=1}^p X_i \to [0,\infty]$ is $\prod_{i=1}^p \mathscr{E}_i$ measurable, $(i_1, \cdots, i_p)$ is any permutation of $(1, \cdots, p)$, then*

$$\int f d\lambda = \int_{X_{i_p}} \cdots \int_{X_{i_1}} f d\mu_{i_1} \cdots d\mu_{i_p} \tag{9.19}$$

The conclusion 9.19 is called Fubini's theorem. More generally

Theorem 9.12.10 *Suppose, in the situation of Theorem 9.12.9 $f \in L^1$ with respect to the measure λ. Then 9.19 continues to hold.*

Proof: It suffices to prove this for f having real values because if this is shown the general case is obtained by taking real and imaginary parts. Since $f \in L^1\left(\prod_{i=1}^p X_i\right)$, $\int |f| d\lambda < \infty$ and so both $\frac{1}{2}(|f| + f)$ and $\frac{1}{2}(|f| - f)$ are in $L^1\left(\prod_{i=1}^p X_i\right)$ and are each nonnegative. Hence from Theorem 9.12.9,

$$\int f d\lambda = \int \left[\frac{1}{2}(|f| + f) - \frac{1}{2}(|f| - f)\right] d\lambda = \int \frac{1}{2}(|f| + f) d\lambda - \int \frac{1}{2}(|f| - f) d\lambda$$

$$= \int \cdots \int \frac{1}{2}(|f| + f) d\mu_{i_1} \cdots d\mu_{i_p} - \int \cdots \int \frac{1}{2}(|f| - f) d\mu_{i_1} \cdots d\mu_{i_p}$$

$$= \int \cdots \int \frac{1}{2}(|f| + f) - \frac{1}{2}(|f| - f) d\mu_{i_1} \cdots d\mu_{i_p} = \int \cdots \int f d\mu_{i_1} \cdots d\mu_{i_p} \ \blacksquare$$

The following corollary is a convenient way to verify the hypotheses of the above theorem.

Corollary 9.12.11 *Suppose f is measurable with respect to $\prod_{i=1}^{p}\mathscr{E}_i$ and suppose for some permutation, $(i_1,\cdots,i_p)$, $\int\cdots\int|f|\,d\mu_{i_1}\cdots d\mu_{i_p}<\infty$. Then $f\in L^1\left(\prod_{i=1}^{p}X_i\right)$.*

Proof: By Theorem 9.12.9, $\int_{\mathbb{R}^p}|f|\,d\lambda=\int\cdots\int|f(\boldsymbol{x})|\,d\mu_{i_1}\cdots d\mu_{i_p}<\infty$ and so f is in $L^1(\mathbb{R}^p)$. ■

You can of course consider the completion of a product measure by using the outer measure approach described earlier. This could be used to get p dimensional Lebesgue measure.

9.13 The Brouwer Fixed Point Theorem

I found this proof of the Brouwer fixed point theorem in Evans [16] and Dunford and Schwartz [14]. The main idea which makes proofs like this work is Lemma 6.11.2 which is stated next for convenience.

Lemma 9.13.1 *Let $\boldsymbol{g}:U\to\mathbb{R}^p$ be C^2 where U is an open subset of $\mathbb{R}^p$. Then it follows that $\sum_{j=1}^{p}\operatorname{cof}(D\boldsymbol{g})_{ij,j}=0$,where here $(D\boldsymbol{g})_{ij}\equiv g_{i,j}\equiv\frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij}=\frac{\partial\det(D\boldsymbol{g})}{\partial g_{i,j}}$.*

Definition 9.13.2 *Let $\boldsymbol{h}$ be a function defined on an open set, $U\subseteq\mathbb{R}^p$. Then $\boldsymbol{h}\in C^k\left(\overline{U}\right)$ if there exists a function $\boldsymbol{g}$ defined on an open set, W containng $\overline{U}$ such that $\boldsymbol{g}=\boldsymbol{h}$ on U and $\boldsymbol{g}$ is $C^k(W)$.*

Lemma 9.13.3 *There does not exist $\boldsymbol{h}\in C^2\left(\overline{B(\mathbf{0},R)}\right)$ with $\boldsymbol{h}:\overline{B(\mathbf{0},R)}\to\partial B(\mathbf{0},R)$ which has the property that $\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{x}$ for all $\boldsymbol{x}\in\partial B(\mathbf{0},R)\equiv\{\boldsymbol{x}:|\boldsymbol{x}|=R\}$ Such a function is called a retract.*

Proof: First note that if $\boldsymbol{h}$ is such a retract, then for all $\boldsymbol{x}\in B(\mathbf{0},R)$, $\det(D\boldsymbol{h}(\boldsymbol{x}))=0$. This is because if $\det(D\boldsymbol{h}(\boldsymbol{x}))\neq0$ for some such $\boldsymbol{x}$, then by the inverse function theorem, $\boldsymbol{h}(B(\boldsymbol{x},\delta))$ is an open set for small enough δ but this would require that this open set is a subset of $\partial B(\mathbf{0},R)$ which is impossible because no open ball is contained in $\partial B(\mathbf{0},R)$. Here and below, let B_R denote $\overline{B(\mathbf{0},R)}$.

Now suppose such an $\boldsymbol{h}$ exists. Let $\lambda\in[0,1]$ and let $\boldsymbol{p}_\lambda(\boldsymbol{x})\equiv\boldsymbol{x}+\lambda(\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{x})$. This function, $\boldsymbol{p}_\lambda$ is a homotopy of the identity map and the retract $\boldsymbol{h}$. Define the function $I(\lambda)$ by $I(\lambda)\equiv\int_{B(\mathbf{0},R)}\det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))\,dx$. Then using the dominated convergence theorem,

$$\begin{aligned}I'(\lambda)&=\int_{B(\mathbf{0},R)}\sum_{i.j}\frac{\partial\det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))}{\partial p_{\lambda i,j}}\frac{\partial p_{\lambda ij}(\boldsymbol{x})}{\partial\lambda}dx\\&=\int_{B(\mathbf{0},R)}\sum_i\sum_j\frac{\partial\det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))}{\partial p_{\lambda i,j}}(h_i(\boldsymbol{x})-x_i)_{,j}\,dx\end{aligned}$$

$$=\int_{B(\mathbf{0},R)}\sum_i\sum_j\operatorname{cof}(D\boldsymbol{p}_\lambda(\boldsymbol{x}))_{ij}(h_i(\boldsymbol{x})-x_i)_{,j}\,dx$$

Now by assumption, $h_i(\boldsymbol{x})=x_i$ on $\partial B(\mathbf{0},R)$ and so one can integrate by parts, in the iterated integrals used to compute $\int_{B(\mathbf{0},R)}$ and write

$$I'(\lambda)=-\sum_i\int_{B(\mathbf{0},R)}\sum_j\operatorname{cof}(D\boldsymbol{p}_\lambda(\boldsymbol{x}))_{ij,j}(h_i(\boldsymbol{x})-x_i)\,dx=0.$$

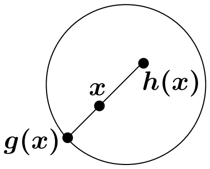
Therefore, $I(\lambda)$ equals a constant. However, $I(0) = m_p(B(\mathbf{0},R)) \neq 0$ and as explained above, $I(1) = 0$. ■

The following is the Brouwer fixed point theorem for C^2 maps.

Lemma 9.13.4 *If $\boldsymbol{h} \in C^2\left(\overline{B(\mathbf{0},R)}\right)$ and $\boldsymbol{h} : \overline{B(\mathbf{0},R)} \to \overline{B(\mathbf{0},R)}$, then $\boldsymbol{h}$ has a fixed point $\boldsymbol{x}$ such that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Suppose the lemma is not true. Then for all $\boldsymbol{x}, |\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})| \neq 0$. Then define $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}) + \frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|} t(\boldsymbol{x})$ where $t(\boldsymbol{x})$ is nonnegative and is chosen such that $\boldsymbol{g}(\boldsymbol{x}) \in \partial B(\mathbf{0},R)$.

This mapping is illustrated in the following picture.


$\boldsymbol{x}$
$\boldsymbol{h}(\boldsymbol{x})$
$\boldsymbol{g}(\boldsymbol{x})$

If $\boldsymbol{x} \to t(\boldsymbol{x})$ is C^2 near $\overline{B(\mathbf{0},R)}$, it will follow $\boldsymbol{g}$ is a C^2 retract onto $\partial B(\mathbf{0},R)$ contrary to Lemma 9.13.3. Thus $t(\boldsymbol{x})$ is the nonnegative solution t to

$$\left|\boldsymbol{h}(\boldsymbol{x}) + \frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|} t(\boldsymbol{x})\right|^2 = |\boldsymbol{h}(\boldsymbol{x})|^2 + 2\left(\boldsymbol{h}(\boldsymbol{x}), \frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right) t + t^2 = R^2 \tag{9.20}$$

then by the quadratic formula,

$$t(\boldsymbol{x}) = -\left(\boldsymbol{h}(\boldsymbol{x}), \frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right) + \sqrt{\left(\boldsymbol{h}(\boldsymbol{x}), \frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right)^2 + \left(R^2 - |\boldsymbol{h}(\boldsymbol{x})|^2\right)}$$

Is $\boldsymbol{x} \to t(\boldsymbol{x})$ a function in C^2? If what is under the radical is positive, then this is so because $s \to \sqrt{s}$ is smooth for $s > 0$. In fact, this is the case here. The inside of the radical is positive if $R > |\boldsymbol{h}(\boldsymbol{x})|$. If $|\boldsymbol{h}(\boldsymbol{x})| = R$, it is still positive because in this case, the angle between $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})$ cannot be $\pi/2$ because of the shape of the ball. This shows that $\boldsymbol{x} \to t(\boldsymbol{x})$ is the composition of C^2 functions and is therefore C^2. Thus this $\boldsymbol{g}(\boldsymbol{x})$ is a C^2 retract and by the above lemma, there isn't one. ■

Now it is easy to prove the Brouwer fixed point theorem. The following theorem is the Brouwer fixed point theorem for a ball.

Theorem 9.13.5 *Let B_R be the above closed ball and let $\boldsymbol{f} : B_R \to B_R$ be continuous. Then there exists $\boldsymbol{x} \in B_R$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Let $\boldsymbol{f}_k(\boldsymbol{x}) \equiv \frac{\boldsymbol{f}(\boldsymbol{x})}{1+k^{-1}}$. Thus

$$\begin{aligned}
\|\boldsymbol{f}_k - \boldsymbol{f}\| &= \max_{\boldsymbol{x} \in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x})}{1+(1/k)} - \boldsymbol{f}(\boldsymbol{x})\right|\right\} = \max_{\boldsymbol{x} \in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})(1+(1/k))}{1+(1/k)}\right|\right\} \\
&= \max_{\boldsymbol{x} \in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x})(1/k)}{1+(1/k)}\right|\right\} \leq \frac{R}{1+k}
\end{aligned}$$

Letting $\|\boldsymbol{h}\| \equiv \max\{|\boldsymbol{h}(\boldsymbol{x})| : \boldsymbol{x} \in B_R\}$, It follows from the Weierstrass approximation theorem, there exists a function whose components are polynomials $\boldsymbol{g}_k$ such that $\|\boldsymbol{g}_k - \boldsymbol{f}_k\| < \frac{R}{k+1}$. Then if $\boldsymbol{x} \in B_R$, it follows

$$|\boldsymbol{g}_k(\boldsymbol{x})| \leq |\boldsymbol{g}_k(\boldsymbol{x}) - \boldsymbol{f}_k(\boldsymbol{x})| + |\boldsymbol{f}_k(\boldsymbol{x})| < \frac{R}{1+k} + \frac{kR}{1+k} = R$$

and so $\boldsymbol{g}_k$ maps B_R to B_R. By Lemma 9.13.4 each of these $\boldsymbol{g}_k$ has a fixed point $\boldsymbol{x}_k$ such that $\boldsymbol{g}_k(\boldsymbol{x}_k) = \boldsymbol{x}_k$. The sequence of points, $\{\boldsymbol{x}_k\}$ is contained in the compact set, B_R and so there exists a convergent subsequence still denoted by $\{\boldsymbol{x}_k\}$ which converges to a point $\boldsymbol{x} \in B_R$. Then from uniform convergence of $\boldsymbol{g}_k$ to $\boldsymbol{f}$, $\boldsymbol{f}(\boldsymbol{x}) = \lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}_k) = \lim_{k\to\infty} \boldsymbol{g}_k(\boldsymbol{x}_k) = \lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$. ■

The ball does not have to be centered at $\mathbf{0}$. If $\boldsymbol{f} : B(\boldsymbol{a}, R) \to B(\boldsymbol{a}, R)$ is continuous, then $\boldsymbol{y} \to \boldsymbol{f}(\boldsymbol{y} + \boldsymbol{a}) - \boldsymbol{a}$ maps $B(\mathbf{0}, R)$ to $B(\mathbf{0}, R)$ has a fixed point $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{y} + \boldsymbol{a}) - \boldsymbol{a}$ so $\boldsymbol{f}(\boldsymbol{y} + \boldsymbol{a}) = \boldsymbol{y} + \boldsymbol{a}$.

Corollary 9.13.6 *A continuous function mapping a closed ball having center at $\boldsymbol{a}$ to itself has a fixed point.*

Definition 9.13.7 *A set A is a retract of a set B if $A \subseteq B$, and there is a continuous map $\boldsymbol{h} : B \to A$ such that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in A$ and $\boldsymbol{h}$ is onto. B has the fixed point property means that whenever $\boldsymbol{g}$ is continuous and $\boldsymbol{g} : B \to B$, it follows that $\boldsymbol{g}$ has a fixed point.*

Proposition 9.13.8 *Let A be a retract of B and suppose B has the fixed point property. Then so does A.*

Proof: Suppose $\boldsymbol{f} : A \to A$. Let $\boldsymbol{h}$ be the retract of B onto A. Then $\boldsymbol{f} \circ \boldsymbol{h} : B \to B$ is continuous. Thus, it has a fixed point $\boldsymbol{x} \in B$ so $\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{x}$. However, $\boldsymbol{h}(\boldsymbol{x}) \in A$ and $\boldsymbol{f} : A \to A$ so in fact, $\boldsymbol{x} \in A$. Now $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ and so $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$. ■

Although we won't use this, every convex compact subset K of $\mathbb{R}^p$ is a retract of all of $\mathbb{R}^p$ obtained by using the projection map. See Problems beginning with 8 on Page 137. In particular, K is a retract of a large closed ball containing K, which ball has the fixed point property. Therefore, K also has the fixed point property. This shows the following which is a convenient formulation of the Brouwer fixed point theorem. However, Proposition 9.13.8 is more general. You can likely imagine sets which are retracts which might not be convex.

Theorem 9.13.9 *Every convex closed and bounded subset of $\mathbb{R}^p$ has the fixed point property.*

As an application of the Brouwer fixed point theorem is the following lemma. It says roughly that if a continuous function does not move points near $\boldsymbol{p}$ very far, then the image of a ball centered at $\boldsymbol{p}$ contains a slightly smaller open ball.

Lemma 9.13.10 *Let $\boldsymbol{f}$ be continuous and map $\overline{B(\boldsymbol{p}, r)} \subseteq \mathbb{R}^n$ to $\mathbb{R}^n$. Suppose that for all $\boldsymbol{x} \in \overline{B(\boldsymbol{p}, r)}$, $|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x}| < \varepsilon r$. Then it follows that $\boldsymbol{f}\left(\overline{B(\boldsymbol{p}, r)}\right) \supseteq B(\boldsymbol{p}, (1-\varepsilon) r)$*

Proof: This is from the Brouwer fixed point theorem. Consider for $\boldsymbol{y} \in B(\boldsymbol{p}, (1-\varepsilon) r)$,

$$\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{y}$$

Then $\boldsymbol{h}$ is continuous and for $\boldsymbol{x} \in \overline{B(\boldsymbol{p}, r)}$,

$$|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{p}| = |\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{y} - \boldsymbol{p}| < \varepsilon r + |\boldsymbol{y} - \boldsymbol{p}| < \varepsilon r + (1 - \varepsilon) r = r$$

Hence $\boldsymbol{h} : \overline{B(\boldsymbol{p}, r)} \to \overline{B(\boldsymbol{p}, r)}$ and so it has a fixed point $\boldsymbol{x}$ by Corollary 9.13.6. Thus

$$\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{y} = \boldsymbol{x}$$

so $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$. ■

9.14 Invariance of Domain

As an application of the inverse function theorem is a simple proof of the important invariance of domain theorem which says that continuous and one to one functions defined on an open set in $\mathbb{R}^n$ with values in $\mathbb{R}^n$ take open sets to open sets. You know that this is true for functions of one variable because a one to one continuous function must be either strictly increasing or strictly decreasing. However, the n dimensional version isn't at all obvious but is just as important if you want to consider manifolds with boundary for example. The need for this theorem occurs in many other places as well in addition to being extremely interesting for its own sake. The inverse function theorem gives conditions under which a differentiable function maps open sets to open sets.

The notation $\|\boldsymbol{f}\|_K$ means $\sup_{\boldsymbol{x} \in K} |\boldsymbol{f}(\boldsymbol{x})|$. If you have a continuous function $\boldsymbol{h}$ defined on a compact set K, then the Stone Weierstrass theorem implies you can uniformly approximate it with a polynomial $\boldsymbol{g}$. That is $\|\boldsymbol{h} - \boldsymbol{g}\|_K$ is small. The following lemma says that you can also have $\boldsymbol{g}(\boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{z})$ and $D\boldsymbol{g}(\boldsymbol{z})^{-1}$ exists so that near $\boldsymbol{z}$, the function $\boldsymbol{g}$ will map open sets to open sets as claimed by the inverse function theorem. First is a little observation about approximating.

Lemma 9.14.1 *Suppose* $\det(A) = 0$. *Then for all sufficiently small nonzero* ε,

$$\det(A + \varepsilon I) \neq 0$$

Proof: First suppose A is a $p \times p$ matrix. Suppose also that $\det(A) = 0$. Thus, the constant term of $\det(\lambda I - A)$ is 0. Consider $\varepsilon I + A \equiv A_\varepsilon$ for small real ε. The characteristic polynomial of A_ε is

$$\det(\lambda I - A_\varepsilon) = \det((\lambda - \varepsilon) I - A)$$

This is of the form

$$(\lambda - \varepsilon)^p + a_{p-1}(\lambda - \varepsilon)^{p-1} + \cdots + (\lambda - \varepsilon)^m a_m$$

where the a_j are the coefficients in the characteristic equation for A and m is the largest such that $a_m \neq 0$. The constant term of this characteristic polynomial for A_ε must be nonzero for all ε small enough because it is of the form

$$(-1)^m \varepsilon^m a_m + (\text{higher order terms in } \varepsilon)$$

which shows that $\varepsilon I + A$ is invertible for all ε small enough but nonzero. ■

Lemma 9.14.2 *Let* K *be a compact set in* $\mathbb{R}^n$ *and let* $\boldsymbol{h} : K \to \mathbb{R}^n$ *be continuous,* $\boldsymbol{z} \in K$ *is fixed. Let* $\delta > 0$. *Then there exists a polynomial* $\boldsymbol{g}$ *(each component a polynomial) such that*

$$\|\boldsymbol{g} - \boldsymbol{h}\|_K < \delta,\ \boldsymbol{g}(\boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{z}),\ D\boldsymbol{g}(\boldsymbol{z})^{-1} \text{ exists}$$

Proof: By the Weierstrass approximation theorem, Corollary 5.7.8, or Theorem 5.9.5, there exists a polynomial $\hat{g}$ such that $\|\hat{g}-h\|_K < \frac{\delta}{3}$. Then define for $y \in K$ $g(y) \equiv \hat{g}(y) + h(z) - \hat{g}(z)$ Then $g(z) = \hat{g}(z) + h(z) - \hat{g}(z) = h(z)$. Also

$$\begin{aligned} |g(y)-h(y)| &\leq |(\hat{g}(y)+h(z)-\hat{g}(z))-h(y)| \\ &\leq |\hat{g}(y)-h(y)|+|h(z)-\hat{g}(z)| < \frac{2\delta}{3} \end{aligned}$$

and so since y was arbitrary, $\|g-h\|_K \leq \frac{2\delta}{3} < \delta$. If $Dg(z)^{-1}$ exists, then this is what is wanted. If not, use Lemma 9.14.1 and note that for all η small enough, you could replace g with $y \to g(y) + \eta(y-z)$ and it will still be the case that $\|g-h\|_K < \delta$ along with $g(z) = h(z)$ but now $Dg(z)^{-1}$ exists. Simply use the modified g. ■

The main result is essentially the following lemma which combines the conclusions of the above.

Lemma 9.14.3 *Let $f : \overline{B(p,r)} \to \mathbb{R}^n$ where the ball is also in $\mathbb{R}^n$. Let f be one to one, f continuous. Then there exists $\delta > 0$ such that $f\left(\overline{B(p,r)}\right) \supseteq B(f(p),\delta)$. In other words, $f(p)$ is an interior point of $f\left(\overline{B(p,r)}\right)$.*

Proof: Since $f\left(\overline{B(p,r)}\right)$ is compact, it follows that $f^{-1} : f\left(\overline{B(p,r)}\right) \to \overline{B(p,r)}$ is continuous. By Lemma 9.14.2, there exists a polynomial $g : f\left(\overline{B(p,r)}\right) \to \mathbb{R}^n$ such that

$$\begin{aligned} \|g-f^{-1}\|_{f\left(\overline{B(p,r)}\right)} &< \varepsilon r,\ \varepsilon < 1,\ \ Dg(f(p))^{-1} \\ \text{exists, and } g(f(p)) &= f^{-1}(f(p)) = p \end{aligned}$$

From the first inequality in the above,

$$|g(f(x)) - x| = \left|g(f(x)) - f^{-1}(f(x))\right| \leq \|g-f^{-1}\|_{f\left(\overline{B(p,r)}\right)} < \varepsilon r$$

By Lemma 9.13.10,

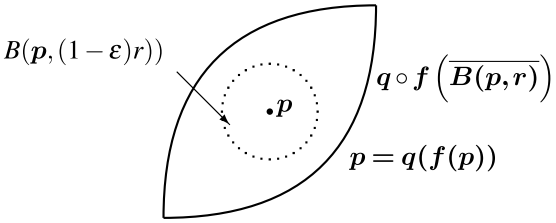
$$g \circ f\left(\overline{B(p,r)}\right) \supseteq B(p,(1-\varepsilon)r) = B(g(f(p)),(1-\varepsilon)r)$$

Since $Dg(f(p))^{-1}$ exists, it follows from the inverse function theorem that g^{-1} also exists and that g, g^{-1} are open maps on small open sets containing $f(p)$ and p respectively. Thus there exists $\eta < (1-\varepsilon)r$ such that g^{-1} is an open map on $B(p,\eta) \subseteq B(p,(1-\varepsilon)r)$. Thus

$$g \circ f\left(\overline{B(p,r)}\right) \supseteq B(p,(1-\varepsilon)r) \supseteq B(p,\eta)$$

So do $g^{-1\text{‘}}$ to both ends. Then you have $g^{-1}(p) = f(p)$ is in the open set $g^{-1}(B(p,\eta))$. Thus

$$f\left(\overline{B(p,r)}\right) \supseteq g^{-1}(B(p,\eta)) \supseteq B\left(g^{-1}(p),\delta\right) = B(f(p),\delta) \ ■$$



With this lemma, the invariance of domain theorem comes right away. This remarkable theorem states that if $\boldsymbol{f}: U \to \mathbb{R}^n$ for U an open set in $\mathbb{R}^n$ and if $\boldsymbol{f}$ is one to one and continuous, then $\boldsymbol{f}(U)$ is also an open set in $\mathbb{R}^n$.

Theorem 9.14.4 *Let U be an open set in $\mathbb{R}^n$ and let $\boldsymbol{f}: U \to \mathbb{R}^n$ be one to one and continuous. Then $\boldsymbol{f}(U)$ is also an open subset in $\mathbb{R}^n$.*

Proof: It suffices to show that if $\boldsymbol{p} \in U$ then $\boldsymbol{f}(\boldsymbol{p})$ is an interior point of $\boldsymbol{f}(U)$. Let $\overline{B(\boldsymbol{p},r)} \subseteq U$. By Lemma 9.14.3, $\boldsymbol{f}(U) \supseteq \boldsymbol{f}\left(\overline{B(\boldsymbol{p},r)}\right) \supseteq B(\boldsymbol{f}(\boldsymbol{p}),\delta)$ so $\boldsymbol{f}(\boldsymbol{p})$ is indeed an interior point of $\boldsymbol{f}(U)$. ■

The inverse mapping theorem assumed quite a bit about the mapping. In particular it assumed that the mapping had a continuous derivative. The following version of the inverse function theorem seems very interesting because it only needs an invertible derivative at a point.

Corollary 9.14.5 *Let U be an open set in $\mathbb{R}^p$ and let $\boldsymbol{f}: U \to \mathbb{R}^p$ be one to one and continuous. Then, $\boldsymbol{f}^{-1}$ is also continuous on the open set $\boldsymbol{f}(U)$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_1 \in U$ and if $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists for $\boldsymbol{x}_1 \in U$, then it follows that $D\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1)) = D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$.*

Proof: $|\cdot|$ will be a norm on $\mathbb{R}^p$, whichever is desired. If you like, let it be the Euclidean norm. $\|\cdot\|$ will be the operator norm. The first part of the conclusion of this corollary is from invariance of domain.

From the assumption that $D\boldsymbol{f}(\boldsymbol{x}_1)$ and $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists,

$$\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1) = \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right) - \boldsymbol{f}(\boldsymbol{x}_1) = D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right) + \boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)$$

Since $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists, $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)) = \boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1 + \boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)$ by continuity, if $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ is small enough, then $\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$ is small enough that in the above, $\left|\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)\right| < \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$. Hence, if $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ is sufficiently small, then from the triangle inequality of the form $|p - q| \geq ||p| - |q||$,

$$\left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\| |(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1))| \geq \left|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1))\right|$$

$$\geq \left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right| - \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right| = \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$$

$$|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)| \geq \left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\|^{-1} \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$$

It follows that for $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ small enough,

$$\left|\frac{\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)}{\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)}\right| \leq \left|\frac{\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)}{\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1}\right| \frac{2}{\left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\|^{-1}}$$

Then, using continuity of the inverse function again, it follows that if $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ is possibly still smaller, then $\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1$ is sufficiently small that the right side of the above inequality is no larger than ε. Since ε is arbitrary, it follows

$$\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right) = \boldsymbol{o}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1))$$

Now from differentiability of $\boldsymbol{f}$ at $\boldsymbol{x}_1$,

$$\begin{aligned}\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1) &= \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)-\boldsymbol{f}(\boldsymbol{x}_1)=D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)+\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)\\ &= D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))\\ &= D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))\right)+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))\end{aligned}$$

Therefore, solving for $\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))$,

$$\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))=D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))$$

From the definition of the derivative, this shows that $D\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))=D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$. ■

9.15 Jensen's Inequality

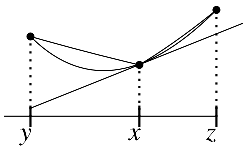
When you have $\phi:\mathbb{R}\to\mathbb{R}$ is convex, then secant lines lie above the graph of ϕ. Say $x<w<z$ so $w=\lambda z+(1-\lambda)x$ for some $\lambda\in(0,1)$. Then refering to the following picture,

$$\frac{\phi(w)-\phi(x)}{w-x}\leq\frac{\lambda\phi(z)+(1-\lambda)\phi(x)-\phi(x)}{(\lambda z+(1-\lambda)x)-x}=\frac{\lambda(\phi(z)-\phi(x))}{\lambda(z-x)}=\frac{\phi(z)-\phi(x)}{z-x}$$

For $y<w<x$ so $w=\lambda y+(1-\lambda)x$. Since $w-x<0$,

$$\frac{\phi(w)-\phi(x)}{w-x}\geq\frac{\lambda\phi(y)+(1-\lambda)\phi(x)-\phi(x)}{\lambda(y-x)}=\frac{\phi(y)-\phi(x)}{y-x}$$

Since x is arbitrary, this has shown that slopes of secant lines of the graph of ϕ over intervals increase as the intervals move to the right.



Lemma 9.15.1 *If $\phi:\mathbb{R}\to\mathbb{R}$ is convex, then ϕ is continuous. Also, if ϕ is convex, $\mu(\Omega)=1$, and $f,\phi(f):\Omega\to\mathbb{R}$ are in $L^1(\Omega)$, then $\phi(\int_\Omega f\,du)\leq\int_\Omega\phi(f)d\mu$.*

Proof: Let $\lambda\equiv\lim_{w\to x+}\frac{\phi(w)-\phi(x)}{w-x}$. Those slopes of secant lines are decreasing and so this limit exists. Then in the picture, for $w\in(x,z)$, $\phi(x)+\lambda(w-x)\leq\phi(w)\leq\phi(x)+\left(\frac{\phi(z)-\phi(x)}{z-x}\right)(w-x)$ and so ϕ is continuous from the right. A similar argument shows ϕ is continuous from the left. In particular, letting $\mu\equiv\lim_{w\to x-}\frac{\phi(x)-\phi(w)}{x-w}\leq\lambda$ because each of these slopes is smaller than the slopes whose inf gives λ. Then this shows that for $w\in(y,x)$, $\frac{\phi(w)-\phi(x)}{w-x}\leq\lambda$ so $\phi(w)-\phi(x)\geq\lambda(w-x)$ and so $\phi(w)\geq\phi(x)+\lambda(w-x)$ and for these ω, $\frac{\phi(x)-\phi(w)}{x-w}\geq\frac{\phi(x)-\phi(y)}{x-y}$ so $\phi(w)\leq\phi(x)+\left(\frac{\phi(x)-\phi(y)}{x-y}\right)(w-x)$ so one obtains continuity from the left. This has also shown that for w not equal to $x,\phi(w)\geq\phi(x)+\lambda(w-x)$ or in other words, $\phi(x)\leq\phi(w)+\lambda(x-w)$ .Letting $x=\int_\Omega fd\mu$, and using the λ whose existence was just established, for each ω,

$$\phi\left(\int_\Omega fd\mu\right)\leq\phi(f(\omega))+\lambda\left(\int_\Omega fd\mu-f(\omega)\right)$$

Do $\int_\Omega d\mu$ to both sides and use $\mu(\Omega) = 1$. Thus

$$\phi\left(\int_\Omega f d\mu\right) \leq \int_\Omega \phi(f)\, d\mu + \lambda\left(\int_\Omega f d\mu - \int_\Omega f d\mu\right) = \int_\Omega \phi(f)\, d\mu.$$

There are no difficulties with measurability because ϕ is continuous. ■

Corollary 9.15.2 *In the situation of Lemma 9.15.1 where* $\mu(\Omega) = 1$, *suppose* f *has values in* $[0,\infty)$ *and is measurable. Also suppose* ϕ *is convex and increasing on* $[0,\infty)$. *Then* $\phi(\int_\Omega f\, du) \leq \int_\Omega \phi(f) d\mu$.

Proof: Let $f_n(\omega) = f(\omega)$ if $f(\omega) \leq n$ and let $f_n(\omega) = n$ if $f(\omega) \geq n$. Then both $f_n, \phi(f_n)$ are in $L^1(\Omega)$. Therefore, the above holds and $\phi(\int_\Omega f_n du) \leq \int_\Omega \phi(f_n) d\mu$. Let $n \to \infty$ and use the monotone convergence theorem. ■

9.16 Faddeyev's Lemma

This next lemma is due to Faddeyev. I found it in [35].

Lemma 9.16.1 *Let* f, g *be nonnegative measurable nonnegative functions on a measure space* (Ω, μ). *Then* $\int fg d\mu = \int_0^\infty \int_{[g>t]} f d\mu dt = \int_0^\infty \int_0^\infty \mu([f > s] \cap [g > t])\, dsdt$.

Proof: First suppose $g = a\mathscr{X}_E$ where E is measurable, $a > 0$. Now $[g > t] = \emptyset$ if $t \geq a$ and it equals $\mathscr{X}_E$ if $t < a$. Thus the right side equals $\int_0^a \int_E f d\mu dt = \int_0^a \int \mathscr{X}_E f d\mu = \int a\mathscr{X}_E f d\mu$ which equals the left side. Thus the first equation is true if $g = a\mathscr{X}_E$. Similar reasoning shows that when you have g a nonnegative simple function, $g = \sum_{i=1}^n a_i \mathscr{X}_{E_i}$ where we can arrange to have $\{a_i\}$ increasing, the first equation still holds. Now by the monotone convergence theorem, this yields the desired result for the first equation.

To get the second equal sign, note that

$$\begin{aligned} \int_0^\infty \int_{[g>t]} f d\mu dt &= \int_0^\infty \int \mathscr{X}_{[g>t]} f d\mu dt = \int_0^\infty \int_0^\infty \mu\left(\left[\mathscr{X}_{[g>t]} f > s\right]\right) dsdt \\ &= \int_0^\infty \int_0^\infty \mu([f > s] \cap [g > t])\, dsdt \blacksquare \end{aligned}$$

9.17 Exercises

1. Let $\Omega = \mathbb{N} = \{1, 2, \cdots\}$. Let $\mathscr{F} = \mathscr{P}(\mathbb{N})$, the set of all subsets of $\mathbb{N}$, and let $\mu(S) =$ number of elements in S. Thus $\mu(\{1\}) = 1 = \mu(\{2\})$, $\mu(\{1,2\}) = 2$, etc. In this case, all functions are measurable. For a nonnegative function, f defined on $\mathbb{N}$, show $\int_\mathbb{N} f d\mu = \sum_{k=1}^\infty f(k)$. What do the monotone convergence and dominated convergence theorems say about this example?

2. For the measure space of Problem 1, give an example of a sequence of nonnegative measurable functions $\{f_n\}$ converging pointwise to a function f, such that inequality is obtained in Fatou's lemma.

3. If $(\Omega, \mathscr{F}, \mu)$ is a measure space and $f \geq 0$ is measurable, show that if $g(\omega) = f(\omega)$ a.e. ω and $g \geq 0$, then $\int g d\mu = \int f d\mu$. Show that if $f, g \in L^1(\Omega)$ and $g(\omega) = f(\omega)$ a.e. then $\int g d\mu = \int f d\mu$.

4. Let $\{f_n\}, f$ be measurable functions with values in $\mathbb{C}$. $\{f_n\}$ converges in measure if $\lim_{n\to\infty}\mu(x\in\Omega : |f(x)-f_n(x)|\geq\varepsilon)=0$ for each fixed $\varepsilon>0$. Prove the theorem of F. Riesz. If f_n converges to f in measure, then there exists a subsequence $\{f_{n_k}\}$ which converges to f a.e. In case μ is a probability measure, this is called convergence in probability. It does not imply pointwise convergence but does imply that there is a subsequence which converges pointwise off a set of measure zero. **Hint:** Choose n_1 such that $\mu(x : |f(x)-f_{n_1}(x)|\geq 1)<1/2$. Choose $n_2>n_1$ such that $\mu(x : |f(x)-f_{n_2}(x)|\geq 1/2)<1/2^2$ $n_3>n_2$ such that $\mu(x : |f(x)-f_{n_3}(x)|\geq 1/3)<1/2^3$, etc. Now consider what it means for $f_{n_k}(x)$ to fail to converge to $f(x)$. Use the Borel Cantelli Lemma 8.2.5 on Page 184.

5. $(X,\mathscr{F},\mu)$ is said to be a **regular** measure space if $\mathscr{F}\supseteq\mathscr{B}(X)$ and for every $F\in\mathscr{F}$,

$$\begin{aligned}\mu(F) &= \sup\{\mu(K) : K \text{ compact, } K\subseteq F\}\\ \mu(F) &= \inf\{\mu(V) : V \text{ open, } V\supseteq F\}\end{aligned}$$

Let $(X,\mathscr{F},\mu)$ be a regular measure space. For example, it could be $\mathbb{R}^p$ with Lebesgue measure, shown later. Why do we care about a measure space being regular? This problem will show why. Suppose that closures of balls are compact as in the case of $\mathbb{R}^p$.

   (a) Let $\mu(E)<\infty$. By regularity, there exists $K\subseteq E\subseteq V$ where K is compact and V is open such that $\mu(V\setminus K)<\varepsilon$. Show there exists W open such that $K\subseteq\bar{W}\subseteq V$ and $\bar{W}$ is compact. Now show there exists a function h such that h has values in $[0,1], h(x)=1$ for $x\in K$, and $h(x)$ equals 0 off W. **Hint:** You might consider Problem 10 on Page 197.

   (b) Show that $\int|\mathscr{X}_E-h|\,d\mu<\varepsilon$

   (c) Next suppose $s=\sum_{i=1}^n c_i\mathscr{X}_{E_i}$ is a nonnegative simple function where each $\mu(E_i)<\infty$. Show there exists a continuous nonnegative function h which equals zero off some compact set such that $\int|s-h|\,d\mu<\varepsilon$

   (d) Now suppose $f\geq 0$ and $f\in L^1(\Omega)$. Show that there exists $h\geq 0$ which is continuous and equals zero off a compact set such that $\int|f-h|\,d\mu<\varepsilon$

   (e) If $f\in L^1(\Omega)$ with complex values, show the conclusion in the above part of this problem is the same.

6. Let $(\Omega,\mathscr{F},\mu)$ be a measure space and suppose $f,g:\Omega\to(-\infty,\infty]$ are measurable. Prove the sets $\{\omega : f(\omega)<g(\omega)\}$ and $\{\omega : f(\omega)=g(\omega)\}$ are measurable. **Hint:** The easy way to do this is to write

$$\{\omega : f(\omega)<g(\omega)\}=\cup_{r\in\mathbb{Q}}[f<r]\cap[g>r].$$

Note that $l(x,y)=x-y$ is not continuous on $(-\infty,\infty]$ so the obvious idea doesn't work. Here $[g>r]$ signifies $\{\omega : g(\omega)>r\}$.

7. Let $\{f_n\}$ be a sequence of real or complex valued measurable functions. Let

$$S=\{\omega : \{f_n(\omega)\}\text{ converges}\}.$$

Show S is measurable. **Hint:** You might try to exhibit the set where f_n converges in terms of countable unions and intersections using the definition of a Cauchy sequence.

8. Suppose $u_n(t)$ is a differentiable function for $t \in (a,b)$ and suppose that for $t \in (a,b)$, $|u_n(t)|$, $|u_n'(t)| < K_n$ where $\sum_{n=1}^{\infty} K_n < \infty$. Show$(\sum_{n=1}^{\infty} u_n(t))' = \sum_{n=1}^{\infty} u_n'(t)$.

   Hint: This is an exercise in the use of the dominated convergence theorem and the mean value theorem.

9. Suppose $\{f_n\}$ is a sequence of nonnegative measurable functions defined on a measure space, $(\Omega, \mathscr{S}, \mu)$. Show that $\int \sum_{k=1}^{\infty} f_k d\mu = \sum_{k=1}^{\infty} \int f_k d\mu$. **Hint:** Use the monotone convergence theorem along with the fact the integral is linear.

10. Explain why for each $t > 0, x \to e^{-tx}$ is a function in $L^1(\mathbb{R})$ and $\int_0^{\infty} e^{-tx} dx = \frac{1}{t}$. Thus
$$\int_0^R \frac{\sin(t)}{t} dt = \int_0^R \int_0^{\infty} \sin(t) e^{-tx} dx dt$$
Now explain why you can change the order of integration in the above iterated integral. Then compute what you get. Next pass to a limit as $R \to \infty$ and show $\int_0^{\infty} \frac{\sin(t)}{t} dt = \frac{1}{2}\pi$. This is a very important integral. Note that the thing on the left is an improper integral. $\sin(t)/t$ is not Lebesgue integrable because it is not absolutely integrable. That is $\int_0^{\infty} \left|\frac{\sin t}{t}\right| dm = \infty$. It is important to understand that the Lebesgue theory of integration only applies to nonnegative functions and those which are absolutely integrable.

11. Let the rational numbers in $[0,1]$ be $\{r_k\}_{k=1}^{\infty}$ and define
$$f_n(t) = \begin{cases} 1 \text{ if } t \in \{r_1, \cdots, r_n\} \\ 0 \text{ if } t \notin \{r_1, \cdots, r_n\} \end{cases}$$
Show that $\lim_{n\to\infty} f_n(t) = f(t)$ where f is one on the rational numbers and 0 on the irrational numbers. Explain why each f_n is Riemann integrable but f is not. However, each f_n is actually a simple function and its Lebesgue and Riemann integral is equal to 0. Apply the monotone convergence theorem to conclude that f is Lebesgue integrable and in fact, $\int f dm = 0$.

12. Show $\lim_{n\to\infty} \frac{n}{2^n} \sum_{k=1}^{n} \frac{2^k}{k} = 2$. This problem was shown to me by Shane Tang, a former student. It is a nice exercise in dominated convergence theorem if you massage it a little. **Hint:**
$$\frac{n}{2^n} \sum_{k=1}^{n} \frac{2^k}{k} = \sum_{k=1}^{n} 2^{k-n} \frac{n}{k} = \sum_{l=0}^{n-1} 2^{-l} \frac{n}{n-l} = \sum_{l=0}^{n-1} 2^{-l} \left(1 + \frac{l}{n-l}\right) \leq \sum_{l}^{n-1} 2^{-l}(1+l)$$

13. Give an example of a sequence of functions $\{f_n\}, f_n \geq 0$ and a function $f \geq 0$ such that $f(x) = \liminf_{n\to\infty} f_n(x)$ but $\int f dm < \liminf_{n\to\infty} \int f_n dm$ so you get strict inequality in Fatou's lemma.

14. Let f be a nonnegative Riemann integrable function defined on $[a,b]$. Thus there is a unique number between all the upper sums and lower sums. First explain why, if $a_i \geq 0, \int \sum_{i=1}^{n} a_i \mathscr{X}_{[t_i, t_{i-1})}(t) dm = \sum_i a_i (t_i - t_{i-1})$. Explain why there exists an increasing sequence of Borel measurable functions $\{g_n\}$ converging to a Borel measurable function g, and a decreasing sequence of functions $\{h_n\}$ which are also Borel measurable converging to a Borel measurable function h such that $g_n \leq f \leq h_n$,
$$\int g_n dm \text{ equals a lower sum, } \int h_n dm \text{ equals an upper sum}$$

and $\int (h-g)\, dm = 0$. Explain why $\{x : f(x) \neq g(x)\}$ is a set of measure zero. Then explain why f is measurable and $\int_a^b f(x)\, dx = \int f dm$ so that the Riemann integral gives the same answer as the Lebesgue integral. Here m is one dimensional Lebesgue measure discussed earlier. Now you know how to compute a Lebesgue integral for reasonable functions. In case of a multiple integral, you would use the iterated integrals to do it.

15. Let λ, μ be finite measures. We say $\lambda \ll \mu$ if whenever $\mu(E) = 0$ it follows that $\lambda(E) = 0$. Show that if $\lambda \ll \mu$, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mu(E) < \delta$, then $\lambda(E) < \varepsilon$.

16. If λ is a signed measure with values in $\mathbb{R}$ so that when $\{E_i\}$ are disjoint, $\sum_i \lambda(E_i)$ converges, show that the infinite series converges absolutely also.

17. Suppose $\nu \ll \mu$ where these are finite measures so there exists $h \geq 0$ and measurable such that $\nu(E) = \int_E h d\mu$ by the Radon Nikodym theorem. Show that if f is measurable and non-negative, then $\int f d\nu = \int f h d\mu$. **Hint:** It holds if f is χ_E and so it holds for a simple function. Now consider a sequence of simple functions increasing to f and use the monotone convergence theorem.

18. Use Jensen's inequality in Corollary 9.15.2 to show that if f is nonnegative and measurable, then for $p > 1$ show that whenever μ is a finite measure, then if $f^p \in L^1(\Omega)$ it follows that $f \in L^1(\Omega)$. Give an example to show that this is not necessarily true if $\mu(\Omega) = \infty$. **Hint:** For the second part, you might consider $\Omega = \mathbb{N}$, the σ algebra the set of all subsets, and $\mu(S)$ equal to the number of elements in S. Maybe $f(n) = 1/n$.

Chapter 10

Regular Measures

10.1 Measures and Regularity

Regular measures have already been discussed. In this section are a few general results which are surprising. The new information involves the possibility that closed balls may not be compact.

Definition 10.1.1 *A Polish space is a complete separable metric space. For a Polish space E or more generally a metric space or even a general topological space, $\mathscr{B}(E)$ denotes the* ***Borel sets*** *of E. This is defined to be the smallest σ algebra which contains the open sets. Thus it contains all open sets and closed sets and compact sets and many others.*

For example, $\mathbb{R}$ is a Polish space as is any separable Banach space. **Amazing things** can be said about finite measures on the Borel sets of a Polish space. First the case of a finite measure on a metric space will be considered.

It is best to not attempt to describe a generic Borel set. Always work with the definition that it is the smallest σ algebra containing the open sets. Attempts to give an explicit description of a "typical" Borel set tend to lead nowhere because there are so many things which can be done.You can take countable unions and complements and then countable intersections of what you get and then another countable union followed by complements and on and on. You just can't get a good useable description in this way. However, it is easy to see that something like $\left(\cap_{i=1}^{\infty}\cup_{j=i}^{\infty}E_j\right)^C$ is a Borel set if the E_j are. This is useful. This said, you can look at Hewitt and Stromberg [23] in their discussion of why there are more Lebesgue measurable sets than Borel measurable sets to see the kind of technicalities which result by describing Borel sets. This is an extremely significant result based on describing Borel sets, so it can be done.

For finite measures, defined on the Borel sets $\mathscr{B}(X)$ of a metric space X, the first definition of regularity is automatic. These are always outer and inner regular provided inner regularity refers to closed sets. Note that if $A \supseteq B$ then $A \setminus B = B^C \setminus A^C$.

Lemma 10.1.2 *Let μ be a finite measure defined on a σ algebra $\mathscr{F} \supseteq \mathscr{B}(X)$ where X is a metric space. Then the following hold.*

1. *μ is regular on $\mathscr{B}(X)$ meaning 8.16, 8.17 whenever $F \in \mathscr{B}(X)$.*
2. *μ is outer regular satisfying 8.17 on sets of $\mathscr{F}$ if and only if it is inner regular satisfying 8.16 on sets of $\mathscr{F}$.*
3. *If μ is either inner or outer regular on sets of $\mathscr{F}$ then if E is any set of $\mathscr{F}$, there exist F an F_σ set and G a G_δ set such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$.*

Proof: 1.) First note every open set is the countable union of closed sets and every closed set is the countable intersection of open sets. Here is why. Let V be an open set and let

$$K_k \equiv \left\{x \in V : \text{dist}\left(x, V^C\right) \geq 1/k\right\}.$$

Then clearly the union of the K_k equals V. Thus

$$\mu(V) = \sup\{\mu(K) : K \subseteq V \text{ and } K \text{ is closed}\}.$$

If U is open and contains V, then $\mu(U) \geq \mu(V)$ and so

$$\mu(V) \leq \inf\{\mu(U) : U \supseteq V,\ U \text{ open}\} \leq \mu(V) \text{ since } V \subseteq V.$$

Thus μ is inner and outer regular on open sets. In what follows, K will be closed and V will be open.

Let $\mathscr{K}$ be the open sets. This is a π system since it is closed with respect to finite intersections. Let

$$\mathscr{G} \equiv \{E \in \mathscr{B}(X) : \mu \text{ is inner and outer regular on } E\} \text{ so } \mathscr{G} \supseteq \mathscr{K}.$$

For $E \in \mathscr{G}$, let $V \supseteq E \supseteq K$ such that $\mu(V \setminus K) = \mu(V \setminus E) + \mu(E \setminus K) < \varepsilon$. Thus $K^C \supseteq E^C$ and so $\mu\left(K^C \setminus E^C\right) = \mu(E \setminus K) < \varepsilon$. Thus μ is outer regular on E^C because

$$\mu\left(K^C\right) = \mu\left(E^C\right) + \mu\left(K^C \setminus E^C\right) < \mu\left(E^C\right) + \varepsilon,\ K^C \supseteq E^C$$

Also, $E^C \supseteq V^C$ and $\mu\left(E^C \setminus V^C\right) = \mu(V \setminus E) < \varepsilon$ so μ is inner regular on E^C and so $\mathscr{G}$ is closed for complements. If the sets of $\mathscr{G}$ $\{E_i\}$ are disjoint, let $K_i \subseteq E_i \subseteq V_i$ with $\mu(V_i \setminus K_i) < \varepsilon 2^{-i}$. Then for $E \equiv \cup_i E_i$,and choosing m sufficiently large,

$$\mu(E) = \sum_i \mu(E_i) \leq \sum_{i=1}^m \mu(E_i) + \varepsilon \leq \sum_{i=1}^m \mu(K_i) + 2\varepsilon = \mu\left(\cup_{i=1}^m K_i\right) + 2\varepsilon$$

and so μ is inner regular on $E \equiv \cup_i E_i$. It remains to show that μ is outer regular on E. Letting $V \equiv \cup_i V_i$,

$$\mu(V \setminus E) \leq \mu\left(\cup_i (V_i \setminus E_i)\right) \leq \sum_i \varepsilon 2^{-i} = \varepsilon.$$

Hence μ is outer regular on E since $\mu(V) = \mu(E) + \mu(V \setminus E) \leq \mu(E) + \varepsilon$ and $V \supseteq E$.

By Dynkin's lemma, $\mathscr{G} = \sigma(\mathscr{K}) \equiv \mathscr{B}(X)$.

2.) Suppose that μ is outer regular on sets of $\mathscr{F} \supseteq \mathscr{B}(X)$. Letting $E \in \mathscr{F}$, by outer regularity, there exists an open set $V \supseteq E^C$ such that $\mu(V) - \mu\left(E^C\right) < \varepsilon$. Since μ is finite, $\varepsilon > \mu(V) - \mu\left(E^C\right) = \mu\left(V \setminus E^C\right) = \mu\left(E \setminus V^C\right) = \mu(E) - \mu\left(V^C\right)$ and V^C is a closed set contained in E. Therefore, if 8.17 holds, then so does 8.16. The converse is proved in the same way. There is $K \subseteq E^C$ with $\varepsilon > \mu\left(E^C \setminus K\right) = \mu\left(K^C \setminus E\right)$ showing outer regularity from inner regularity.

3.) The last claim is obtained by letting $G = \cap_n V_n$ where V_n is open, contains E, $V_n \supseteq V_{n+1}$, and $\mu(V_n) < \mu(E) + \frac{1}{n}$ and K_n, increasing closed sets contained in E such that $\mu(E) < \mu(K_n) + \frac{1}{n}$. Then let $F \equiv \cup F_n$ and $G \equiv \cap_n V_n$. Then $F \subseteq E \subseteq G$ and $\mu(G \setminus F) \leq \mu(V_n \setminus K_n) < 2/n$. ■

Next is a lemma which allows the replacement of closed with compact in the definition of inner regular.

Lemma 10.1.3 *Let μ be a finite measure on a σ algebra containing $\mathscr{B}(X)$, the Borel sets of X, a separable complete metric space, Polish space. Then if C is a closed set,*

$$\mu(C) = \sup\{\mu(K) : K \subseteq C \text{ and } K \text{ is compact.}\}$$

It follows that for a finite measure on $\mathscr{B}(X)$ where X is a Polish space, μ is inner regular in the sense that for all $F \in \mathscr{B}(X)$,$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\}$

Proof: Let $\{a_k\}$ be a countable dense subset of C. Thus $\cup_{k=1}^{\infty} B\left(a_k, \frac{1}{n}\right) \supseteq C$. Therefore, there exists m_n such that

$$\mu\left(C \setminus \cup_{k=1}^{m_n} \overline{B\left(a_k, \frac{1}{n}\right)}\right) \equiv \mu\left(C \setminus C_n\right) < \frac{\varepsilon}{2^n}, \quad \cup_{k=1}^{m_n} \overline{B\left(a_k, \frac{1}{n}\right)} \equiv C_n.$$

Now let $K = C \cap \left(\cap_{n=1}^{\infty} C_n\right)$. Then K is a subset of C_n for each n and so for each $\varepsilon > 0$ there exists an ε net for K since C_n has a $1/n$ net, namely $a_1, \cdots, a_{m_n}$. Since K is closed, it is complete and so it is also compact since it is complete and totally bounded, Theorem 3.5.8. Now $\mu\left(C \setminus K\right) \leq \mu\left(\cup_{n=1}^{\infty}\left(C \setminus C_n\right)\right) < \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon$. Thus $\mu\left(C\right)$ can be approximated by $\mu\left(K\right)$ for K a compact subset of C. The last claim follows from Lemma 10.1.2. ■

The next theorem is the main result. It says that if the measure is outer regular and μ is σ finite then there is an approximation for $E \in \mathscr{F}$ in terms of F_σ and G_δ sets in which the F_σ set is a countable union of compact sets. Also μ is inner and outer regular on $\mathscr{F}$.

Next is a very interesting result on approximations in the context of a regular measure on a metric space in which the closed balls are compact, like $\mathbb{R}^p$.

Proposition 10.1.4 *Suppose (X, d) is a metric space in which the closed balls are compact and X is a countable union of closed balls. Also suppose $(X, \mathscr{F}, \mu)$ is a complete measure space, $\mathscr{F}$ contains the Borel sets, and that μ is regular and finite on measurable subsets of finite balls. Then*

1. *For each $E \in \mathscr{F}$, there is an F_σ set F and a G_δ set G such that $F \subseteq E \subseteq G$ and $\mu\left(G \setminus F\right) = 0$.*

2. *Also if $f \geq 0$ is $\mathscr{F}$ measurable, then there exists $g \leq f$ such that g is Borel measurable and $g = f$ a.e.*

   and $h \geq f$ such that h is Borel measurable and $h = f$ a.e.

3. *If $E \in \mathscr{F}$ is a bounded set contained in a ball $B(x_0, r) = V$, then there exists a sequence of continuous functions in $C_c(V)$ $\{h_n\}$ having values in $[0,1]$ and a set of measure zero N such that for $x \notin N, h_n(x) \to \mathscr{X}_E(x)$. Also $\int |h_n - \mathscr{X}_E|\, d\mu \to 0$. Letting $\tilde{N}$ be a G_δ set of measure zero containing $N, h_n \mathscr{X}_{\tilde{N}^C} \to \mathscr{X}_F$ where $F \subseteq E$ and $\mu\left(E \setminus F\right) = 0$.*

4. *If $f \in L^1(X, \mathscr{F}, \mu)$, there exists $g \in C_c(X)$, such that $\int_X |f - g|\, d\mu < \varepsilon$. There also exists a sequence of functions in $C_c(X)$ $\{g_n\}$ which converges pointwise to f.*

Proof: 1. follows from Theorem 8.7.4.

2. If f is measurable and nonnegative, from Theorem 8.1.6 there is an increasing sequence of simple functions s_n such that $\lim_{n\to\infty} s_n(x) = f(x)$. Say $s_n(x) \equiv \sum_{k=1}^{m_n} c_k^n \mathscr{X}_{E_k^n}(x)$. Let $m_p\left(E_k^n \setminus F_k^n\right) = 0$ where F_k^n is an F_σ set. Replace E_k^n with F_k^n and let $\tilde{s}_n$ be the resulting simple function. Let $g(x) \equiv \lim_{n\to\infty} \tilde{s}_n(x)$. Then g is Borel measurable and $g \leq f$ and $g = f$ except for a set of measure zero, the union of the sets where s_n is not equal to $\tilde{s}_n$. As to the other claim, let $h_n(x) \equiv \sum_{k=1}^{\infty} \mathscr{X}_{A_{kn}}(x) \frac{k}{2^n}$ where A_{kn} is a G_δ set containing $f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)$ for which $\mu\left(A_{kn} \setminus f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)\right) \equiv \mu\left(D_{kn}\right) = 0$. If $N = \cup_{k,n} D_{kn}$, then N is a set of measure zero. On N^C, $h_n(x) \to f(x)$. Let $h(x) = \liminf_{n\to\infty} h_n(x)$. Note that $\mathscr{X}_{A_{kn}}(x) \frac{k}{2^n} \geq \mathscr{X}_{f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)}(x) \frac{k}{2^n}$ and so $h_n(x) \geq h(x)$ and $\liminf_{n\to\infty} h_n(x)$ is Borel measurable because each h_n is.

3. Let $K_n \subseteq E \subseteq V_n$ with K_n compact and V_n open such that $V_n \subseteq B(x_0, r)$ and also that $\mu(V_n \setminus K_n) < 2^{-(n+1)}$. Then from Lemma 3.12.4, there is h_n with $K_n \prec h_n \prec V_n$. Then $\int |h_n - \mathscr{X}_E| \, d\mu < 2^{-n}$ and so

$$\mu\left(|h_n - \mathscr{X}_E| > \left(\frac{2}{3}\right)^n\right) < \left(\left(\frac{3}{2}\right)^n \int_{\left[|h_n - \mathscr{X}_E| > \left(\frac{2}{3}\right)^n\right]} |h_n - \mathscr{X}_E| \, d\mu\right) \leq \left(\frac{3}{4}\right)^n$$

By Lemma 8.2.5 there is a set of measure zero N such that if $x \notin N$, it is in only finitely many of the sets $\left[|h_n - \mathscr{X}_E| > \left(\frac{2}{3}\right)^n\right]$. Thus on N^C, eventually, for all k large enough, $|h_k - \mathscr{X}_E| \leq \left(\frac{2}{3}\right)^k$ so $h_k(x) \to \mathscr{X}_E(x)$ off N. The assertion about convergence of the integrals follows from the dominated convergence theorem and the fact that each h_n is nonnegative, bounded by 1, ($K_n \prec h_n \prec V_n$) and is 0 off some ball. In the last claim, it only remains to verify that $h_n \mathscr{X}_{\tilde{N}^C}$ converges to an indicator function because each $h_n \mathscr{X}_{\tilde{N}^C}$ is Borel measurable. ($\tilde{N} \supseteq N$ and $\tilde{N}$ is a Borel set and $\mu\left(\tilde{N} \setminus N\right) = 0$) Thus its limit will also be Borel measurable. However, $h_n \mathscr{X}_{\tilde{N}^C}$ converges to 1 on $E \cap \tilde{N}^C, 0$ on $E^C \cap \tilde{N}^C$ and 0 on $\tilde{N}$. Thus $E \cap \tilde{N}^C = F$ and $h_n \mathscr{X}_{\tilde{N}^C}(x) \to \mathscr{X}_F$ where $F \subseteq E$ and $\mu(E \setminus F) \leq \mu\left(\tilde{N}\right) = 0$.

4. It suffices to assume $f \geq 0$ because you can consider the positive and negative parts of the real and imaginary parts of f and reduce to this case. Let $f_n(x) \equiv \mathscr{X}_{B(x_0,n)}(x) f(x)$. Then by the dominated convergence theorem, if n is large enough, $\int |f - f_n| \, d\mu < \varepsilon$. There is a nonnegative simple function $s \leq f_n$ such that $\int |f_n - s| \, d\mu < \varepsilon$. This follows from picking k large enough in an increasing sequence of simple functions $\{s_k\}$ converging to f_n and the dominated convergence theorem. Say $s(x) = \sum_{k=1}^m c_k \mathscr{X}_{E_k}(x)$. Then let $K_k \subseteq E_k \subseteq V_k$ where K_k, V_k are compact and open respectively and $\sum_{k=1}^m c_k \mu(V_k \setminus K_k) < \varepsilon$. By Lemma 3.12.4, there exists h_k with $K_k \prec h_k \prec V_k$. Then

$$\begin{aligned} \int \left| \sum_{k=1}^m c_k \mathscr{X}_{E_k}(x) - \sum_{k=1}^m c_k h_k(x) \right| d\mu &\leq \sum_k c_k \int \left| \mathscr{X}_{E_k}(x) - h_k(x) \right| dx \\ &< 2 \sum_k c_k \mu(V_k \setminus K_k) < 2\varepsilon \end{aligned}$$

Let $g \equiv \sum_{k=1}^m c_k h_k(x)$. Thus $\int |s - g| \, d\mu \leq 2\varepsilon$. Then

$$\int |f - g| \, d\mu \leq \int |f - f_n| \, d\mu + \int |f_n - s| \, d\mu + \int |s - g| \, d\mu < 4\varepsilon$$

Since ε is arbitrary, this proves the first part of 4. For the second part, let $g_n \in C_c(X)$ such that $\int |f - g_n| \, d\mu < 2^{-n}$. Let $A_n \equiv \left\{x : |f - g_n| > \left(\frac{2}{3}\right)^n\right\}$. Then

$$\mu(A_n) \leq \left(\frac{3}{2}\right)^n \int_{A_n} |f - g_n| \, d\mu \leq \left(\frac{3}{4}\right)^n.$$

Thus, if N is all x in infinitely many A_n, then by the Borel Cantelli lemma, $\mu(N) = 0$ and if $x \notin N$, then x is in only finitely many A_n and so for all n large enough, $|f(x) - g_n(x)| \leq \left(\frac{2}{3}\right)^n$. ■

10.2 Fundamental Theorem of Calculus

In this section the Vitali covering theorem, Proposition 4.5.3 will be used to give a generalization of the fundamental theorem of calculus. Let f be in $L^1(\mathbb{R}^p)$ where the measure is Lebesgue measure as discussed above.

Let $Mf : \mathbb{R}^p \to [0,\infty]$ by

$$Mf(\boldsymbol{x}) \equiv \sup_{r\leq 1} \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f| \, dm_p \text{ if } \boldsymbol{x} \notin Z.$$

We denote as $\|f\|_1$ the integral $\int_\Omega |f| \, dm_p$.

The special points described in the following theorem are called Lebesgue points. Also $\overline{m_p}$ will denote the outer measure determined by Lebesgue measure. See Proposition 8.4.2. $\overline{m_p}(E) \equiv \inf\left\{\overline{m_p}(F) : F \text{ is measurable and } F \supseteq E\right\}$.

Theorem 10.2.1 *Let m_p be p dimensional Lebesgue measure measure and let $f \in L^1(\mathbb{R}^p, m_p)$.($\int_\Omega |f| \, dm_p < \infty$). Then for m_p a.e.$\boldsymbol{x}$,*

$$\lim_{r\to 0} \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})| \, dm_p(y) = 0$$

Proof: First consider the following claim which is called a weak type estimate.

Claim 1: The following inequality holds for N_p the constant of the Vitali covering theorem, Proposition 4.5.3.

$$\overline{m_p}([Mf > \varepsilon]) \leq 5^p \varepsilon^{-1} \|f\|_1$$

Proof: For each $\boldsymbol{x} \in [Mf > \varepsilon]$ there exists a ball $B_{\boldsymbol{x}} = B(\boldsymbol{x}, r_{\boldsymbol{x}})$ with $0 < r_{\boldsymbol{x}} \leq 1$ and

$$m_p(B_{\boldsymbol{x}})^{-1} \int_{B(\boldsymbol{x},r_{\boldsymbol{x}})} |f| \, dm_p > \varepsilon. \tag{10.1}$$

Let $\mathscr{F}$ be this collection of balls. By the Vitali covering theorem, there is a collection of disjoint balls $\mathscr{G}$ such that if each ball in $\mathscr{G}$ is enlarged making the center the same but the radius 5 times as large, then the corresponding collection of enlarged balls covers $[Mf > \varepsilon]$. By separability, $\mathscr{G}$ is countable, say $\{B_i\}_{i=1}^\infty$ and the enlarged balls will be denoted as $\hat{B}_i$. Then from 10.1,

$$\overline{m_p}([Mf > \varepsilon]) \leq \sum_i m_p(\hat{B}_i) \leq 5^p \sum_i m_p(B_i) \leq \frac{5^p}{\varepsilon} \sum_i \int_{B_i} |f| \, dm_p \leq 5^p \varepsilon^{-1} \|f\|_1$$

This proves **claim 1.**

Claim 2: If $g \in C_c(\mathbb{R}^p)$, then

$$\lim_{r\to 0} \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |g(\boldsymbol{y}) - g(\boldsymbol{x})| \, dm_p(y) = 0$$

Proof: Since g is continuous at $\boldsymbol{x}$, whenever r is small enough,

$$\frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |g(\boldsymbol{y}) - g(\boldsymbol{x})| \, dm_p(y) \leq \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} \varepsilon \, dm_p(y) = \varepsilon.$$

This proves the claim.

Now let $g \in C_c(\mathbb{R}^p)$. Then from the above observations about continuous functions in **Claim 2,**

$$\overline{m_p}\left(\left[\boldsymbol{x} : \limsup_{r\to 0} \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})| \, dm_p(y) > \varepsilon\right]\right) \tag{10.2}$$

$$\begin{aligned}
\leq \quad & \overline{m_p}\left(\left[\boldsymbol{x} : \limsup_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-g\left(\boldsymbol{y}\right)\right| dm_p\left(y\right) > \frac{\varepsilon}{2}\right]\right) \\
& +\overline{m_p}\left(\left[\boldsymbol{x} : \limsup_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|g\left(\boldsymbol{y}\right)-g\left(\boldsymbol{x}\right)\right| dm_p\left(y\right) > \frac{\varepsilon}{2}\right]\right) \\
& +\overline{m_p}\left(\left[\boldsymbol{x} \ : \left|g\left(\boldsymbol{x}\right)-f\left(\boldsymbol{x}\right)\right| > \frac{\varepsilon}{2}\right]\right). \\
& \qquad \leq \overline{m_p}\left(\left[M\left(f-g\right) > \frac{\varepsilon}{2}\right]\right) + m_p\left(\left[\left|f-g\right| > \frac{\varepsilon}{2}\right]\right) \qquad (10.3)
\end{aligned}$$

Now

$$\left\|f-g\right\|_1 \geq \int_{\left[\left|f-g\right|>\frac{\varepsilon}{2}\right]} \left|f-g\right| dm_p \geq \frac{\varepsilon}{2} m_p\left(\left[\left|f-g\right| > \frac{\varepsilon}{2}\right]\right)$$

and so using Claim 1 and 10.3, 10.2 is dominated by $\left(\frac{2}{\varepsilon}+\frac{5^p}{\varepsilon}\right)\int\left|f-g\right| dm_p$. But by Proposition 10.1.4, g can be chosen to make the above as small as desired. Hence 10.2 is 0.

$$\begin{aligned}
& \overline{m_p}\left(\left[\limsup_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right| dm_p\left(y\right) > 0\right]\right) \\
\leq \quad & \sum_{k=1}^{\infty} \overline{m_p}\left(\left[\limsup_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right| dm_p\left(y\right) > \frac{1}{k}\right]\right) = 0
\end{aligned}$$

By completeness of m_p this implies

$$\left[\limsup_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right| dm_p\left(y\right) > 0\right]$$

is a set of m_p measure zero. ■

The following corollary is the main result referred to as the Lebesgue Differentiation theorem.

Definition 10.2.2 *$f \in L^1_{loc}\left(\mathbb{R}^p, m_p\right)$ means $f\mathscr{X}_B$ is in $L^1\left(\mathbb{R}^n, m_p\right)$ whenever B is a ball.*

Corollary 10.2.3 *If $f \in L^1_{loc}\left(\mathbb{R}^p, m_p\right)$, then for a.e.$\boldsymbol{x}$,*

$$\lim_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right| dm_p\left(y\right) = 0\ . \qquad (10.4)$$

In particular, for a.e.$\boldsymbol{x}$,

$$\lim_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} f\left(\boldsymbol{y}\right) dm_p\left(y\right) = f\left(\boldsymbol{x}\right)$$

Proof: If f is replaced by $f\mathscr{X}_{B(\mathbf{0},k)}$ then the conclusion 10.4 holds for all $\boldsymbol{x} \notin F_k$ where F_k is a set of m_p measure 0. Letting $k = 1,2,\cdots$, and $F \equiv \cup_{k=1}^{\infty} F_k$, it follows that F is a set of measure zero and for any $\boldsymbol{x} \notin F$, and $k \in \{1,2,\cdots\}$, 10.4 holds if f is replaced by $f\mathscr{X}_{B(\mathbf{0},k)}$. Picking any such $\boldsymbol{x}$, and letting $k > |\boldsymbol{x}|+1$, this shows

$$\lim_{r\to 0} \frac{1}{m_p\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)} \left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right| dm_p\left(y\right)$$

$$= \lim_{r \to 0} \frac{1}{m_p (B(\boldsymbol{x}, r))} \int_{B(\boldsymbol{x},r)} \left| f \mathscr{X}_{B(\mathbf{0},k)} (\boldsymbol{y}) - f \mathscr{X}_{B(\mathbf{0},k)} (\boldsymbol{x}) \right| dm_p (y) = 0.$$

The last claim holds because

$$\begin{aligned} & \left| f(\boldsymbol{x}) - \frac{1}{m_p (B(\boldsymbol{x}, r))} \int_{B(\boldsymbol{x},r)} f(\boldsymbol{y}) \, dm_p (y) \right| \\ \leq \; & \frac{1}{m_p (B(\boldsymbol{x}, r))} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})| \, dm_p (y) \; \blacksquare \end{aligned}$$

Definition 10.2.4 *Let E be a measurable set. Then $\boldsymbol{x} \in E$ is called a point of density if*

$$\lim_{r \to 0} \frac{m_p (B(\boldsymbol{x}, r) \cap E)}{m_p (B(\boldsymbol{x}, r))} = 1$$

Proposition 10.2.5 *Let E be a measurable set. Then m_p a.e. $\boldsymbol{x} \in E$ is a point of density.*

Proof: This follows from letting $f(\boldsymbol{x}) = \mathscr{X}_E (\boldsymbol{x})$ in Corollary 10.2.3. ■

10.3 Change of Variables, Linear Maps

This is about changing the variables for linear maps where $\mathscr{F}_p$ denotes the Lebesgue measurable sets.

Theorem 10.3.1 *In case $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^p$ is Lipschitz, satisfying the Lipschitz condition $\|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{y})\| \leq K \|\boldsymbol{x} - \boldsymbol{y}\|$, then if T is a set for which $m_p(T) = 0$, it follows that $m_p (\boldsymbol{h}(T)) = 0$. Also if $E \in \mathscr{F}_p$, then $\boldsymbol{h}(E) \in \mathscr{F}_p$.*

Proof: By the Lipschitz condition, $\|\boldsymbol{h}(\boldsymbol{x} + \boldsymbol{v}) - \boldsymbol{h}(\boldsymbol{x})\| \leq K \|\boldsymbol{v}\|$ and you can simply let $T \subseteq V$ where $m_p (V) < \varepsilon / (K^p 5^p)$. Then there is a countable disjoint sequence of balls $\{B_i\}$ such that $\{\hat{B}_i\}$ covers T and each ball B_i is contained in V each having radius no more than 1. Then the Lipschitz condition implies $\boldsymbol{h}(\hat{B}_i) \subseteq B(\boldsymbol{h}(\boldsymbol{x}_i), 5K)$ and so

$$\bar{m}_p (\boldsymbol{h}(T)) \leq \sum_{i=1}^{\infty} m_p (\boldsymbol{h}(\hat{B}_i)) \leq 5^p K^p \sum_{i=1}^{\infty} m_p (B_i) \leq K^p 5^p m_p (V) < \varepsilon$$

Since ε is arbitrary, this shows that $\boldsymbol{h}(T)$ is measurable and $m_p (\boldsymbol{h}(T)) = 0$.

Now let $E \in \mathscr{F}_p$, $m_p (E) < \infty$. Then by of the measure and Theorem 8.7.4, there exists F which is the countable union of compact sets such that $E = F \cup N$ where N is a set of measure zero. Then from the first part, $\boldsymbol{h}(E \setminus F) \subseteq \boldsymbol{h}(N)$ and this set on the right has measure zero and so by completeness of the measure, $\boldsymbol{h}(E \setminus F) \in \mathscr{F}_p$ and so $\boldsymbol{h}(E) = \boldsymbol{h}(E \setminus F) \cup \boldsymbol{h}(F) \in \mathscr{F}_p$ because $F = \cup_k K_k$, each K_k compact. Hence $\boldsymbol{h}(F) = \cup_k \boldsymbol{h}(K_k)$ which is the countable union of compact sets, a Borel set, due to the continuity of $\boldsymbol{h}$. For arbitrary E, $\boldsymbol{h}(E) = \cup_{k=1}^{\infty} \boldsymbol{h}(E \cap B(\mathbf{0}, k)) \in \mathscr{F}_p$. ■

Of course an example of a Lipschitz map is a linear map. $\|A\boldsymbol{x} - A\boldsymbol{y}\| = \|A(\boldsymbol{x} - \boldsymbol{y})\| \leq \|A\| \|\boldsymbol{x} - \boldsymbol{y}\|$. Therefore, if A is linear and E is Lebesgue measurable, then $A(E)$ is also Lebesgue measurable. This is convenient.

Lemma 10.3.2 *Every open set U in $\mathbb{R}^p$ is a countable disjoint union of half open boxes of the form $Q \equiv \prod_{i=1}^{p} [a_i, a_i + 2^{-k})$ where $a_i = l 2^{-k}$ for l some integer.*

Proof: It is clear that there exists $\mathscr{Q}_k$ a countable disjoint collection of these half open boxes each of sides of length 2^{-k} whose union is all of $\mathbb{R}^p$. Let $\mathscr{B}_1$ be those sets of $\mathscr{Q}_1$ which are contained in U, if any. Having chosen $\mathscr{B}_{k-1}$, let $\mathscr{B}_k$ consist of those sets of $\mathscr{Q}_k$ which are contained in U such that none of these are contained in $\mathscr{B}_{k-1}$. Then $\cup_{k=1}^{\infty}\mathscr{B}_k$ is a countable collection of disjoint boxes of the right sort whose union is U. This is because if R is a box of $\mathscr{Q}_k$ and $\hat{R}$ is a box of $\mathscr{Q}_{k-1}$, then either $R \subseteq \hat{R}$ or $R \cap \hat{R} = \emptyset$. If $p \in U$ then it is ultimately contained in some $\mathscr{B}_k$ for k as small as possible because p is at a positive distance from U^C. ■

Corollary 10.3.3 *If D is a diagonal matrix having nonnegative eigenvalues, and U is an open set, it follows that $m_p(DU) = \det(D) m_p(U)$.*

Proof: The multiplication by D just scales each side of the boxes whose disjoint union is U, multiplying the side in the i^{th} direction by the i^{th} diagonal element. Thus if R is one of the boxes, $m_p(DR) = \det(D) m_p(R)$. The desired result follows from adding these together. ■

I will write dx or dy instead of $dm_p(x)$ or $dm_p(y)$ to save on notation.

Theorem 10.3.4 *Let $E \in \mathscr{F}_p$ and let A be a $p \times p$ matrix. Then $A(E)$ is Lebesgue measurable and $m_p(A(E)) = |\det(A)| m_p(E)$. Also, if E is any Lebesgue measurable set, then $\int \mathscr{X}_{A(E)}(\boldsymbol{y})\,dy = \int \mathscr{X}_E(\boldsymbol{x}) |\det(A)|\,dx$.*

Proof: First note that if $C(\boldsymbol{x}, r) \equiv \{\boldsymbol{y} \in \mathbb{R}^p : |\boldsymbol{y} - \boldsymbol{x}| = r\}$, then $m_p(C(\boldsymbol{x}, r)) = 0$. This follows from translation invariance and Corollary 10.3.3 applied to diagonal D having diagonal entries $r(1+\varepsilon)$ and one with diagonal entries $r(1-\varepsilon)$ to obtain that for arbitrary $\varepsilon > 0$,

$$\begin{aligned} m_p(C(\mathbf{0}, r)) &\leq m_p(B(\mathbf{0}, (1+\varepsilon) r) \setminus B(\mathbf{0}, (1-\varepsilon) r)) \\ &= m_p(B(\mathbf{0}, r)) \left[((1+\varepsilon) r)^p - ((1-\varepsilon) r)^p\right] \end{aligned}$$

Here $|\cdot|$ is the Euclidean norm so all orthogonal transormations acting on a ball centered at $\mathbf{0}$ leave the ball unchanged. Now let U be an open set, then by Theorem 4.5.6, there are disjoint open balls $\{B_i\}_{i=1}^{\infty}$ such that $U = (\cup_i B_i) \cup N$ where $m_n(N) = 0$.

From the right polar decomposition, Theorem 1.5.5 and the fact that one can diagonalize a symmetric matrix S, $A = RS = RQ^* DQ$ where R and Q are orthogonal matrices and D is a diagonal matrix with all nonnegative diagonal entries. Thus, if B is an open ball centered at $\mathbf{0}$,

$$\begin{aligned} m_p(A(B)) &= m_p(RQ^* DQ(B)) = m_p(RQ^* D(B)) \\ &= |\det(R)| |\det(Q^*)| \det(D) m_p(B) = |\det(A)| m_p(B) \end{aligned}$$

By continuity of translation, the same holds if B has a center at some other point than $\mathbf{0}$. It follows that $m_p(A(U)) = \sum_i m_p(AB_i) = \sum_i |\det(A)| m_p(B_i) = |\det(A)| m_p(U)$. Now let $\mathscr{K}$ be the open sets and $\mathfrak{S}$ be those Borel sets E such that $m_p(A(E \cap B(\mathbf{0}, n))) = |\det(A)| m_p(E \cap B(\mathbf{0}, n))$. It is routine to verify that $\mathfrak{S}$ is closed with respect to countable disjoint unions and complements. Therefore, $\mathfrak{S} = \sigma(\mathscr{K})$ and so this holds for all Borel E. Letting $n \to \infty$, it follows that for all E Borel, $m_p(A(E)) = |\det(A)| m_p(E)$.

If E is only Lebesgue measurable, then by regularity and Proposition 10.1.4, there exists G and F, G_δ and F_σ sets respectively such that $F \subseteq E \subseteq G$ and $m_p(G) = m_p(E) = m_p(F)$.

Then $AF \subseteq AE \subseteq AG$ and and for $\bar{m}_p$ the outer measure determined by m_p,

$$\begin{aligned} |\det(A)| m_p(F) &= m_p(AF) \leq m_p(AE) \leq m_p(AG) \\ &= \det(A) m_p(G) = |\det(A)| m_p(E) = |\det(A)| m_p(F) \end{aligned}$$

Thus all inequalities are equal signs. ■

Theorem 10.3.5 *Let $f \geq 0$ and suppose it is Lebesgue measurable. Then if A is a $p \times p$ matrix,*

$$\int \mathscr{X}_{A(\mathbb{R}^p)}(\boldsymbol{y}) f(\boldsymbol{y}) dm_p(y) = \int f(A\boldsymbol{x}) |\det(A)| dm_p(x). \tag{10.5}$$

Proof: From Theorem 10.3.4, the equation is true if $\det(A) = 0$. It follows that it suffices to consider only the case where A^{-1} exists. First suppose $f(\boldsymbol{y}) = \mathscr{X}_E(\boldsymbol{y})$ where E is a Lebesgue measurable set. In this case, $A(\mathbb{R}^n) = \mathbb{R}^n$. Then from Theorem 10.3.4

$$\int \mathscr{X}_{A(\mathbb{R}^p)}(\boldsymbol{y}) f(\boldsymbol{y}) dy = \int \mathscr{X}_E(\boldsymbol{y}) dy = m_p(E) = |\det(A)| m_p\left(A^{-1}E\right)$$

$$= \int_{\mathbb{R}^n} |\det(A)| \mathscr{X}_{A^{-1}E}(\boldsymbol{x}) dx = \int_{\mathbb{R}^n} |\det(A)| \mathscr{X}_E(A\boldsymbol{x}) dx = \int f(A\boldsymbol{x}) |\det(A)| dx$$

It follows from this that 10.5 holds whenever $\boldsymbol{f}$ is a nonnegative simple function. Finally, the general result follows from approximating the Lebesgue measurable function with nonnegative simple functions using Theorem 8.1.6 and then applying the monotone convergence theorem. ■

This is now a very good change of variables formula for a linear transformation. Next this is extended to differentiable functions.

10.4 Differentiable Functions and Measurability

To begin with, certain kinds of functions map measurable sets to measurable sets. It was shown earlier, Theorem 10.3.1, that Lipschitz functions do this. So do differentiable functions.

In this part of the argument it is convenient to take all balls with respect to the norm on $\mathbb{R}^p$ given by $\|\boldsymbol{x}\| = \max\{|x_k| : k = 1, 2, \cdots, p\}$. Thus from the definition of this norm, $B(\boldsymbol{x}, r)$ is the open box, $\prod_{k=1}^p (x_k - r, x_k + r)$ and so $m_p(B(\boldsymbol{x}, r)) = (2r)^p = 2^p r^p$. Also for a linear transformation $A \in \mathscr{L}(\mathbb{R}^p, \mathbb{R}^p)$, I will continue to use $\|A\| \equiv \sup_{\|\boldsymbol{x}\| \leq 1} \|A\boldsymbol{x}\|$.

Lemma 10.4.1 *Let $T \subseteq U$, where U is open, $\boldsymbol{h}$ is continuous, and let $\boldsymbol{h}$ be differentiable at each $\boldsymbol{x} \in T$ and suppose that $m_p(T) = 0$, then $m_p(\boldsymbol{h}(T)) = 0$.*

Proof: For $k \in \mathbb{N}$, let $T_k \equiv \{\boldsymbol{x} \in T : \|D\boldsymbol{h}(\boldsymbol{x})\| < k\}$ and let $\varepsilon > 0$ be given. Since T_k is a subset of a set of measure zero, it is measurable, but we don't need to pay much attention to this fact. Now by outer regularity, there exists an open set V, containing T_k which is contained in U such that $m_p(V) < \varepsilon$. Let $\boldsymbol{x} \in T_k$. Then by differentiability, $\boldsymbol{h}(\boldsymbol{x} + \boldsymbol{v}) = \boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})$ and so there exist arbitrarily small $r_{\boldsymbol{x}} < 1$ such that $B(\boldsymbol{x}, 5r_{\boldsymbol{x}}) \subseteq V$ and whenever $\|\boldsymbol{v}\| \leq 5r_{\boldsymbol{x}}$, $\|\boldsymbol{o}(\boldsymbol{v})\| < \frac{1}{5}\|\boldsymbol{v}\|$. Thus, from the Vitali covering theorem, Theorem 4.5.3,

$$\begin{aligned} \boldsymbol{h}(B(\boldsymbol{x}, 5r_{\boldsymbol{x}})) &\subseteq D\boldsymbol{h}(\boldsymbol{x})(B(\boldsymbol{0}, 5r_{\boldsymbol{x}})) + \boldsymbol{h}(\boldsymbol{x}) + B(\boldsymbol{0}, r_{\boldsymbol{x}}) \subseteq B(\boldsymbol{0}, k5r_{\boldsymbol{x}}) + \\ +B(\boldsymbol{0}, r_{\boldsymbol{x}}) + \boldsymbol{h}(\boldsymbol{x}) &\subseteq B(\boldsymbol{h}(\boldsymbol{x}), (5k+1) r_x) \subseteq B(\boldsymbol{h}(\boldsymbol{x}), 6kr_{\boldsymbol{x}}) \end{aligned}$$

From the Vitali covering theorem, there exists a countable disjoint sequence of these balls, $\{B(\boldsymbol{x}_i, r_i)\}_{i=1}^{\infty}$ such that $\{B(\boldsymbol{x}_i, 5r_i)\}_{i=1}^{\infty} = \left\{\widehat{B}_i\right\}_{i=1}^{\infty}$ covers T_k. Then letting $\overline{m_p}$ denote the outer measure determined by m_p,

$$\overline{m_p}(\boldsymbol{h}(T_k)) \leq \overline{m_p}\left(\boldsymbol{h}\left(\cup_{i=1}^{\infty}\widehat{B}_i\right)\right) \leq \sum_{i=1}^{\infty}\overline{m_p}\left(\boldsymbol{h}\left(\widehat{B}_i\right)\right)$$

$$\leq \sum_{i=1}^{\infty} m_p\left(B\left(\boldsymbol{h}(\boldsymbol{x}_i), 6kr_{\boldsymbol{x}_i}\right)\right) = \sum_{i=1}^{\infty} m_p\left(B\left(\boldsymbol{x}_i, 6kr_{\boldsymbol{x}_i}\right)\right)$$

$$= (6k)^p \sum_{i=1}^{\infty} m_p\left(B\left(\boldsymbol{x}_i, r_{\boldsymbol{x}_i}\right)\right) \leq (6k)^p m_p(V) \leq (6k)^p \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows $m_p(\boldsymbol{h}(T_k)) = \overline{m_p}(\boldsymbol{h}(T_k)) = 0$. Now $m_p(\boldsymbol{h}(T)) = \lim_{k\to\infty} m_p(\boldsymbol{h}(T_k)) = 0$. ■

Lemma 10.4.2 *Let $\boldsymbol{h}$ be continuous on U and let $\boldsymbol{h}$ be differentiable on $T \subseteq U$. If S is a Lebesgue measurable subset of T, then $\boldsymbol{h}(S)$ is Lebesgue measurable.*

Proof: By Theorem 8.8.2 there exists F which is a countable union of compact sets $F = \cup_{k=1}^{\infty} K_k$ such that $F \subseteq S$, $m_p(S \setminus F) = 0$. Then $\boldsymbol{h}(F) = \cup_k \boldsymbol{h}(K_k) \in \mathscr{B}(\mathbb{R}^p)$ because the continuous image of a compact set is compact. Also, $\boldsymbol{h}(S \setminus F)$ is a set of measure zero by Lemma 10.4.1 and so $\boldsymbol{h}(S) = \boldsymbol{h}(F) \cup \boldsymbol{h}(S \setminus F) \in \mathscr{F}_p$ because it is the union of two sets which are in $\mathscr{F}_p$. ■

In particular, this proves the following theorem from a different point of view to that done before, using $\boldsymbol{x} \to A\boldsymbol{x}$ being differentiable rather than $\boldsymbol{x} \to A\boldsymbol{x}$ being Lipschitz. Later on, is a theorem which says that Lipschitz implies differentiable a.e. However, it is also good to note that if $\boldsymbol{h}$ has a derivative on an open set U, it does not follow that $\boldsymbol{h}$ is Lipschitz.

I will also use the following fundamental assertion, Sard's lemma.

Lemma 10.4.3 *(Sard) Let U be an open set in $\mathbb{R}^p$. Let $\boldsymbol{h} : U \to \mathbb{R}^p$ be continuous and let $\boldsymbol{h}$ be differentiable on $A \subseteq U$. Let $Z \equiv \{\boldsymbol{x} \in A : \det D\boldsymbol{h}(\boldsymbol{x}) = 0\}$. Then $m_p(\boldsymbol{h}(Z)) = 0$.*

Proof: Suppose first that A is bounded. Let $\varepsilon > 0$ be given. Also let $V \supseteq Z$ with $V \subseteq U$ open, and $m_p(Z) + \varepsilon > m_p(V)$. Now let $\boldsymbol{x} \in Z$. Then since $\boldsymbol{h}$ is differentiable at $\boldsymbol{x}$, there exists $\delta_{\boldsymbol{x}} > 0$ such that if $r < \delta_{\boldsymbol{x}}$, then $B(\boldsymbol{x}, r) \subseteq V$ and also,

$$\boldsymbol{h}(B(\boldsymbol{x},r)) \subseteq \boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})(B(\boldsymbol{0},r)) + B(\boldsymbol{0},r\eta),\ \eta < 1.$$

Regard $D\boldsymbol{h}(\boldsymbol{x})$ as an $n \times n$ matrix, the matrix of the linear transformation $D\boldsymbol{h}(\boldsymbol{x})$ with respect to the usual coordinates. Since $\boldsymbol{x} \in Z$, it follows that there exists an invertible matrix M such that $MD\boldsymbol{h}(\boldsymbol{x})$ is in row reduced echelon form with a row of zeros on the bottom. Therefore, using Theorem 10.3.4 about taking out the determinant of a transformation,

$$\begin{aligned} m_p(\boldsymbol{h}(B(\boldsymbol{x},r))) &= \left|\det\left(M^{-1}\right)\right| m_p(M(\boldsymbol{h}(B(\boldsymbol{x},r)))) \\ &\leq \left|\det\left(M^{-1}\right)\right| m_p(M(D\boldsymbol{h}(\boldsymbol{x}))(B(\boldsymbol{0},r)) + MB(\boldsymbol{0},r\eta)) \end{aligned}$$

$$\begin{aligned} &\leq \left|\det\left(M^{-1}\right)\right| \alpha_{p-1} \|M(D\boldsymbol{h}(\boldsymbol{x}))\|^{p-1}(2r+2\eta r)^{p-1}\|M\| 2r\eta \\ &\leq C\left(\|M\|, \left|\det\left(M^{-1}\right)\right|, \|D\boldsymbol{h}(\boldsymbol{x})\|\right) 4^{p-1} r^p 2\eta \end{aligned}$$

Here α_n is the volume of the unit ball in $\mathbb{R}^n$. This is because $M(D\boldsymbol{h}(\boldsymbol{x}))(B(\mathbf{0},r)) + MB(\mathbf{0},r\eta)$ in the third line up is contained in a cylinder, the base in $\mathbb{R}^{p-1}$ which has radius $\|M(D\boldsymbol{h}(\boldsymbol{x}))\|(2r+2\eta r)$ and height $\|M\|2r\eta$. Thus its measure is no more than $\int_{\mathbb{R}^{p-1}} \int_{-\|Mr\eta\|}^{\|Mr\eta\|} dx_p dm_{p-1}$.Then letting $\delta_{\boldsymbol{x}}$ be still smaller if necessary, corresponding to sufficiently small η ,

$$m_p(\boldsymbol{h}(B(\boldsymbol{x},r))) \leq \varepsilon m_p(B(\boldsymbol{x},r)).$$

The balls of this form constitute a Vitali cover of Z. Hence, by the covering theorem Corollary 4.5.6, there exists $\{B_i\}_{i=1}^{\infty}, B_i = B_i(\boldsymbol{x}_i, r_i)$, a collection of disjoint balls, each of which is contained in V, such that $m_p(\boldsymbol{h}(B_i)) \leq \varepsilon m_p(B_i)$ and $m_p(Z \setminus \cup_i B_i) = 0$. Hence from Lemma 10.4.1,

$$m_p(\boldsymbol{h}(Z) \setminus \cup_i \boldsymbol{h}(B_i)) \leq m_p(\boldsymbol{h}(Z \setminus \cup_i B_i)) = 0$$

Therefore,

$$m_p(\boldsymbol{h}(Z)) \leq \sum_i m_p(\boldsymbol{h}(B_i)) \leq \varepsilon \sum_i m_p(B_i) \leq \varepsilon(m_p(V)) \leq \varepsilon(m_p(Z) + \varepsilon).$$

Since ε is arbitrary, this shows $m_p(\boldsymbol{h}(Z)) = 0$. What if A is not bounded? Then consider $Z_n = Z \cap B(\mathbf{0}, n) \subseteq A \cap B(\mathbf{0}, n)$. From what was just shown, $\boldsymbol{h}(Z_n)$ has measure 0 and so it follows that $\boldsymbol{h}(Z)$ also does, being the countable union of sets of measure zero. ■

10.5 Change of Variables, Nonlinear Maps

This preparation leads to a good change of variables formula. First is a lemma which is likely familiar by now.

Lemma 10.5.1 *Let $\boldsymbol{h} : \Omega \to \mathbb{R}^p$ where $(\Omega, \mathscr{F})$ is a measurable space and suppose $\boldsymbol{h}$ is continuous. Then $\boldsymbol{h}^{-1}(B) \in \mathscr{F}$ whenenver B is a Borel set.*

Proof: Measurability applied to components of $\boldsymbol{h}$ shows that $\boldsymbol{h}^{-1}(U) \in \mathscr{F}$ whenever U is an open set. If $\mathscr{G}$ is consists of the subsets G of $\mathbb{R}^p$ for which $\boldsymbol{h}^{-1}(G) \in \mathscr{F}$, then $\mathscr{G}$ is a σ algebra and $\mathscr{G}$ contains the open sets. ■

Definition 10.5.2 *Let $\boldsymbol{h} : U \to \boldsymbol{h}(U)$ be continuous, U open, and let $H \subseteq U$ be measurable and $\boldsymbol{h}$ is one to one and differentiable on H. Define $\lambda(F) \equiv m_p(\boldsymbol{h}(F \cap H))$.*

Lemma 10.5.3 *λ is a well defined measure on measurable subsets of U and $\lambda \ll m_p$.*

Proof: Since the E_i are disjoint and $\boldsymbol{h}$ is one to one. $\lambda(\cup_i E_i) \equiv m_p(\boldsymbol{h}(\cup_i E_i \cap H)) = \sum_i m_p(\boldsymbol{h}(E_i \cap H)) = \sum_i \lambda(E_i)$. If $m_p(E) = 0$, then $\lambda(E) \equiv m_p(\boldsymbol{h}(E \cap H)) = 0$ because of Lemma 10.4.1. ■

Since $\lambda \ll m_p$, it follows from the Radon Nikodym theorem of Corollary 9.11.11 that there exists $g \in L^1_{loc}(U)$ such that for F a measurable subset of U,

$$\lambda(F) = m_p(\boldsymbol{h}(F \cap H)) = \int_F g dm_p \tag{10.6}$$

where $g = 0$ off H. To see that this corollary applies, note that both λ and m_p are finite on compact sets and that every open set is a countable union of compact sets.

Now let F be a Borel set so that $\boldsymbol{h}^{-1}(F)\cap H$ is measurable and plays the role of F in the above. Then

$$\lambda\left(\boldsymbol{h}^{-1}(F)\right)\equiv m_p\left(\boldsymbol{h}\left(\boldsymbol{h}^{-1}(F)\cap H\right)\right)$$

$$=\int_U \mathscr{X}_{\boldsymbol{h}^{-1}(F)\cap H}(\boldsymbol{x})\,g(\boldsymbol{x})\,dm_p(x)=\int_H \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\,dm_p(x)$$

Thus also for s a Borel measurable nonnegative simple function,

$$\int_{\boldsymbol{h}(H)} s(\boldsymbol{y})\,dm_p(y)=\int_H s(\boldsymbol{h}(\boldsymbol{x}))(\boldsymbol{x})\,g(\boldsymbol{x})\,dm_p(x)$$

Using a sequence of nonnegative simple functions to approximate a nonnegative Borel measurable f, we obtain from the monotone convergence theorem that

$$\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\,dm_p(y)=\int_H f(\boldsymbol{h}(\boldsymbol{x}))(\boldsymbol{x})\,g(\boldsymbol{x})\,dm_p(x)$$

If f is only Lebesgue measurable, then there are nonnegative Borel measurable functions k,l such that $k(\boldsymbol{y})\leq f(\boldsymbol{y})\leq l(\boldsymbol{y})$ with equality holding off a set of m_p measure zero. Then $k(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\leq f(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\leq l(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})$ and the two on the ends are Lebesgue measurable which forces the function in the center to also be Lebesgue measurable by completeness of Lebesgue measure because

$$\begin{aligned}\int_H l(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})-k(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\,dm_p &= \int_{\boldsymbol{h}(H)} l(\boldsymbol{y})\,dm_p-\int_{\boldsymbol{h}(H)} k(\boldsymbol{y})\,dm_p\\ &= \int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\,dm_p-\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\,dm_p=0\end{aligned}$$

Thus $l(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})-k(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})=0$ a.e. Then for f nonnegative and Lebesgue measurable,

$$\int_H f(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\,dm_p=\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\,dm_p.$$

This shows the following lemma.

Lemma 10.5.4 *Let $\boldsymbol{h}:U\to\boldsymbol{h}(U)$ be continuous, U open, and let $H\subseteq U$ be measurable and $\boldsymbol{h}$ is one to one and differentiable on H. Then there exists nonnegative measurable $g\in L^1_{loc}$ such that whenever f is nonnegative and Lebesgue measurable,*

$$\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\,dm_p=\int_H f(\boldsymbol{h}(\boldsymbol{x}))\,g(\boldsymbol{x})\,dm_p$$

where all necessary measurability is obtained.

It remains to identify g.

Lemma 10.5.5 *For a.e. $\boldsymbol{x}$, satisfying $|\det D\boldsymbol{h}(\boldsymbol{x})|>0$, and r small enough,*

$$\begin{aligned}D\boldsymbol{h}(\boldsymbol{x})\,B(\boldsymbol{0},(1-\varepsilon)r) &\subseteq \boldsymbol{h}(B(\boldsymbol{x},r))\subseteq\boldsymbol{h}\left(\overline{B(\boldsymbol{x},r)}\right)\subseteq D\boldsymbol{h}(\boldsymbol{x})\,\overline{B(\boldsymbol{0},(1+\varepsilon)r)},\\ \frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\boldsymbol{x},r))} &\in \left[|\det D\boldsymbol{h}(\boldsymbol{x})|\,(1-\varepsilon)^p,|\det D\boldsymbol{h}(\boldsymbol{x})|\,(1+\varepsilon)^p\right]\\ \lim_{r\to 0}\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\boldsymbol{x},r))} &= |\det D\boldsymbol{h}(\boldsymbol{x})|\end{aligned}$$

Proof: For r small enough,

$$\begin{aligned}\boldsymbol{h}(B(\boldsymbol{x},r)) &\subseteq \boldsymbol{h}(\boldsymbol{x})+D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r)+D\boldsymbol{h}(\boldsymbol{x})D\boldsymbol{h}(\boldsymbol{x})^{-1}B(\mathbf{0},\varepsilon r)\\ &\subseteq \boldsymbol{h}(\boldsymbol{x})+D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r)+D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},\varepsilon r)\\ &\subseteq \boldsymbol{h}(\boldsymbol{x})+D\boldsymbol{h}(\boldsymbol{x})(B(\mathbf{0},(1+\varepsilon)r))\end{aligned}$$

and so $m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))\leq|\det(D\boldsymbol{h}(\boldsymbol{x}))|\,m_p(B(\mathbf{0},(1+\varepsilon)r))$. Also,

$$\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v})=\boldsymbol{h}(\boldsymbol{x})+D\boldsymbol{h}(\boldsymbol{x})\boldsymbol{v}+D\boldsymbol{h}(\boldsymbol{x})D\boldsymbol{h}(\boldsymbol{x})^{-1}\boldsymbol{o}(\boldsymbol{v})$$

and so $\left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}(\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{h}(\boldsymbol{x}))-\boldsymbol{v}\right\|=\left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}\boldsymbol{o}(\boldsymbol{v})\right\|=\|\boldsymbol{o}(\boldsymbol{v})\|$. Thus if r is chosen sufficiently small, it follows that for $\boldsymbol{v}\in B(\mathbf{0},r)$

$$\left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}(\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{h}(\boldsymbol{x}))-\boldsymbol{v}\right\|<\varepsilon r$$

and so, from Lemma 9.13.10, $B(\mathbf{0},(1-\varepsilon)r)\subseteq D\boldsymbol{h}(\boldsymbol{x})^{-1}\left(\boldsymbol{h}\left(\boldsymbol{x}+\overline{B(\mathbf{0},r)}\right)-\boldsymbol{h}(\boldsymbol{x})\right)$.

$$\boldsymbol{h}\left(\overline{B(\boldsymbol{x},r)}\right)=\boldsymbol{h}\left(\boldsymbol{x}+\overline{B(\mathbf{0},r)}\right)-\boldsymbol{h}(\boldsymbol{x})\supseteq D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},(1-\varepsilon)r)$$

Therefore, since $m_p(B(\boldsymbol{x},r))=m_p\left(\overline{B(\boldsymbol{x},r)}\right)$,

$$|\det(D\boldsymbol{h}(\boldsymbol{x}))|\,m_p(B(\mathbf{0},(1-\varepsilon)r))=|\det(D\boldsymbol{h}(\boldsymbol{x}))|(1-\varepsilon)^p r^p\alpha_p\leq m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))$$

so for r small enough,

$$\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\mathbf{0},(1+\varepsilon)r))}\leq|\det(D\boldsymbol{h}(\boldsymbol{x}))|\leq\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\mathbf{0},(1-\varepsilon)r))}$$

The claim follows from this since $\varepsilon>0$ is arbitrary. ■

Lemma 10.5.6 *For a.e.* $\boldsymbol{x}$ *with* $|\det D\boldsymbol{h}(\boldsymbol{x})|>0,\lim_{r\to0}\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)\cap H))}{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}=\frac{g(\boldsymbol{x})}{|\det D\boldsymbol{h}(\boldsymbol{x})|}$.

Proof: Using the result of Lemma 10.5.5, for a.e. $\boldsymbol{x}$ satisfying $|\det D\boldsymbol{h}(\boldsymbol{x})|>0$, if r small enough, then

$$m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))\in\left[|\det D\boldsymbol{h}(\boldsymbol{x})|\,m_p(B(\boldsymbol{x},r))(1-\varepsilon)^p,|\det D\boldsymbol{h}(\boldsymbol{x})|\,m_p(B(\boldsymbol{x},r))(1+\varepsilon)^p\right]$$

Therefore, for $Q_r\equiv\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)\cap H))}{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}\geq\frac{1}{|\det D\boldsymbol{h}(\boldsymbol{x})|m_p(B(\boldsymbol{x},r))(1+\varepsilon)^p}\int_{B(\boldsymbol{x},r)}g\,dm_p$ so

$$\begin{aligned}\frac{1}{m_p(B(\mathbf{0},r))(1+\varepsilon)^p}\int_{B(\boldsymbol{x},r)}\frac{g}{|\det D\boldsymbol{h}(\boldsymbol{x})|}dm_p&\leq Q_r\\ \leq\frac{1}{m_p(B(\mathbf{0},r))(1-\varepsilon)^p}\int_{B(\boldsymbol{x},r)}\frac{g}{|\det D\boldsymbol{h}(\boldsymbol{x})|}dm_p&\end{aligned}$$

and so for Lebesgue points of g, a.e. $\boldsymbol{x}$ with $|\det D\boldsymbol{h}(\boldsymbol{x})|\neq0$,

$$\frac{1}{(1+\varepsilon)^p}\leq\frac{g(\boldsymbol{x})}{|\det D\boldsymbol{h}(\boldsymbol{x})|}\leq\frac{1}{(1-\varepsilon)^p}$$

Then for such $\boldsymbol{x}$, $\frac{1}{(1+\varepsilon)^p}\frac{g}{|\det D\boldsymbol{h}(\boldsymbol{x})|}\leq\liminf_{r\to0}Q_r\leq\limsup_{r\to0}Q_r,\leq\frac{1}{(1-\varepsilon)^p}\frac{g}{|\det D\boldsymbol{h}(\boldsymbol{x})|}$ so, since ε is arbitrary, $\lim_{r\to0}Q_r=\frac{g(\boldsymbol{x})}{|\det D\boldsymbol{h}(\boldsymbol{x})|}$. ■

Lemma 10.5.7 *For a.e.* $\boldsymbol{x} \in H, g(\boldsymbol{x}) = |\det D\boldsymbol{h}(\boldsymbol{x})|$.

Proof: First consider $\boldsymbol{x}$ such that $|\det(D\boldsymbol{h}(\boldsymbol{x}))| \neq 0$. Then by Lemmas 10.5.5 and 10.5.6

$$\begin{aligned}\lim_{r\to 0}\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)\cap H))}{m_p(B(\boldsymbol{x},r))} &= \lim_{r\to 0}\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)\cap H))}{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}\frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\boldsymbol{x},r))} \\ &= \frac{g(\boldsymbol{x})}{|\det D\boldsymbol{h}(\boldsymbol{x})|}|\det D\boldsymbol{h}(\boldsymbol{x})| = g(\boldsymbol{x})\end{aligned}$$

for a.e. $\boldsymbol{x}$ where $|\det(D\boldsymbol{h}(\boldsymbol{x}))| \neq 0$.

If $|\det D\boldsymbol{h}(\boldsymbol{x})| = 0$ then for r small enough,

$$\frac{1}{m_p(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)} g dm_p = \frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)\cap H))}{m_p(B(\boldsymbol{x},r))}$$

$$\leq \frac{m_p(\boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + B(\mathbf{0},\varepsilon r))}{m_p(B(\boldsymbol{x},r))} = \frac{m_p(D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + B(\mathbf{0},\varepsilon r))}{m_p(B(\boldsymbol{x},r))}$$

Now $D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + B(\mathbf{0},\varepsilon r)$ has finite diameter and lies in a $p-1$ dimensional subset. Therefore, from Theorem 10.3.4 on linear mappings, there is an orthogonal matrix Q preserving all distances such that

$$|\det Q|\, m_p(D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + B(\mathbf{0},\varepsilon r)) = m_p(QD\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + B(\mathbf{0},\varepsilon r))$$

where $QD\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r)$ lies in a ball in $\mathbb{R}^{p-1}$ of some radius $\hat{r} = \|D\boldsymbol{h}(\boldsymbol{x})\| r,$. Thus the set on the right side is contained in a cylinder of radius $\hat{r} + \varepsilon r$ and height $2r\varepsilon$ so its measure is no more than $\alpha_{p-1}(\hat{r} + r\varepsilon)^{p-1} 2\varepsilon r$ for $\alpha_{p-1} = m_{p-1}(B(\mathbf{0},1))$. Thus,

$$\begin{aligned}\frac{1}{m_p(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)} g dm_p &\leq \frac{(\|D\boldsymbol{h}(\boldsymbol{x})\| + 1)^p \alpha_{p-1}(r + r\varepsilon)^{p-1} 2\varepsilon r}{\alpha_p r^p} \\ &= 2(\|D\boldsymbol{h}(\boldsymbol{x})\| + 1)^p \frac{\alpha_{p-1}}{\alpha_p}(1+\varepsilon)^{p-1}\varepsilon\end{aligned}$$

Since ε is arbitrary, for every Lebesgue point where $|\det D\boldsymbol{h}(\boldsymbol{x})| = 0$, it follows $g = 0 = |\det D\boldsymbol{h}(\boldsymbol{x})|$. ■

Here is the change of variables formula which follows from Lemma 10.5.4 now that g has been identified.

Theorem 10.5.8 *Let U be an open set and let $\boldsymbol{h}: U \to \boldsymbol{h}(U)$ be continuous and one to one and differentiable on the measurable $H \subseteq U$. Then if $f \geq 0$ is Lebesgue measurable,*

$$\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\, dm_p = \int_H f(\boldsymbol{h}(\boldsymbol{x}))\,|\det(D\boldsymbol{h}(\boldsymbol{x}))|\, dm_p$$

10.6 Mappings Which are Not One to One

Now suppose $\boldsymbol{h}: U \to V = \boldsymbol{h}(U)$ and $\boldsymbol{h}$ is only C^1, not necessarily one to one. Note that I am using C^1, not just differentiable. This makes it convenient to use the inverse function theorem. You can get more generality if you work harder. See my book "Real and Abstract Analysis" for example. For

$$U_+ \equiv \{\boldsymbol{x} \in U : |\det D\boldsymbol{h}(x)| > 0\}$$

and Z the set where $|\det D\boldsymbol{h}(\boldsymbol{x})| = 0$, Lemma 10.4.3 implies $m_p(\boldsymbol{h}(Z)) = 0$. For $\boldsymbol{x} \in U_+$, the inverse function theorem implies there exists an open set $B_{\boldsymbol{x}} \subseteq U_+$, such that $\boldsymbol{h}$ is one to one on $B_{\boldsymbol{x}}$.

Let $\{B_i\}$ be a countable subset of $\{B_{\boldsymbol{x}}\}_{\boldsymbol{x}\in U_+}$ such that $U_+ = \cup_{i=1}^{\infty} B_i$. Let $E_1 = B_1$. If $E_1, \cdots, E_k$ have been chosen, $E_{k+1} = B_{k+1} \setminus \cup_{i=1}^{k} E_i$. Thus

$$\cup_{i=1}^{\infty} E_i = U_+,\ \ \boldsymbol{h} \text{ is one to one on } E_i,\ \ E_i \cap E_j = \emptyset,$$

and each E_i is a Borel set contained in the open set B_i. Now define

$$n(\boldsymbol{y}) \equiv \sum_{i=1}^{\infty} \mathscr{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) + \mathscr{X}_{\boldsymbol{h}(Z)}(\boldsymbol{y}).$$

The sets $\boldsymbol{h}(E_i), \boldsymbol{h}(Z)$ are measurable by Proposition 10.4.1. Thus $n(\cdot)$ is measurable.

Lemma 10.6.1 *Let $F \subseteq \boldsymbol{h}(U)$ be measurable. Then*

$$\int_{\boldsymbol{h}(U)} n(\boldsymbol{y}) \mathscr{X}_F(\boldsymbol{y}) dm_p = \int_U \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p.$$

Proof: Using Lemma 10.4.3 and the Monotone Convergence Theorem

$$\begin{aligned}
\int_{\boldsymbol{h}(U)} n(\boldsymbol{y}) \mathscr{X}_F(\boldsymbol{y}) dm_p &= \int_{\boldsymbol{h}(U)} \left(\sum_{i=1}^{\infty} \mathscr{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) + \overbrace{\mathscr{X}_{\boldsymbol{h}(Z)}(\boldsymbol{y})}^{m_p(\boldsymbol{h}(Z))=0} \right) \mathscr{X}_F(\boldsymbol{y}) dm_p \\
&= \sum_{i=1}^{\infty} \int_{\boldsymbol{h}(U)} \mathscr{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) \mathscr{X}_F(\boldsymbol{y}) dm_p
\end{aligned}$$

$$= \sum_{i=1}^{\infty} \int_{\boldsymbol{h}(B_i)} \mathscr{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) \mathscr{X}_F(\boldsymbol{y}) dm_p = \sum_{i=1}^{\infty} \int_{B_i} \mathscr{X}_{E_i}(\boldsymbol{x}) \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p$$

$$\begin{aligned}
&= \sum_{i=1}^{\infty} \int_U \mathscr{X}_{E_i}(\boldsymbol{x}) \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p \\
&= \int_U \sum_{i=1}^{\infty} \mathscr{X}_{E_i}(\boldsymbol{x}) \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p
\end{aligned}$$

$$= \int_{U_+} \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p = \int_U \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p. \blacksquare$$

Definition 10.6.2 *For $\boldsymbol{y} \in \boldsymbol{h}(U)$, define a function, #, according to the formula*

$$\#(\boldsymbol{y}) \equiv \textit{number of elements in } \boldsymbol{h}^{-1}(\boldsymbol{y}).$$

Observe that

$$\#(\boldsymbol{y}) = n(\boldsymbol{y}) \quad \text{a.e.} \tag{10.7}$$

because $n(\boldsymbol{y}) = \#(\boldsymbol{y})$ if $\boldsymbol{y} \notin \boldsymbol{h}(Z)$, a set of measure 0. Therefore, # is a measurable function because of completeness of Lebesgue measure.

Theorem 10.6.3 *Let $g \geq 0$, g measurable, and let $\boldsymbol{h}$ be $C^1(U)$. Then*

$$\int_{\boldsymbol{h}(U)} \#(\boldsymbol{y}) g(\boldsymbol{y}) dm_p = \int_U g(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p. \tag{10.8}$$

In fact, you can have E some Borel measurable subset of U and conclude that

$$\int_{\boldsymbol{h}(E)} \#(\boldsymbol{y}) g(\boldsymbol{y}) dm_p = \int_E g(\boldsymbol{h}(\boldsymbol{x})) |\det D\boldsymbol{h}(\boldsymbol{x})| dm_p$$

Proof: From 10.7 and Lemma 10.6.1, 10.8 holds for all g, a nonnegative simple function. Approximating an arbitrary measurable nonnegative function g, with an increasing pointwise convergent sequence of simple functions and using the monotone convergence theorem, yields 10.8 for an arbitrary nonnegative measurable function g. To get the last claim, simply replace g with $g\mathscr{X}_{\boldsymbol{h}(E)}$ in the first formula. ■

10.7 Mollifiers and Density of Smooth Functions

Definition 10.7.1 *Let U be an open subset of $\mathbb{R}^n$. $C_c^\infty(U)$ is the vector space of all infinitely differentiable functions which equal zero for all $\boldsymbol{x}$ outside of some compact set contained in U. Similarly, $C_c^m(U)$ is the vector space of all functions which are m times continuously differentiable and whose support is a compact subset of U.*

Example 10.7.2 *Let $U = B(\boldsymbol{z}, 2r)$*

$$\psi(\boldsymbol{x}) = \begin{cases} \exp\left[\left(|\boldsymbol{x}-\boldsymbol{z}|^2 - r^2\right)^{-1}\right] \text{ if } |\boldsymbol{x}-\boldsymbol{z}| < r, \\ 0 \text{ if } |\boldsymbol{x}-\boldsymbol{z}| \geq r. \end{cases}$$

Then a little work shows $\psi \in C_c^\infty(U)$. The following also is easily obtained.

Lemma 10.7.3 *Let U be any open set. Then $C_c^\infty(U) \neq \emptyset$.*

Proof: Pick $\boldsymbol{z} \in U$ and let r be small enough that $B(\boldsymbol{z}, 2r) \subseteq U$. Then let

$$\psi \in C_c^\infty(B(\boldsymbol{z}, 2r)) \subseteq C_c^\infty(U)$$

be the function of the above example.

Definition 10.7.4 *Let $U = \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| < 1\}$. A sequence $\{\psi_m\} \subseteq C_c^\infty(U)$ is called a mollifier [1] if $\psi_m(\boldsymbol{x}) \geq 0$, $\psi_m(\boldsymbol{x}) = 0$, if $|\boldsymbol{x}| \leq \frac{1}{m}$,and $\int \psi_m(\boldsymbol{x}) = 1$. Sometimes it may be written as $\{\psi_\varepsilon\}$ where ψ_ε satisfies the above conditions except $\psi_\varepsilon(\boldsymbol{x}) = 0$ if $|\boldsymbol{x}| \geq \varepsilon$. In other words, ε takes the place of $1/m$ and in everything that follows $\varepsilon \to 0$ instead of $m \to \infty$.*

As before, $\int f(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{y})$ will mean $\boldsymbol{x}$ is fixed and the function $\boldsymbol{y} \to f(\boldsymbol{x}, \boldsymbol{y})$ is being integrated. To make the notation more familiar, dx is written instead of $dm_n(x)$.

[1] This is sometimes called an approximate identity if the differentiability is not included.

Example 10.7.5 *Let $\psi \in C_c^\infty(B(0,1))$ with $\psi(\boldsymbol{x}) \geq 0$ and $\int \psi dm = 1$. Let $\psi_m(\boldsymbol{x}) = c_m \psi(m\boldsymbol{x})$ where c_m is chosen in such a way that $\int \psi_m dm = 1$. By the change of variables theorem $c_m = m^n$. Also ψ_m is zero off $B(\mathbf{0}, 1/m)$.*

Definition 10.7.6 *A function f, is said to be in $L^1_{loc}(\mathbb{R}^n, \mu)$ if f is μ measurable and if $|f| \mathscr{X}_K \in L^1(\mathbb{R}^n, \mu)$ for every compact set K. Here μ is a regular, complete measure on $\mathbb{R}^n$. Usually $\mu = m_n$, Lebesgue measure. When this is so, write $L^1_{loc}(\mathbb{R}^n)$, etc. If $f \in L^1_{loc}(\mathbb{R}^n, \mu)$, and $g \in C_c(\mathbb{R}^n)$, $f * g(\boldsymbol{x}) \equiv \int f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) d\mu$.*

The following lemma will be useful in what follows. It says that one of these very un-regular functions in $L^1_{loc}(\mathbb{R}^n, \mu)$ is smoothed out by convolving with a mollifier.

Lemma 10.7.7 *Let $f \in L^1_{loc}(\mathbb{R}^n, \mu)$, and $g \in C_c^\infty(\mathbb{R}^n)$. Then $f * g$ is an infinitely differentiable function. Here μ is a Radon measure on $\mathbb{R}^n$. In case f is continuous with compact support* $\operatorname{spt}(f)$, *and if ψ_m is a mollifier as described above, then* $\operatorname{spt}(f * \psi_m) \subseteq \operatorname{spt}(f) + B(\mathbf{0}, 1/m)$. *Also $\|f - f * \psi_m\| \to 0$.*

Proof: Consider the difference quotient for calculating a partial derivative of $f * g$.

$$\frac{f * g(\boldsymbol{x} + t\boldsymbol{e}_j) - f * g(\boldsymbol{x})}{t} = \int f(\boldsymbol{y}) \frac{g(\boldsymbol{x} + t\boldsymbol{e}_j - \boldsymbol{y}) - g(\boldsymbol{x} - \boldsymbol{y})}{t} d\mu(y).$$

Using the fact that $g \in C_c^\infty(\mathbb{R}^n)$, the quotient $\frac{g(\boldsymbol{x}+t\boldsymbol{e}_j - \boldsymbol{y}) - g(\boldsymbol{x}-\boldsymbol{y})}{t}$ is uniformly bounded. To see this easily, use Theorem 6.5.2 on Page 149 to get the existence of a constant, M depending on $\max\{||Dg(\boldsymbol{x})|| : \boldsymbol{x} \in \mathbb{R}^n\}$ such that $|g(\boldsymbol{x} + t\boldsymbol{e}_j - \boldsymbol{y}) - g(\boldsymbol{x} - \boldsymbol{y})| \leq M|t|$ for any choice of $\boldsymbol{x}$ and $\boldsymbol{y}$. Therefore, there exists a dominating function for the integrand of the above integral which is of the form $C|f(\boldsymbol{y})| \mathscr{X}_K$ where K is a compact set depending on the support of g. It follows the limit of the difference quotient above passes inside the integral as $t \to 0$ and $\frac{\partial}{\partial x_j}(f * g)(\boldsymbol{x}) = \int f(\boldsymbol{y}) \frac{\partial}{\partial x_j} g(\boldsymbol{x} - \boldsymbol{y}) d\mu(y)$. Now letting $\frac{\partial}{\partial x_j} g$ play the role of g in the above argument, partial derivatives of all orders exist. A similar use of the dominated convergence theorem shows all these partial derivatives are also continuous.

For the last claim, it is clear that $\operatorname{spt}(f * \psi_m) \subseteq \operatorname{spt}(f) + B(\mathbf{0}, 1/m)$ since off $\operatorname{spt}(f) + B(\mathbf{0}, 1/m)$ the integral for $f * \psi_m$ will be 0. To verify the last claim, let $\varepsilon > 0$ be given. By uniform continuity of f, $|f(\boldsymbol{x}) - f(\boldsymbol{x} - \boldsymbol{y})| < \varepsilon$ whenever $|\boldsymbol{y}|$ is sufficiently small. Therefore,

$$\begin{aligned} |f(\boldsymbol{x}) - f * \psi_m(\boldsymbol{x})| &= \left| \int (f(\boldsymbol{x}) - f(\boldsymbol{x} - \boldsymbol{y})) \psi_m(\boldsymbol{y}) d\mu(y) \right| \\ &\leq \int_{B(\mathbf{0}, 1/m)} |f(\boldsymbol{x}) - f(\boldsymbol{x} - \boldsymbol{y})| \psi_m(\boldsymbol{y}) d\mu(y) < \varepsilon \int \psi_m d\mu = \varepsilon \end{aligned}$$

whenever m is large enough. ■

Another thing should probably be mentioned. If you have had a course in complex analysis, you may be wondering whether these infinitely differentiable functions having compact support have anything to do with analytic functions which also have infinitely many derivatives. The answer is no! Recall that if an analytic function has a limit point in the set of zeros then it is identically equal to zero. Thus these functions in $C_c^\infty(\mathbb{R}^n)$ are not analytic. This is a strictly real analysis phenomenon and has absolutely nothing to do with the theory of functions of a complex variable.

10.8 Smooth Partitions of Unity

Partitions of unity were discussed earlier. Here the idea of a smooth partition of unity is considered. The earlier general result on metric space is Theorem 3.12.5 on Page 84. Recall the following notation.

Notation 10.8.1 *I will write $\phi \prec V$ to symbolize $\phi \in C_c(V)$, ϕ has values in $[0,1]$, and ϕ has compact support in V. I will write $K \prec \phi \prec V$ for K compact and V open to symbolize ϕ is 1 on K and ϕ has values in $[0,1]$ with compact support contained in V.*

Definition 10.8.2 *A collection of sets $\mathscr{H}$ is called locally finite if for every $\boldsymbol{x}$, there exists $r>0$ such that $B(\boldsymbol{x},r)$ has nonempty intersection with only finitely many sets of $\mathscr{H}$. Of course every finite collection of sets is locally finite. This is the case of most interest in this book but the more general notion is interesting.*

The thing about locally finite collection of sets is that the closure of their union equals the union of their closures. This is clearly true of a finite collection.

Lemma 10.8.3 *Let $\mathscr{H}$ be a locally finite collection of sets of a normed vector space V. Then*

$$\overline{\cup\mathscr{H}} = \cup\left\{\overline{H} : H \in \mathscr{H}\right\}.$$

Proof: It is obvious $\supseteq$ holds in the above claim. It remains to go the other way. Suppose then that $\boldsymbol{p}$ is a limit point of $\cup\mathscr{H}$ and $\boldsymbol{p} \notin \cup\mathscr{H}$. There exists $r>0$ such that $B(\boldsymbol{p},r)$ has nonempty intersection with only finitely many sets of $\mathscr{H}$ say these are $H_1,\cdots,H_m$. Then I claim $\boldsymbol{p}$ must be a limit point of one of these. If this is not so, there would exist r' such that $0<r'<r$ with $B(\boldsymbol{p},r')$ having empty intersection with each of these H_i. But then $\boldsymbol{p}$ would fail to be a limit point of $\cup\mathscr{H}$. Therefore, $\boldsymbol{p}$ is contained in the right side. It is clear $\cup\mathscr{H}$ is contained in the right side and so This proves the lemma. ■

A good example to consider is the rational numbers each being a set in $\mathbb{R}$. This is **not** a locally finite collection of sets and you note that $\overline{\mathbb{Q}} = \mathbb{R} \neq \cup\{\overline{x} : x \in \mathbb{Q}\}$. By contrast, $\mathbb{Z}$ is a locally finite collection of sets, the sets consisting of individual integers. The closure of $\mathbb{Z}$ is equal to $\mathbb{Z}$ because $\mathbb{Z}$ has no limit points so it contains them all.

Lemma 10.8.4 *Let K be a closed set in $\mathbb{R}^p$ and let $\{V_i\}_{i=1}^{\infty}$ be a locally finite sequence of **bounded** open sets whose union contains K. Then there exist functions, $\psi_i \in C_c^{\infty}(V_i)$ such that for all $\boldsymbol{x} \in K, 1 = \sum_{i=1}^{\infty}\psi_i(\boldsymbol{x})$ and the function $f(\boldsymbol{x})$ given by $f(\boldsymbol{x}) = \sum_{i=1}^{\infty}\psi_i(\boldsymbol{x})$ is in $C^{\infty}(\mathbb{R}^p)$.*

Proof: Let $K_1 = K \setminus \cup_{i=2}^{\infty} V_i$. Thus K_1 is compact because it is a closed subset of a bounded set and $K_1 \subseteq V_1$. Let W_1 be an open set having compact closure which satisfies

$$K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$$

Thus $W_1, V_2, \cdots$ covers K and $\overline{W}_1 \subseteq V_1$. Suppose $W_1,\cdots,W_r$ have been defined such that $\overline{W_i} \subseteq V_i$ for each i, and $W_1,\cdots,W_r,V_{r+1},\cdots$ covers K. Then let

$$K_{r+1} \equiv K \setminus \left(\left(\cup_{i=r+2}^{\infty} V_i\right) \cup \left(\cup_{j=1}^{r} W_j\right)\right).$$

It follows K_{r+1} is compact because $K_{r+1} \subseteq V_{r+1}$. Let W_{r+1} satisfy

$$K_{r+1} \subseteq W_{r+1} \subseteq \overline{W}_{r+1} \subseteq V_{r+1},\ \overline{W}_{r+1} \text{ is compact}$$

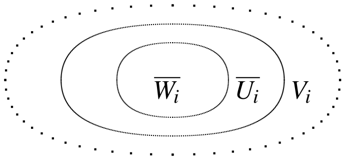
Continuing this way defines a sequence of open sets $\{W_i\}_{i=1}^{\infty}$ having compact closures with the property

$$\overline{W_i} \subseteq V_i,\ K \subseteq \cup_{i=1}^{\infty} W_i.$$

Note $\{W_i\}_{i=1}^{\infty}$ is locally finite because the original sequence, $\{V_i\}_{i=1}^{\infty}$ was locally finite. Now let U_i be open sets which satisfy

$$\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i,\ \overline{U}_i \text{ is compact.}$$

Similarly, $\{U_i\}_{i=1}^{\infty}$ is locally finite.



Now the local finiteness implies $\overline{\cup_{i=1}^{\infty} W_i} = \cup_{i=1}^{\infty} \overline{W_i}$. Define ϕ_i and γ, continuous having compact support such that

$$\overline{U}_i \prec \phi_i \prec V_i,\ \cup_{i=1}^{\infty} \overline{W}_i \prec \gamma \prec \cup_{i=1}^{\infty} U_i.$$

by convolving each of these with a mollifier, we can use Lemma 10.7.7 to preserve the above and also have each of these functions infinitely differentiable. Now define

$$\psi_i(\boldsymbol{x}) = \begin{cases} \gamma(\boldsymbol{x})\phi_i(\boldsymbol{x})/\sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \text{ if } \sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \neq 0, \\ 0 \text{ if } \sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) = 0. \end{cases}$$

All of these infinite sums are really finite sums because of the local finiteness of the $\{V_i\}$. Thus for $\boldsymbol{y}$ near a given $\boldsymbol{x}$, all $\phi_j(\boldsymbol{y})$ are zero. Therefore, all continuity and differentiability of the individual ϕ_j is retained by the "infinite" sum.

If $\boldsymbol{x}$ is such that $\sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) = 0$, then $\boldsymbol{x} \notin \cup_{i=1}^{\infty} \overline{U_i}$ because ϕ_i equals one on $\overline{U_i}$. Consequently $\gamma(\boldsymbol{y}) = 0$ for all $\boldsymbol{y}$ near $\boldsymbol{x}$ thanks to the fact that $\cup_{i=1}^{\infty} \overline{U_i}$ is closed and so $\psi_i(\boldsymbol{y}) = 0$ for all $\boldsymbol{y}$ near $\boldsymbol{x}$. Hence ψ_i is infinitely differentiable at such $\boldsymbol{x}$. If $\sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \neq 0$, this situation persists near $\boldsymbol{x}$ because each ϕ_j is continuous and so ψ_i is infinitely differentiable at such points also. Therefore ψ_i is infinitely differentiable. If $\boldsymbol{x} \in K$, then $\gamma(\boldsymbol{x}) = 1$ and so $\sum_{j=1}^{\infty} \psi_j(\boldsymbol{x}) = 1$. Clearly $0 \leq \psi_i(\boldsymbol{x}) \leq 1$ and $\operatorname{spt}(\psi_j) \subseteq V_j$. ■

The functions, $\{\psi_i\}$ are called a C^{∞} partition of unity. The following is very useful.

Corollary 10.8.5 *In the context of Lemma 10.8.4, if H is a compact subset of V_i for some V_i there exists a partition of unity such that $\psi_i(x) = 1$ for all $x \in H$ in addition to the conclusion of Lemma 10.8.4.*

Proof: Keep V_i the same but replace all the V_j with $\widetilde{V_j} \equiv V_j \setminus H$. Now in the proof above, applied to this modified collection of open sets, if $j \neq i, \phi_j(x) = 0$ whenever $x \in H$. Therefore, $\psi_i(x) = 1$ on H. ■

If K is compact, we can always reduce to a finite cover and so we obtain the following:

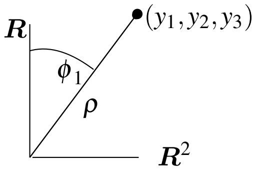
Theorem 10.8.6 *Let K be a compact set in $\mathbb{R}^n$ and let $\{U_i\}_{i=1}^{\infty}$ be an open cover of K. Then there exist functions, $\psi_k \in C_c^{\infty}(U_i)$ such that $\psi_i \prec U_i$ and for all $\boldsymbol{x} \in K$, it follows that $\sum_{i=1}^{\infty} \psi_i(\boldsymbol{x}) = 1$. If K_1 is a compact subset of U_1 there exist such functions such that also $\psi_1(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in K_1$.*

10.9 Spherical Coordinates in p Dimensions

Sometimes there is a need to deal with spherical coordinates in more than three dimensions. In this section, this concept is defined and formulas are derived for these coordinate systems. Recall polar coordinates are of the form

$$\begin{aligned} y_1 &= \rho \cos \theta \\ y_2 &= \rho \sin \theta \end{aligned}$$

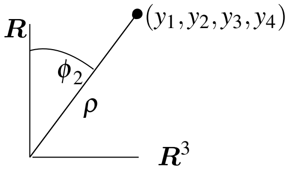
where $\rho > 0$ and $\theta \in \mathbb{R}$. Thus these transformation equations are not one to one but they are one to one on $(0,\infty) \times [0,2\pi)$. Here I am writing ρ in place of r to emphasize a pattern which is about to emerge. I will consider polar coordinates as spherical coordinates in two dimensions. I will also simply refer to such coordinate systems as polar coordinates regardless of the dimension. This is also the reason I am writing y_1 and y_2 instead of the more usual x and y. Now consider what happens when you go to three dimensions. The situation is depicted in the following picture.



From this picture, you see that $y_3 = \rho \cos \phi_1$. Also the distance between (y_1, y_2) and $(0,0)$ is $\rho \sin(\phi_1)$. Therefore, using polar coordinates to write (y_1, y_2) in terms of θ and this distance,

$$\begin{aligned} y_1 &= \rho \sin \phi_1 \cos \theta, \\ y_2 &= \rho \sin \phi_1 \sin \theta, \\ y_3 &= \rho \cos \phi_1. \end{aligned}$$

where $\phi_1 \in \mathbb{R}$ and the transformations are one to one if ϕ_1 is restricted to be in $[0,\pi]$. What was done is to replace ρ with $\rho \sin \phi_1$ and then to add in $y_3 = \rho \cos \phi_1$. Having done this, there is no reason to stop with three dimensions. Consider the following picture:



From this picture, you see that $y_4 = \rho \cos \phi_2$. Also the distance from (y_1, y_2, y_3) to $(0,0,0)$ is $\rho \sin(\phi_2)$. Therefore, using polar coordinates to write (y_1, y_2, y_3) in terms of θ, ϕ_1, and this distance,

$$\begin{aligned} y_1 &= \rho \sin \phi_2 \sin \phi_1 \cos \theta, \\ y_2 &= \rho \sin \phi_2 \sin \phi_1 \sin \theta, \\ y_3 &= \rho \sin \phi_2 \cos \phi_1, \\ y_4 &= \rho \cos \phi_2 \end{aligned}$$

where $\phi_2 \in \mathbb{R}$ and the transformations will be one to one if $\phi_2, \phi_1 \in (0,\pi)$, $\theta \in (0,2\pi)$, $\rho \in (0,\infty)$.

Continuing this way, given spherical coordinates in $\mathbb{R}^p$, to get the spherical coordinates in $\mathbb{R}^{p+1}$, you let $y_{p+1} = \rho \cos \phi_{p-1}$ and then replace every occurance of ρ with $\rho \sin \phi_{p-1}$ to obtain $y_1, \cdots, y_p$ in terms of $\phi_1, \phi_2, \cdots, \phi_{p-1}, \theta$, and ρ.

It is always the case that ρ measures the distance from the point in $\mathbb{R}^p$ to the origin in $\mathbb{R}^p$, $\mathbf{0}$. Each $\phi_i \in \mathbb{R}$ and the transformations will be one to one if each $\phi_i \in (0,\pi)$, and $\theta \in (0,2\pi)$. Denote by $\boldsymbol{h}_p\left(\rho,\vec{\phi},\theta\right)$ the above transformation.

It can be shown using math induction and geometric reasoning that these coordinates map $\prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)\times(0,\infty)$ one to one onto an open subset of $\mathbb{R}^p$ which is everything except for the set of measure zero $\Psi_p(N)$ where N results from having some ϕ_i equal to 0 or π or for $\rho=0$ or for θ equal to either 2π or 0. Each of these are sets of Lebesgue measure zero and so their union is also a set of measure zero. You can see that $\boldsymbol{h}_p\left(\prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)\times(0,\infty)\right)$ omits the union of the coordinate axes except for maybe one of them. This is not important to the integral because it is just a set of measure zero.

Theorem 10.9.1 *Let $\boldsymbol{y}=\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)$ be the spherical coordinate transformations in $\mathbb{R}^p$. Then letting $A=\prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)$, it follows $\boldsymbol{h}$ maps $A\times(0,\infty)$ one to one onto all of $\mathbb{R}^p$ except a set of measure zero given by $\boldsymbol{h}_p(N)$ where N is the set of measure zero*

$$\left(\bar{A}\times[0,\infty)\right)\setminus\left(A\times(0,\infty)\right)$$

Also $\left|\det D\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right|$ will always be of the form

$$\left|\det D\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right|=\rho^{p-1}\Phi\left(\vec{\phi},\theta\right). \tag{10.9}$$

where Φ is a continuous function of $\vec{\phi}$ and θ.[2] *Then if f is nonnegative and Lebesgue measurable,*

$$\int_{\mathbb{R}^p} f(\boldsymbol{y})\,dm_p=\int_{\boldsymbol{h}_p(A)} f(\boldsymbol{y})\,dm_p=\int_A f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p \tag{10.10}$$

Furthermore whenever f is Borel measurable and nonnegative, one can apply Fubini's theorem and write

$$\int_{\mathbb{R}^p} f(\boldsymbol{y})\,dy=\int_0^\infty \rho^{p-1}\int_A f\left(\boldsymbol{h}\left(\vec{\phi},\theta,\rho\right)\right)\Phi\left(\vec{\phi},\theta\right)d\vec{\phi}\,d\theta\,d\rho \tag{10.11}$$

where here $d\vec{\phi}d\theta$ denotes dm_{p-1} on A. The same formulas hold if $f\in L^1(\mathbb{R}^p)$.

Proof: Formula 10.9 is obvious from the definition of the spherical coordinates because in the matrix of the derivative, there will be a ρ in $p-1$ columns. The first claim is also clear from the definition and math induction or from the geometry of the above description. It remains to verify 10.10 and 10.11. It is clear $\boldsymbol{h}_p$ maps $\bar{A}\times[0,\infty)$ onto $\mathbb{R}^p$. Since $\boldsymbol{h}_p$ is differentiable, it maps sets of measure zero to sets of measure zero. Then

$$\mathbb{R}^p=\boldsymbol{h}_p(N\cup A\times(0,\infty))=\boldsymbol{h}_p(N)\cup\boldsymbol{h}_p(A\times(0,\infty)),$$

the union of a set of measure zero with $\boldsymbol{h}_p(A\times(0,\infty))$. Therefore, from the change of variables formula,

$$\begin{aligned}\int_{\mathbb{R}^p} f(\boldsymbol{y})\,dm_p &= \int_{\boldsymbol{h}_p(A\times(0,\infty))} f(\boldsymbol{y})\,dm_p\\ &= \int_{A\times(0,\infty)} f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p\end{aligned}$$

[2] Actually it is only a function of the first but this is not important in what follows.

which proves 10.10. This formula continues to hold if f is in $L^1(\mathbb{R}^p)$ by consideration of positive and negative parts of real and imaginary parts.

Finally, if $f \geq 0$ or in $L^1(\mathbb{R}^n)$ and is Borel measurable, the Borel sets denoted as $\mathscr{B}(\mathbb{R}^p)$ then one can write the following. From the definition of m_p

$$\begin{aligned}
&\int_{A\times(0,\infty)} f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p \\
= &\int_{(0,\infty)}\int_A f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_{p-1}dm \\
= &\int_{(0,\infty)}\rho^{p-1}\int_A f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\Phi\left(\vec{\phi},\theta\right)dm_{p-1}dm
\end{aligned}$$

Now the claim about $f \in L^1$ follows routinely from considering the positive and negative parts of the real and imaginary parts of f in the usual way. ■

Note that the above equals $\int_{\bar{A}\times[0,\infty)} f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p$ and the iterated integral is also equal to

$$\int_{[0,\infty)}\rho^{p-1}\int_{\bar{A}} f\left(\boldsymbol{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\Phi\left(\vec{\phi},\theta\right)dm_{p-1}dm$$

because the difference is just a set of measure zero.

Notation 10.9.2 *Often this is written differently. Note that from the spherical coordinate formulas,* $f\left(\boldsymbol{h}\left(\vec{\phi},\theta,\rho\right)\right) = f(\rho\omega)$ *where* $|\omega| = 1$. *Letting* S^{p-1} *denote the unit sphere,* $\{\omega \in \mathbb{R}^p : |\omega| = 1\}$, *the inside integral in the above formula is sometimes written as*

$$\int_{S^{p-1}} f(\rho\omega)\,d\sigma$$

where σ *is a measure on* S^{p-1}. *See "Real and Abstract Analysis" for another description of this measure. It isn't an important issue here. Either 10.11 or the formula*

$$\int_0^\infty \rho^{p-1}\left(\int_{S^{p-1}} f(\rho\omega)\,d\sigma\right)d\rho$$

will be referred to as polar coordinates and is very useful in establishing estimates. Here $\sigma\left(S^{p-1}\right) \equiv \int_A \Phi\left(\vec{\phi},\theta\right)dm_{p-1}$.

Example 10.9.3 *For what values of s is the integral* $\int_{B(\mathbf{0},R)}\left(1+|\boldsymbol{x}|^2\right)^s dy$ *bounded independent of R? Here* $B(\mathbf{0},R)$ *is the ball,* $\{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| \leq R\}$.

I think you can see immediately that s must be negative but exactly how negative? It turns out it depends on p and using polar coordinates, you can find just exactly what is needed. From the polar coordinates formula above,

$$\begin{aligned}
\int_{B(\mathbf{0},R)}\left(1+|\boldsymbol{x}|^2\right)^s dy &= \int_0^R\int_{S^{p-1}}\left(1+\rho^2\right)^s\rho^{p-1}d\sigma d\rho \\
&= C_p\int_0^R\left(1+\rho^2\right)^s\rho^{p-1}d\rho
\end{aligned}$$

Now the very hard problem has been reduced to considering an easy one variable problem of finding when $\int_0^R \rho^{p-1}\left(1+\rho^2\right)^s d\rho$ is bounded independent of R. You need $2s+(p-1) < -1$ so you need $s < -p/2$.

10.10 Exercises

1. Use a change of variables to find the volume of the ellipsoid $\frac{x^2}{9}+\frac{y^2}{4}+z^2 \leq 1$. **Hint:** You might let $u=\frac{x}{3}, v, w$ defined similarly and reduce to volume of a ball of radius 1.

2. A random vector $\boldsymbol{X},$ with values in $\mathbb{R}^p$ has a multivariate normal distribution written as $\boldsymbol{X} \sim N_p(\boldsymbol{m}, \Sigma)$ if for all Borel $E \subseteq \mathbb{R}^p$,

$$\lambda_{\boldsymbol{X}}(E)=\int_{\mathbb{R}^p} \mathscr{X}_E(\boldsymbol{x}) \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^* \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dm_p$$

Here Σ is a positive definite symmetric matrix. Recall that $\lambda_{\boldsymbol{X}}(E) \equiv P(\boldsymbol{X} \in E)$. Using the change of variables formula, show that $\lambda_{\boldsymbol{X}}$ defined above is a probability measure. One thing you must show is that

$$\int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^* \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dm_p = 1$$

Hint: To do this, you might use the fact from linear algebra that $\Sigma = Q^* D Q$ where D is a diagonal matrix and Q is an orthogonal matrix. Thus $\Sigma^{-1} = Q^* D^{-1} Q$. Maybe you could first let $\boldsymbol{y} = D^{-1/2} Q(\boldsymbol{x}-\boldsymbol{m})$ and change the variables. Note that the change of variables formula works fine when the open sets are all of $\mathbb{R}^p$. You don't need to confine your attention to finite open sets which would be the case with Riemann integrals which are only defined on bounded sets.

3. Consider the positive integers $\mathbb{N}$. Explain why every function defined on these is continuous. Here the distance function is just the absolute value of the difference. What are the functions in $C_c(\mathbb{N})$? Define $L(f)$ as $\sum_i f(i)$ for every f in $C_c(\mathbb{N})$. From the Riesz representation theorem for positive linear functionals, what is the measure which results in this way? What if $L(f) = \sum_i a_i f(i)$ where $a_i \geq 0$?

4. Let g be an increasing function on $\mathbb{R}$. Consider the Rieman Stieltjes integral $L(f) \equiv \int_{\mathbb{R}} f dg$ where $f \in C_c(\mathbb{R})$. These integrals are described in "Analysis of Functions of One Variable. Use the Riesz representation theorem to obtain a measure ν which is regular and complete and Borel representing this functional. Recall that in these Stieltjes integrals, one typically has one of f, g continuous and the other increasing or of bounded variation. In this case, g is of bounded variation on any interval and the Riesz representation theorem shows we can give meaning to an integral in which f could be Borel measurable or worse. The next several problems will develop the fundamentals of the L^p spaces.

5. Show that for $a, b \geq 0$ and $p > 1$, $ab \leq \frac{a^p}{p}+\frac{b^q}{q}$ where q is defined by $\frac{1}{p}+\frac{1}{q}=1$. **Hint:** Show that $p-1 = p/q$ and $q-1 = q/p$.It is obvious if $b=0$ so fix $b>0$ and use calculus procedures on $a \to \frac{a^p}{p}+\frac{b^q}{q}-ab$.

6. $\uparrow$Let $(\Omega, \mathscr{F}, \mu)$ be any measure space. Show that if f, g are nonnegative measurable functions, then it is always the case that $\int fg d\mu \leq \left(\int f^p d\mu\right)^{1/p} \left(\int g^q d\mu\right)^{1/q}$ where here $p, q > 1$ and $1/p + 1/q = 1$. **Hint:** If either of the terms on the right is 0, then there is nothing to show. If either of the terms on the right is ∞, there is also nothing to show by using the above inequality. Assume these are finite and positive. Denoting

them as $\|f\|_p$ and $\|g\|_q$ respectively, consider the above inequality in $\int \frac{f}{\|f\|_p} \frac{g}{\|g\|_q} d\mu$. This is Holder's inequality. When does equality hold? See previous problem to determine this.

7. ↑ $L^p(\Omega,\mu)$ consists of those measurable functions f such that $|f|^p$ is integrable. Show using Holder's inequality that if $\mu(\Omega) < \infty$ that if $1 < p < q$, then $L^q(\Omega) \subseteq L^p(\Omega)$. Give an example which shows that sometimes the opposite inclusion holds if $\mu(\Omega) = \infty$.

8. ↑ Show that if $\|f\|_p, \|g\|_p < \infty$, then $\|f+g\|_p \leq \|f\|_p + \|g\|_p$. **Hint:** Show
$$\int |f+g|^p d\mu \leq \int |f+g|^{p/q} |f| d\mu + \int |f+g|^{p/q} |g| d\mu$$
Now apply Holder's inequality.

9. ↑If we regard $f = g$ when the two differ only on a set of measure zero, explain why $\|f\|_p$ is a norm. The collection of these functions with this convention for $\|\cdot\|_p$ is called $L^p(\Omega,\mu)$

10. ↑Now suppose $\{f_n\}$ is a Cauchy sequence in L^p. That is: For every $\varepsilon > 0$ there exists n_ε such that if $m,n > n_\varepsilon$ then $\|f_m - f_n\|_p < \varepsilon$. Show there exists $f \in L^p$ such that $\lim_{n\to\infty} \|f - f_n\|_p = 0$. **Hint:** Recall from Theorem 3.2.2 on Page 65 that we only need to obtain a subsequence which converges. Here is how you can get such a subsequence. Pick a subsequence $\{f_{n_k}\}$ denoted as $\{g_k\}$ for short such that $\|g_k - g_{k+1}\|_p^p < 4^{-k}$. Now let $E_k \equiv \{\omega : |g_k(\omega) - g_{k+1}(\omega)|^p > 2^{-k}\}$. Explain why
$$4^{-k} \geq \int_{E_k} |g_k(\omega) - g_{k+1}(\omega)|^p d\mu \geq 2^{-k}\mu(E_k),\ \mu(E_k) < 2^{-k}$$
Now recall the Borel Cantelli lemma, Lemma 8.2.5, there is a set of measure zero N such that if $\omega \notin N$ then ω is in only finitely many of the E_k. Thus for such ω $\{g_k(\omega)\}$ is a Cauchy sequence which converges to some $f(\omega)$. Now explain using Fatou's lemma how $\|g_k - f\|_p \to 0$ and that $f \in L^p$. First argue that $\|f\|_p \leq \liminf_{k\to\infty} \|g_k\|_p < \infty$ because the $\|g_k\|_p$ are bounded due to the fact that $\{g_k\}$ is a Cauchy sequence. Next pick m such that if $k,l > m$, then $\|g_k - g_l\| < \varepsilon$. Use Fatou's lemma again to obtain that for $k > m$, $\|g_k - f\|_p < \varepsilon$ which was to be shown.

11. ↑Show that the simple functions are dense in L^p. **Hint:** Consider positive and negative parts of $f \in L^p$ and use Theorem 8.1.6 about pointwise limits of simple functions along with the dominated convergence theorem or some such thing.

12. ↑In case the measure space is $(X,\mathscr{F},\mu)$ where μ is regular and Borel and X is a Polish space, show that $C_c(X)$ is dense in $L^p(X)$.

13. ↑In the situation of $(\mathbb{R}^n,\mathscr{F}_p,m_p)$, Lebesgue measure, define $f_{\boldsymbol{y}}(\boldsymbol{x}) \equiv f(\boldsymbol{x}-\boldsymbol{y})$. Show $\lim_{\boldsymbol{y}\to\mathbf{0}} \|f_{\boldsymbol{y}} - f\|_p = 0$. This is very important and is called continuity of translation in L^p. **Hint:** Let $g \in C_c(\mathbb{R}^n)$. Then
$$\|f_{\boldsymbol{y}} - f\|_p \leq \|f_{\boldsymbol{y}} - g_{\boldsymbol{y}}\|_p + \|g_{\boldsymbol{y}} - g\|_p + \|g - f\|_p$$
Now from the above problem, pick $g \in C_c(\mathbb{R}^n)$ such that $\|f - g\|_p = \|f_{\boldsymbol{y}} - g_{\boldsymbol{y}}\|_p < \varepsilon$. This comes from a change of variables exercise of using translation invariance of the measure. Now if $\boldsymbol{y}$ is small enough, the right side is no more than 3ε.

14. ↑Let $\{\psi_m\}$ be a mollifier. Explain why if $f \in L^p(\mathbb{R}^n)$ then $f \in L^1_{loc}(\mathbb{R}^n)$. You might use Holder's inequality to show this. Thus from Lemma 10.7.7, $f * \psi_m$ is infinitely differentiable. Now fill in the details of the following:

$$\left|\int f(\boldsymbol{x}-\boldsymbol{y})\psi_m(\boldsymbol{y})\,dm_n - f(\boldsymbol{x})\right|^p = \left|\int (f(\boldsymbol{x}-\boldsymbol{y})-f(\boldsymbol{x}))\psi_m(\boldsymbol{y})\,dm_n\right|^p$$
$$\leq \left(\int |f(\boldsymbol{x}-\boldsymbol{y})-f(\boldsymbol{x})|\,\psi_m(\boldsymbol{y})\,dm_n\right)^p \leq \int |f(\boldsymbol{x}-\boldsymbol{y})-f(\boldsymbol{x})|^p\,\psi_m(\boldsymbol{y})\,dm_n(y)$$

Then

$$\begin{aligned}\|f*\psi_m - f\|_p^p \leq \int\int |f(\boldsymbol{x}-\boldsymbol{y})-f(\boldsymbol{x})|^p\,\psi_m(\boldsymbol{y})\,dm_n(y)\,dm_n(x) \\ = \int \psi_m(\boldsymbol{y})\int |f(\boldsymbol{x}-\boldsymbol{y})-f(\boldsymbol{x})|^p\,dm_n(x)\,dm_n(y) \\ = \int_{B(\mathbf{0},1/m)} \psi_m(\boldsymbol{y})\,\|f_{\boldsymbol{y}}-f\|_p^p\,dm_n(y) \leq \varepsilon\end{aligned}$$

whenever m large enough. You will want to use Jensen's inequality at the third relation in the top. Indeed, $\psi_m dm_n$ is a probability measure. This shows that even though the functions in L^p might be discontinuous everywhere, the space of infinitely differentiable functions is dense in L^p.

15. Minkowski's inequality is very useful. Fill in the details for finite measure spaces and $f \geq 0$ product measurable. Recall $1/p + 1/q = 1$ where $p > 1$.

$$\begin{aligned}\int_Y\left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda &= \int_Y J(y)^{p/q}\int_X |f(x,y)|\,d\mu d\lambda \\ &= \int_X\int_Y |f(x,y)|\,J(y)^{p/q}\,d\lambda d\mu\end{aligned}$$

Use Holder's inequality on the inside integral to get

$$\leq \left(\int_Y\left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda\right)^{1/q}\int_X\left(\int_Y |f(x,y)|^p\,d\lambda\right)^{1/p} d\mu$$

Now divide both sides by $\left(\int_Y\left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda\right)^{1/q}$. This gives the Minkowski inequality

$$\left(\int_Y\left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda\right)^{1/p} \leq \int_X\left(\int_Y |f(x,y)|^p\,d\lambda\right)^{1/p} d\mu$$

Extend to σ finite measure spaces.

16. You have a measure space $(\Omega,\mathscr{F},P)$ where P is a probability measure on $\mathscr{F}$. Then you also have a measurable function $X:\Omega\to Z$ where Z is some metric space. Thus $X^{-1}(U)\in\mathscr{F}$ whenever U is open. Now define a measure on $\mathscr{B}(Z)$ denoted by λ_X and defined by $\lambda_X(E) = P(\{\omega : X(\omega)\in E\})$. Explain why this yields a well defined probability measure on $\mathscr{B}(Z)$ which is regular.

$$\begin{aligned}\lambda_X(F) &= \sup\{\lambda_X(K) : K \text{ compact, } K\subseteq F\} \\ \lambda_X(F) &= \inf\{\lambda_X(V) : V \text{ open, } V\supseteq F\}\end{aligned}$$

This is called the distribution measure.

Chapter 11

Integration on Manifolds

Till now, integrals have mostly pertained to measurable subsets of $\mathbb{R}^p$ and not something like a surface contained in a higher dimensional space. This is what is considered in this chapter. First is an abstract description of manifolds and then an interesting application of the representation theorem for positive linear functionals is used to give a measure on a manifold. This is the higher dimensional version of arc length for a smooth curve seen in calculus.

Definition 11.0.1 *Let S be a nonempty set in a metric space (X,d). ∂S is the set of points x, if any with the property that $B(x,r)$ contains points of S and points of $X \setminus S$ for each $r > 0$. The interior of S consists of the union of all open subsets of S.*

Lemma 11.0.2 *Let U be a nonempty open set in a metric space (X,d). $\partial U = \bar{U} \setminus U$.*

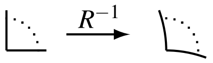
Proof: If $x \in \partial U$, then x can't be in U because some ball containing x is contained in U. However, it must be in $\bar{U}$ because if not, some ball containing x would contain no points of $\bar{U}$ since $\bar{U}$ is closed.

If $x \in \bar{U} \setminus U$ then if some ball containing x fails to contain other points which are in U then that ball would show $x \notin \bar{U}$. Hence every ball containing x must contain points of U. However, x itself is not in U and so $x \in \partial U$. ■

11.1 Manifolds

Definition 11.1.1 *An essential part of the definition of a manifold is the idea of a relatively open set defined next. Recall that a homeomorphism is a one to one, onto, continuous mapping from one metric space to another which has continuous inverse. A half space will be of the form $\{\boldsymbol{x} : x_i \geq a_i\}$ or $\{\boldsymbol{x} : x_i \leq a_i\}$.*

Definition 11.1.2 *Let X be a metric space and let $\Omega \subseteq X$. Then a set U is called a relatively open set or open in Ω if it is the intersection of an open set of X with Ω. Thus Ω is a metric space with respect to the distance $d(x,y)$ inherited from X and all considerations such as limit points, closures, limits, closed sets, open sets etc. in this metric space are taken with respect to this metric. Continuity is also defined in terms of this metric on Ω inherited from X. Ω is a p dimensional manifold with boundary if there is a locally finite cover $\{U_i\}$ (here it will be a finite cover) of sets open in Ω such that each U_i is homeomorphic to a set open in H where H is a half space or some finite intersection of such half spaces. Denote the open sets and homeomorphisms by (U_i, R_i). The collection of these is called an atlas. Thus R_iU_i is a set open in H_{R_i} where H_{R_i} is described above. Note that it could be a closed box. Then a point x is called a boundary point if and only if R_ix is a boundary point of the interior of some H_{R_i} for some i.*



I will be assuming that we can replace "locally finite" with finite in the above definition. This would happen, for example if Ω were compact, but this is not necessary. First I need to verify that the idea of $\partial\Omega$ is well defined.

Lemma 11.1.3 *$\partial\Omega$ is well defined in the sense that the statement that $\boldsymbol{x}$ is a boundary point does not depend on which chart is considered.*

Proof: Suppose x is not a boundary point with respect to the chart (U,R) but is a boundary point with respect to (V,S). Then $U\cap V$ is open in Ω so $Rx\in B\subseteq R(U\cap V)$ where $R(U\cap V)$ is open in H_R and B is an open ball contained in $R(U\cap V)$. But then, by Theorem 9.14.4, $S\circ R^{-1}(B)$ is open in $\mathbb{R}^p$ and contains Sx so x is not a boundary point with respect to (V,S) after all. ■

Definition 11.1.4 *Let $V\subseteq\mathbb{R}^q$. $C^k\left(\overline{V};\mathbb{R}^p\right)$ is the set of functions which are restrictions to V of some function defined on $\mathbb{R}^q$ which has k continuous derivatives which has values in $\mathbb{R}^p$. When $k=0$, it means the restriction to V of continuous functions. A function is in $D\left(\overline{V};\mathbb{R}^p\right)$ if it is the restriction to V of a differentiable function defined on $\mathbb{R}^q$. A Lipschitz function $\boldsymbol{f}$ is one which satisfies $\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{y})\|\leq K\|\boldsymbol{x}-\boldsymbol{y}\|$.*

Thus, if $\boldsymbol{f}\in C^k\left(\overline{V};\mathbb{R}^q\right)$ or $D\left(\overline{V};\mathbb{R}^p\right)$, we can consider it defined on $\overline{V}$ and not just on V. This is the way one can generalize a one sided derivative of a function defined on a closed interval.

Lemma 11.1.5 *Suppose A is a $m\times n$ matrix in which $m>n$ and A is one to one. Then $\|\boldsymbol{v}\|\equiv|A\boldsymbol{v}|$ is a norm on $\mathbb{R}^n$ equivalent to the usual norm.*

Proof: All the algebraic properties of the norm are obvious. If $\|\boldsymbol{v}\|=0$ then $|A\boldsymbol{v}|=0$ and since A is one to one, it follows $\boldsymbol{v}=0$ also. Now recall that all norms on $\mathbb{R}^n$ are equivalent. ■

We have in mind, from now on that our manifold will be a compact subset of $\mathbb{R}^q$ for some $q\geq p$.

Proposition 11.1.6 *Suppose in the atlas for a manifold with boundary Ω it is also the case that each chart $(U,\boldsymbol{R})$ has $\boldsymbol{R}^{-1}\in C^1\left(\overline{\boldsymbol{R}(U)}\right)$ and $D\boldsymbol{R}^{-1}(\boldsymbol{x})$ is one to one on $\overline{\boldsymbol{R}(U)}$. Then for two charts $(U,\boldsymbol{R})$ and $(V,\boldsymbol{S})$, it will be the case that $\boldsymbol{S}\circ\boldsymbol{R}^{-1}:\boldsymbol{R}(U\cap V)\to\boldsymbol{S}(V)$ will be also $C^1\left(\overline{\boldsymbol{R}(U\cap V)}\right)$.*

Proof: Then

$$\begin{aligned} D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{h}+\boldsymbol{o}(\boldsymbol{h}) &= \boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})-\boldsymbol{R}^{-1}(\boldsymbol{x}) \\ &= \boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)\right)-\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right) \quad (11.1) \end{aligned}$$

$$\begin{aligned} = {} & D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right) \\ & +\boldsymbol{o}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right) \quad (11.2) \end{aligned}$$

By continuity of $\boldsymbol{R}^{-1},\boldsymbol{S}$, if $\boldsymbol{h}$ is small enough, which will always be assumed,

$$\begin{aligned} & \left|\boldsymbol{o}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right| \\ \leq {} & \frac{\alpha}{2}\left|\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right| \end{aligned}$$

where here there is $\alpha > 0$ such that

$$\begin{aligned} & \left|D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right| \\ \geq\ & \alpha\left|\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right| \end{aligned}$$

thanks to the assumption that $D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)$ is one to one. Thus from 11.2

$$\frac{\alpha}{2}\left|\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right| \leq \left|D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{h}+\boldsymbol{o}(\boldsymbol{h})\right| \tag{11.3}$$

Now

$$\begin{aligned} & \frac{\left|\boldsymbol{o}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right|}{|\boldsymbol{h}|} \\ \leq\ & \frac{\left|\boldsymbol{o}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\right|}{\left|\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right|}\frac{\left|\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right|}{|\boldsymbol{h}|} \end{aligned}$$

From 11.3, the second factor in the above is bounded. Now continuity of $\boldsymbol{S}\circ\boldsymbol{R}^{-1}$ implies that as $\boldsymbol{h}\to\boldsymbol{0},$ the first factor also converges to 0. Thus

$$\boldsymbol{o}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})\right)-\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)=\boldsymbol{o}(\boldsymbol{h})$$

Returning to 11.2,

$$D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{h}+\boldsymbol{o}(\boldsymbol{h})=D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x}+\boldsymbol{h})-\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x})\right)$$

Thus if $\boldsymbol{h}=t\boldsymbol{v},$

$$\begin{aligned} & \lim_{t\to 0} D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)\left(\frac{\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x})\right)}{t}\right) \\ =\ & D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{v}+\lim_{t\to 0}\frac{\boldsymbol{o}(t\boldsymbol{v})}{t}=D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{v} \end{aligned}$$

By the above lemma, $\lim_{t\to 0}\frac{\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{S}\circ\boldsymbol{R}^{-1}(\boldsymbol{x})\right)}{t}=D_{\boldsymbol{v}}\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}\right)(\boldsymbol{x})$ exists. Also

$$D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)D_{\boldsymbol{v}}\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}\right)(\boldsymbol{x})=D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{v}$$

Let $A(\boldsymbol{x})\equiv D\boldsymbol{S}^{-1}\left(\boldsymbol{S}\left(\boldsymbol{R}^{-1}(\boldsymbol{x})\right)\right)$. Then A^*A is invertible and $\boldsymbol{x}\to A(\boldsymbol{x})$ is continuous. Then

$$\begin{aligned} A(\boldsymbol{x})^*A(\boldsymbol{x})D_{\boldsymbol{v}}\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}\right)(\boldsymbol{x}) &= A(\boldsymbol{x})^*D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{v} \\ D_{\boldsymbol{v}}\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}\right)(\boldsymbol{x}) &= \left(A(\boldsymbol{x})^*A(\boldsymbol{x})\right)^{-1}A(\boldsymbol{x})^*D\boldsymbol{R}^{-1}(\boldsymbol{x})\boldsymbol{v} \end{aligned}$$

so $D_{\boldsymbol{v}}\left(\boldsymbol{S}\circ\boldsymbol{R}^{-1}\right)(\boldsymbol{x})$ is continuous. It follows from Theorem 6.6.1 that $\boldsymbol{S}\circ\boldsymbol{R}^{-1}$ is a function in $C^1\left(\boldsymbol{R}\left(\overline{U\cap V}\right)\right)$ because the Gateaux derivatives exist and are continuous. ■

Saying $D\boldsymbol{R}^{-1}(\boldsymbol{x})$ is one to one is the analog of the situation in calculus with a smooth curve in which we assume the derivative is non zero and that the parametrization has continuous derivative.

I will assume in what follows that Ω is a compact subset of $\mathbb{R}^q$, $q\geq p$. You could get by with less using Stone's theorem about paracompactness but this is enough for what will be used here.

Definition 11.1.7 *A compact subset Ω of $\mathbb{R}^q$ will be called a differentiable p dimensional manifold with boundary if it is a C^0 manifold and also has some differentiable structure about to be described. Ω is a differentiable manifold if $\boldsymbol{R}_j \circ \boldsymbol{R}_i^{-1}$ is differentiable on $\boldsymbol{R}_i (U_j \cap U_i)$. This is implied by the condition of Proposition 11.1.6. If, in addition to this, it has an atlas $(U_i, \boldsymbol{R}_i)$ such that all partial derivatives are continuous and for all $\boldsymbol{x}$*

$$\det \left(D\boldsymbol{R}_i^{-1} (\boldsymbol{R}_i (\boldsymbol{x}))\right)^* \left(D\boldsymbol{R}_i^{-1} (\boldsymbol{R}_i (\boldsymbol{x}))\right) \neq 0$$

then it is called a smooth manifold. This condition is like the one for a smooth curve in calculus in which the derivative does not vanish. If, in addition "differentiable" is replaced with C^k meaning the first k derivatives exist and are continuous, then it will be a smooth C^k manifold with boundary.

Next is the concept of an oriented manifold. Orientation can be defined for general C^0 manifolds using the topological degree, but the reason for considering this, at least here, involves some sort of differentiability.

Definition 11.1.8 *A differentiable manifold Ω with boundary is called orientable if there exists an atlas, $\{(U_r, \boldsymbol{R}_r)\}_{r=1}^m$, such that whenever $U_i \cap U_j \neq \emptyset$,*

$$\det \left(D \left(\boldsymbol{R}_j \circ \boldsymbol{R}_i^{-1}\right)\right) (\boldsymbol{u}) \geq 0 \textit{ for all } \boldsymbol{u} \in \boldsymbol{R}_i (U_i \cap U_j) \tag{11.4}$$

An atlas satisfying 11.4 is called an oriented atlas. Also the following notation is often used with the convention that $\boldsymbol{v} = \boldsymbol{R}_i \circ \boldsymbol{R}_j^{-1} (\boldsymbol{u})$

$$\frac{\partial (v_1 \cdots v_p)}{\partial (u_1 \cdots u_p)} \equiv \det D \left(\boldsymbol{R}_i \circ \boldsymbol{R}_j^{-1}\right) (\boldsymbol{u})$$

In this case, another atlas will be called an equivalent atlas $(V_i, \boldsymbol{S}_i)$ if

$$\det \left(D \left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)\right) (\boldsymbol{u}) \geq 0 \textit{ for all } \boldsymbol{u} \in \boldsymbol{R}_i (U_i \cap V_j)$$

You can verify using the chain rule that this condition does indeed define an equivalence relation. Thus an oriented manifold would consist of a metric space along with an equivalence class of atlases. You could also define a piecewise smooth manifold as the union of finitely many smooth manifolds which have intersection only at boundary points.

Orientation is about the order in which the variables are listed or the way the positive coordinate axes point relative to each other. When you have an $n \times n$ matrix, you can always write its row reduced echelon form as a product of elementary matrices, some of which are permutation matrices or involve changing the direction by multiplying by a negative scalar, which also changes orientation the others having positive determinant. If there are an odd number of switches or multiplication by a negative scalar, you get the determinant is nonpositive. If an even number, the determinant is non-negative. This is why we use the determinant to keep track of orientation in the above definition.

Example 11.1.9 *Let $f : \mathbb{R}^{p+1} \to \mathbb{R}$ is C^1 and suppose and that $Df (\boldsymbol{x}) \neq \boldsymbol{0}$ for all $\boldsymbol{x}$ contained in the set $\{\boldsymbol{x} : f (\boldsymbol{x}) = 0\}$. Then if $\{\boldsymbol{x} : f (\boldsymbol{x}) = 0\}$ is nonempty, it is a C^1 manifold thanks to an application of the implicit function theorem.*

Note that this includes $S^{p-1}, \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| = 1\}$ and lots of other things like $x^4 + y^2 + z^4 = 1$ and so forth. The details are left as an exercise.

Recall from calculus how you can get pointy places in a space curve when the derivative of the parametrization is allowed to vanish. Here this would correspond to some $D\boldsymbol{R}_i^{-1}(\boldsymbol{u})$ not being one to one which is the same as having $D\left(\boldsymbol{R}_i \circ \boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)$ having zero determinant.

In the above, it is not assumed that $D\boldsymbol{R}^{-1}$ is one to one. This can be used to include the concept of a higher dimensional version of a piecewise smooth curve. Suppose, for example you have $Q_1 \equiv [-1,0] \times \prod_{i=2}^{p} [a_i, b_i], Q_2 \equiv [0,1] \times \prod_{i=2}^{p} [a_i, b_i]$ so there are two boxes joined along a common side. Let $\boldsymbol{R}_{1i}^{-1}, \boldsymbol{R}_{2j}^{-1}$ be as described above on these boxes and that $\boldsymbol{R}_{1i}^{-1}$ and $\boldsymbol{R}_{2j}^{-1}$ are continuous along the common face. We assume the union of $\boldsymbol{R}_{ri}^{-1}(U_{ri}), r = 1,2$ is a smooth manifold so that $D\boldsymbol{R}_{ri}^{-1}$ exists on Q_r. Maybe $D\boldsymbol{R}_{1i}^{-1}$, $D\boldsymbol{R}_{2j}^{-1}$ are one to one on Q_1, Q_2 but on the common face, there is a difference in $D_1\boldsymbol{R}_{1i}^{-1}$, $D_1\boldsymbol{R}_{2j}^{-1}$ at a point on that face. Thus, if the restriction of $\boldsymbol{R}_i^{-1}$ to Q_r is $\boldsymbol{R}_{ri}^{-1}$ then $\boldsymbol{R}_i^{-1}$ is not differentiable at points on this face. However, we could change the parametrization at the expense of allowing $D\boldsymbol{R}_{ri}^{-1}$ to equal zero on the common face which will result in $\boldsymbol{R}_i^{-1}$ being differentiable. One simply replaces $\boldsymbol{x} \to \boldsymbol{R}_{ri}^{-1}(x_1, ..., x_p)$ with $\boldsymbol{x} \to \boldsymbol{R}_{ri}^{-1}\left(x_1^3, x_2, ..., x_p\right)$. This could be generalized to strings of boxes, successive pairs intersecting along a face thereby obtaining a higher dimensional notion of "piecewise smooth" as a case where the determinant of $D\boldsymbol{R}_i^{-1}$ is allowed to vanish. This is why it is useful in what follows to have a change of variables formula which does not require the non-vanishing of the determinant of the derivative of the transformation. This is the higher dimensional notion of pointy places occuring in space curves at points where the derivative vanishes. Note that the resulting union of the two smooth manifolds would end up being orientable if $\det\left(D\left(\boldsymbol{R}_{1j} \circ \boldsymbol{R}_{2i}^{-1}\right)\right)(\boldsymbol{u}) > 0$ for all pertinent $\boldsymbol{u}$ on the common face. Here we would take the partial derivative D_1 from the appropriate side in the chain rule. This is all very fussy but is mentioned to illustrate that in order to include piecewise smooth manifolds it suffices to only require that an atlas be differentiable. Thanks to Theorem 10.3.1 edges of a differentiable manifold can be ignored in the development of the area measure on a manifold if they result from some lower dimensional curve in $\mathbb{R}^p$ or more generally a set of measure zero in $\mathbb{R}^p$. In this regard, see the rank theorem, Theorem 7.6.3 which identifies this as happening when $D\boldsymbol{R}_i^{-1}$ has smaller rank.

11.2 The Area Measure on a Manifold

Next the "surface measure" on a manifold is given. In what follows, the manifold will be a compact subset of $\mathbb{R}^q$. This has nothing to do with orientation. It will involve the following definition. To motivate this definition, recall the way you found the length of a curve in calculus where $t \in [a,b]$. It was $\int_a^b |\boldsymbol{r}'(t)|\, dt = \int_a^b \det\left(D\boldsymbol{r}(t)^* D\boldsymbol{r}(t)\right)^{1/2} dt$. where $\boldsymbol{r}(t)$ is a parametrization for the curve. Think of $dl = \det\left(D\boldsymbol{r}(t)^* D\boldsymbol{r}(t)\right)^{1/2} dt$ and you sum these to get the length.

Definition 11.2.1 *Let $(U_i, \boldsymbol{R}_i)$ be an atlas for a p dimensional differentiable manifold with boundary Ω. Also let $\{\psi_i\}_{i=1}^r$ be a partition of unity from Theorem 3.12.5* spt $\psi_i \subseteq U_i$. *Then for* $f \in C_c(\Omega)$, *define*

$$Lf \equiv \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du$$

Here du signifies $dm_p(\boldsymbol{u})$ *and*

$$J_i(\boldsymbol{u}) \equiv \left(\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{u})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\right)^{1/2}$$

I need to show that the same thing is obtained if another atlas and/or partition of unity is used. This is an application of the change of variables theorem.

Theorem 11.2.2 *The functional L is well defined in the sense that if another atlas is used, then for* $f \in C_c(\Omega)$, *the same value is obtained for* Lf.

Proof: Let the other atlas be $\left\{(V_j, \boldsymbol{S}_j)\right\}_{j=1}^{s}$ where $\boldsymbol{v} \in V_j$ and $\boldsymbol{S}_j$ has the same properties as the $\boldsymbol{R}_i$. Then $\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u}) = \boldsymbol{v}$ so $\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = \boldsymbol{S}_j^{-1}(\boldsymbol{v})$ and so $\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = \boldsymbol{S}_j^{-1}\left(\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)$ implying $D\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = D\boldsymbol{S}_j^{-1}(\boldsymbol{v})\, D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})$. Therefore,

$$J_i(\boldsymbol{u}) = \left(\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{u})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\right)^{1/2}$$

$$= \left(\det\left(\overbrace{D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)^*(\boldsymbol{u})}^{p\times p}\overbrace{D\boldsymbol{S}_j^{-1}(\boldsymbol{v})^* D\boldsymbol{S}_j^{-1}(\boldsymbol{v})}^{(p\times q)(q\times p)}\overbrace{D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})}^{p\times p}\right)\right)^{1/2}$$

$$= \left[\det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)^*(\boldsymbol{u})\right) \det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)\right]^{1/2} J_j(\boldsymbol{v})$$

$$= \left|\det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)\right| J_j(\boldsymbol{v}) \tag{11.5}$$

Similarly

$$J_j(\boldsymbol{v}) = \left|\det\left(D\left(\boldsymbol{R}_i \circ \boldsymbol{S}_j^{-1}\right)(\boldsymbol{v})\right)\right| J_i(\boldsymbol{u}). \tag{11.6}$$

Let $\hat{L}$ go with this new atlas. Thus

$$\hat{L}(f) \equiv \sum_{j=1}^{s} \int_{\boldsymbol{S}_j(V_j)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv \tag{11.7}$$

where η_j is a partition of unity associated with the sets V_j as described above. Now letting ψ_i be the partition of unity for the U_i, $\boldsymbol{v} = \boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}(\boldsymbol{u})$ for $\boldsymbol{u} \in \boldsymbol{R}_i(V_j \cap U_i)$.

$$\int_{\boldsymbol{S}_j(V_j)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv$$

$$= \sum_{i=1}^{r} \int_{\boldsymbol{S}_j(V_j \cap U_i)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \psi_i\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv$$

By Lemma 10.4.1, the assumptions of differentiability imply that the boundary points of Ω are always mapped to a set of measure zero so these can be neglected if desired. Now $\boldsymbol{S}_j(V_j\cap U_i)=\boldsymbol{S}_j\circ\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i(V_j\cap U_i))$ and so using 11.6, the above expression equals

$$\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(V_j\cap U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\cdot$$
$$\left|\det\left(D\left(\boldsymbol{R}_i\circ\boldsymbol{S}_j^{-1}\right)(\boldsymbol{v})\right)\right|J_i(\boldsymbol{u})\left|\det D\left(\boldsymbol{S}_j\circ\boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right|du$$

Now $I=\left(\boldsymbol{R}_i\circ\boldsymbol{S}_j^{-1}\right)\circ\left(\boldsymbol{S}_j\circ\boldsymbol{R}_i^{-1}\right)$ and so the chain rule implies that the product of the two Jacobians is 1. Hence 11.7 equals

$$\begin{aligned}
&\sum_{j=1}^{s}\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(V_j\cap U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du\\
=&\sum_{i=1}^{r}\sum_{j=1}^{s}\int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du\\
=&\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\sum_{j=1}^{s}\eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du\\
=&\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du=L(f)
\end{aligned}$$

Thus L is a well defined positive linear functional. ■

Definition 11.2.3 *By the representation theorem for positive linear functionals, Theorem 8.8.2, there exists a complete Radon measure σ_p defined on the Borel sets of Ω such that $Lf=\int_\Omega fd\sigma_p$. Then σ_p is what is meant by the measure on the differentiable manifold Ω.*

If O is an open set in Ω, what is $\sigma_p(O)$? Let $f_n\uparrow\mathscr{X}_O$ where f_n is continuous. Then by the monotone convergence theorem,

$$\begin{aligned}
\sigma_p(O) &= \lim_{n\to\infty}L(f_n)=\lim_{n\to\infty}\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(U_i)} f_n\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du\\
&= \lim_{n\to\infty}\sum_{i=1}^{r}\int_{\boldsymbol{R}_i(U_i\cap O)} f_n\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du\\
&= \sum_{i=1}^{r}\int_{\boldsymbol{R}_i(U_i\cap O)} \mathscr{X}_O\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du.
\end{aligned}$$

If K is a compact subset of some U_i, then use Corollary 10.8.5 to obtain a partition of unity which has $\psi_i=1$ on K so that all other ψ_j equal 0. Then

$$\int_\Omega\mathscr{X}_K d\sigma_p=\int_{\boldsymbol{R}_i(U_i)}\mathscr{X}_K\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du$$

It then follows from regularity of the measure and the monotone convergence theorem that if E is any measurable set contained in U_i, you can replace K in the above with E. In general, this implies that for nonnegative measurable f, having support in U_i,

$$\int_\Omega fd\sigma_p=\int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)J_i(\boldsymbol{u})\,du$$

Indeed, $\partial\Omega$ is a closed subset of Ω and so $\mathscr{X}_{\partial\Omega}$ is measurable. That part of the boundary contained in U_i would then involve a Lebesgue integral over a set of measure zero. This shows the following proposition.

Proposition 11.2.4 *Let Ω be a differentiable manifold as discussed above and let σ_p be the measure on the manifold defined above. Then $\sigma_p(\partial\Omega) = 0$.*

Note that it suffices in the above to assume only that $\boldsymbol{DR}_i^{-1}(\boldsymbol{u})$ exists for a.e. $\boldsymbol{u}$.

11.3 Divergence Theorem

The divergence theorem considered here will feature an open set in $\mathbb{R}^p$ whose boundary has a particular form. For convenience, if $\boldsymbol{x} \in \mathbb{R}^p, \hat{\boldsymbol{x}}_i \equiv \begin{pmatrix} x_1 & \cdots & x_{i-1} & x_{i+1} & \cdots & x_p \end{pmatrix}^T$.

Definition 11.3.1 *Let $U \subseteq \mathbb{R}^p$ satisfy the following conditions. There exist open boxes, $Q_1, \cdots, Q_N$, $Q_i = \prod_{j=1}^p \left(a_j^i, b_j^i\right)$ such that $\partial U \equiv \overline{U} \setminus U$ is contained in their union. Also, there exists an open set, Q_0 such that $Q_0 \subseteq \overline{Q_0} \subseteq U$ and $\overline{U} \subseteq Q_0 \cup Q_1 \cup \cdots \cup Q_N$. Assume for each Q_i, there exists k and a function g_i such that $U \cap Q_i$ is of the form*

$$\left\{ \begin{array}{c} \boldsymbol{x} : (x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \in \prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \\ \prod_{j=k+1}^p \left(a_j^i, b_j^i\right) \text{ and } a_k^i < x_k < g_i(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \end{array} \right\} \tag{11.8}$$

or else of the form

$$\left\{ \begin{array}{c} \boldsymbol{x} : (x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \in \prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \\ \prod_{j=k+1}^p \left(a_j^i, b_j^i\right) \text{ and } g_i(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) < x_k < b_j^i \end{array} \right\} \tag{11.9}$$

The function, g_i is differentiable and has a measurable partial derivatives on

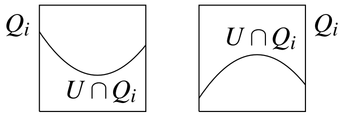
$$A_i \subseteq \prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \prod_{j=k+1}^{p} \left(a_j^i, b_j^i\right) \equiv \hat{Q}_k$$

where

$$m_{p-1}\left(\prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \prod_{j=k+1}^{p} \left(a_j^i, b_j^i\right) \setminus A_i \right) = 0.$$

and we assume there is a constant C such that for all i and j, $\left|\frac{\partial g_i}{\partial x_j}\right| \leq C$ off A_i and that each g_i is Lipschitz. Thus there are no measurability issues by Theorem 10.3.1.

To illustrate the above here is a picture.



Recall from calculus that if $z - g(\hat{\boldsymbol{x}}) = 0$ then to get a normal vector to the level surface, it will be $\pm$ the gradient.

Lemma 11.3.2 *Let $\alpha_1, \cdots, \alpha_p$ be real numbers and let $A(\alpha_1, \cdots, \alpha_p)$ be the matrix which has $1+\alpha_i^2$ in the ii^{th} slot and $\alpha_i \alpha_j$ in the ij^{th} slot when $i \neq j$. Then* $\det A = 1 + \sum_{i=1}^p \alpha_i^2$.

Proof of the claim: The matrix, $A(\alpha_1, \cdots, \alpha_p)$ is of the form

$$A(\alpha_1, \cdots, \alpha_p) = \begin{pmatrix} 1+\alpha_1^2 & \alpha_1\alpha_2 & \cdots & \alpha_1\alpha_p \\ \alpha_1\alpha_2 & 1+\alpha_2^2 & & \alpha_2\alpha_p \\ \vdots & & \ddots & \vdots \\ \alpha_1\alpha_p & \alpha_2\alpha_p & \cdots & 1+\alpha_p^2 \end{pmatrix}$$

Now consider the product of a matrix and its transpose, $B^T B$ below.

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & \alpha_1 \\ 0 & 1 & & 0 & \alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & \alpha_p \\ -\alpha_1 & -\alpha_2 & \cdots & -\alpha_p & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 & -\alpha_1 \\ 0 & 1 & & 0 & -\alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & -\alpha_p \\ \alpha_1 & \alpha_2 & \cdots & \alpha_p & 1 \end{pmatrix} \tag{11.10}$$

This product equals a matrix of the form

$$\begin{pmatrix} A(\alpha_1, \cdots, \alpha_p) & \mathbf{0} \\ \mathbf{0} & 1+\sum_{i=1}^p \alpha_i^2 \end{pmatrix}$$

Therefore, $\left(1+\sum_{i=1}^p \alpha_i^2\right) \det(A(\alpha_1, \cdots, \alpha_p)) = \det(B)^2 = \det\left(B^T\right)^2$. However, using row operations,

$$\det B^T = \det \begin{pmatrix} 1 & 0 & \cdots & 0 & \alpha_1 \\ 0 & 1 & & 0 & \alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & \alpha_p \\ 0 & 0 & \cdots & 0 & 1+\sum_{i=1}^p \alpha_i^2 \end{pmatrix} = 1 + \sum_{i=1}^p \alpha_i^2$$

and therefore,

$$\left(1+\sum_{i=1}^p \alpha_i^2\right) \det(A(\alpha_1, \cdots, \alpha_p)) = \left(1+\sum_{i=1}^p \alpha_i^2\right)^2$$

which shows $\det(A(\alpha_1, \cdots, \alpha_p)) = \left(1+\sum_{i=1}^p \alpha_i^2\right)$. ■

Now consider the case of σ on ∂U. The maps will be of the form

$$\hat{\boldsymbol{x}} \in Q_k \to \begin{pmatrix} x_1 & \cdots & x_{i-1} & g(\hat{\boldsymbol{x}}_i) & x_{i+1} & \cdots & x_p \end{pmatrix}^T = \boldsymbol{h}(\hat{\boldsymbol{x}}_i)$$

I need to describe $\det\left(D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)^* D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)\right)^{1/2} \equiv J(\hat{\boldsymbol{x}})$.

Consider an example sufficient to see what happens in general in which $p=3$ and $i=2$. Then in this case, $J(\hat{\boldsymbol{x}})$ will be the square root of the determinant of

$$\begin{pmatrix} 1 & g_{x_1} & 0 \\ 0 & g_{x_3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ g_{x_1} & g_{x_3} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} g_{x_1}^2+1 & g_{x_1}g_{x_3} \\ g_{x_1}g_{x_3} & g_{x_3}^2+1 \end{pmatrix}.$$

One can verify that this is just a special case in which $\boldsymbol{Dh}(\hat{\boldsymbol{x}}_i)^* \boldsymbol{Dh}(\hat{\boldsymbol{x}}_i)$ will be of the form considered in Lemma 11.3.2. Thus by this lemma, $J(\boldsymbol{x}) = \sqrt{1+\sum_{k\neq i} g_{,x_k}^2}$.

Then if $U\cap Q$ is of the form in 11.8 or in 11.9 one can identify the unit exterior normal to the surface either on the top or the bottom of $U\cap Q$ from beginning calculus. These are respectively

$$\boldsymbol{n} = \frac{\begin{pmatrix} -g_{,x_1} & \cdots & -g_{,x_{p-1}} & \cdots & 1 \end{pmatrix}^T}{\sqrt{1+\sum_{k=1}^{p-1} g_{,x_k}^2}}, \frac{\begin{pmatrix} g_{,x_1} & \cdots & g_{,x_{p-1}} & \cdots & -1 \end{pmatrix}^T}{\sqrt{1+\sum_{k=1}^{p-1} g_{,x_k}^2}}$$

The first pointing up away from U and the second pointing down away from U.

If you simply assume g_k is differentiable, there is no problem in Definition 11.3.1. One can show with Rademacher's theorem that it suffices to assume these functions are Lipschitz continuous.

In the following proof, I will regard $f(x_1,x_2,\ldots,x_p)$ as a function of the listed variables.

Definition 11.3.3 *Let $\boldsymbol{F}\in C^1\left(\overline{U};\mathbb{R}^p\right)$ and the rectangular coordinates are denoted as $\boldsymbol{x}=(x_1,\ldots,x_p)$. Then the divergence of $\boldsymbol{F}$ written as* $\operatorname{div}(\boldsymbol{F})$ *is defined as $\sum_i \frac{\partial F_i}{\partial x_i} \equiv \sum_i F_{i,i}$. It is also written as $\nabla\cdot\boldsymbol{F}$.*

Theorem 11.3.4 *Let U be a bounded open set in $\mathbb{R}^p$ satisfying the conditions of Definition 11.3.1 and let $\boldsymbol{F}\in C^1\left(\overline{U};\mathbb{R}^p\right)$. Then*

$$\int_U \sum_{i=1}^p F_{i,i}(\boldsymbol{x})\,dm_p = \int_{\partial U} \boldsymbol{F}\cdot\boldsymbol{n}\,d\sigma_{p-1}$$

where $\boldsymbol{n}$ is the unit exterior normal to U just described. $\sum_{i=1}^p F_{i,i}(\boldsymbol{x})$ is denoted $\nabla\cdot\boldsymbol{F}$. Thus this is written as $\int_U \nabla\cdot\boldsymbol{F}dm_p = \int_{\partial U}\boldsymbol{F}\cdot\boldsymbol{n}d\sigma_{p-1}$. Sometimes you see $\operatorname{div}(\boldsymbol{F})$ *in place of $\nabla\cdot\boldsymbol{F}$.*

Proof: Let $\operatorname{spt}(\psi_i)$ be a compact subset of Q_i and $\sum_{i=0}^N \psi_i = 1$ on $\bar{U}$ and each ψ_i is infinitely differentiable. This partition of unity exists by Lemma 10.8.4. There is an explicit description of the unit outer normal for each point of the boundary of U described above in either of the two cases described in Definition 11.3.1 and illustrated in the above picture. Then

$$\begin{aligned}\int_U \sum_i F_{i,i}(\boldsymbol{x})\,dm_p &= \int_U \sum_i \sum_{k=0}^N (\psi_k F)_{i,i}(\boldsymbol{x})\,dm_p = \sum_i \sum_{k=0}^N \int_U (\psi_k F)_{i,i}(\boldsymbol{x})\,dm_p \\ &= \sum_{k=0}^N \int_{Q_k} \sum_i (\psi_k F)_{i,i}(\boldsymbol{x})\,dm_p \qquad (11.11)\end{aligned}$$

Now consider one of the terms in the above. For the sake of simplicity assume $k=p$ so that the special direction corresponds to x_p. Also, I will assume that the function $g(\hat{\boldsymbol{x}})$ is on the top, so it is like the left picture in the above. A similar argument works if $g(\hat{\boldsymbol{x}})$ were on the bottom. Either way we can specify a unit exterior normal a.e. I will omit the subscript on g_k, Q_k, and ψ_k.

Case that $i<p$: Pick $i<p$. Letting $\hat{Q}$ be $(x_1,\ldots,x_{p-1})$ where $\boldsymbol{x}\in Q$, For any i,

$$\int_Q (\psi_k F)_{i,i}\,dm_p = \int_{\hat{Q}} \int_{-\infty}^{g(x_1,\ldots,x_{p-1})} (\psi_k F_i)_{,i}\,dx_p d\hat{x} = \int_{\hat{Q}} \int_{-\infty}^{0} D_i(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))\,dy d\hat{x} \qquad (11.12)$$

Now for $i < p$, that in the integrand is not $\frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}}))$. Indeed, by the chain rule,

$$\frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}})) = D_i(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}})) + D_p(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}}))\frac{\partial g(\hat{\boldsymbol{x}})}{\partial x_i}$$

Since $\operatorname{spt}(\psi) \subseteq Q$, it follows that 11.12 reduces to

$$\int_{-\infty}^{0}\int_{\hat{Q}}\frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}}))\,d\hat{x}dy - \int_{\hat{Q}}\int_{-\infty}^{0} D_p(\psi F_i)(\hat{\boldsymbol{x}}, y + g(\hat{\boldsymbol{x}}))\frac{\partial g(\hat{\boldsymbol{x}})}{\partial x_i}dyd\hat{x}$$

$$= 0 - \int_{\hat{Q}}(\psi F_i)(\hat{\boldsymbol{x}}, g(\hat{\boldsymbol{x}}))\,d\hat{x}$$

Case that $i = p$: In this case, 11.12 becomes $\int_{\hat{Q}}(\psi F_p)(\hat{\boldsymbol{x}}, g(\hat{\boldsymbol{x}}))\,d\hat{x}$. Recall how it was just shown that the unit normal is $\frac{\left(-g_{x_1}, \ldots, -g_{x_{p-1}}, 1\right)}{\sqrt{\sum_{i=1}^{p-1} g_{x_k}^2 + 1}}$ and $d\sigma = \sqrt{\sum_{i=1}^{p-1} g_{x_k}^2 + 1}dm_{p-1}$. Then the above reduces to $\int_{\partial(Q\cap U)}(\psi \boldsymbol{F})\cdot \boldsymbol{n}d\sigma$. The same result will hold for all the Q_i. The sign changes if in the situation of 11.9. As to $Q_0, \int_{Q_0}\sum_i(\psi_0 F)_{i,i}(\boldsymbol{x})\,dm_p = 0$ because $\operatorname{spt}(\psi_0) \subseteq Q_0$. Returning to 11.11, it follows that

$$\int_U \sum_i F_{i,i}(\boldsymbol{x})\,dm_p = \sum_{k=0}^{N}\int_{Q_k}\sum_i(\psi_k F)_{i,i}(\boldsymbol{x})\,dm_p = \sum_{k=0}^{N}\int_{Q_k}\sum_i(\psi_k F)_{i,i}(\boldsymbol{x})\,dm_p$$

$$= \sum_{k=1}^{N}\int_{\partial(Q_k\cap U)}(\psi_k \boldsymbol{F})\cdot \boldsymbol{n}d\sigma = \sum_{k=1}^{N}\int_{\partial U}(\psi_k \boldsymbol{F})\cdot \boldsymbol{n}d\sigma$$

$$= \int_{\partial U}\left(\sum_{k=0}^{N}\psi_k\right)\boldsymbol{F}\cdot \boldsymbol{n}d\sigma = \int_{\partial U}\boldsymbol{F}\cdot \boldsymbol{n}d\sigma \;\blacksquare$$

Definition 11.3.5 *The expression* $\sum_{i=1}^{p} F_{i,i}(\boldsymbol{x})$ *is called* $\operatorname{div}(\boldsymbol{F})$. *It is defined above in terms of the coordinates with respect to a fixed orthonormal basis* $(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p)$. *However, it does not depend on such a particular choice for coordinates.*

If you had some other orthonormal basis $(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_p)$ and if $(y_1, \cdots, y_p)$ are the coordinates of a point $\boldsymbol{z}$ with respect to this other orthonormal system, then there is an orthogonal matrix Q such that $\boldsymbol{y} = Q\boldsymbol{x}$ for $\boldsymbol{y}$ the coordinate vector for the new basis and $\boldsymbol{x}$ the coordinate vector for the old basis. Then

$$J_i(\boldsymbol{x}) \equiv \left(\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{x})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{x})\right)\right)^{1/2} = \left(\det\left(\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{y})Q\right)^* D\boldsymbol{R}_i^{-1}(\boldsymbol{y})Q\right)\right)^{1/2}$$

$$= \left(\det\left(Q^* D\boldsymbol{R}_i^{-1}(\boldsymbol{y})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{y})Q\right)\right)^{1/2} = \left(\det\left(Q^* D\boldsymbol{R}_i^{-1}(\boldsymbol{y})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{y})Q\right)\right)^{1/2} = J_i(\boldsymbol{y})$$

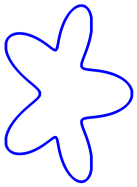
so the two definitions of $d\sigma$ will be the same with either set of coordinates.

List the $\boldsymbol{v}_i$ in the order which will give $\det(Q) = 1$. That is to say, the two bases have the same orientation.The insistence that $\det Q = 1$ will ensure that the unit normal vectors defined as above will point away from U. Thus we could take the divergence with respect to coordinates of any orthonormal basis having the same orientation. Note that for a.e. geometric point $\boldsymbol{z}$

$$\operatorname{div}(\boldsymbol{F})(\boldsymbol{z}) = \lim_{r\to 0} \frac{1}{m_p(B(\boldsymbol{z},r))} \int_{B(\boldsymbol{z},r)} \operatorname{div}(\boldsymbol{F})\, dm_p = \lim_{r\to 0} \frac{1}{m_p(B(\boldsymbol{z},r))} \int_{\partial B(\boldsymbol{z},r)} \boldsymbol{F}\cdot\boldsymbol{n}\, d\sigma_{p-1}$$

the first equal sign from the fundamental theorem of calculus and the last expression on the right being independent of the choice of basis. This implies that we could have generalized the kind of region to be one for which the little rectangles are allowed to be slanted. Creases and pointy places in the manifold can result from places where some $J_i(\boldsymbol{x}) = 0$, due to some $D\boldsymbol{R}_i^{-1}$ not being one to one, but this will not matter because in the definition of the surface measure this will be a set of measure zero on the manifold. The change of variables formula which was so important in the above argument is unaffected by these creases.

Globally the region could be quite complicated. As an example in two dimensions, it might look like this:



Corollary 11.3.6 *If the divergence is computed with respect to $\boldsymbol{y}$ where $\boldsymbol{y} = Q\boldsymbol{x}$ for Q orthogonal with determinant 1, and each box used in the argument of Theorem 11.3.4 is taken with respect to such a new basis $(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_p)$, then one still obtains $\int_U \operatorname{div}(\boldsymbol{F})\, dm_p = \int_{\partial U} \boldsymbol{F}\cdot\boldsymbol{n}\, d\sigma_{p-1}$.*

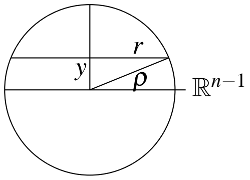
11.4 Volumes of Balls in $\mathbb{R}^p$

This short section will give an explicit description of surface area given in Section 10.9.

Recall, $B(\boldsymbol{x},r)$ denotes the set of all $\boldsymbol{y} \in \mathbb{R}^p$ such that $|\boldsymbol{y}-\boldsymbol{x}| < r$. By the change of variables formula for multiple integrals or simple geometric reasoning, all balls of radius r have the same volume. Furthermore, simple reasoning or change of variables formula will show that the volume of the ball of radius r equals $\alpha_p r^p$ where α_p will denote the volume of the unit ball in $\mathbb{R}^p$. With the divergence theorem, it is now easy to give a simple relationship between the surface area of the ball of radius r and the volume. Let $d\alpha_{p-1}$ be the area measure above. By the divergence theorem, $\int_{B(\mathbf{0},r)} \overset{p}{\operatorname{div}}(\boldsymbol{x}) dx = \int_{\partial B(\mathbf{0},r)} \boldsymbol{x}\cdot\frac{\boldsymbol{x}}{|\boldsymbol{x}|} d\alpha_{p-1}$ because the unit outward normal on $\partial B(\mathbf{0},r)$ is $\frac{\boldsymbol{x}}{|\boldsymbol{x}|}$. Therefore, $p\alpha_p r^p = r\alpha_{p-1}(\partial B(\mathbf{0},r))$ and so $\alpha_{p-1}(\partial B(\mathbf{0},r)) = p\alpha_p r^{p-1}$.

Let ω_p denote the area of the sphere $S^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| = 1\}$. I just showed that $\omega_p = p\alpha_p$.

I want to find α_p now.



Taking slices at height y as shown and using that these slices have $p-1$ dimensional area equal to $\alpha_{p-1} r^{p-1}$, it follows $\alpha_p \rho^p = 2\int_0^\rho \alpha_{p-1}\left(\rho^2 - y^2\right)^{(p-1)/2} dy$ since the r at a given y is $\sqrt{\rho^2 - y^2}$. In the integral, change variables, letting $y = \rho\cos\theta$. Then $\alpha_p \rho^p = 2\rho^p \alpha_{p-1}\int_0^{\pi/2} \sin^p(\theta)\, d\theta$. It follows that

$$\alpha_p = 2\alpha_{p-1}\int_0^{\pi/2} \sin^p(\theta)\, d\theta. \tag{11.13}$$

From this we find a formula for α_p.

First note that $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-t}t^{-1/2}dt = \int_0^\infty e^{-u^2}u^{-1}2udu = 2\int_0^\infty e^{-u^2} = \sqrt{\pi}$ from elementary calculus using polar coordinates and change of variables.

Theorem 11.4.1 *$\alpha_p = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}$ where Γ denotes the gamma function, defined for $\alpha > 0$ by $\Gamma(\alpha) \equiv \int_0^\infty e^{-t}t^{\alpha-1}dt$.*

Proof: Let $p = 1$ first. Then $\alpha_1 = \pi = \frac{\pi^{1/2}}{\Gamma\left(\frac{1}{2}+1\right)}$ because $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ so the right side is $\frac{\pi^{1/2}}{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)} = 2$ which is indeed the one dimensional area of the unit ball in one dimension. Similarly it is true for $p = 2,3$. Assume true for $p \geq 3$. Then using 11.13 and induction,

$$\alpha_{p+1} = 2\overset{\alpha_p}{\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}}\int_0^{\pi/2} \sin^{p+1}(\theta)\, d\theta$$

Using an integration by parts, this equals $2\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\int_0^{\pi/2}\sin^{p-1}(\theta)\, d\theta$. By 11.13 and induction this is

$$\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\frac{\alpha_{p-1}}{\alpha_{p-2}} = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\frac{\frac{\pi^{(p-1)/2}}{\Gamma\left(\frac{p-1}{2}+1\right)}}{\frac{\pi^{(p-2)/2}}{\Gamma\left(\frac{p-2}{2}+1\right)}} = \frac{2\pi^{(p+1)/2}\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p}{2}+1\right)\Gamma\left(\frac{p-1}{2}+1\right)}\frac{p/2}{p+1}$$

$$= \frac{2\pi^{(p+1)/2}\Gamma\left(\frac{p}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)\Gamma\left(\frac{p-1}{2}+1\right)}\frac{1}{p+1} = \frac{\pi^{(p+1)/2}}{\Gamma\left(\frac{p+1}{2}\right)}\frac{1}{\frac{p+1}{2}} = \frac{\pi^{(p+1)/2}}{\Gamma\left(\frac{p+1}{2}+1\right)} \blacksquare$$

There is a general treatment of Stoke's theorem which involves differential forms. This is developed in my book on the web site, "Real and Abstract Analysis". I have given a fairly general version of the divergence theorem above. However, I have chosen here to deal with the versions of these theorems which are of most use in applications. These are known as Green's theorem and Stokes theorem in Calculus and they are the original forms of these theorems, not that algebraic extravaganza involving differential forms which is in the other book. I prefer seeing this other approach in terms of the area formula and Hausdorff measures which are not topics in this book.

11.5 Space Curves and Line Integrals

Here is a short discussion of line integrals and space curves.

Definition 11.5.1 *Let $[a,b)$ be a half open interval and let $\mathbf{r} : [a,b) \to \mathbb{R}^q$ be one to one and continuous such that $\mathbf{r} \in C\left(\overline{[a,b)};\mathbb{R}^q\right)$ and $\mathbf{r}^{-1} : \mathbf{r}([a,b)) \to [a,b)$ is continuous.*

If $\boldsymbol{r}$ is the restriction of a C^1 function to $[a,b)$ then it is called a C^1 curve. It is called a simple closed curve if $\lim_{t\to b}\boldsymbol{r}(t)=\boldsymbol{r}(a)$ *and we can define* $\boldsymbol{r}(b)\equiv\boldsymbol{r}(a)$. *This $\boldsymbol{r}$ plays the role of $\boldsymbol{R}^{-1}$ in the above and so C is a one dimensional manifold in $\mathbb{R}^q$. Here we have just one chart in the atlas. If $\hat{\boldsymbol{r}}:[c,d)$ is defined similarly mapping to C such that $\hat{\boldsymbol{r}}^{-1}\circ\boldsymbol{r}$ is increasing, this would be an equivalent atlas and so from the above general presentation, if this transition map is differentiable we could use either one to describe the one dimensional measure on C. Also, since $\hat{\boldsymbol{r}}^{-1}\circ\boldsymbol{r}$ is one to one, this function is either increasing or decreasing. To say that its derivative is nonnegative is to say that the two go over C in the same direction and deliver the same orientation. These functions are called parametrizations in this context. It is piecewise smooth if $\boldsymbol{r}'\neq\boldsymbol{0}$ on a succession of non-overlapping intervals whose union is $[a,b)$ but possibly at the ends of these intervals the derivative from left and right are different.*

Suppose you have $a<b<c$ and $\boldsymbol{r}_1'(t)\neq 0$ on $[a,b]$, $\boldsymbol{r}_2'(t)\neq 0$ on $[b,c]$ but $\boldsymbol{r}_1'(b)\neq\boldsymbol{r}_2'(b)$ although $\boldsymbol{r}_1=\boldsymbol{r}_2$ at b. Then consider $\hat{\boldsymbol{r}}(t)=\begin{cases}\boldsymbol{r}_1\left(b+(t-b)^3\frac{1}{(b-a)^2}\right), t\in[a,b]\\ \boldsymbol{r}_2\left(b+(t-b)^3\frac{1}{(c-b)^2}\right), t\in[b,c]\end{cases}$. Then $\hat{\boldsymbol{r}}(t)$ moves from $\boldsymbol{r}_1(a)$ to $\hat{\boldsymbol{r}}(b)$ in the same direction as $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ and is differentiable on all of $[a,c]$ although $\hat{\boldsymbol{r}}'(b)=0$. Thus this piecewise smooth curve can be expressed as a C^1 curve not smooth because of the vanishing of the derivative at b.

Lemma 11.5.2 *If $\boldsymbol{r}:[a,b)\to\mathbb{R}^q$ is a piecewise smooth curve smooth on successive non-overlapping intervals. Then there exists a C^1 space curve which has the same orientation, parametrizing the curve.*

Therefore a C^1 curve, as defined above, includes the case of curves which are piecewise smooth because modifying the parameter we can have finitely many t with $\boldsymbol{r}'(t)=0$ to account for pointed places which have a discontinuity in the derivative from either side, allowing the inclusion of piecewise smooth curves as a special case. Then as a case of the above theory if f is Borel measurable or continuous defined on C, then for a particular orientation of C,

$$\int_C f\,d\sigma=\int_a^b f(\boldsymbol{r}(t))\left(\det\left(\boldsymbol{r}'(t)^*\boldsymbol{r}'(t)\right)\right)^{1/2}dt=\int_C f(\boldsymbol{r}(t))\left|\boldsymbol{r}'(t)\right|dt$$

where $\boldsymbol{r}'(t)\cdot\boldsymbol{r}'(t)=\det\left(\boldsymbol{r}'(t)^*\boldsymbol{r}'(t)\right)$. Indeed, $\boldsymbol{r}'(t)^*\boldsymbol{r}'(t)$ is just a nonnegative number. Now a unit tangent vector to the curve consistent with its orientation is $\boldsymbol{r}'(t)/|\boldsymbol{r}'(t)|$ and at a point of C the same unit tangent vector would be obtained from another equivalent parametrization. For a piecewise smooth curve, there will be finitely many points where this tangent vector will not be defined but this is of no importance here since the interest is in an integral. For $\boldsymbol{f}(\boldsymbol{x})$ a vector valued function for each $\boldsymbol{x}\in C$ where the components are Borel measurable, the line integral $\int_C\boldsymbol{f}\cdot d\boldsymbol{r}$ is defined as

$$\int_a^b\boldsymbol{f}(\boldsymbol{r}(t))\cdot\frac{\boldsymbol{r}'(t)}{|\boldsymbol{r}'(t)|}\left|\boldsymbol{r}'(t)\right|dt=\int_a^b\boldsymbol{f}(\boldsymbol{r}(t))\cdot\boldsymbol{r}'(t)\,dt$$

and the above discussion shows that this is independent of the choice of parametrization, having the same orientation.

Notation 11.5.3 *It is customary to write the line integral $\int_C \boldsymbol{f}\cdot d\boldsymbol{r}$ as*

$$\int_C f_1(\boldsymbol{x})\,dx_1 + f_2(\boldsymbol{x})\,dx_2 + \cdots + f_q(\boldsymbol{x})\,dx_q$$

which is called differential form notation.

11.6 Exercises

1. A random vector $\boldsymbol{X},$ with values in $\mathbb{R}^p$ has a multivariate normal distribution written as $\boldsymbol{X} \sim N_p(\boldsymbol{m},\Sigma)$ if for all Borel $E \subseteq \mathbb{R}^p$,
$$\lambda_{\boldsymbol{X}}(E) = \int_{\mathbb{R}^p} \mathscr{X}_E(\boldsymbol{x}) \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dm_p$$
Here Σ is a positive definite symmetric matrix. Recall that $\lambda_{\boldsymbol{X}}(E) \equiv P(\boldsymbol{X} \in E)$. Using the change of variables formula, show that $\lambda_{\boldsymbol{X}}$ defined above is a probability measure. One thing you must show is that
$$\int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dm_p = 1$$
Hint: To do this, you might use the fact from linear algebra that $\Sigma = Q^*DQ$ where D is a diagonal matrix and Q is an orthogonal matrix. Thus $\Sigma^{-1} = Q^*D^{-1}Q$. Maybe you could first let $\boldsymbol{y} = D^{-1/2}Q(\boldsymbol{x}-\boldsymbol{m})$ and change the variables. Note that the change of variables formula works fine when the open sets are all of $\mathbb{R}^p$. You don't need to confine your attention to finite open sets which would be the case with Riemann integrals which are only defined on bounded sets.

2. Consider the surface $z = x^2$ for $(x,y) \in (0,1)\times(0,1)$. Find the area of this surface. **Hint:** You can make do with just one chart in this case. Let $\boldsymbol{R}^{-1}(x,y) = \left(x,y,x^2\right)^T, (x,y) \in (0,1)\times(0,1)$. Then
$$D\boldsymbol{R}^{-1} = \begin{pmatrix} 1 & 0 & 2x \\ 0 & 1 & 0 \end{pmatrix}^T$$
It follows that $D\boldsymbol{R}^{-1*}D\boldsymbol{R}^{-1} = \begin{pmatrix} 4x^2+1 & 0 \\ 0 & 1 \end{pmatrix}$.

3. A parametrization for most of the sphere of radius $a > 0$ in three dimensions is $x = a\sin(\phi)\cos(\theta), y = a\sin(\phi)\sin(\theta), z = a\cos(\phi)$. where we will let $\phi \in (0,\pi), \theta \in (0,2\pi)$ so there is just one chart involved. As mentioned earlier, this includes all of the sphere except for the line of longitude corresponding to $\theta = 0$. Find a formula for the area of this sphere. Again, we are making do with a single chart.

4. Let V be such that the divergence theorem holds. Show that $\int_V \nabla\cdot(v\nabla u)\,dV = \int_{\partial V} v\frac{\partial u}{\partial n}\,dA$ where $\boldsymbol{n}$ is the exterior normal. Here $\frac{\partial u}{\partial n} \equiv \nabla u\cdot\boldsymbol{n}$.

5. To prove the divergence theorem, it was shown first that the spacial partial derivative in the volume integral could be exchanged for multiplication by an appropriate component of the exterior normal. This problem starts with the divergence theorem and goes the other direction. Assuming the divergence theorem, holds for a region V, show that $\int_{\partial V}\boldsymbol{n}u\,dA = \int_V \nabla u\,dV$. Note this implies $\int_V \frac{\partial u}{\partial x}\,dV = \int_{\partial V} n_1 u\,dA$.

6. Fick's law for diffusion states the flux of a diffusing species, $\boldsymbol{J}$ is proportional to the gradient of the concentration c. Write this law getting the sign right for the constant of proportionality and derive an equation similar to the heat equation for the concentration c. Typically, c is the concentration of some sort of pollutant or a chemical.

7. Sometimes people consider diffusion in materials which are not homogeneous. This means that $\boldsymbol{J} = -K\nabla c$ where K is a 3×3 matrix and c is called the concentration. Thus in terms of components, $J_i = -\sum_j K_{ij}\frac{\partial c}{\partial x_j}$. Here c is the concentration which means the amount of pollutant or whatever is diffusing in a volume is obtained by integrating c over the volume. Derive a formula for a nonhomogeneous model of diffusion based on the above.

8. Let V be a ball and suppose $\nabla^2 u = f$ in V while $u = g$ on ∂V. Show that there is at most one solution to this boundary value problem which is C^2 in V and continuous on V with its boundary. **Hint:** You might consider $w = u - v$ where u and v are solutions to the problem. Then use the result of Problem 4 and the identity $w\nabla^2 w = \nabla\cdot(w\nabla w) - \nabla w\cdot\nabla w$ to conclude $\nabla w = 0$. Then show this implies w must be a constant by considering $h(t) = w(t\boldsymbol{x} + (1-t)\boldsymbol{y})$ and showing h is a constant.

9. Show that $\int_{\partial V}\nabla\times\boldsymbol{v}\cdot\boldsymbol{n}\,dA = 0$ where V is a region for which the divergence theorem holds and $\boldsymbol{v}$ is a C^2 vector field.

10. Let $\boldsymbol{F}(x,y,z) = (x,y,z)$ be a vector field in $\mathbb{R}^3$ and let V be a three dimensional shape and let $\boldsymbol{n} = (n_1,n_2,n_3)$. Show that $\int_{\partial V}(xn_1 + yn_2 + zn_3)\,dA = 3\times$ volume of V.

11. Let $\boldsymbol{F} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$ and let V denote the tetrahedron formed by the planes, $x = 0, y = 0, z = 0$, and $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{5}z = 1$. Verify the divergence theorem for this example.

12. Suppose $f : U \to \mathbb{R}$ is continuous where U is some open set and for all $B \subseteq U$ where B is a ball, $\int_B f(\boldsymbol{x})\,dV = 0$. Show that this implies $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}\in U$.

13. Let U denote the box centered at $(0,0,0)$ with sides parallel to the coordinate planes which has width 4, length 2 and height 3. Find the flux integral $\int_{\partial U}\boldsymbol{F}\cdot\boldsymbol{n}dS$ where $\boldsymbol{F} = (x+3, 2y, 3z)$. **Hint:** If you like, you might want to use the divergence theorem.

14. Find the flux out of the cylinder whose base is $x^2 + y^2 \le 1$ which has height 2 of the vector field $\boldsymbol{F} = \left(xy, zy, z^2 + x\right)$. The flux is the surface integral in the divergence theorem.

15. Find the flux out of the ball of radius 4 centered at $\mathbf{0}$ of the vector field $\boldsymbol{F} = (x, zy, z + x)$.

16. In one dimension, the heat equation is of the form $u_t = \alpha u_{xx}$. Show that $u(x,t) = e^{-\alpha n^2 t}\sin(nx)$ satisfies the heat equation

17. The contour integral for C, an oriented piecewise smooth in $\mathbb{C} = \mathbb{R}^2$, written as $\int_C f(z)\,dz$ is defined as $\int_a^b f(z(t))\,z'(t)\,dt$ where we can take $z(t)$ to be a C^1 curve which might vanish at finitely many points. Here $f(z) = u(x,y) + iv(x,y)$ for $z = x + iy$. Show $f(z)z' = ux' - vy' + i(vx' + uy')$. Show that the real part of this contour integral is $\int_C(u,-v)\cdot d\boldsymbol{r}$ and the imaginary part is $\int_C(v,u)\cdot d\boldsymbol{r}$. Thus, contour integrals, important in complex analysis, reduce to the consideration of line integrals in $\mathbb{R}^2$.

18. You have a ball B in three dimensions and there is a material of some sort having density ρ moving through this ball. Then if $\boldsymbol{v}$ is the velocity of the material, it being a function of $\boldsymbol{x}$ and t, the rate at which the material leaves B is $\int_{\partial B} \rho \boldsymbol{v} \cdot \boldsymbol{n} d\sigma$ where $\boldsymbol{n}$ is the exterior normal. Thus the rate at which material enters B is $-\int_{\partial B} \rho \boldsymbol{v} \cdot \boldsymbol{n} d\sigma$. This must equal $\frac{d}{dt} \int_B \rho dm_3$ m_3 being Lebesque measure. Explain why it is reasonable to expect that $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0$. Here $\nabla \cdot$ signifies the divergence. In general, $\nabla \cdot \boldsymbol{f}$ means $\sum_{i=1}^n \frac{\partial f_i}{\partial x_i}$. Thus $\nabla \cdot (\rho \boldsymbol{v})$ means $\sum_{i=1}^3 (\rho \boldsymbol{v})_{i,i}$. This is called mass balance. See "Calculus of One and Many Variables" to see many other examples of similar uses of the divergence theorem to physical models.

19. Let V be such that the divergence theorem holds. Show that for $\nabla^2 = \Delta$
$$\int_V (v\Delta u - u\Delta v)\, dV = \int_{\partial V} \left(v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n} \right) d\sigma$$
where $\boldsymbol{n}$ is the exterior normal and $\frac{\partial u}{\partial n}$ is defined in Problem 4. Here $\nabla^2 u \equiv \sum_i u_{,x_i x_i} \equiv \Delta u$. **Hint:** Show that $\nabla \cdot g\boldsymbol{f} = \nabla g \cdot \boldsymbol{f} + g\nabla \cdot \boldsymbol{f}$. Use for $g = v$ and $\boldsymbol{f} = \nabla u$ so $\nabla \cdot \boldsymbol{f} = \nabla \cdot \nabla v = \Delta v$.

11.7 Harmonic Functions

I am going to give a brief presentation on harmonic functions. To begin with, this features a ball of radius r centered at $\mathbf{0}$. I will refer to this ball as U to save notation. Later U will be allowed to be more general. Δ_y will indicate partial derivatives are taken with respect to the y_i components of $\boldsymbol{y}$. Also, for $\boldsymbol{x}, \boldsymbol{y} \in U$, I will use the following definition. For more on these topics see [16].

Definition 11.7.1 $r^{\boldsymbol{x}}(\boldsymbol{y}) \equiv \begin{cases} |\boldsymbol{y}-\boldsymbol{x}|^{-(n-2)} \text{ for } n \geq 3 \\ \ln|\boldsymbol{y}-\boldsymbol{x}| \text{ for } n = 2 \end{cases}$ *, and also define*
$$\psi^{\boldsymbol{x}}(\boldsymbol{y}) \equiv \begin{cases} \left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right|^{-(n-2)}, \ r^{-(n-2)} \text{ for } \boldsymbol{x} = \mathbf{0} \text{ if } n \geq 3 \\ \ln\left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right|, \ \ln(r) \text{ if } \boldsymbol{x} = \mathbf{0} \text{ if } n = 2 \end{cases}$$

Lemma 11.7.2 *The following hold.*

1. *When* $|\boldsymbol{y}| = r$ *and* $\boldsymbol{x} \neq \mathbf{0}, \psi^{\boldsymbol{x}}(\boldsymbol{y}) = r^{\boldsymbol{x}}(\boldsymbol{y})$.
2. $\left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right| \neq 0$ *if* $|\boldsymbol{y}| < r, \boldsymbol{x} \neq \mathbf{0}$.
3. $\lim_{\boldsymbol{x}\to\mathbf{0}} \left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right| = r$.
4. *If a is a real number and* $\boldsymbol{b}$ *a vector,* $\nabla_y |a\boldsymbol{y} + \boldsymbol{b}| = a\frac{a\boldsymbol{y}+\boldsymbol{b}}{|a\boldsymbol{y}+\boldsymbol{b}|}$.
5. $\Delta\psi^{\boldsymbol{x}} = 0 = \Delta r^{\boldsymbol{x}}$.

Proof: 1.) For 1., $\left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right|^2 = |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 - 2(\boldsymbol{x}\cdot\boldsymbol{y}) = |\boldsymbol{x}-\boldsymbol{y}|^2$ because $|\boldsymbol{y}| = r$.Thus $\psi^{\boldsymbol{x}}(\boldsymbol{y}) = r^{\boldsymbol{x}}(\boldsymbol{y})$.

2.) If $\left| \frac{\boldsymbol{y}|\boldsymbol{x}|}{r} - \frac{r\boldsymbol{x}}{|\boldsymbol{x}|} \right| = 0$ then $\boldsymbol{y}|\boldsymbol{x}|^2 = r^2\boldsymbol{x}$ which cannot happen if $|\boldsymbol{y}| < r, \boldsymbol{x} \neq \mathbf{0}$.

3.) $\lim_{\boldsymbol{x}\to\boldsymbol{0}}\left|\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|^2=\lim_{\boldsymbol{x}\to\boldsymbol{0}}\left(\frac{|\boldsymbol{x}|^2}{r^2}|\boldsymbol{y}|^2+r^2-2(\boldsymbol{x}\cdot\boldsymbol{y})\right)=r^2.$

4.)$|a\boldsymbol{y}+\boldsymbol{b}|=\left(a^2|\boldsymbol{y}|^2+2a(\boldsymbol{y}\cdot\boldsymbol{b})+|\boldsymbol{b}|^2\right)^{1/2}$ and so

$$\nabla_y|a\boldsymbol{y}+\boldsymbol{b}|=\frac{1}{2}\left(|a\boldsymbol{y}+\boldsymbol{b}|^2\right)^{-1/2}\left(2a^2y_1+2ab_1,...,2a^2y_n+2ab_n\right)$$

$$=\frac{1}{|a\boldsymbol{y}+\boldsymbol{b}|}\left(a^2y_1+ab_1,...,a^2y_n+ab_n\right)=\frac{1}{|a\boldsymbol{y}+\boldsymbol{b}|}\left(a^2\boldsymbol{y}+a\boldsymbol{b}\right)=a\frac{a\boldsymbol{y}+\boldsymbol{b}}{|a\boldsymbol{y}+\boldsymbol{b}|}$$

5.) Say $n>2$. The other case will be similar. From 4.) and the chain rule,

$$\nabla_y\left(|a\boldsymbol{y}+\boldsymbol{b}|^{-(n-2)}\right)=-(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-(n-1)}\nabla_y|a\boldsymbol{y}+\boldsymbol{b}|$$

$$=-(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-(n-1)}a\frac{a\boldsymbol{y}+\boldsymbol{b}}{|a\boldsymbol{y}+\boldsymbol{b}|}=-(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-n}a(a\boldsymbol{y}+\boldsymbol{b})$$

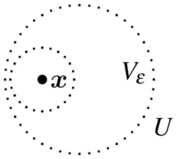
Then $\nabla_y\cdot\nabla_y\left(|a\boldsymbol{y}+\boldsymbol{b}|^{-(n-2)}\right)=$

$$n(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-(n+1)}\nabla_y|a\boldsymbol{y}+\boldsymbol{b}|\cdot a(a\boldsymbol{y}+\boldsymbol{b})-a(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-n}\nabla_y\cdot(a\boldsymbol{y}+\boldsymbol{b})$$

$$\begin{aligned}&= n(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-(n+1)}a\frac{a\boldsymbol{y}+\boldsymbol{b}}{|a\boldsymbol{y}+\boldsymbol{b}|}\cdot a(a\boldsymbol{y}+\boldsymbol{b})-a(n-2)|a\boldsymbol{y}+\boldsymbol{b}|^{-n}an\\ &= n(n-2)a^2|a\boldsymbol{y}+\boldsymbol{b}|^{-(n+1)}\frac{|a\boldsymbol{y}+\boldsymbol{b}|^2}{|a\boldsymbol{y}+\boldsymbol{b}|}-a^2(n-2)n|a\boldsymbol{y}+\boldsymbol{b}|^{-n}=0\end{aligned}$$

In case $n=2$ it works similarly. Thus if $\boldsymbol{x}\neq\boldsymbol{0},\Delta\psi^{\boldsymbol{x}}=0=\Delta r^{\boldsymbol{x}}$. In case $\boldsymbol{x}=\boldsymbol{0}$ and $\boldsymbol{b}$ might not be defined, there is nothing to show because $\psi^{\boldsymbol{0}}$ is a constant. This shows 5.) since both functions fit into the above situation. ■

Now let $B(\boldsymbol{x},\varepsilon)\equiv B_\varepsilon$ be a small ball inside U and let V_ε be the region between them.



Note that when $|\boldsymbol{y}|=r$,

$$\left|\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|^2=\frac{r^2|\boldsymbol{x}|^2}{r^2}+\frac{r^2|\boldsymbol{x}|^2}{|\boldsymbol{x}|^2}-2\boldsymbol{y}\cdot\boldsymbol{x}=|\boldsymbol{y}-\boldsymbol{x}|^2.$$

Lemma 11.7.3 *Let $v(\boldsymbol{y})\equiv r^{\boldsymbol{x}}(\boldsymbol{y})-\psi^{\boldsymbol{x}}(\boldsymbol{y})$. Thus $v=0$ on ∂U. Then $\nabla_y v\cdot\boldsymbol{n}\equiv\frac{\partial v}{\partial n}=\frac{(n-2)}{r}\frac{|\boldsymbol{x}|^2-r^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}$ if $n>2$ and $\frac{\partial v}{\partial n}=\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^2}$ if $n=2$. This is on $\partial B(\boldsymbol{0},r)$. $\left|\frac{\partial\psi^{\boldsymbol{x}}}{\partial n}\right|$ is uniformly bounded on ∂B_ε for all ε small enough.*

Proof: Say $n>2$ first. Then from Lemma 11.7.2 above, for $|\boldsymbol{y}|=r, \nabla_y v(\boldsymbol{y})=$

$$
\begin{aligned}
& -(n-2)\,|\boldsymbol{y}-\boldsymbol{x}|^{-(n-1)}\,\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|}-\left(-(n-2)\,|\boldsymbol{y}-\boldsymbol{x}|^{-(n-1)}\,\frac{|\boldsymbol{x}|}{r}\,\frac{\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}}{\left|\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|}\right) \\
= \;& -(n-2)\,\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}-\left(-(n-2)\,\frac{\frac{\boldsymbol{y}|\boldsymbol{x}|^2}{r^2}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}\right) \\
= \;& (n-2)\left(\frac{\frac{\boldsymbol{y}|\boldsymbol{x}|^2}{r^2}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}-\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}\right)=(n-2)\left(\frac{\frac{\boldsymbol{y}|\boldsymbol{x}|^2}{r^2}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}-\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}\right)
\end{aligned}
$$

The unit outer normal is $\frac{\boldsymbol{y}}{|\boldsymbol{y}|}=\frac{\boldsymbol{y}}{r}$ on ∂U and so dotting with this we get

$$
\begin{aligned}
\frac{\partial v}{\partial n} &= (n-2)\left(\frac{\frac{r^2|\boldsymbol{x}|^2}{r^3}-\frac{1}{r}\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}-\frac{r-\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{r}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}\right)=(n-2)\,\frac{1}{r}\left(\frac{\frac{r^2|\boldsymbol{x}|^2}{r^2}}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}-\frac{r^2}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}\right) \\
&= (n-2)\,\frac{1}{r}\,\frac{|\boldsymbol{x}|^2-r^2}{|\boldsymbol{y}-\boldsymbol{x}|^{n}}
\end{aligned}
$$

The case where $\boldsymbol{x}=\boldsymbol{0}$, $r_n^{\boldsymbol{0}}(\boldsymbol{y})-\psi^{\boldsymbol{0}}(\boldsymbol{y})=|\boldsymbol{y}|^{-(n-2)}-r^{-n-2}$. Thus

$$\nabla_y v(\boldsymbol{y})=-(n-2)\,|\boldsymbol{y}|^{-(n-1)}\,\boldsymbol{y}$$

so taking the dot product with $\boldsymbol{y}/r$ gives $\frac{\partial v}{\partial n}=\frac{-(n-2)}{r|\boldsymbol{y}|^{n}}$ which is the desired formula in case $\boldsymbol{x}=\boldsymbol{0}$.

In case $n=2$, It works exactly the same but in this case you get $\frac{dv}{dn}=\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^2}$.

Now consider the claim about $\frac{\partial\psi^{\boldsymbol{x}}}{\partial n}$ on ∂B_ε. Here $\boldsymbol{n}=\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|}$. In case $\boldsymbol{x}=\boldsymbol{0}$ there is nothing to show because all partials equal 0 in this case since $\psi^{\boldsymbol{0}}=r^{n-2}$ or $\ln(r)$. So assume $\boldsymbol{x}\neq\boldsymbol{0}$. First let $n>2$. From Lemma 11.7.2,

$$
\begin{aligned}
\nabla\psi^{\boldsymbol{x}}(\boldsymbol{y}) &= -(n-2)\left|\frac{\boldsymbol{y}\,|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|^{-(n-1)}\frac{|\boldsymbol{x}|}{r}\,\frac{\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}}{\left|\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|} \\
|\nabla\psi^{\boldsymbol{x}}(\boldsymbol{y})| &\leq \frac{(n-2)\,|\boldsymbol{x}|}{r}\left|\frac{\boldsymbol{y}\,|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|^{-(n-1)}
\end{aligned}
$$

Now $\boldsymbol{y}\to\left|\frac{\boldsymbol{y}|\boldsymbol{x}|}{r}-\frac{r\boldsymbol{x}}{|\boldsymbol{x}|}\right|^{-(n-1)}$ is continuous, bounded, and nonzero on $\bar{B}_\varepsilon$ for all ε sufficiently small and so $\left|\frac{\partial\psi^{\boldsymbol{x}}}{\partial n}\right|$ is uniformly bounded on ∂B_ε for all ε small enough. It works the same for $n=2$. ■

Next I want to represent solutions to $\Delta u=0, u=g$ on $\partial B(\boldsymbol{0},r)$ where g is some continuous function and $u\in C^2\left(\overline{U}\right)$. This will use Problem 19 on Page 277 on V_ε also the above lemmas.

$$\lim_{\varepsilon\to 0}\int_{V_\varepsilon}(u\Delta v-v\Delta u)\,dm_n=\lim_{\varepsilon\to 0}\int_{\partial V_\varepsilon}u\frac{\partial v}{\partial n}-v\frac{\partial u}{\partial n}d\sigma \tag{11.14}$$

$$= \int_{\partial U} g\frac{\partial v}{\partial n} - \overset{=0}{v}\frac{\partial u}{\partial n} d\sigma - \lim_{\varepsilon \to 0}\int_{\partial B_\varepsilon}\left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right)d\sigma$$

$$0 = \int_{\partial U} g\frac{\partial v}{\partial n}d\sigma - \lim_{\varepsilon \to 0}\int_{\partial B_\varepsilon}\left(u\frac{\partial r_n}{\partial n} - r_n\frac{\partial u}{\partial n}\right)d\sigma$$

Now $r^{\boldsymbol{x}}\frac{\partial u}{\partial n}$ is bounded uniformly on ∂B_ε for small ε and $\frac{\partial r^{\boldsymbol{x}}}{\partial n} = \frac{-(n-2)}{|\boldsymbol{y}-\boldsymbol{x}|^{n-1}}\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|}\cdot\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|} = \frac{-(n-2)}{|\boldsymbol{y}-\boldsymbol{x}|^{n-1}}$

$$\begin{aligned} 0 &= \int_{\partial U} g\frac{\partial v}{\partial n}d\sigma + \lim_{\varepsilon \to 0}\omega_{n-1}\int_{\partial B_\varepsilon} u\frac{(n-2)}{\omega_{n-1}\varepsilon^{n-1}}d\sigma \\ &= \int_{\partial U} g\frac{\partial v}{\partial n}d\sigma + (n-2)\,\omega_{n-1}u(\boldsymbol{x}) \qquad (11.15) \end{aligned}$$

$$u(\boldsymbol{x}) = \frac{1}{-(n-2)\,\omega_{n-1}}\int_{\partial U} g(\boldsymbol{y})\frac{(n-2)}{r}\frac{|\boldsymbol{x}|^2 - r^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma = \int_{\partial U} g(\boldsymbol{y})\frac{1}{r}\frac{r^2 - |\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma$$

In case $n = 2$, $u(\boldsymbol{x}) = \int_{\partial U} g(\boldsymbol{y})\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^2}d\sigma$.

Theorem 11.7.4 *Let $u \in C^2\left(\overline{B(\mathbf{0},r)}\right)$ satisfy $\Delta u = 0$ and $u = g$ on $\partial B(\mathbf{0},r)$, and also $B(\mathbf{0},r) \subseteq \mathbb{R}^n$. If $n \geq 2$, $u(\boldsymbol{x}) = \frac{1}{\omega_{n-1}}\int_{\partial B(\mathbf{0},r)} g(\boldsymbol{y})\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma(y)$. If $u(\boldsymbol{x})$ is given by this formula, then in fact $\Delta u = 0$ and $u = g$ on $\partial B(\mathbf{0},r)$ in the sense that $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} u(\boldsymbol{x}) = g(\boldsymbol{x}_0)$ and u is infinitely differentiable since all partial derivatives exist.*

Proof: I know a solution to $\Delta u = 0, u = 1$ on $B(\mathbf{0},r)$, namely $u = 1$. It follows from what was just shown that if $n > 2, 1 = \frac{1}{\omega_{n-1}}\int_{\partial B(\mathbf{0},r)}\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma(y)$ and if $n = 2$, then it follows that $1 = \frac{1}{2\pi}\int_{\partial B(\mathbf{0},r)}\frac{1}{r}\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^2}d\sigma(y)$.

Let $n > 2$. It is similar if $n = 2$. Let $\boldsymbol{x}_0 \in \partial B(\mathbf{0},r)$ and $\boldsymbol{x}$ will be close to $\boldsymbol{x}_0$ and it is desired to show that $|g(\boldsymbol{x}_0) - u(\boldsymbol{x})|$ is small.

$$\begin{aligned} |g(\boldsymbol{x}_0) - u(\boldsymbol{x})| &\leq \frac{1}{\omega_{n-1}r}\int_{\partial B(\mathbf{0},r)}|g(\boldsymbol{y}) - g(\boldsymbol{x}_0)|\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(\boldsymbol{y}) \\ &\leq \frac{1}{\omega_{n-1}r}\int_{[|\boldsymbol{y}-\boldsymbol{x}_0|<\delta]}|g(\boldsymbol{y}) - g(\boldsymbol{x}_0)|\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(\boldsymbol{y}) + \\ &\quad \frac{1}{\omega_{n-1}r}\int_{[|\boldsymbol{y}-\boldsymbol{x}_0|\geq\delta]}|g(\boldsymbol{y}) - g(\boldsymbol{x}_0)|\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(\boldsymbol{y}) \quad (11.16) \end{aligned}$$

Now since g is continuous, there is a constant M such that $|g(\boldsymbol{y})| < M$ for all $\boldsymbol{y}$. Therefore,

$$\begin{aligned} |g(\boldsymbol{x}_0) - u(\boldsymbol{x})| &< \frac{2M}{\omega_{n-1}r}\int_{[|\boldsymbol{y}-\boldsymbol{x}_0|\geq\delta]}\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(\boldsymbol{y}) + \\ &\quad \frac{1}{\omega_{n-1}r}\int_{[|\boldsymbol{y}-\boldsymbol{x}_0|<\delta]}|g(\boldsymbol{y}) - g(\boldsymbol{x}_0)|\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(\boldsymbol{y}) \end{aligned}$$

By continuity of g the second term is no more than ε if δ is chosen small enough. Now having picked δ, the first term converges to 0 as $\boldsymbol{x} \to \boldsymbol{x}_0$ by an application the dominated convergence theorem. Letting $\boldsymbol{x}_n \to \boldsymbol{x}_0$, then eventually $|\boldsymbol{y} - \boldsymbol{x}_n| > \delta/2$ and so the integrands are bounded for all n large enough. Since ε is arbitrary, this shows $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} u(\boldsymbol{x}) = g(\boldsymbol{x}_0)$.

As to $\Delta u = 0$, this will follow from $\boldsymbol{x} \to \frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}$ being harmonic. That $\Delta u = 0$ in U follows from the observation that the difference quotients used to compute the partial derivatives converge uniformly in $\boldsymbol{y} \in \partial U$ for any given $\boldsymbol{x} \in U$. To see this note that for $\boldsymbol{y} \in \partial U$, the partial derivatives of the expression, $\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}$ taken with respect to the k^{th} variable x_k are uniformly bounded and continuous for $\boldsymbol{y} \in \partial U$ and $\boldsymbol{x} \in U$. This continues to hold for higher order partial derivatives also. Therefore you can take the differential operator inside the integral, using the dominated convergence theorem, and write

$$\Delta_x \frac{1}{\omega_n r}\int_{\partial U} g(\boldsymbol{y})\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma(y) = \frac{1}{\omega_n r}\int_{\partial U} g(\boldsymbol{y})\Delta_x\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right)d\sigma(y) = 0.$$

It involves a computation to verify that $\Delta_x\left(\frac{r^2-|\boldsymbol{x}|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n}\right) = 0$ for $\boldsymbol{x} \neq \boldsymbol{y}$. ■

The Laplace equation and boundary conditions described above is called the Dirichlet problem.

Here is a remarkable result on harmonic functions.

Theorem 11.7.5 *(Liouville's theorem) If u is bounded and harmonic on $\mathbb{R}^n$, then u is constant.*

Proof: From Theorem 11.7.4 when $n > 2$,

$$\frac{r^2-|\boldsymbol{x}|^2}{\omega_n r}\int_{\partial B(\mathbf{0},r)} u(\boldsymbol{y})|\boldsymbol{y}-\boldsymbol{x}|^{-n}d\sigma(y) = u(\boldsymbol{x}).$$

Now, as mentioned, we can take partial derivatives inside the integral.

$$\frac{\partial u(\boldsymbol{x})}{\partial x_k} = \frac{-2x_k}{\omega_n r}\int_{\partial B(\mathbf{0},r)}\frac{u(\boldsymbol{y})}{|\boldsymbol{y}-\boldsymbol{x}|^n}d\sigma(y) + \frac{r^2-|\boldsymbol{x}|^2}{\omega_n r}\int_{\partial B(\mathbf{0},r)}(-n)\frac{(x_k-y_k)}{|\boldsymbol{y}-\boldsymbol{x}|^{(n+2)}}d\sigma(y)$$

Therefore, letting $|u(\boldsymbol{y})| \leq M$ for all $\boldsymbol{y} \in \mathbb{R}^n$,

$$\left|\frac{\partial u(\boldsymbol{x})}{\partial x_k}\right| \leq \frac{2|\boldsymbol{x}|}{\omega_n r}\int_{\partial B(\boldsymbol{x}_0,r)}\frac{M}{(r-|\boldsymbol{x}|)^n}d\sigma(y) + \frac{\left(r^2-|\boldsymbol{x}|^2\right)M}{\omega_n r}\int_{\partial B(\mathbf{0},r)}\frac{1}{(r-|\boldsymbol{x}|)^{n+1}}d\sigma(y)$$

$$= \frac{2|\boldsymbol{x}|}{\omega_n r}\frac{M}{(r-|\boldsymbol{x}|)^n}\omega_n r^{n-1} + \frac{\left(r^2-|\boldsymbol{x}|^2\right)M}{\omega_n r}\frac{1}{(r-|\boldsymbol{x}|)^{n+1}}\omega_n r^{n-1}$$

and these terms converge to 0 as $r \to \infty$. Since the inequality holds for all $r > |\boldsymbol{x}|$, it follows $\frac{\partial u(\boldsymbol{x})}{\partial x_k} = 0$. Similarly all the other partial derivatives equal zero as well and so u is a constant by using the mean value inequality Theorem 6.5.2. It works the same way for $n = 2$. ■

What about $\Delta u = 0$ on $B(\boldsymbol{x}_0, r)$ and $u = g$ on $\partial B(\boldsymbol{x}_0, r)$? Consider $\Delta w = 0$ on $B(\mathbf{0}, r)$ and $w(\boldsymbol{y}) = g(\boldsymbol{y} + \boldsymbol{x}_0)$ at $\hat{\boldsymbol{y}} \in \partial B(\mathbf{0}, r)$. The above shows that if $n \geq 2$,

$$w(\boldsymbol{z}) = \frac{1}{\omega_{n-1}}\int_{\partial B(\mathbf{0},r)} g(\hat{\boldsymbol{y}} + \boldsymbol{x}_0)\frac{1}{r}\frac{r^2-|\boldsymbol{z}|^2}{|\hat{\boldsymbol{y}}-\boldsymbol{z}|^n}d\sigma(\hat{y})$$

for $\boldsymbol{z} \in B(\mathbf{0}, r)$. So let $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{x}_0, \hat{\boldsymbol{y}} + \boldsymbol{x}_0 = \boldsymbol{y}$ Then let

$$\begin{aligned} u(\boldsymbol{x}) &\equiv w(\boldsymbol{x} - \boldsymbol{x}_0) = \frac{1}{\omega_{n-1}} \int_{\partial B(\boldsymbol{x}_0, r)} g(\boldsymbol{y}) \frac{1}{r} \frac{r^2 - |\boldsymbol{x} - \boldsymbol{x}_0|^2}{|\boldsymbol{y} - \boldsymbol{x}|^n} d\sigma(y) \\ &= \frac{r^2 - |\boldsymbol{x} - \boldsymbol{x}_0|^2}{r\omega_{n-1}} \int_{\partial B(\boldsymbol{x}_0, r)} g(\boldsymbol{y}) \frac{1}{|\boldsymbol{y} - \boldsymbol{x}|^n} d\sigma(y) \qquad (11.17) \end{aligned}$$

In particular, if u is harmonic on all of $\mathbb{R}^n$,

$$u(\boldsymbol{x}_0) = \frac{r}{\omega_{n-1}} \int_{\partial B(\boldsymbol{x}_0, r)} u(\boldsymbol{y}) \frac{1}{|\boldsymbol{y} - \boldsymbol{x}_0|^n} d\sigma(y) = \frac{1}{\omega_{n-1} r^{n-1}} \int_{\partial B(\boldsymbol{x}_0, r)} u(\boldsymbol{y}) d\sigma(y)$$

So a harmonic function at a point is always equal to the average of its values around any sphere centered at that point.

Corollary 11.7.6 *If u is harmonic, then $u(\boldsymbol{x})$ is equal to the average of its boundary values around any sphere centered at $\boldsymbol{x}$.*

This implies the following theorem.

Theorem 11.7.7 *Let U be an open connected set in $\mathbb{R}^n$ and let u be continuous on $\bar{U}$ and for all $\boldsymbol{z} \in U$, $u(\boldsymbol{z}) \leq \frac{1}{\omega_{n-1} r^{n-1}} \int_{\partial B(\boldsymbol{z}, r)} u(\boldsymbol{y}) d\sigma(y)$ for every $r > 0$ sufficiently small. Then if $u(\boldsymbol{z}) = \sup\{u(\boldsymbol{x}) : \boldsymbol{x} \in U\}$ for some $\boldsymbol{z} \in U$, it follows that u equals a constant.*

Proof: Let $\boldsymbol{z} \in B(\boldsymbol{z}, r) \subseteq U$. Then $u(\boldsymbol{z}) \leq \frac{1}{\omega_{n-1} r^{n-1}} \int_{\partial B(\boldsymbol{z}, r)} u(\boldsymbol{y}) d\sigma(y)$ and so

$$0 \geq \frac{1}{\omega_{n-1} r^{n-1}} \int_{\partial B(\boldsymbol{z}, r)} (u(\boldsymbol{z}) - u(\boldsymbol{y})) d\sigma(y) \geq 0,\ u(\boldsymbol{z}) \geq u(\boldsymbol{y})$$

Then $u(\boldsymbol{y}) = u(\boldsymbol{z})$ for all $\boldsymbol{y} \in \partial B(\boldsymbol{z}, r)$ since otherwise the inequality could not hold. Since this holds for all r it follows that u is identically equal to $u(\boldsymbol{z})$ on some open disk containing $\boldsymbol{z}$. Thus if $S \equiv \{\boldsymbol{x} : u(\boldsymbol{x}) = u(\boldsymbol{z})\}$, S is open and it is also closed. Therefore, $U \setminus S$ must be empty since otherwise U is not connected and so u is constant. ■

Definition 11.7.8 *Let U be an open set and let u be a function defined on $\bar{U}$. Then u is subharmonic if it is continuous on $\bar{U}$ and for all $\boldsymbol{x} \in U, u(\boldsymbol{x}) \leq \frac{1}{\omega_{n-1} r^{n-1}} \int_{\partial B(\boldsymbol{x}, r)} u(\boldsymbol{y}) d\sigma$ whenever $r > 0$ is small enough.*

Proposition 11.7.9 *(Maximum principle)Let U be a bounded open set and u is subharmonic on U and continuous on $\bar{U}$. Then u achieves its maximum on ∂U.*

Proof: Apply 11.7.7 Theorem 11.7.7 to V a connected component of U. Recall these are connected open sets, Theorem 3.11.12. If u achieves its maximum at some point of V then it is constant on V and by continuity, it achieves its maximum on $\partial V \subseteq \partial U$. ■

11.8 The Dirichlet Problem in General

Here is a general result about the supremum of continuous functions.

Lemma 11.8.1 *Let $\{f_n\}$ be continuous real valued functions and $f(\boldsymbol{x}) \equiv \sup_k \{f_k(\boldsymbol{x})\}$. Then f is lower semicontinuous.*

Proof: I need to show that if $\boldsymbol{x}_n \to \boldsymbol{x}$, then $f(\boldsymbol{x}) \leq \liminf_{n\to\infty} f(\boldsymbol{x}_n)$. If not, then there exists, for some $\varepsilon > 0$ a subsequence, still denoted by n such that $\lim_{n\to\infty} f(\boldsymbol{x}_n) < f(\boldsymbol{x}) - \varepsilon$. But then there is f_k such that $f_k(\boldsymbol{x}) > f(\boldsymbol{x}) - \frac{\varepsilon}{2}$ and so by continuity of f_k, $f_k(\boldsymbol{x}) = \lim_{n\to\infty} f_k(\boldsymbol{x}_n) \leq \lim_{n\to\infty} f(\boldsymbol{x}_n) < f_k(\boldsymbol{x}) - \frac{\varepsilon}{2}$, a contradiction. ■

This lemma is about finitely many subharmonic functions.

Lemma 11.8.2 *Let U be an open set and let $u_1, u_2, \cdots, u_p$ be subharmonic functions defined on U. Then letting $v \equiv \max(u_1, u_2, \cdots, u_p)$, it follows that v is also subharmonic.*

Proof: Let $\boldsymbol{x} \in U$. Then whenever r is small enough to satisfy the subharmonic condition for each u_i.

$$\begin{aligned} v(\boldsymbol{x}) &= \max(u_1(\boldsymbol{x}), u_2(\boldsymbol{x}), \cdots, u_p(\boldsymbol{x})) \\ &\leq \max\left(\frac{1}{\omega_{n-1}r^{n-1}} \int_{\partial B(\boldsymbol{x},r)} u_1(\boldsymbol{y})\, d\sigma(y), \cdots, \frac{1}{\omega_{n-1}r^{n-1}} \int_{\partial B(\boldsymbol{x},r)} u_p(\boldsymbol{y})\, d\sigma(y)\right) \end{aligned}$$

$$\leq \frac{1}{\omega_{n-1}r^{n-1}} \int_{\partial B(\boldsymbol{x},r)} \max(u_1, u_2, \cdots, u_p)(\boldsymbol{y})\, d\sigma(y) = \frac{1}{\omega_{n-1}r^{n-1}} \int_{\partial B(\boldsymbol{x},r)} v(\boldsymbol{y})\, d\sigma(y).$$

This proves the lemma. ■

Definition 11.8.3 *Let U be an open set and let u be subharmonic on U. Then for $\overline{B(\boldsymbol{x}_0,r)} \subseteq U$ define*

$$u_{\boldsymbol{x}_0,r}(\boldsymbol{x}) \equiv \begin{cases} u(\boldsymbol{x}) \text{ if } \boldsymbol{x} \notin B(\boldsymbol{x}_0, r) \\ \frac{1}{\omega_{n-1}r} \int_{\partial B(\boldsymbol{x}_0,r)} u(\boldsymbol{y}) \frac{r^2 - |\boldsymbol{x}-\boldsymbol{x}_0|^2}{|\boldsymbol{y}-\boldsymbol{x}|^n} d\sigma(\boldsymbol{y}) \text{ if } \boldsymbol{x} \in B(\boldsymbol{x}_0, r) \end{cases}$$

Thus $u_{\boldsymbol{x}_0,r}$ is harmonic on $B(\boldsymbol{x}_0, r)$, and equals to u off $B(\boldsymbol{x}_0, r)$. This is because there exists a harmonic function w whose boundary values on $\partial B(\boldsymbol{x}_0, r)$ are given by $u(\boldsymbol{y})$ and it equals $u_{\boldsymbol{x}_0,r}$ at $\boldsymbol{x} \in B(\boldsymbol{x}_0, r)$. The wonderful thing about this is that $u_{\boldsymbol{x}_0,r}$ is still subharmonic on all of U. Also note that, from Corollary 11.7.6 on Page 282, every harmonic function is subharmonic.

Lemma 11.8.4 *Let U be an open set and $\overline{B(\boldsymbol{x}_0,r)} \subseteq U$ as in the above definition where u is subharmonic. Then $u_{\boldsymbol{x}_0,r}$ is subharmonic on U and $u \leq u_{\boldsymbol{x}_0,r}$.*

Proof: First I show that $u \leq u_{\boldsymbol{x}_0,r}$. This follows from the maximum principle. Indeed, the function $u - u_{\boldsymbol{x}_0,r}$ is subharmonic on $B(\boldsymbol{x}_0, r)$ and equals zero on $\partial B(\boldsymbol{x}_0, r)$. Thus for all ρ small enough,

$$u(\boldsymbol{z}) - u_{\boldsymbol{x}_0r}(\boldsymbol{z}) \leq \frac{1}{\omega\rho^{n-1}} \int_{\partial B(\boldsymbol{z},\rho)} u(\boldsymbol{y}) - u_{\boldsymbol{x}_0,r}(\boldsymbol{y})\, d\sigma(\boldsymbol{y})$$

thanks to the mean value property of harmonic functions, Corollary 11.7.6. Since this is true for all small ρ, it follows from continuity that $u(\boldsymbol{z}) - u_{\boldsymbol{x}_0,r}(\boldsymbol{z}) \leq 0$.The two functions are equal off $B(\boldsymbol{x}_0, r)$. Thus for such $\boldsymbol{z}$,

$$u_{\boldsymbol{x}_0,r}(\boldsymbol{z}) = u(\boldsymbol{z}) \leq \frac{1}{\omega\rho^{n-1}} \int_{\partial B(\boldsymbol{z},\rho)} u(\boldsymbol{y})\, d\sigma(y) \leq \frac{1}{\omega\rho^{n-1}} \int_{\partial B(\boldsymbol{z},\rho)} u_{\boldsymbol{x}_0,r}(\boldsymbol{y})\, d\sigma(y)$$

The second inequality is because there may be points of $B(\boldsymbol{x}_0,r)$ in $\partial B(\boldsymbol{z},\rho)$ if $\boldsymbol{z} \in \partial B(\boldsymbol{x}_0, r)$. ■

Definition 11.8.5 *For U a bounded open set and $g \in C(\partial U)$, define*

$$w_g(\boldsymbol{x}) \equiv \sup\{u(\boldsymbol{x}) : u \in S_g\}$$

where S_g consists of those functions u which are subharmonic with $u(\boldsymbol{y}) \leq g(\boldsymbol{y})$ for all $\boldsymbol{y} \in \partial U$ and $u(\boldsymbol{y}) \geq \min\{g(\boldsymbol{y}) : \boldsymbol{y} \in \partial U\} \equiv m$.

Note that $S_g \neq \emptyset$ because $u(\boldsymbol{x}) \equiv m$ is in S_g. Also all functions in S_g have values between m and $\max\{g(\boldsymbol{y}) : \boldsymbol{y} \in \partial U\}$. The fundamental result is the following amazing result.

Proposition 11.8.6 *Let U be a bounded open set and let $g \in C(\partial U)$. Then $w_g \in S_g$ and in addition to this, w_g is harmonic.*

Proof: Let $\overline{B(\boldsymbol{x}_0, 2r)} \subseteq U$ and let $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ denote a countable dense subset of $\overline{B(\boldsymbol{x}_0, r)}$. Let $\{u_{1k}\}$ denote a sequence of functions of S_g with the property that $\lim_{k\to\infty} u_{1k}(\boldsymbol{x}_l) = w_g(\boldsymbol{x}_l)$. By Lemma 11.8.4, it can be assumed each u_{lk} is a harmonic function in $B(\boldsymbol{x}_0, 2r)$ and continuous on $\overline{B(\boldsymbol{x}_0, r)}$ since otherwise, you could use the process of replacing u with $u_{\boldsymbol{x}_0,2r}$. Now define $w_k = (\max(u_{1k}, \cdots, u_{kk}))_{\boldsymbol{x}_0,2r}$. Then each $w_k \in S_g$, each w_k is harmonic in $B(\boldsymbol{x}_0, 2r)$, and for each $\boldsymbol{x}_l, \lim_{k\to\infty} w_k(\boldsymbol{x}_l) = w_g(\boldsymbol{x}_l)$.

From the representation theorem for harmonic functions, 11.17, if $\boldsymbol{x} \in \overline{B(\boldsymbol{x}_0, r)}$

$$w_k(\boldsymbol{x}) = \frac{1}{\omega_{n-1} 2r} \int_{\partial B(\boldsymbol{x}_0, 2r)} w_k(\boldsymbol{y}) \frac{r^2 - |\boldsymbol{x} - \boldsymbol{x}_0|^2}{|\boldsymbol{y} - \boldsymbol{x}|^n} d\sigma(\boldsymbol{y}) \tag{11.18}$$

and so there exists a constant C which is independent of k such that for all $i = 1, 2, \cdots, n$ and $\boldsymbol{x} \in \overline{B(\boldsymbol{x}_0, r)}$, $\left|\frac{\partial w_k(\boldsymbol{x})}{\partial x_i}\right| \leq C$. Indeed, you could differentiate under the integral sign with respect to the x_i. The w_k are all bounded functions thanks to the maximum principle Proposition 11.7.9. Therefore, this set of functions, $\{w_k\}$ is equicontinuous on $\overline{B(\boldsymbol{x}_0, r)}$ as well as being uniformly bounded, thanks to the mean value inequality, and so by the Ascoli Arzela theorem, Theorem 3.10.5, it has a subsequence which converges uniformly on $\overline{B(\boldsymbol{x}_0, r)}$ to a continuous function I will denote by w which has the property that for all k, $w(\boldsymbol{x}_k) = w_g(\boldsymbol{x}_k)$. Also since each w_k is harmonic,

$$w_k(\boldsymbol{x}) = \frac{1}{\omega_{n-1} r} \int_{\partial B(\boldsymbol{x}_0, r)} w_k(\boldsymbol{y}) \frac{r^2 - |\boldsymbol{x} - \boldsymbol{x}_0|^2}{|\boldsymbol{y} - \boldsymbol{x}|^n} d\sigma(y) \tag{11.19}$$

Passing to the limit in 11.19 using the uniform convergence, it follows

$$w(\boldsymbol{x}) = \frac{1}{\omega_{n-1} r} \int_{\partial B(\boldsymbol{x}_0, r)} w(\boldsymbol{y}) \frac{r^2 - |\boldsymbol{x} - \boldsymbol{x}_0|^2}{|\boldsymbol{y} - \boldsymbol{x}|^n} d\sigma(y) \tag{11.20}$$

which shows that w is also harmonic. I have shown that $w = w_g$ on a dense set. Also, it follows by definition of w_g that $w(\boldsymbol{x}) \leq w_g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \overline{B(\boldsymbol{x}_0, r)}$. It remains to verify these two functions are in fact equal. By Lemma 11.8.1 w_g is lower semicontinuous on U. Let $\boldsymbol{x} \in \overline{B(\boldsymbol{x}_0, r)}$ and pick $\boldsymbol{x}_{k_l} \to \boldsymbol{x}$ where $\{\boldsymbol{x}_{k_l}\}$ is a subsequence of the dense set, $\{\boldsymbol{x}_k\}$. Then

$$w_g(\boldsymbol{x}) \geq w(\boldsymbol{x}) = \liminf_{l\to\infty} w(\boldsymbol{x}_{k_l}) = \liminf_{l\to\infty} w_g(\boldsymbol{x}_{k_l}) \geq w_g(\boldsymbol{x}).$$

This proves $w = w_g$ and since w is harmonic, so is w_g. ■

It remains to consider whether the boundary values are obtained. This requires an additional assumption on the set U.

Definition 11.8.7 *A bounded open set U has the barrier condition at $\boldsymbol{z} \in \partial U$, if there exists a continuous on ∂U function, $b_{\boldsymbol{z}}$ called a barrier function which has the property that $b_{\boldsymbol{z}}$ is subharmonic on U, $b_{\boldsymbol{z}}(\boldsymbol{z}) = 0$, and for all $\boldsymbol{x} \in \partial U \setminus \{\boldsymbol{z}\}, b_{\boldsymbol{z}}(\boldsymbol{x}) < 0$.*

The main result is the following remarkable theorem.

Theorem 11.8.8 *Let U be a bounded open set which has the barrier condition at $\boldsymbol{z} \in \partial U$ and let $g \in C(\partial U)$. Then the function w_g, defined above, is in $C^2(U)$ and satisfies $\Delta w_g = 0$ in U, $\lim_{\boldsymbol{x}\to\boldsymbol{z}} w_g(\boldsymbol{x}) = g(\boldsymbol{z})$.*

Proof: From Proposition 11.8.6 it follows $\Delta w_g = 0$. Let $\boldsymbol{z} \in \partial U$ and let $b_{\boldsymbol{z}}$ be the barrier function at $\boldsymbol{z}$. Note that $S_g \neq \emptyset$ because $v \in S_g$ where $v(\boldsymbol{x}) \equiv m \equiv \min\{g(\boldsymbol{x}) : \boldsymbol{x} \in \partial U\}$.

Claim: For K large enough, $g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x}) \leq g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial U$ and $g(\boldsymbol{z}) + \varepsilon - Kb_{\boldsymbol{z}}(\boldsymbol{x}) \geq g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial U$.

Proof of claim: If $\boldsymbol{x}$ is close enough to $\boldsymbol{z}$ that $g(\boldsymbol{x}) - g(\boldsymbol{z}) + \varepsilon > 0$,say $|\boldsymbol{x} - \boldsymbol{z}| < \delta$, then it does not matter what $K > 0$ is picked. You will have $\frac{g(\boldsymbol{x})-g(\boldsymbol{z})+\varepsilon}{K} \geq b_{\boldsymbol{z}}(\boldsymbol{x})$ because $b_{\boldsymbol{z}}(\boldsymbol{x}) \leq 0$. On the other hand if $|\boldsymbol{x} - \boldsymbol{z}| \geq \delta$ then $\max\{b_{\boldsymbol{z}}(\boldsymbol{x}) : \boldsymbol{x} \in \partial U \setminus B(\boldsymbol{z}, \delta)\} < 0$. Thus, for $K > 0$ sufficiently large, if $\boldsymbol{x} \in \partial U \setminus B(\boldsymbol{z}, \delta), b_{\boldsymbol{z}}(\boldsymbol{x}) \leq \frac{g(\boldsymbol{x})-g(\boldsymbol{z})+\varepsilon}{K}$. Thus for K this large, $b_{\boldsymbol{z}}(\boldsymbol{x}) \leq \frac{g(\boldsymbol{x})-g(\boldsymbol{z})+\varepsilon}{K}$ for all $\boldsymbol{x} \in \partial U$ which gives the first claim.

Next, if $\boldsymbol{x} \in \partial U$ is close enough to $\boldsymbol{z}$ that $g(\boldsymbol{x}) - g(\boldsymbol{z}) - \varepsilon < 0$, say if $|\boldsymbol{x} - \boldsymbol{z}| < \delta$ then for any $K > 0$, $\frac{g(\boldsymbol{x})-g(\boldsymbol{z})-\varepsilon}{-K} \geq b_{\boldsymbol{z}}(\boldsymbol{x})$ because the left is positive while the right is ≤ 0. If $|\boldsymbol{x} - \boldsymbol{z}| \geq \delta$, then $\max\{b_{\boldsymbol{z}}(\boldsymbol{x}) : \boldsymbol{x} \in \partial U \setminus B(\boldsymbol{z}, \delta)\} < 0$ and so if K is large enough and positive, then $\frac{g(\boldsymbol{x})-g(\boldsymbol{z})-\varepsilon}{-K} \geq b_{\boldsymbol{z}}(\boldsymbol{x})$ and so there is K large enough that for all $\boldsymbol{x} \in \partial U, g(\boldsymbol{x}) - g(\boldsymbol{z}) - \varepsilon \leq -Kb_{\boldsymbol{z}}(\boldsymbol{x})$ and so for all $\boldsymbol{x} \in \partial U, \varepsilon \geq g(\boldsymbol{x}) - g(\boldsymbol{z}) + Kb_{\boldsymbol{z}}(\boldsymbol{x})$. This proves the claim.

We have two subharmonic functions of $\boldsymbol{x} \in U$,

$$\overset{\leq -g(\boldsymbol{x}) \text{ if } \boldsymbol{x}\in\partial U}{-g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x})} \text{ and } \overset{\leq g(\boldsymbol{x})\text{if } \boldsymbol{x}\in\partial U}{g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x})} \tag{11.21}$$

For $\boldsymbol{x} \in \partial U$, the first $\leq -g(\boldsymbol{x})$ and the second $\leq g(\boldsymbol{x})$ from the above claim. Let $u \in S_g$. Then $\boldsymbol{x} \to u(\boldsymbol{x}) + (-g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x}))$ is subharmonic and when $\boldsymbol{x} \in \partial U$, it is no more than $g(\boldsymbol{x}) - g(\boldsymbol{x}) = 0$. By the maximum principle, Proposition 11.7.9, for $\boldsymbol{x} \in U$,

$$u(\boldsymbol{x}) + (-g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x})) \leq 0$$

It follows that $w_g(\boldsymbol{x}) \leq g(\boldsymbol{z}) + \varepsilon - Kb_{\boldsymbol{z}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$. Now consider the second in 11.21 $g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x}) \leq g(\boldsymbol{x})$ for $\boldsymbol{x} \in \partial U$ and so $g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x}) \leq w_g(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$ by definition of w_g. Thus for $\boldsymbol{x} \in U$,

$$g(\boldsymbol{z}) - \varepsilon + Kb_{\boldsymbol{z}}(\boldsymbol{x}) \leq w_g(\boldsymbol{x}) \leq g(\boldsymbol{z}) + \varepsilon - Kb_{\boldsymbol{z}}(\boldsymbol{x})$$

It follows that if $\boldsymbol{x}_n \to \boldsymbol{z}$, then from the above and continuity of $b_{\boldsymbol{z}}(\boldsymbol{x})$,

$$g(\boldsymbol{z}) - \varepsilon \leq \liminf_{n\to\infty} w_g(\boldsymbol{x}_n) \leq \limsup_{n\to\infty} w_g(\boldsymbol{x}_n) \leq g(\boldsymbol{z}) + \varepsilon.$$

Since ε is arbitrary, this shows $\lim_{n\to\infty} w_g(\boldsymbol{x}_n)$ exists and equals $g(\boldsymbol{z})$. ■

How can you recognize that a point on the boundary of a bounded open set U has the barrier condition? One way would be to check the following condition.

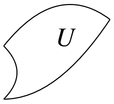
Condition 11.8.9 *For $\boldsymbol{z} \in \partial U$, there exists $\boldsymbol{x}_{\boldsymbol{z}} \notin \overline{U}$ such that $|\boldsymbol{x}_{\boldsymbol{z}} - \boldsymbol{z}| < |\boldsymbol{x}_{\boldsymbol{z}} - \boldsymbol{y}|$ for every $\boldsymbol{y} \in \partial U \setminus \{\boldsymbol{z}\}$.*

It says that there is a point $\boldsymbol{x}_{\boldsymbol{z}} \notin \bar{U}$ and a ball B centered at $\boldsymbol{x}_{\boldsymbol{z}}$ with $\partial B \cap \partial U = \{\boldsymbol{z}\}$.

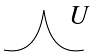
Proposition 11.8.10 *Suppose Condition 11.8.9 holds. Then ∂U satisfies the barrier condition at such $\boldsymbol{z} \in \partial U$. Consequently for such a $\boldsymbol{z}$, $\lim_{\boldsymbol{x} \to \boldsymbol{z}} w_g(\boldsymbol{x}) = g(\boldsymbol{z})$ where $\Delta w_g = 0$ and w_g is given above.*

Proof: For $n \geq 3$, let $b_{\boldsymbol{z}}(\boldsymbol{y}) \equiv r^{\boldsymbol{x}_{\boldsymbol{z}}}(\boldsymbol{y} - \boldsymbol{x}_{\boldsymbol{z}}) - r^{\boldsymbol{x}_{\boldsymbol{z}}}(\boldsymbol{z} - \boldsymbol{x}_{\boldsymbol{z}})$, $r^{\boldsymbol{x}}$ in Lemma 11.7.2. Then $b_{\boldsymbol{z}}(\boldsymbol{z}) = 0$ and if $\boldsymbol{y} \in \partial U$ with $\boldsymbol{y} \neq \boldsymbol{z}$, then clearly $b_{\boldsymbol{z}}(\boldsymbol{y}) < 0$. ■

Here is a picture of a domain which has a barrier at each point of the boundary.



You might try to think of some examples which won't satisfy the above condition. Maybe an inward pointing cusp would give such an example. Let $\boldsymbol{z}$ be the point of the cusp.



You might also consider $U = B(\boldsymbol{0},1) \setminus \{\text{positive } z \text{ axis}\}$ in $\mathbb{R}^3$. However, for many ordinary regions, the above condition would hold for all points, and perhaps fail to hold only at finitely many exceptional points or points in a set of σ measure zero.

11.9 Exercises

1. Suppose U is an open bounded set in $\mathbb{R}^p$, $u \in C^2(U) \cap C\left(\overline{U}\right)$, and $\Delta u \geq 0$ in U. Then
$$\max\left\{u(\boldsymbol{x}) : \boldsymbol{x} \in \overline{U}\right\} = \max\left\{u(\boldsymbol{x}) : \boldsymbol{x} \in \partial U\right\}.$$
Here $\Delta u \equiv \sum_{i=1}^{p} u_{x_i x_i}$. This is called the weak maximum principle. **Hint:** Suppose not. Then $\boldsymbol{u}(\boldsymbol{x}_0)$ is the maximum at $\boldsymbol{x}_0 \in U$. Letting $u_\varepsilon(\boldsymbol{x}) \equiv \varepsilon|\boldsymbol{x}|^2 + u(\boldsymbol{x})$, it follows that if $\varepsilon > 0$ is small enough, u_ε also has its maximum at an interior point, say $\boldsymbol{x}_\varepsilon$. For some $x_i, u_{x_i x_i}(\boldsymbol{x}_\varepsilon) \geq 0$ so $u_{\varepsilon x_i x_i}(\boldsymbol{x}_\varepsilon) > 0$. A C^2 function is harmonic if $\Delta u = 0$. Show that if a C^2 function is harmonic on a bounded open set U, continuous on $\bar{U}$, then if it equals 0 on ∂U, it must be 0 on U. Show that this proves uniqueness for the Dirichlet problem which is to find harmonic u on U with given boundary values.

2. Show that $\Delta u \geq 0$ implies u is subharmonic. **Hint:** Let $v \equiv |x - x_0|^{-(n-2)} - r^{-(n-2)}$, for $n > 2$, so $v = 0$ on $\partial B(\boldsymbol{x}_0, r)$ and $\Delta v = 0$. Modify using ln if $n = 2$. Then consider identity $\int_V (u\Delta v - v\Delta u)\, dm_n = \int_{\partial V} u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n} d\sigma$ for $V = B \setminus B_\varepsilon$.

3. For $n > 2$ show that $\boldsymbol{x} \to |a\boldsymbol{x} - \boldsymbol{b}|^k$ is harmonic away from $a^{-1}\boldsymbol{b}$ if and only if $k = -(n-2)$. What of the case where $n = 2$?

Chapter 12

Theorems Involving Line Integrals

12.1 Green's Theorem

Green's theorem is an important theorem which relates line integrals to integrals over a surface in the plane. It can be used to establish the seemingly more general Stoke's theorem but is interesting for it's own sake. Historically, theorems like it were important in the development of complex analysis.

Here is a proof of Green's theorem from the divergence theorem. This discussion is also in "Calculus of One and Many Variables" which discusses line integrals or see Section 11.5 for a short treatment based on the general case of differentiable manifolds.

Theorem 12.1.1 *(Green's Theorem) Let U be an open set in the plane for which the divergence theorem holds and let*

$$\boldsymbol{F}(x,y) = (P(x,y), Q(x,y))$$

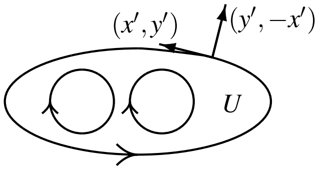
be a C^1 vector field defined near U. Then

$$\int_{\partial U} \boldsymbol{F} \cdot d\boldsymbol{R} = \int_U \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dm_2.$$

In the other notation,

$$\int_{\partial U} Pdx + Qdy = \int_U (Q_x - P_y)\, dm_2$$

Proof: Suppose the divergence theorem holds for U. Consider the following picture.



Counter clockwise motion around the curve is determined by imagining you stand upright with your left hand over U and walk in the direction you are facing. Your right hand will then be in the direction of the outer normal. The tangent vector in direction of motion is (x', y') is as shown. The unit **exterior normal** is a multiple of

$$\left(x', y', 0\right) \times (0,0,1) = \left(y', -x', 0\right).$$

This would be the case at all the points where the unit exterior normal exists.

Now let $\boldsymbol{G}(x,y) = (Q(x,y), -P(x,y))$. Also note the area (length) element on the bounding curve ∂U is $\sqrt{(x')^2 + (y')^2}dt$. Then by the divergence theorem,

$$\int_U (Q_x - P_y)\, dm_2 = \int_U \operatorname{div}(\boldsymbol{G})\, dm_2 = \int_{\partial U} \boldsymbol{G} \cdot \boldsymbol{n} d\sigma =$$

$$\sum_{i=1}^{m}\int_{a_i}^{b_i}\left(Q\left(x_i\left(t\right),y_i\left(t\right)\right),-P\left(x_i\left(t\right),y_i\left(t\right)\right)\right)\frac{1}{\sqrt{\left(x_i'\right)^2+\left(y_i'\right)^2}}\left(y_i',-x_i'\right)\overbrace{\sqrt{\left(x_i'\right)^2+\left(y_i'\right)^2}dt}^{dS}$$

$$=\sum_{i=1}^{m}\int_{a_i}^{b_i}\left(Q\left(x_i\left(t\right),y_i\left(t\right)\right),-P\left(x_i\left(t\right),y_i\left(t\right)\right)\right)\cdot\left(y_i',-x_i'\right)dt$$

$$=\sum_{i=1}^{m}\int_{a_i}^{b_i}Q\left(x_i\left(t\right),y_i\left(t\right)\right)y_i'\left(t\right)+P\left(x_i\left(t\right),y_i\left(t\right)\right)x_i'\left(t\right)dt\equiv\int_{\partial U}Pdx+Qdy$$

This proves Green's theorem from the divergence theorem. ■

Proposition 12.1.2 *Let U be an open set in $\mathbb{R}^2$ for which Green's theorem holds. Then Area of $U=\int_{\partial U}\boldsymbol{F}\cdot d\boldsymbol{R}$ where $\boldsymbol{F}(x,y)=\frac{1}{2}(-y,x),(0,x)$, or $(-y,0)$.*

Proof: This follows immediately from Green's theorem. ■

Example 12.1.3 *Use Proposition 12.1.2 to find the area of the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}\leq 1$.*

You can parameterize the boundary of this ellipse as $x=a\cos t,\ y=b\sin t,\ t\in[0,2\pi]$. Then from Proposition 12.1.2,

$$\text{Area equals }=\frac{1}{2}\int_0^{2\pi}(-b\sin t,a\cos t)\cdot(-a\sin t,b\cos t)\,dt=\frac{1}{2}\int_0^{2\pi}(ab)\,dt=\pi ab.$$

Example 12.1.4 *Find $\int_{\partial U}\boldsymbol{F}\cdot d\boldsymbol{R}$ where U is the set $\left\{(x,y):x^2+3y^2\leq 9\right\}$ and $\boldsymbol{F}(x,y)=(y,-x)$.*

One way to do this is to parameterize the boundary of U and then compute the line integral directly. It is easier to use Green's theorem. The desired line integral equals $\int_U((-1)-1)\,dA=-2\int_U dA$.Now U is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

Example 12.1.5 *Find $\int_{\partial U}\boldsymbol{F}\cdot d\boldsymbol{R}$ where U is the set $\{(x,y):2\leq x\leq 4,0\leq y\leq 3\}$ and*

$$\boldsymbol{F}(x,y)=\left(x\sin y,y^3\cos x\right)$$

From Green's theorem this line integral equals

$$\int_2^4\int_0^3\left(-y^3\sin x-x\cos y\right)dydx=\frac{81}{4}\cos 4-6\sin 3-\frac{81}{4}\cos 2.$$

This is much easier than computing the line integral because you don't have to break the boundary in pieces and consider each separately.

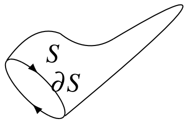
Example 12.1.6 *Find $\int_{\partial U}\boldsymbol{F}\cdot d\boldsymbol{R}$ where U is the set $\{(x,y):2\leq x\leq 4,x\leq y\leq 4\}$ and*

$$\boldsymbol{F}(x,y)=(x\sin y,y\sin x)$$

From Green's theorem, this line integral equals $\int_2^4\int_x^4(y\cos x-x\cos y)\,dydx=4\cos 2-8\cos 4-8\sin 2-4\sin 4$.

12.2 Stokes Theorem from Green's Theorem

Stoke's theorem is a generalization of Green's theorem which relates the integral over a surface to the integral around the boundary of the surface. These terms are a little different from what occurs in $\mathbb{R}^2$. To describe this, consider a sock. The surface is the sock and its boundary will be the edge of the opening of the sock in which you place your foot. Another way to think of this is to imagine a region in $\mathbb{R}^2$ of the sort discussed above for Green's theorem. Suppose it is on a sheet of rubber and the sheet of rubber is stretched in three dimensions. The boundary of the resulting surface is the result of the stretching applied to the boundary of the original region in $\mathbb{R}^2$. Here is a picture describing the situation.



Recall the following definition of the curl of a vector field. Why do we even consider it?

Definition 12.2.1 *Let $\boldsymbol{F}(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$ be a C^1 vector field defined on an open set V in $\mathbb{R}^3$. Then*

$$\nabla \times \boldsymbol{F} \equiv \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} \equiv \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\boldsymbol{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\boldsymbol{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\boldsymbol{k}.$$

This is also called curl$(\boldsymbol{F})$ *and written as indicated, $\nabla \times \boldsymbol{F}$.*

The following lemma gives the fundamental identity which will be used in the proof of Stoke's theorem.

Lemma 12.2.2 *Let $\boldsymbol{R} : U \to V \subseteq \mathbb{R}^3$ where U is an open subset of $\mathbb{R}^2$ and V is an open subset of $\mathbb{R}^3$. Suppose $\boldsymbol{R}$ is C^2 and let $\boldsymbol{F}$ be a C^1 vector field defined in V.*

$$(\boldsymbol{R}_u \times \boldsymbol{R}_v) \cdot (\nabla \times \boldsymbol{F})(\boldsymbol{R}(u,v)) = ((\boldsymbol{F} \circ \boldsymbol{R})_u \cdot \boldsymbol{R}_v - (\boldsymbol{F} \circ \boldsymbol{R})_v \cdot \boldsymbol{R}_u)(u,v). \tag{12.1}$$

Proof: Start with the left side and let $x_i = R_i(u,v)$ for short.

$$(\boldsymbol{R}_u \times \boldsymbol{R}_v) \cdot (\nabla \times \boldsymbol{F})(\boldsymbol{R}(u,v)) = \varepsilon_{ijk} x_{ju} x_{kv} \varepsilon_{irs} \frac{\partial F_s}{\partial x_r} = (\delta_{jr}\delta_{ks} - \delta_{js}\delta_{kr}) x_{ju} x_{kv} \frac{\partial F_s}{\partial x_r}$$

$$= x_{ju} x_{kv} \frac{\partial F_k}{\partial x_j} - x_{ju} x_{kv} \frac{\partial F_j}{\partial x_k} = \boldsymbol{R}_v \cdot \frac{\partial(\boldsymbol{F} \circ \boldsymbol{R})}{\partial u} - \boldsymbol{R}_u \cdot \frac{\partial(\boldsymbol{F} \circ \boldsymbol{R})}{\partial v}$$

which proves 12.1. ■

The proof of Stoke's theorem given next follows [10]. First, it is convenient to give a definition.

Definition 12.2.3 *A vector valued function $\boldsymbol{R} : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is said to be in $C^k\left(\overline{U}, \mathbb{R}^n\right)$ if it is the restriction to $\overline{U}$ of a vector valued function which is defined on $\mathbb{R}^m$ and is C^k. That is, this function has continuous partial derivatives up to order k.*

Theorem 12.2.4 *(Stoke's Theorem) Let U be any region in $\mathbb{R}^2$ for which the conclusion of Green's theorem holds and let $\boldsymbol{R} \in C^2\left(\overline{U}, \mathbb{R}^3\right)$ be a one to one function satisfying $|(\boldsymbol{R}_u \times \boldsymbol{R}_v)(u,v)| \neq 0$ for all $(u,v) \in U$ and let S denote the surface*

$$S \equiv \{\boldsymbol{R}(u,v) : (u,v) \in U\}, \ \partial S \equiv \{\boldsymbol{R}(u,v) : (u,v) \in \partial U\}$$

where the orientation on ∂S is consistent with the counter clockwise orientation on ∂U (U is on the left as you walk around ∂U). Then for $\boldsymbol{F}$ a C^1 vector field defined near S,

$$\int_{\partial S} \boldsymbol{F} \cdot d\boldsymbol{R} = \int_S \operatorname{curl}(\boldsymbol{F}) \cdot \boldsymbol{n} d\sigma$$

where $\boldsymbol{n}$ is the normal to S defined by $\boldsymbol{n} \equiv \frac{\boldsymbol{R}_u \times \boldsymbol{R}_v}{|\boldsymbol{R}_u \times \boldsymbol{R}_v|}$.

Proof: Letting C be an oriented part of ∂U having parametrization, $\boldsymbol{r}(t) \equiv (u(t), v(t))$ for $t \in [\alpha, \beta]$ and letting $\boldsymbol{R}(C)$ denote the oriented part of ∂S corresponding to C, $\int_{\boldsymbol{R}(C)} \boldsymbol{F} \cdot d\boldsymbol{R} =$

$$\begin{aligned}
&= \int_\alpha^\beta \boldsymbol{F}(\boldsymbol{R}(u(t), v(t))) \cdot \left(\boldsymbol{R}_u u'(t) + \boldsymbol{R}_v v'(t)\right) dt \\
&= \int_\alpha^\beta \boldsymbol{F}(\boldsymbol{R}(u(t), v(t))) \boldsymbol{R}_u(u(t), v(t)) u'(t) dt \\
&\quad + \int_\alpha^\beta \boldsymbol{F}(\boldsymbol{R}(u(t), v(t))) \boldsymbol{R}_v(u(t), v(t)) v'(t) dt \\
&= \int_C ((\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_u, (\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_v) \cdot d\boldsymbol{r}.
\end{aligned}$$

Since this holds for each such piece of ∂U, it follows

$$\int_{\partial S} \boldsymbol{F} \cdot d\,\boldsymbol{R} = \int_{\partial U} ((\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_u, (\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_v) \cdot d\boldsymbol{r}.$$

By the assumption that the conclusion of Green's theorem holds for U, this equals

$$\begin{aligned}
&\int_U \left[((\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_v)_u - ((\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_u)_v\right] dm_2 \\
= &\int_U \left[(\boldsymbol{F} \circ \boldsymbol{R})_u \cdot \boldsymbol{R}_v + (\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_{vu} - (\boldsymbol{F} \circ \boldsymbol{R}) \cdot \boldsymbol{R}_{uv} - (\boldsymbol{F} \circ \boldsymbol{R})_v \cdot \boldsymbol{R}_u\right] dm_2 \\
= &\int_U \left[(\boldsymbol{F} \circ \boldsymbol{R})_u \cdot \boldsymbol{R}_v - (\boldsymbol{F} \circ \boldsymbol{R})_v \cdot \boldsymbol{R}_u\right] dm_2
\end{aligned}$$

the last step holding by equality of mixed partial derivatives, a result of the assumption that $\boldsymbol{R}$ is C^2. Now by Lemma 12.2.2, this equals

$$\int_U (\boldsymbol{R}_u \times \boldsymbol{R}_v) \cdot (\nabla \times \boldsymbol{F}) dm_2 = \int_U \nabla \times \boldsymbol{F} \cdot (\boldsymbol{R}_u \times \boldsymbol{R}_v) dm_2 = \int_S \nabla \times \boldsymbol{F} \cdot \boldsymbol{n} d\sigma$$

because $d\sigma = |(\boldsymbol{R}_u \times \boldsymbol{R}_v)| dm_2$ and $\boldsymbol{n} = \frac{(\boldsymbol{R}_u \times \boldsymbol{R}_v)}{|(\boldsymbol{R}_u \times \boldsymbol{R}_v)|}$. Thus

$$(\boldsymbol{R}_u \times \boldsymbol{R}_v) dm_2 = \frac{(\boldsymbol{R}_u \times \boldsymbol{R}_v)}{|(\boldsymbol{R}_u \times \boldsymbol{R}_v)|} |(\boldsymbol{R}_u \times \boldsymbol{R}_v)| dm_2 = \boldsymbol{n} d\sigma.$$

This proves Stoke's theorem. ■

Note that there is no mention made in the final result that $\boldsymbol{R}$ is C^2. Therefore, it is not surprising that versions of this theorem are valid in which this assumption is not present. It is possible to obtain extremely general versions of Stoke's theorem.

12.2.1 The Normal and the Orientation

Stoke's theorem as just presented needs no apology. However, it is helpful in applications to have some additional geometric insight.

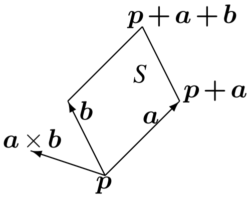
To begin with, suppose the surface S of interest is a parallelogram in $\mathbb{R}^3$ determined by the two vectors $\boldsymbol{a},\boldsymbol{b}$. Thus $S=\boldsymbol{R}(Q)$ where $Q=[0,1]\times[0,1]$ is the unit square and for $(u,v)\in Q$,

$$\boldsymbol{R}(u,v)\equiv u\boldsymbol{a}+v\boldsymbol{b}+\boldsymbol{p},$$

the point $\boldsymbol{p}$ being a corner of the parallelogram S. Then orient ∂S consistent with the counter clockwise orientation on ∂Q. Thus, following this orientation on S you go from $\boldsymbol{p}$ to $\boldsymbol{p}+\boldsymbol{a}$ to $\boldsymbol{p}+\boldsymbol{a}+\boldsymbol{b}$ to $\boldsymbol{p}+\boldsymbol{b}$ to $\boldsymbol{p}$. Then Stoke's theorem implies that with this orientation on ∂S,

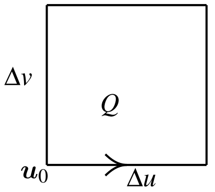
$$\int_{\partial S}\boldsymbol{F}\cdot d\boldsymbol{R}=\int_S\nabla\times\boldsymbol{F}\cdot\boldsymbol{n}ds$$

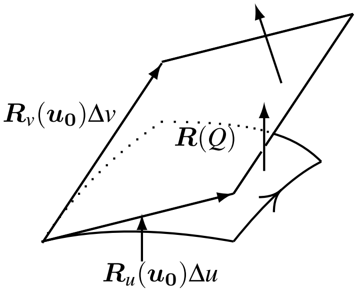
where $\boldsymbol{n}=\boldsymbol{R}_u\times\boldsymbol{R}_v/\left|\boldsymbol{R}_u\times\boldsymbol{R}_v\right|=\boldsymbol{a}\times\boldsymbol{b}/\left|\boldsymbol{a}\times\boldsymbol{b}\right|$. Now recall $\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}\times\boldsymbol{b}$ forms a right hand system.


$\boldsymbol{p}+\boldsymbol{a}+\boldsymbol{b}$
S
$\boldsymbol{p}+\boldsymbol{a}$
$\boldsymbol{b}$
$\boldsymbol{a}$
$\boldsymbol{a}\times\boldsymbol{b}$
$\boldsymbol{p}$

Thus, if you were walking around ∂S in the direction of the orientation with your left hand over the surface S, the normal vector $\boldsymbol{a}\times\boldsymbol{b}$ would be pointing in the direction of your head.

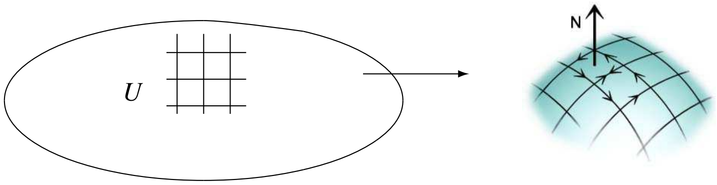
More generally, if S is a surface which is not necessarily a parallelogram but is instead as described in Theorem 12.2.4, you could consider a **small** rectangle Q contained in U and orient the boundary of $\boldsymbol{R}(Q)$ consistent with the counter clockwise orientation on ∂Q. Then if Q is small enough, as you walk around $\partial\boldsymbol{R}(Q)$ in the direction of the described orientation with your left hand over $\boldsymbol{R}(Q)$, your head points roughly in the direction of $\boldsymbol{R}_u\times\boldsymbol{R}_v$.



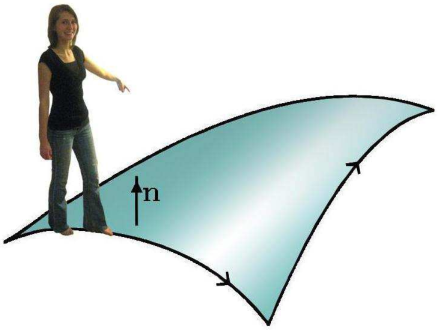

$\boldsymbol{R}_v(\boldsymbol{u_0})\Delta v$
$\boldsymbol{R}(Q)$
$\boldsymbol{R}_u(\boldsymbol{u_0})\Delta u$

As explained above, this is true of the tangent parallelogram, and by continuity of $\boldsymbol{R}_v,\boldsymbol{R}_u$, the normals to the surface $\boldsymbol{R}(Q)$ $\boldsymbol{R}_u\times\boldsymbol{R}_v(\boldsymbol{u})$ for $\boldsymbol{u}\in Q$ will still point roughly in the same direction as your head if you walk in the indicated direction over $\partial\boldsymbol{R}(Q)$, meaning the angle between the vector from your feet to your head and the vector $\boldsymbol{R}_u\times\boldsymbol{R}_v(\boldsymbol{u})$ is less than $\pi/2$.

You can imagine filling U with such non-overlapping regions Q_i. Then orienting $\partial\boldsymbol{R}(Q_i)$ consistent with the counter clockwise orientation on Q_i, and adding the resulting line integrals, the line integrals over the common sides cancel as indicated in the following picture and the result is the line integral over ∂S.



Thus there is a simple relation between the field of normal vectors on S and the orientation of ∂S. It is simply this. If you walk along ∂S in the direction mandated by the orientation, with your left hand over the surface, the nearby normal vectors in Stoke's theorem will point roughly in the direction of your head.

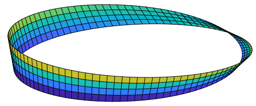


This also illustrates that you can **define** an orientation for ∂S by specifying a field of unit normal vectors for the surface, which varies continuously over the surface, and require that the motion over the boundary of the surface is such that your head points roughly in the direction of nearby normal vectors as you walk along the boundary with your left hand over S. The existence of such a continuous field of normal vectors is what constitutes an **orientable** surface.

12.2.2 The Mobeus Band

It turns out there are more general formulations of Stoke's theorem than what is presented above. However, it is always necessary for the surface S to be **orientable**. This means it is possible to obtain a vector field of unit normals to the surface which is a continuous function of position on S.

An example of a surface which is not orientable is the famous Mobeus band, obtained by taking a long rectangular piece of paper and gluing the ends together after putting a twist in it. Here is a picture of one.



There is something quite interesting about this Mobeus band and this is that it can be written parametrically with a simple parameter domain. The picture above is a maple graph of the parametrically defined surface

$$\boldsymbol{R}(\theta,v)\equiv\left\{\begin{array}{l} x=4\cos\theta+v\cos\frac{\theta}{2} \\ y=4\sin\theta+v\cos\frac{\theta}{2}, \\ z=v\sin\frac{\theta}{2}\end{array}\right. \qquad \theta\in[0,2\pi], v\in[-1,1].$$

An obvious question is why the normal vector $\boldsymbol{R}_{,\theta}\times\boldsymbol{R}_{,v}/\left|\boldsymbol{R}_{,\theta}\times\boldsymbol{R}_{,v}\right|$ is not a continuous function of position on S. You can see easily that it is a continuous function of both θ and v. However, the map, $\boldsymbol{R}$ is not one to one. In fact, $\boldsymbol{R}(0,0)=\boldsymbol{R}(2\pi,0)$. Therefore, near this point on S, there are two different values for the above normal vector. In fact, a tedious computation will show that this normal vector is

$$\frac{\left(4\sin\frac{1}{2}\theta\cos\theta-\frac{1}{2}v, 4\sin\frac{1}{2}\theta\sin\theta+\frac{1}{2}v, -8\cos^2\frac{1}{2}\theta\sin\frac{1}{2}\theta-8\cos^3\frac{1}{2}\theta+4\cos\frac{1}{2}\theta\right)}{D}$$

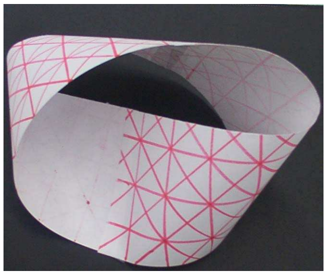
where

$$\begin{aligned} D = {} & 16\sin^2\left(\frac{\theta}{2}\right)+\frac{v^2}{2}+4\sin\left(\frac{\theta}{2}\right)v(\sin\theta-\cos\theta) \\ & +4^3\cos^2\left(\frac{\theta}{2}\right)\left(\cos\left(\frac{1}{2}\theta\right)\sin\left(\frac{1}{2}\theta\right)+\cos^2\left(\frac{1}{2}\theta\right)-\frac{1}{2}\right)^2 \end{aligned}$$

and you can verify that the denominator will not vanish. Letting $v=0$ and $\theta=0$ and 2π yields the two vectors $(0,0,-1),(0,0,1)$ so there is a discontinuity. This is why I was careful to say in the statement of Stoke's theorem given above that $\boldsymbol{R}$ is one to one.

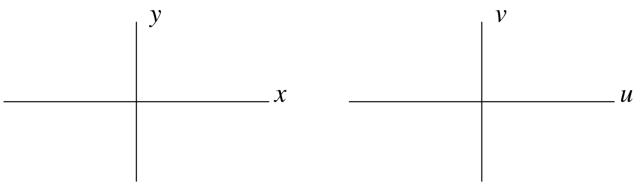
The Mobeus band has some usefulness. In old machine shops the equipment was run by a belt which was given a twist to spread the surface wear on the belt over twice the area.

The above explanation shows that $\boldsymbol{R}_{,\theta}\times\boldsymbol{R}_{,v}/\left|\boldsymbol{R}_{,\theta}\times\boldsymbol{R}_{,v}\right|$ fails to deliver an orientation for the Mobeus band. However, this does not answer the question whether there is some orientation for it other than this one. In fact there is none. You can see this by looking at the first of the two pictures below or by making one and tracing it with a pencil. There is only one side to the Mobeus band. An oriented surface must have two sides, one side identified by the given unit normal which varies continuously over the surface and the other side identified by the negative of this normal. The second picture below was taken by Ouyang when he was at meetings in Paris and saw it at a museum.



12.3 A General Green's Theorem

Now suppose U is a region in the uv plane for which Green's theorem holds and that $V\equiv\boldsymbol{R}(U)$ where $\boldsymbol{R}$ is $C^2\left(\overline{U},\mathbb{R}^2\right)$ and is one to one, $\boldsymbol{R}_u\times\boldsymbol{R}_v\neq\mathbf{0}$. Here, to be specific, the u,v axes are oriented as the x,y axes respectively.



Also let $\boldsymbol{F}(x,y,z) = (P(x,y), Q(x,y), 0)$ be a C^1 vector field defined near V. Note that F does not depend on z. Therefore, $\nabla \times \boldsymbol{F}(x,y) = (Q_x(x,y) - P_y(x,y))\boldsymbol{k}$.You can check this from the definition. Also

$$\boldsymbol{R}(u,v) = \begin{pmatrix} x(u,v) \\ y(u,v) \end{pmatrix}$$

and so, from the definition of $\boldsymbol{R}_u \times \boldsymbol{R}_v$, the desired unit normal vector to V is $\frac{x_u y_v - x_v y_u}{|x_u y_v - x_v y_u|}\boldsymbol{k}$. Suppose $x_u y_v - x_v y_u > 0$. Then the unit normal is $\boldsymbol{k}$. Then Stoke's theorem applied to this special case yields

$$\int_{\partial V} \boldsymbol{F} \cdot d\boldsymbol{R} = \int_U (Q_x(x(u,v), y(u,v)) - P_y(x(u,v), y(u,v)))\boldsymbol{k} \cdot \boldsymbol{k} \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} d\sigma$$

Now by the change of variables formula, this equals $\int_V (Q_x(x,y) - P_y(x,y))\,d\sigma$. This is just Green's theorem for V. Thus if U is a region for which Green's theorem holds and if V is another region, $V = \boldsymbol{R}(U)$, where $|\boldsymbol{R}_u \times \boldsymbol{R}_v| \neq 0$, $\boldsymbol{R}$ is one to one, and twice continuously differentiable with $\boldsymbol{R}_u \times \boldsymbol{R}_v$ in the direction of $\boldsymbol{k}$, then Green's theorem holds for V also.

This verifies the following theorem.

Theorem 12.3.1 *(Green's Theorem) Let V be an open set in the plane for which the divergence theorem holds and $\boldsymbol{F}(x,y) = (P(x,y), Q(x,y))$ be a C^1 vector field defined near V. Then if V is oriented counter clockwise,*

$$\int_{\partial V} \boldsymbol{F} \cdot d\boldsymbol{R} = \int_V \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) d\sigma. \tag{12.2}$$

In particular, if there exists U for which the divergence theorem holds and $V = \boldsymbol{R}(U)$ where $\boldsymbol{R} : U \to V$ is $C^2(\overline{U}, \mathbb{R}^2)$ such that $|\boldsymbol{R}_x \times \boldsymbol{R}_y| \neq 0$ and $\boldsymbol{R}_x \times \boldsymbol{R}_y$ is in the direction of $\boldsymbol{k}$, then 12.2 is valid where the orientation around ∂V is consistent with the orientation around U.

This is a very general version of Green's theorem which will include most if not all of what will be of interest. However, there are more general versions of this important theorem. [1] The exercises will present a development of the main topics in complex analysis. For more, see "Analysis of Functions of Complex and Many Variables" or "Analysis of Functions of one Variable". Here I am celebrating the role of Green's theorem more than in the latter of the two books. See the listed books for residue theory.

[1] For a general version see the advanced calculus book by Apostol. Also see my book on calculus of real and complex variables. The general versions involve the concept of a rectifiable Jordan curve. You need to be able to take the area integral and to take the line integral around the boundary. This general version of this theorem appeared in 1951. Green lived in the early 1800's.

12.4 Exercises

1. Show using Green's theorem that the area enclosed by a closed C^1 curve is $\int_C xdy$.

2. Use the above problem to find the area of the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}\leq 1$. Here a,b are positive constants.

3. Let $\boldsymbol{p}_i=(x_i,y_i)$ and consider the polygon $\boldsymbol{p}_0\boldsymbol{p}_1\boldsymbol{p}_2\cdots\boldsymbol{p}_n$ meaning the polygonal curve going from $\boldsymbol{p}_0$ to $\boldsymbol{p}_1$ etc. till you get to $\boldsymbol{p}_n=\boldsymbol{p}_0$ and suppose this polygon forms a curve which does not cross itself and the area is on the left as you walk over the curve with you head pointing in the direction of $\boldsymbol{k}$. Use Green's theorem to obtain an easy to use formula for the area of this polygon. Use $\int_C xdy$ is the area and obtain a simple description for the line integrals from $\boldsymbol{p}_i$ to $\boldsymbol{p}_{i+1}$ then add these together.

4. Using the chain rule, show the following: Suppose C is a piecewise smooth curve which goes from $\boldsymbol{p}$ to $\boldsymbol{q}$. Also suppose that $\boldsymbol{F}(\boldsymbol{x})=\nabla\phi(\boldsymbol{x})$. Then $\int_C \boldsymbol{F}\cdot d\boldsymbol{R}=\phi(\boldsymbol{q})-\phi(\boldsymbol{p})$. Such vector fields are called conservative. **Hint:** To make easier, you could use Lemma 11.5.2 to consider a single parameter domain.

5. Recall that a connected open set is arcwise connected. Show that between any two points in such a connected open set, the is a piecewise smooth curve.

6. Use the above problem to show that in a connected open set U, if the line integral $\int_C \boldsymbol{F}\cdot d\boldsymbol{R}$ joining any two points does not depend on the particular piecewise smooth curve joining them then there exists ϕ a scalar function defined on U such that $\nabla\phi=\boldsymbol{F}$. Thus a vector valued function $\boldsymbol{F}$ defined on U is conservative if and only if the line integrals are path independent.

7. Let U be an open connected set in $\mathbb{R}^2=\mathbb{C}$ and consider the points as complex numbers. That is $x+iy$ means (x,y) and the usual conventions for multiplication hold in which $i^2=-1$. Let $f:U\to\mathbb{C}$. Then f is said to be analytic if $f'(z)\equiv\lim_{h\to 0}\frac{f(z+h)-f(z)}{h}$ exists and is a continuous function of z. Show that for $f(z)=u(x,y)+iv(x,y)$ with u,v respectively the real and imaginary parts of $f(z)$ that f is analytic if and only if the partial derivatives of u,v are continuous and the Cauchy Riemann equations hold, $u_x=v_y,u_y=-v_x$. **Hint:** You should let $h=t$ and $h=it$ for t real and see what happens. Both choices must yield the derivative. The first would be u_x+iv_x the second something similar and these would need to be the same.

8. Using Problem 7 show that u,v are both harmonic. That is $\Delta u=\Delta v=0$. Use Theorem 11.7.5 to verify that if f is bounded and analytic on all of $\mathbb{C}$ then f is constant.

9. Show that many of the usual differentiation formulas hold. For example, the chain rule and product rule and quotient rule. Show $(z^n)'=nz^{n-1}$ for n an integer. Show that polynomials are analytic.

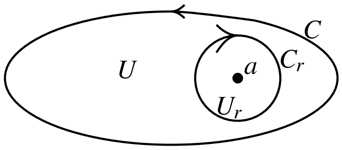
10. Using Problem 17 on Page 276 which defines contour integrals, show that for f an analytic function defined on an open set $V\subseteq\mathbb{C}=\mathbb{R}^2$ and U a region contained in V such that U is a region for which the divergence theorem holds, $\int_C f(z)\,dz=0$ where C is the oriented ∂U. Recall this contour integral is just $\int_a^b f(z(t))\,z'(t)\,dt$ where $z(\cdot):[a,b)\to\mathbb{C}$ is a closed curve and z is a C^1 differentiable map whose derivative might vanish at finitely many points. **Hint:** Show that for u,v the real and

imaginary parts of f, $f(z)z' = ux' - vy' + i(vx' + uy')$. Show that the real part of this contour integral is $\int_C u dx - v dy$ and the imaginary part is $\int_C v dx + u dy$. Apply Greens theorem and Problem 7 to conclude that this contour integral is 0. This is the Cauchy integral theorem which is from Cauchy in the early 1800's and is the foundation for complex analysis. This is roughly the way Cauchy did it.

11. Suppose f is analytic, explain why $\bar{f}$ will usually not be analytic. $\bar{f}(z) = u(x,y) - iv(x,y)$ where $f(z) = u(x,y) + iv(x,y)$.

12. Let C be a piecewise smooth oriented curve in $\mathbb{C}$ and let $f : C \to \mathbb{C}$ be continuous and bounded so that $|f(z)| \leq M$ for some M. Show that $|\int_C f(z)\, dz| \leq ML$ where L is the length of this curve. For C an oriented curve, let $-C$ be the same set of points but oriented in the opposite direction. Explain how $\int_C f dz = -\int_{-C} f dz$. Go right to the definition, $-C$ involves t going from b to a where the interval for the parameter is $[a,b]$.

13. Consider the following picture which illustrates a region for Green's theorem U and inside a small disk of radius r called U_r centered at a which also is a region for Green's theorem. The boundaries of these two are oriented as shown.



Show that the small circle is parametrized by $a + re^{-it}$ for $t \in [0, 2\pi]$. Then justify the following for $f(z) = u(x,y) + iv(x,y)$, analytic on an open set containing $\bar{U}$. First of all show $z \to \frac{f(z)}{z-a}$ is analytic on the region between U and U_r. Next verify that this region is one which works for Green's theorem.

$$\int_C \frac{f(z)}{z-a} dz + \int_{C_r} \frac{f(z)}{z-a} dz = 0$$

Then show that $\lim_{r\to 0} \left| \int_{C_r} \frac{f(z) - \left(f(a) + f'(a)(z-a)\right)}{z-a} dz \right| = 0$ using the differentiability of f and the estimate of Problem 12. Thus $\int_C \frac{f(z)}{z-a} dz + \int_{C_r} \frac{f(a)+f'(a)(z-a)}{z-a} dz = e(r)$ where $\lim_{r\to 0} e(r) = 0$. Now show that $\lim_{r\to 0} \int_{C_r} \frac{f(a)+f'(a)(z-a)}{z-a} dz = -2\pi i f(a)$. Explain why $f(a) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z-a} dz$. This is the famous Cauchy integral formula.

14. Use whatever convergence theorem is useful to show that in the above situation,

$$f^{(n)}(a) = \frac{n!}{2\pi i} \int_C \frac{f(z)}{(z-a)^{n+1}} dz$$

Also show that if $f(z)$ is analytic on all of $\mathbb{C}$ (entire) and $f'(z) = 0$ for all z then $f(z)$ is a constant. Show using the formula of this problem and the estimate of Problem 12 that if f is bounded and entire, then f is a constant. This is Liouville's theorem.

15. The easiest proof of the fundamental theorem of algebra which states that every nonconstant polynomial having complex coefficients has a zero comes from the above Liouville's theorem. If $p(z)$ is a nonconstant polynomial with no zeros, explain why

$1/p(z)$ is analytic and bounded, thus yielding a contradiction to $p(z)$ being nonconstant thanks to Liouville's theorem. **Hint:** If $p(z) = a_n z^n + q(z)$ with $q(z)$ a polynomial with all exponents less than n then $\lim_{|z|\to\infty} \frac{1}{|p(z)|} = \lim_{|z|\to\infty} \frac{1}{|a_n||z|^n} \frac{|z|^n}{|p(z)|} = 0$ and $1/p(z)$ is continuous if $p(z)$ never is 0.

16. Let $f : \mathbb{C} \to \mathbb{C}$ be entire (has a derivative on all of $\mathbb{C}$) and suppose that

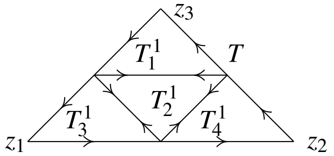
$$\max\{|f(z)| : |z| \leq R\} \leq CR^k.$$

Then show that $f(z)$ is actually a polynomial of degree k. **Hint:** Recall the formula for the derivative in terms of the Cauchy integral in Problem 14. Nothing like this holds for functions of a real variable.

17. Suppose you have a sequence of functions $\{f_n\}$ analytic on an open set U. If they converge uniformly to a function f, show that f is also analytic on U. This is totally opposed to what takes place in real analysis.

18. Suppose $F'(z) = f(z)$. Show that if γ is an oriented piecewise smooth curve from z_0 to z, then $\int_\gamma f(z)\,dz = F(z) - F(z_0)$. In particular, if γ is a closed curve, $\int_\gamma f(z)\,dz = 0$. This F is called a primitive.

19. Suppose f has a derivative on a convex set open set U. Let T be a triangular region along with its boundary ∂T. Orient this boundary counter clockwise. Show $\int_T f(z)\,dz = 0$.

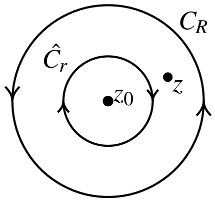


To do this, suppose $|\int_T f(z)\,dz| = \alpha > 0$. Cut up the triangle into four pieces as shown above. Then $\left|\int_{T_j^1} f(z)\,dz\right| \geq \alpha/4$ for one of those triangles. Do the same thing for it, and pick $\left|\int_{T_i^2} f(z)\,dz\right| \geq \alpha/4^2$ each time making the diameter of the new triangles half the diameter of the one before. This yields a nested sequence of compact sets $\{T^n\}$ such that $\operatorname{diam}(T^n) \leq C2^{-n}$. Let z be a point of the intersecton of all of these. Then for $w \in T^n, f(w) = f(z) + f'(z)(w-z) + o(w-z)$ and for all n large enough, $|o(w-z)| < \varepsilon C2^{-n}$. Now explain why $\int_{\partial T^n}(f(z) + f'(z)(w-z))\,dw = 0$. Now explain why $\frac{\alpha}{4^n} \leq \varepsilon C(2^{-n})(2^{-n})$ and so $\alpha \leq C\varepsilon$ a contradiction if ε is chosen small enough. Now pick a point of U called z_0. Let z_0z denote the straight line segment from z_0 to z. Let $F(z) \equiv \int_{z_0z} f(w)\,dw$. Show that $F'(z) = f(z)$. From the Liouville problem above, the same procedure will show that $F^{(n)}(z) = \frac{n!}{2\pi i}\int_C \frac{F(w)dw}{(w-z)^{n+1}}$ which is a continuous function of z. Here C is a circle enclosing z_0. Thus f' is continuous as are all of its derivatives. Note how different this is than the situation in real analysis. If the derivative exists on an open set, then it has to be continuous along with all of the higher order derivatives.

20. In calculus, if you have a continuous function f there is F an antiderivative. Show that this is not true for functions of a complex variable unless f is analytic.

21. Show the Weierstrass approximation theorem will not work to approximate arbitrary continuous functions of a complex variable. **Hint:** You might use the Cauchy integral formula and the above problems to show that continuity is not enough.

22. The region between two circles centered at z_0 is called an annulus.



In the above picture $\hat{C}_r$ is oriented so that the area is on left hand as you walk around with head pointed up from the plane and C_R is also oriented this way. Let C_r be the same circle as $\hat{C}_r$ but oriented counter clockwise so the area **inside** this circle is on the left. Then using the Cauchy integral formula of Problem 13

$$f(z) = \frac{1}{2\pi i}\left(\int_{C_R} \frac{f(w)}{w-z}dw - \int_{C_r} \frac{f(w)}{w-z}dw\right)$$

Then show using the formula for the sum of geometric series that for z in the annulus

$$f(z) = \frac{1}{2\pi i}\left(\int_{C_R} \sum_{k=0}^{\infty} \frac{f(w)}{(w-z_0)^{k+1}} (z-z_0)^k dw + \int_{C_r} \sum_{k=0}^{\infty} \frac{f(w)(w-z_0)^k}{(z-z_0)^{k+1}}\right) dw$$

To get this, do the following. On the second integral, $\frac{f(w)}{w-z} = -\frac{f(w)}{(z-z_0)\left(1-\frac{w-z_0}{z-z_0}\right)}$ and something similar for the first integral. Use formula for sum of geometric series. Explain why for fixed z in the annulus convergence is uniform in both terms so

$$f(z) = \sum_{k=0}^{\infty} \frac{1}{2\pi i} \int_{C_R} \frac{f(w)}{(w-z_0)^{k+1}} dw (z-z_0)^k + \sum_{k=0}^{\infty} \frac{1}{2\pi i} \int_{C_r} f(w)(w-z_0)^k \frac{1}{(z-z_0)^{k+1}}$$

In case f is analytic on all of the inside of C_R, the C_r disappears and so by Problem 14 the above reduces to $f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z-z_0)^k$ showing that if f is analytic near a point z_0 then its power series expansion converges to the function on the largest open disk which does not contain a point where f fails to be analytic. In fact one could define analytic functions as being those correctly given by their power series but this approach is not done here. When the second sum is a finite sum, we say that f has a pole at z_0. Otherwise z_0 is called an essential singularity.

23. Show that if $f(z) = \sum_{k=0}^{\infty} a_k (z-z_0)^k$ for $|z-z_0| \leq r$, then f is analytic on $B(z_0, \hat{r})$ for $\hat{r} < r$. **Hint:** You might show uniform convergence and then use the Cauchy integral formula.

24. If f is analytic on an open connected set U, and if Z is the set of zeros of f, show that f is identically zero if and only if the set of zeros has a limit point **in** U. **Hint:** Suppose z is a limit point of Z. Then by continuity $f(z) = 0$ so the set of zeroes is closed. On the other hand, if z is a limit point of Z then $f(w) = a_m (w-z)^m g(w)$ where

$g(z) \neq 0$ and a_m is the first nonzero coefficient in the power series expansion about z. But you could pick $w_n \to z$ where $f(w_n) = 0$ thus $a_m = 0$ after all. Thus all coefficients of the power series of f at $z = 0$ so $f = 0$ in an open ball containing z. Take a piecewise smooth curve γ from z to some $\hat{z} \in U$. Let $T = \sup\{t : f(\gamma(t)) = 0\}$. Then repeat the argument to show that $f(\hat{z}) = 0$.

25. Using the above problem, show that if you want e^z to be analytic and agree with e^x whenever $z = x \in \mathbb{R}$ then you have no choice but to define $e^z \equiv e^x(\cos(y) + i\sin(y))$. Give similar treatments for $\sin(z)$ and $\cos(z)$. Explain why $\sin(z), \cos(z)$ cannot be bounded using Liouville's theorem. Also verify that all the usual trig identities will continue to hold for arbitrary complex z.

26. Let r be a ray, straight line starting at 0 and proceeding in one direction from there, and let a be an angle associated with this ray r. For $z \notin r$ let $\arg(z) = \theta$ where $z = |z| e^{i\theta}$, for $\theta \in (a - \pi, a + \pi)$. Then defining $\log(z) \equiv \ln(|z|) + i\arg(z)$, this delievers a "branch" of the logarithm associated with this ray. Show that if $a = -\pi$, this is the principle branch and in this case $\log(x) = \ln(x)$ for $x > 0$. In any case, show that $e^{\log(z)} = z$ and that, from geometric reasoning, $z \to \log(z)$ is continuous and satisfies the usual functional equation for logarithms, $\log(z + w) = \log(z) + \log(w)$. Note that $z + w$ stays in the complement of r if both z, w are.

27. Show that if f is analytic and one to one on a connected open set U with $f'(z) \neq 0$ on U and f^{-1} is continuous, then f^{-1} will also be analytic on the connected open set $f(U)$. This will show that the above log function is analytic and the usual calculus theorem holds $\left(f^{-1}\right)'(f(z)) f'(z) = 1$. **Hint:** You could make this an exercise in using the inverse function theorem and the Cauchy Riemann equations. You could prove f^{-1} is continuous if desired, using either invariance of domain or inverse function theorem.

28. Suppose $g(z_0) \neq 0$ where g is analytic near z_0. Then if $m \in \mathbb{N}$, is given. Show that on some ball $B(z_0, r)$, there is an analytic function $\phi(z)$ such that $g(z) = \phi(z)^m$. **Hint:** Pick a ray from 0 which misses $g(z_0)$. Then pick r small enough that $g(B(z_0, r))$ does not intersect this ray. Let $\log(z)$ be a branch of the logarithm associated with this ray as described. Then let $\phi(z) = e^{\frac{1}{m}\log(z)}$. This must be analytic because it is the composition of analytic functions.

29. Suppose f is analytic on $\hat{B} \equiv B(z_0, r) \setminus \{z_0\}$. Show that f can be defined at z_0 such that the resulting function is analytic on $B(z_0, r)$ if and only if $\lim_{z \to z_0}(z - z_0) f(z) = 0$. Such a z_0 is called a removable singularity. **Hint:** One direction is obvious. For the other, consider $h(z) \equiv (z - z_0)^2 f(z), h(z_0) \equiv 0$. Argue that h is analytic on $B(z_0, r)$ and in fact, $h'(z_0) = 0$. Thus $h(z)$ has a power series. Note $\frac{h(z) - h(z_0)}{z - z_0} = (z - z_0) f(z)$.

30. Use the above problem here. Suppose f is analytic near z_0 and $f(\hat{B})$ is not dense in $\mathbb{C}$. This means that there is w and $\delta > 0$ such that $B(w, \delta)$ has no points of $f(\hat{B})$. Then, the Casorati Weierstrass theorem says that $f(z) = g(z) + \sum_{k=1}^{m} \frac{b_k}{(z - z_0)^k}$ for some $m < \infty$. That is, f has a pole at z_0. Show this is the case. Note that this implies the surprising result that if $f(z) = g(z) + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$ for z near z_0, then $f(\hat{B})$ is dense in the plane. In this second case, z_0 is called an isolated essential singularity. The Picard theorems say something even more dramatic, that $f(\hat{B})$ actually is all

of $\mathbb{C}$ except for one exception and in addition, other than this single exception, for $w \in \mathbb{C}$, there are infinitely many $z \in \hat{B}$ such that $f(z) = w$. This theorem is for a more advanced presentation of complex analysis than this short introduction. **Hint for Casorati Weierstrass:** If $f\left(\hat{B}\right)$ is not dense, then there exists w and $\delta > 0$ such that $B(w,\delta) \cap f\left(\hat{B}\right) = \emptyset$. Consider $\frac{1}{f(z)-w}$. This is analytic near z_0 and show that $\lim_{z\to z_0}(z-z_0)\frac{1}{f(z)-w} = 0$. Thus there is $h(z)$ analytic which equals $\frac{1}{f(z)-w}$ near z_0 but which also makes sense at z_0. Consider two cases, $h(z_0) = 0$ and $h(z_0) \neq 0$. In the second case, $f(z) - w = \frac{1}{h(z)}$ which is analytic near 0. Now consider the first case. The zero of h at z_0 has some multiplicity m. Otherwise, you would have $h = 0$ on some ball having z_0 as center.

31. Let C be an oriented closed piecewise smooth curve. Then for $z \notin C$, $n(C,z) \equiv \frac{1}{2\pi i}\int_C \frac{1}{w-z}dw$ is an integer called the winding number. To show this, let $\gamma : [0,2\pi] \to C$ be a C^1 parametrization for C where maybe $\gamma'(t) = 0$ for some finite number of $t, \gamma(2\pi) \equiv \lim_{t\to 2\pi}\gamma(t)$. Then define $F(t) \equiv \int_0^t \frac{\gamma'(s)}{\gamma(s)-z}ds$. Show
$$\left(e^{-F(t)}(\gamma(t)-z)\right)' = \frac{-\gamma'(t)}{\gamma(t)-z}e^{-F(t)}(\gamma(t)-z) + e^{-F(t)}\gamma'(t) = 0$$
thus $e^{-F(t)}(\gamma(t)-z)$ is a constant. $e^{-F(2\pi)}(\gamma(2\pi)-z) = (\gamma(0)-z)$ so, $-F(2\pi) = -2\pi i n$ for some integer n and so $n(C,z) = n$.

32. Suppose U is a connected open set and $f : U \to \mathbb{C}$ is analytic. Show that $f(U)$ is either a single point or a connected open set. **Hint:** Suppose $f(U)$ is not a single point. Then pick $z_0 \in U$. Then near $z_0, f(z) = f(z_0) + \sum_{k=m}^{\infty} a_k (z-z_0)^k = f(z_0) + g(z)(z-z_0)^m$ where $g(z_0) \neq 0$. Not all a_k can equal zero because if so, you would have $f - f(z_0)$ zero in a set with a limit point and f would be constant contrary to the assumption that it is not. Using Problem 28, for z sufficiently close to z_0, $f(z) = f(z_0) + \left(g(z)^{1/m}(z-z_0)\right)^m \equiv f(z_0) + \phi(z)^m$ where $\phi(z_0) = 0, g(z_0) \neq 0$, and $\phi'(z_0) \neq 0$. Now apply the inverse function theorem and Cauchy Riemann equations to obtain that $f(B(z_0,r))$ is an open set for r small enough. This is called the open mapping theorem.

33. Use the above open mapping theorem to show the maximum modulus theorem. If f is analytic on an open connected, bounded set U and continuous on $\bar{U}$ then $|f|$ achieves its maximum value on the boundary of U.

34. If U is an open connected subset of $\mathbb{C}$ and $f : U \to \mathbb{R}$ is analytic, what can you say about f? **Hint:** You might consider the open mapping theorem.

35. When we count the zeros of an analytic function f we count them according to multiplicity. This means that if z_0 is a zero of f so that for z near $z_0, f(z) = a_m(z-z_0)^m + a_{m+1}(z-z_0)^{m+1} + \ldots$, then we would regard this z_0 as a zero of multiplicity m. Show that if C is the boundary of a ball $B(a,r)$ and if $f(z)$ is analytic on an open connected set containing this ball and f has no zeros on C, then the number of zeros of f in the ball is $\frac{1}{2\pi i}\int_C \frac{f'(z)}{f(z)}dz$ if these zeroes are counted according to multiplicity.

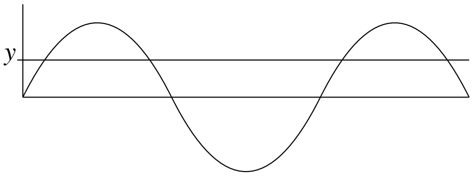
Chapter 13

Degree Theory

This chapter is on the Brouwer degree, a very useful concept with numerous and important applications. The degree can be used to prove some difficult theorems in topology such as the Brouwer fixed point theorem, the Jordan separation theorem, and the invariance of domain theorem. A couple of these big theorems have been presented earlier, but when you have degree theory, they get much easier. Degree theory is also used in bifurcation theory and many other areas in which it is an essential tool. The degree will be developed for $\mathbb{R}^p$ in this book. When this is understood, it is not too difficult to extend to versions of the degree which hold in Banach space. There is more on degree theory in the book by Deimling [11] and much of the presentation here follows this reference. Another more recent book which is really good is [13]. This is a whole book on degree theory.

The original reference for the approach given here, based on analysis, is [24] and dates from 1959. The degree was developed earlier by Brouwer and others using different methods. The more classical approach based on simplices and approximations with these is in [26]. I have given an approach based on singular homology as an appendix in [31].

To give you an idea what the degree is about, consider a real valued C^1 function defined on an interval I, and let $y \in f(I)$ be such that $f'(x) \neq 0$ for all $x \in f^{-1}(y)$. In this case the degree is the sum of the signs of $f'(x)$ for $x \in f^{-1}(y)$, written as $d(f, I, y)$.



In the above picture, $d(f, I, y)$ is 0 because there are two places where the sign is 1 and two where it is -1.

The amazing thing about this is the number you obtain in this simple manner is a specialization of something which is defined for continuous functions and which has nothing to do with differentiability. The reason one can extend the above simple idea to continuous functions is is an integral expression for the degree which is insensitive to homotopy. It is very similar to the winding number of complex analysis. The difference between the two is that with the degree, the integral which ties it all together is taken over the open set while the winding number is taken over the boundary, although proofs of it in the case of the winding number sometimes involve Green's theorem which involves an integral over the open set. I think these analogies are better seen in the other presentation in [31].

In this chapter Ω will refer to a bounded open set.

Definition 13.0.1 *For Ω a bounded open set, denote by $C(\overline{\Omega})$ the set of functions which are restrictions of functions in $C_c(\mathbb{R}^p)$, equivalently $C(\mathbb{R}^p)$ to $\overline{\Omega}$ and by $C^m(\overline{\Omega})$, $m \leq \infty$ the space of restrictions of functions in $C_c^m(\mathbb{R}^p)$, equivalently $C^m(\mathbb{R}^p)$ to $\overline{\Omega}$. If $f \in C(\overline{\Omega})$ the symbol f will also be used to denote a function defined on $\mathbb{R}^p$ equalling f on $\overline{\Omega}$ when convenient. The subscript c indicates that the functions have compact support. The norm in $C(\overline{\Omega})$ is defined as follows.*

$$\|f\|_{\infty,\overline{\Omega}} = \|f\|_\infty \equiv \sup\left\{|f(\boldsymbol{x})| : \boldsymbol{x} \in \overline{\Omega}\right\}.$$

If the functions take values in $\mathbb{R}^p$ write $C^m\left(\overline{\Omega};\mathbb{R}^p\right)$ or $C\left(\overline{\Omega};\mathbb{R}^p\right)$ for these functions if there is no differentiability assumed. The norm on $C\left(\overline{\Omega};\mathbb{R}^p\right)$ is defined in the same way as above,

$$\|\boldsymbol{f}\|_{\infty,\overline{\Omega}} = \|\boldsymbol{f}\|_\infty \equiv \sup\left\{|\boldsymbol{f}(\boldsymbol{x})| : \boldsymbol{x}\in\overline{\Omega}\right\}.$$

If $m=\infty$, the notation means that there are infinitely many derivatives. Also, $C(\Omega;\mathbb{R}^p)$ consists of functions which are continuous on Ω that have values in $\mathbb{R}^p$ and $C^m(\Omega;\mathbb{R}^p)$ denotes the functions which have m continuous derivatives defined on Ω. Also let $\mathscr{P}$ consist of functions $\boldsymbol{f}(\boldsymbol{x})$ such that $f_k(\boldsymbol{x})$ is a polynomial, meaning an element of the algebra of functions generated by $\left\{1,x_1,\cdots,x_p\right\}$. Thus a typical polynomial is of the form $\sum_{i_1\cdots i_p} a(i_1\cdots i_p)x^{i_1}\cdots x^{i_p}$ where the i_j are nonnegative integers and $a(i_1\cdots i_p)$ is a real number.

Some of the theorems are simpler if you base them on the Weierstrass approximation theorem.

Note that, by applying the Tietze extension theorem to the components of the function, one can always extend a function continuous on $\overline{\Omega}$ to all of $\mathbb{R}^p$ so there is no loss of generality in simply regarding functions continuous on $\overline{\Omega}$ as restrictions of functions continuous on $\mathbb{R}^p$. Next is the idea of a regular value.

Definition 13.0.2 *For W an open set in $\mathbb{R}^p$ and $\boldsymbol{g}\in C^1(W;\mathbb{R}^p)$, $\boldsymbol{y}$ is called a regular value of $\boldsymbol{g}$ if whenever $\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})$, $\det(D\boldsymbol{g}(\boldsymbol{x}))\neq 0$. Note that if $\boldsymbol{g}^{-1}(\boldsymbol{y})=\emptyset$, it follows that $\boldsymbol{y}$ is a regular value from this definition. That is, $\boldsymbol{y}$ is a regular value if and only if*

$$\boldsymbol{y}\notin\boldsymbol{g}(\{\boldsymbol{x}\in W : \det D\boldsymbol{g}(\boldsymbol{x})=0\})$$

Denote by $S_{\boldsymbol{g}}$ the set of singular values of $\boldsymbol{g}$, those $\boldsymbol{y}$ such that $\det(D\boldsymbol{g}(\boldsymbol{x}))=0$ for some $\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})$.

Also, $\partial\Omega$ will often be referred to. It is those points with the property that every open set (or open ball) containing the point contains points not in Ω and points in Ω. Then the following simple lemma will be used frequently.

Lemma 13.0.3 *Define ∂U to be those points $\boldsymbol{x}$ with the property that for every $r>0$, $B(\boldsymbol{x},r)$ contains points of U and points of U^C. Then for U an open set, $\partial U=\overline{U}\setminus U$. Let C be a closed subset of $\mathbb{R}^p$ and let $\mathscr{K}$ denote the set of components of $\mathbb{R}^p\setminus C$. Then if K is one of these components, it is open and $\partial K\subseteq C$.*

Proof: Let $\boldsymbol{x}\in\overline{U}\setminus U$. If $B(\boldsymbol{x},r)$ contains no points of U, then $\boldsymbol{x}\notin\overline{U}$. If $B(\boldsymbol{x},r)$ contains no points of U^C, then $\boldsymbol{x}\in U$ and so $\boldsymbol{x}\notin\overline{U}\setminus U$. Therefore, $\overline{U}\setminus U\subseteq\partial U$. Now let $\boldsymbol{x}\in\partial U$. If $\boldsymbol{x}\in U$, then since U is open there is a ball containing $\boldsymbol{x}$ which is contained in U contrary to $\boldsymbol{x}\in\partial U$. Therefore, $\boldsymbol{x}\notin U$. If $\boldsymbol{x}$ is not a limit point of U, then some ball containing $\boldsymbol{x}$ contains no points of U contrary to $\boldsymbol{x}\in\partial U$. Therefore, $\boldsymbol{x}\in\overline{U}\setminus U$ which shows the two sets are equal.

Why is K open for K a component of $\mathbb{R}^p\setminus C$? This follows from Theorem 3.11.12 and results from open balls being connected. Thus if $k\in K$, letting $B(k,r)\subseteq C^C$, it follows $K\cup B(k,r)$ is connected and contained in C^C and therefore is contained in K because K is maximal with respect to being connected and contained in C^C.

Now for K a component of $\mathbb{R}^p\setminus C$, why is $\partial K\subseteq C$? Let $\boldsymbol{x}\in\partial K$. If $\boldsymbol{x}\notin C$, then $\boldsymbol{x}\in K_1$, some component of $\mathbb{R}^p\setminus C$. If $K_1\neq K$ then $\boldsymbol{x}$ cannot be a limit point of K and so it cannot be in ∂K. Therefore, $K = K_1$ but this also is a contradiction because if $\boldsymbol{x} \in \partial K$ then $\boldsymbol{x} \notin K$ thanks to the first part that $\partial U = \overline{U} \setminus U$. ■

Note that for an open set $U \subseteq \mathbb{R}^p$, and $\boldsymbol{h} : \overline{U} \to \mathbb{R}^p$, $\operatorname{dist}(\boldsymbol{h}(\partial U), \boldsymbol{y}) \geq \operatorname{dist}\left(\boldsymbol{h}\left(\overline{U}\right), \boldsymbol{y}\right)$ because $\overline{U} \supseteq \partial U$.

The following lemma will be nice to keep in mind.

Lemma 13.0.4 *$\boldsymbol{f} \in C\left(\overline{\Omega} \times [a,b]; \mathbb{R}^p\right)$ if and only if*

$$t \to \boldsymbol{f}(\cdot, t) \in C\left([a,b]; C\left(\overline{\Omega}; \mathbb{R}^p\right)\right)$$

Also

$$\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} = \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}}\right)$$

Proof: ⇒By uniform continuity, if $\varepsilon > 0$ there is $\delta > 0$ such that if $|t - s| < \delta$, then for all $\boldsymbol{x} \in \overline{\Omega}$, $\|\boldsymbol{f}(\boldsymbol{x}, t) - \boldsymbol{f}(\boldsymbol{x}, s)\| < \frac{\varepsilon}{2}$. It follows that

$$\|\boldsymbol{f}(\cdot, t) - \boldsymbol{f}(\cdot, s)\|_\infty \leq \frac{\varepsilon}{2} < \varepsilon$$

⇐Say $(\boldsymbol{x}_n, t_n) \to (\boldsymbol{x}, t)$. Does it follow that $\boldsymbol{f}(\boldsymbol{x}_n, t_n) \to \boldsymbol{f}(\boldsymbol{x}, t)$?

$$\begin{aligned} \|\boldsymbol{f}(\boldsymbol{x}_n, t_n) - \boldsymbol{f}(\boldsymbol{x}, t)\| &\leq \|\boldsymbol{f}(\boldsymbol{x}_n, t_n) - \boldsymbol{f}(\boldsymbol{x}_n, t)\| + \|\boldsymbol{f}(\boldsymbol{x}_n, t) - \boldsymbol{f}(\boldsymbol{x}, t)\| \\ &\leq \|\boldsymbol{f}(\cdot, t_n) - \boldsymbol{f}(\cdot, t)\|_\infty + \|\boldsymbol{f}(\boldsymbol{x}_n, t) - \boldsymbol{f}(\boldsymbol{x}, t)\| \end{aligned}$$

both terms converge to 0, the first because $\boldsymbol{f}$ is continuous into $C\left(\overline{\Omega}; \mathbb{R}^p\right)$ and the second because $\boldsymbol{x} \to \boldsymbol{f}(\boldsymbol{x}, t)$ is continuous.

The claim about the norms is next. Let $(\boldsymbol{x}, t)$ be such that $\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} < \|\boldsymbol{f}(\boldsymbol{x}, t)\| + \varepsilon$. Then

$$\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} < \|\boldsymbol{f}(\boldsymbol{x}, t)\| + \varepsilon \leq \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}}\right) + \varepsilon$$

and so $\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} \leq \max_{t \in [a,b]} \max \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}}\right)$ because ε is arbitrary. However, the same argument works in the other direction. There exists t such that

$$\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} = \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}}\right)$$

by compactness of the interval. Then by compactness of $\overline{\Omega}$, there is $\boldsymbol{x}$ such that

$$\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} = \|\boldsymbol{f}(\boldsymbol{x}, t)\| \leq \|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]}$$

and so the two norms are the same. ■

13.1 Sard's Lemma and Approximation

First are easy assertions about approximation of continuous functions with smooth ones.

The following is the Weierstrass approximation theorem. It is Corollary 5.5.3 presented earlier.

Corollary 13.1.1 *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 $[a,b]$. The polynomials are of the form*

$$\sum_{k=0}^{m} p_k(t) f\left(l\left(\frac{k}{m}\right)\right) \tag{13.1}$$

where l is a linear one to one and onto map from $[0,1]$ to $[a,b]$ and $p_0(a) = 1$ but $p_k(a) = 0$ if $k \neq 0$, $p_m(b) = 1$ but $p_k(b) = 0$ for $k \neq m$.

Applying the Weierstrass approximation theorem, Theorem 5.7.7 or Theorem 5.9.5 to the components of a vector valued function yields the following Theorem.

Theorem 13.1.2 *If $\boldsymbol{f} \in C(\overline{\Omega};\mathbb{R}^p)$ for Ω a bounded subset of $\mathbb{R}^p$, then for any $\varepsilon > 0$, there exists $\boldsymbol{g} \in C^\infty(\overline{\Omega};\mathbb{R}^p)$ such that $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \varepsilon$.*

Recall Sard's lemma, shown earlier. It is Lemma 10.4.3. I am stating it here for convenience.

Lemma 13.1.3 *(Sard) Let Ω be an open set in $\mathbb{R}^p$ and let $\boldsymbol{h} : \Omega \to \mathbb{R}^p$ be differentiable. Let*

$$S \equiv \{\boldsymbol{x} \in \Omega : \det D\boldsymbol{h}(\boldsymbol{x}) = 0\}.$$

Then $m_p(\boldsymbol{h}(S)) = 0$.

First note that if $\boldsymbol{y} \notin \boldsymbol{g}(\Omega)$, then $\boldsymbol{y} \notin \boldsymbol{g}(\{\boldsymbol{x} \in \Omega : \det D\boldsymbol{g}(\boldsymbol{x}) = 0\})$ so it is a regular value.

Observe that any uncountable set in $\mathbb{R}^p$ has a limit point. To see this, tile $\mathbb{R}^p$ with countably many congruent boxes. One of them has uncountably many points. Now subdivide this into 2^p congruent boxes. One has uncountably many points. Continue subdividing this way to obtain a limit point as the unique point in the intersection of a nested sequence of compact sets whose diameters converge to 0.

Lemma 13.1.4 *Let $\boldsymbol{g} \in C^\infty(\mathbb{R}^p;\mathbb{R}^p)$ and let $\{\boldsymbol{y}_i\}_{i=1}^\infty$ be points of $\mathbb{R}^p$ and let $\eta > 0$. Then there exists $\boldsymbol{e}$ with $\|\boldsymbol{e}\| < \eta$ and $\boldsymbol{y}_i + \boldsymbol{e}$ is a regular value for $\boldsymbol{g}$ for all i.*

Proof: Let $S = \{\boldsymbol{x} \in \mathbb{R}^p : \det D\boldsymbol{g}(\boldsymbol{x}) = 0\}$. By Sard's lemma, $\boldsymbol{g}(S)$ has measure zero. Let $N \equiv \cup_{i=1}^\infty (\boldsymbol{g}(S) - \boldsymbol{y}_i)$. Thus N has measure 0. Pick $\boldsymbol{e} \in B(\boldsymbol{0},\eta) \setminus N$. Then for each $i, \boldsymbol{y}_i + \boldsymbol{e} \notin \boldsymbol{g}(S)$. ■

Next we approximate $\boldsymbol{f}$ with a smooth function $\boldsymbol{g}$ such that each $\boldsymbol{y}_i$ is a regular value of $\boldsymbol{g}$.

Lemma 13.1.5 *Let $\boldsymbol{f} \in C(\overline{\Omega};\mathbb{R}^p)$, Ω a bounded open set, and let $\{\boldsymbol{y}_i\}_{i=1}^\infty$ be points not in $\boldsymbol{f}(\partial\Omega)$ and let $\delta > 0$. Then there exists $\boldsymbol{g} \in C^\infty(\overline{\Omega};\mathbb{R}^p)$ such that $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$ and $\boldsymbol{y}_i$ is a regular value for $\boldsymbol{g}$ for each i. That is,* **if** *$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{y}_i$, then $D\boldsymbol{g}(\boldsymbol{x})^{-1}$ exists. Also, if $\delta < \operatorname{dist}(\boldsymbol{f}(\partial\Omega), \boldsymbol{y})$ for some $\boldsymbol{y}$ a regular value of $\boldsymbol{g} \in C^\infty(\overline{\Omega};\mathbb{R}^p)$, then $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is a finite set of points in Ω. Also, if $\boldsymbol{y}$ is a regular value of $\boldsymbol{g} \in C^\infty(\mathbb{R}^p,\mathbb{R}^p)$, then $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is countable.*

Proof: Pick $\tilde{\boldsymbol{g}} \in C^{\infty}\left(\overline{\Omega};\mathbb{R}^p\right), \|\tilde{\boldsymbol{g}}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$. From Lemma 13.1.4, $\boldsymbol{y}_i+\boldsymbol{e}$ is a regular value for $\tilde{\boldsymbol{g}}$ for each i where $\boldsymbol{e}$ can be chosen as small as desired. Let $\boldsymbol{g}=\tilde{\boldsymbol{g}}-\boldsymbol{e}$ where $\boldsymbol{e}$ is so small that also $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$. Thus $\boldsymbol{y}_i$ is a regular value of $\boldsymbol{g}$ for all i. (same as $\boldsymbol{y}_i+\boldsymbol{e}$ regular value of $\tilde{\boldsymbol{g}}$). This shows the first part.

It remains to verify the last claims. Since $\|\boldsymbol{g}-\boldsymbol{f}\|_{\overline{\Omega},\infty} < \delta$, if $\boldsymbol{x}\in\partial\Omega$, then

$$\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{y}\| \geq \|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y}\|-\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x})\| \geq \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})-\delta > \delta-\delta=0$$

and so $\boldsymbol{y}\notin\boldsymbol{g}(\partial\Omega)$, so if $\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{y}$, then $\boldsymbol{x}\in\Omega$. Thus $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is a compact subset of Ω and so for each $\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})$ there is a ball containing $\boldsymbol{x}$, $B_{\boldsymbol{x}}$ contained in Ω such that there is at most one point in $\boldsymbol{g}^{-1}(\boldsymbol{y})\cap B_{\boldsymbol{x}}$ this by the inverse function theorem. Finitely many of these balls cover $\boldsymbol{g}^{-1}(\boldsymbol{y})$ so this set must be finite and at each point, the determinant of the derivative of $\boldsymbol{g}$ is nonzero. For $\boldsymbol{y}$ a regular value, $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is countable since otherwise, there would be a limit point $\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})$ and $\boldsymbol{g}$ would fail to be one to one near $\boldsymbol{x}$ contradicting the inverse function theorem. ■

Now with this, here is a definition of the degree.

Definition 13.1.6 *Let Ω be a bounded open set in $\mathbb{R}^p$ and let $\boldsymbol{f}:\overline{\Omega}\to\mathbb{R}^p$ be continuous. Let $\boldsymbol{y}\notin\boldsymbol{f}(\partial\Omega)$. Then the degree is defined as follows: Let $\boldsymbol{g}$ be infinitely differentiable,*

$$\|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}} < \delta \equiv \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y}),$$

and $\boldsymbol{y}$ is a regular value of $\boldsymbol{g}$. Then $\boldsymbol{y}\notin\boldsymbol{g}(\partial\Omega)$ and we define

$$d(\boldsymbol{f},\Omega,\boldsymbol{y}) \equiv \sum\left\{\operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}))) : \boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y}), \boldsymbol{x}\in\Omega\right\}$$

where the sum is finite by Lemma 13.1.5, defined to equal 0 if $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is empty.

Note that if $\boldsymbol{g}$ is such an approximation of $\boldsymbol{f}$ then if $\boldsymbol{x}\in\partial\Omega$ and $t\in[0,1]$,

$$\begin{aligned} |t\boldsymbol{g}(\boldsymbol{x})+(1-t)\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y}| &\geq |\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y}|-t\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty} \\ &> \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})-\operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})=0 \end{aligned}$$

Thus $t\,\boldsymbol{g}+(1-t)\boldsymbol{f}$ maps no point of $\partial\Omega$ to $\boldsymbol{y}$. In particular, $\boldsymbol{g}$ maps no point of $\partial\Omega$ to $\boldsymbol{y}$.

Lemma 13.1.7 *The above sum in the definition makes sense for a single $\boldsymbol{g}$ and, assuming this definition of $d(\boldsymbol{f},\Omega,\boldsymbol{y})$ is well defined, then it would follow that if $\boldsymbol{y}\notin\boldsymbol{f}(\Omega)$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y})=0$.*

Proof: As just noted, if $\|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}} < \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})$ then $\boldsymbol{y}\notin\boldsymbol{g}(\partial\Omega)$. In fact

$$\boldsymbol{y}\notin(t\boldsymbol{g}(\boldsymbol{x})+(1-t)\boldsymbol{f}(\boldsymbol{x}))(\partial\Omega)$$

for any $t\in[0,1]$. Thus the sum is a finite sum and makes sense by Lemma 13.1.5. What if $\boldsymbol{y}\notin\boldsymbol{f}(\Omega)$? In this case, assuming the definition is well defined, you could pick $\boldsymbol{g}$ such that $\boldsymbol{y}$ is a regular value for $\boldsymbol{g}$ and also $\|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}} < \operatorname{dist}\left(\boldsymbol{f}\left(\overline{\Omega}\right),\boldsymbol{y}\right)$ so the above definition would say that $d(\boldsymbol{f},\Omega,\boldsymbol{y})=0$ because there would be no terms in the sum. ■

We really need to verify that this definition is well defined, not dependent on which $\boldsymbol{g}$ is chosen. This involves the use of an integral.

Next is an identity. It was Lemma 6.11.2 on Page 159.

Lemma 13.1.8 *Let $\boldsymbol{g} : \Omega \to \mathbb{R}^p$ be C^2 where Ω is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^{p} \operatorname{cof}(D\boldsymbol{g})_{ij,j} = 0,$$

where here $(D\boldsymbol{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij} = \frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}}$.

Next is an integral representation of $\sum\left\{\operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}))) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$ but first is a little lemma about disjoint sets.

Lemma 13.1.9 *Let K be a compact set and C a closed set in $\mathbb{R}^p$ such that $K \cap C = \emptyset$. Then*

$$\operatorname{dist}(K, C) \equiv \inf\{\|\boldsymbol{k} - \boldsymbol{c}\| : \boldsymbol{k} \in K, \boldsymbol{c} \in C\} > 0.$$

Proof: Let $d \equiv \inf\{\|\boldsymbol{k} - \boldsymbol{c}\| : \boldsymbol{k} \in K, \boldsymbol{c} \in C\}$. Let $\{\boldsymbol{k}_i\}, \{\boldsymbol{c}_i\}$ be such that

$$d + \frac{1}{i} > \|\boldsymbol{k}_i - \boldsymbol{c}_i\|.$$

Since K is compact, there is a subsequence still denoted by $\{\boldsymbol{k}_i\}$ such that $\boldsymbol{k}_i \to \boldsymbol{k} \in K$. Then also

$$\|\boldsymbol{c}_i - \boldsymbol{c}_m\| \leq \|\boldsymbol{c}_i - \boldsymbol{k}_i\| + \|\boldsymbol{k}_i - \boldsymbol{k}_m\| + \|\boldsymbol{c}_m - \boldsymbol{k}_m\|$$

If $d = 0$, then as $m, i \to \infty$ it follows $\|\boldsymbol{c}_i - \boldsymbol{c}_m\| \to 0$ and so $\{\boldsymbol{c}_i\}$ is a Cauchy sequence which must converge to some $\boldsymbol{c} \in C$. But then $\|\boldsymbol{c} - \boldsymbol{k}\| = \lim_{i\to\infty} \|\boldsymbol{c}_i - \boldsymbol{k}_i\| = 0$ and so $\boldsymbol{c} = \boldsymbol{k} \in C \cap K$, a contradiction to these sets being disjoint. ■

In particular the distance between a point and a closed set is always positive if the point is not in the closed set. Of course this is obvious even without the above lemma.

Definition 13.1.10 *Let $\boldsymbol{g} \in C^\infty\left(\overline{\Omega}; \mathbb{R}^p\right)$ where Ω is a bounded open set. Also let ϕ_ε be a mollifier.*

$$\phi_\varepsilon \in C_c^\infty(B(\boldsymbol{0}, \varepsilon)),\ \phi_\varepsilon \geq 0,\ \int \phi_\varepsilon dx = 1.$$

The idea is that ε will converge to 0 to get suitable approximations.

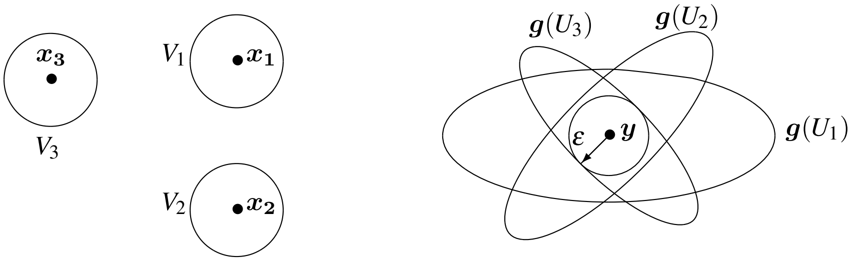
First, here is a technical lemma which will be used to identify the degree with an integral.

Lemma 13.1.11 *Let $\boldsymbol{y} \notin \boldsymbol{g}(\partial\Omega)$ for $\boldsymbol{g} \in C^\infty\left(\overline{\Omega}; \mathbb{R}^p\right)$. Also suppose $\boldsymbol{y}$ is a regular value of $\boldsymbol{g}$. Then for all positive ε small enough,*

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx = \sum\left\{\operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x})) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$$

Proof: First note that the sum is finite from Lemma 13.1.5. It only remains to verify the equation. If $\boldsymbol{y} \notin \boldsymbol{g}(\Omega)$, then for $\varepsilon < \operatorname{dist}\left(\boldsymbol{g}\left(\overline{\Omega}\right), \boldsymbol{y}\right)$, $\phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) = 0$ for all $\boldsymbol{x} \in \Omega$ so both sides equal 0.

I need to show the left side of this equation is constant for ε small enough and equals the right side. By what was just shown, there are finitely many points, $\{\boldsymbol{x}_i\}_{i=1}^m = \boldsymbol{g}^{-1}(\boldsymbol{y})$. By the inverse function theorem, there exist disjoint open sets U_i with $\boldsymbol{x}_i \in U_i$, such that $\boldsymbol{g}$ is one to one on U_i with $\det(D\boldsymbol{g}(\boldsymbol{x}))$ having constant sign on U_i and $\boldsymbol{g}(U_i)$ is an open set containing $\boldsymbol{y}$. Then let ε be small enough that $B(\boldsymbol{y},\varepsilon) \subseteq \cap_{i=1}^m \boldsymbol{g}(U_i)$. Also, $\boldsymbol{y} \notin \boldsymbol{g}\left(\overline{\Omega} \setminus \left(\cup_{i=1}^n U_i\right)\right)$, a compact set. Let ε be still smaller, if necessary, so that $B(\boldsymbol{y},\varepsilon) \cap \boldsymbol{g}\left(\overline{\Omega} \setminus \left(\cup_{i=1}^n U_i\right)\right) = \emptyset$ and let $V_i \equiv \boldsymbol{g}^{-1}(B(\boldsymbol{y},\varepsilon)) \cap U_i$.



Therefore, for any ε this small,

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx = \sum_{i=1}^m \int_{V_i} \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx$$

The reason for this is as follows. The integrand on the left is nonzero only if $\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y} \in B(\boldsymbol{0},\varepsilon)$ which occurs only if $\boldsymbol{g}(\boldsymbol{x}) \in B(\boldsymbol{y},\varepsilon)$ which is the same as $\boldsymbol{x} \in \boldsymbol{g}^{-1}(B(\boldsymbol{y},\varepsilon))$. Therefore, the integrand is nonzero only if $\boldsymbol{x}$ is contained in exactly one of the disjoint sets, V_i. Now using the change of variables theorem, ($\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}, \boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z}) = \boldsymbol{x}$.)

$$= \sum_{i=1}^m \int_{\boldsymbol{g}(V_i)-\boldsymbol{y}} \phi_\varepsilon(\boldsymbol{z}) \det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right| dz \tag{13.2}$$

By the chain rule, $I = D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})$ and so in the above for a single V_i,

$$\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right|$$

$$= \operatorname{sgn}\left(\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right) \left|\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right| \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right|$$

$$= \operatorname{sgn}\left(\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right) = \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x})) = \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)).$$

Therefore, 13.2 reduces to

$$\sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)) \int_{\boldsymbol{g}(V_i)-\boldsymbol{y}} \phi_\varepsilon(\boldsymbol{z})\, dz =$$

$$\sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)) \int_{B(\boldsymbol{0},\varepsilon)} \phi_\varepsilon(\boldsymbol{z})\, dz = \sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)).$$

In case $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \emptyset$, there exists $\varepsilon > 0$ such that $\boldsymbol{g}\left(\overline{\Omega}\right) \cap B(\boldsymbol{y},\varepsilon) = \emptyset$ and so for ε this small,

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx = 0. \blacksquare$$

As noted above, this will end up being $d(\boldsymbol{g},\Omega,\boldsymbol{y})$ in this last case where $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \emptyset$.

Lemma 13.1.12 *Suppose $\boldsymbol{g},\hat{\boldsymbol{g}}$ both satisfy Definition 13.1.6. For δ given there, $\delta =$* $\operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})$,

$$\delta > \|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}},\ \delta > \|\boldsymbol{f}-\hat{\boldsymbol{g}}\|_{\infty,\overline{\Omega}}$$

Then for $t\in[0,1]$ so does $t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}}$. In particular, $\boldsymbol{y}\notin(t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}})(\partial\Omega)$. Also $d(\boldsymbol{f}-\boldsymbol{y},\Omega,\boldsymbol{0}) = d(\boldsymbol{f},\Omega,\boldsymbol{y})$.

Proof: This follows from the fact that $B(\boldsymbol{y},\delta)$ in $\|\cdot\|_{\infty,\overline{\Omega}}$ is convex. From the triangle inequality, if $t\in[0,1]$,

$$\begin{aligned}\|\boldsymbol{f}-(t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}})\|_\infty &\leq t\|\boldsymbol{f}-\boldsymbol{g}\|_\infty+(1-t)\|\boldsymbol{f}-\hat{\boldsymbol{g}}\|_\infty \\ &< t\delta+(1-t)\delta=\delta.\end{aligned}$$

If $\|\boldsymbol{h}-\boldsymbol{f}\|_\infty<\delta$, as was just shown for $\boldsymbol{h}\equiv t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}}$**,** then if $\boldsymbol{x}\in\partial\Omega$,

$$\|\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x})\|\geq\|\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x})\|-\|\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})\|>\operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})-\delta\geq\delta-\delta=0$$

Now consider the last claim. This follows because $\|\boldsymbol{g}-\boldsymbol{f}\|_\infty$ small is the same as $\|\boldsymbol{g}-\boldsymbol{y}-(\boldsymbol{f}-\boldsymbol{y})\|_\infty$ being small. They are the same. Also, $(\boldsymbol{g}-\boldsymbol{y})^{-1}(\boldsymbol{0})=\boldsymbol{g}^{-1}(\boldsymbol{y})$ and $D\boldsymbol{g}(\boldsymbol{x})=D(\boldsymbol{g}-\boldsymbol{y})(\boldsymbol{x})$. ■

Next is an important result on homotopy which is used to show that Definition 13.1.6 is well defined.

Lemma 13.1.13 *If $\boldsymbol{h}$ is in $C^\infty\left(\overline{\Omega}\times[a,b],\mathbb{R}^p\right)$, and $\boldsymbol{0}\notin\boldsymbol{h}(\partial\Omega\times[a,b])$ then for $0<\varepsilon<$* $\operatorname{dist}(\boldsymbol{0},\boldsymbol{h}(\partial\Omega\times[a,b]))$,

$$t\to\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

is constant for $t\in[a,b]$. As a special case, $d(\boldsymbol{f},\Omega,\boldsymbol{y})$ is well defined. Also, if $\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\right)$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y})=0$.

Proof: By continuity of $\boldsymbol{h}$, $\boldsymbol{h}(\partial\Omega\times[a,b])$ is compact and so is at a positive distance from $\boldsymbol{0}$. Let $\varepsilon>0$ be such that for all $t\in[a,b]$,

$$B(\boldsymbol{0},\varepsilon)\cap\boldsymbol{h}(\partial\Omega\times[a,b])=\emptyset \tag{13.3}$$

Define for $t\in(a,b)$, $H(t)\equiv\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$. I will show that $H'(t)=0$ on (a,b). Then, since H is continuous on $[a,b]$, it will follow from the mean value theorem that $H(t)$ is constant on $[a,b]$. If $t\in(a,b)$,

$$H'(t)=\int_\Omega\sum_\alpha\phi_{\varepsilon,\alpha}(\boldsymbol{h}(\boldsymbol{x},t))h_{\alpha,t}(\boldsymbol{x},t)\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

$$+\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\sum_{\alpha,j}\det D_1(\boldsymbol{h}(\boldsymbol{x},t))_{,\alpha j}h_{\alpha,jt}dx\equiv\boldsymbol{A}+\boldsymbol{B}. \tag{13.4}$$

In this formula, the function det is considered as a function of the n^2 entries in the $n\times n$ matrix and the $,\alpha j$ represents the derivative with respect to the αj^{th} entry $h_{\alpha,j}$. Now as in the proof of Lemma 6.11.2 on Page 159, $\det D_1(\boldsymbol{h}(\boldsymbol{x},t))_{,\alpha j}=(\operatorname{cof}D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j}$ and so

$$\boldsymbol{B}=\int_\Omega\sum_\alpha\sum_j\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))(\operatorname{cof}D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j}h_{\alpha,jt}dx.$$

By hypothesis

$$\boldsymbol{x}\to\phi_{\varepsilon}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\left(\text{cof }D_{1}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)_{\alpha j}\text{ for }\boldsymbol{x}\in\Omega$$

is in $C_c^{\infty}(\Omega)$ because if $\boldsymbol{x}\in\partial\Omega$, it follows that for all $t\in[a,b]$, $\boldsymbol{h}(\boldsymbol{x},t)\notin B(\boldsymbol{0},\varepsilon)$ and so $\phi_{\varepsilon}(\boldsymbol{h}(\boldsymbol{x},t))=0$ off some compact set contained in Ω. Therefore, integrate by parts and write

$$\boldsymbol{B}=-\int_{\Omega}\sum_{\alpha}\sum_{j}\frac{\partial}{\partial x_{j}}\left(\phi_{\varepsilon}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)\left(\text{cof }D_{1}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)_{\alpha j}h_{\alpha,t}dx+$$

$$-\int_{\Omega}\sum_{\alpha}\sum_{j}\phi_{\varepsilon}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\left(\text{cof}D\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)_{\alpha j,j}h_{\alpha,t}dx$$

The second term equals zero by Lemma 13.1.8. Simplifying the first term yields

$$\begin{aligned}\boldsymbol{B}&=-\int_{\Omega}\sum_{\alpha}\sum_{j}\sum_{\beta}\phi_{\varepsilon,\beta}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)h_{\beta,j}h_{\alpha,t}\left(\text{cof}D_{1}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)_{\alpha j}dx\\&=-\int_{\Omega}\sum_{\alpha}\sum_{\beta}\phi_{\varepsilon,\beta}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)h_{\alpha,t}\sum_{j}h_{\beta,j}\left(\text{cof}D_{1}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)_{\alpha j}dx\end{aligned}$$

Now the sum on j is the dot product of the β^{th} row with the α^{th} row of the cofactor matrix which equals zero unless $\beta=\alpha$ because it would be a cofactor expansion of a matrix with two equal rows. When $\beta=\alpha$, the sum on j reduces to $\det(D_1(\boldsymbol{h}(\boldsymbol{x},t)))$. Thus $\boldsymbol{B}$ reduces to

$$=-\int_{\Omega}\sum_{\alpha}\phi_{\varepsilon,\alpha}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)h_{\alpha,t}\det\left(D_{1}\left(\boldsymbol{h}\left(\boldsymbol{x},t\right)\right)\right)dx$$

Which is the same thing as $\boldsymbol{A}$, but with the opposite sign. Hence $\boldsymbol{A}+\boldsymbol{B}$ in 13.4 is 0 and $H'(t)=0$ and so H is a constant on $[a,b]$.

Finally consider the last claim. If $\boldsymbol{g},\hat{\boldsymbol{g}}$ both work in the definition for the degree, then consider $\boldsymbol{h}(\boldsymbol{x},t)\equiv t\boldsymbol{g}(\boldsymbol{x})+(1-t)\hat{\boldsymbol{g}}(\boldsymbol{x})-\boldsymbol{y}$ for $t\in[0,1]$. For $\boldsymbol{x}\in\partial\Omega$,

$$\begin{aligned}&\left|t\boldsymbol{g}\left(\boldsymbol{x}\right)+\left(1-t\right)\hat{\boldsymbol{g}}\left(\boldsymbol{x}\right)-\boldsymbol{y}\right|\\=\;&\left|t\left(\boldsymbol{g}\left(\boldsymbol{x}\right)-\boldsymbol{f}\left(\boldsymbol{x}\right)\right)+\left(1-t\right)\left(\hat{\boldsymbol{g}}\left(\boldsymbol{x}\right)-\boldsymbol{f}\left(\boldsymbol{x}\right)\right)+\boldsymbol{f}\left(\boldsymbol{x}\right)-\boldsymbol{y}\right|\end{aligned}$$

$$\begin{aligned}&\geq\left|\boldsymbol{f}\left(\boldsymbol{x}\right)-\boldsymbol{y}\right|-\left|t\left(\boldsymbol{g}\left(\boldsymbol{x}\right)-\boldsymbol{f}\left(\boldsymbol{x}\right)\right)+\left(1-t\right)\left(\hat{\boldsymbol{g}}\left(\boldsymbol{x}\right)-\boldsymbol{f}\left(\boldsymbol{x}\right)\right)\right|\\&\geq\text{dist}\left(\boldsymbol{f}\left(\partial\Omega\right),\boldsymbol{y}\right)-\left(t\left\|\boldsymbol{g}-\boldsymbol{f}\right\|_{\infty}+\left(1-t\right)\left\|\hat{\boldsymbol{g}}-\boldsymbol{f}\right\|_{\infty}\right)\\&>\text{dist}\left(\boldsymbol{f}\left(\partial\Omega\right),\boldsymbol{y}\right)-\left(t\delta+\left(1-t\right)\delta\right)=0\end{aligned}$$

From Lemma 13.1.12, $\boldsymbol{h}$ satisfies what is needed for the first part of this lemma. Namely, $\boldsymbol{0}\notin\boldsymbol{h}(\partial\Omega\times[0,1])$. Then from the first part, if $0<\varepsilon<\text{dist}(\boldsymbol{0},\boldsymbol{h}(\partial\Omega\times[0,1]))$ and ε is also sufficiently small that the second and last equations hold in what follows,

$$d\left(\boldsymbol{f},\Omega,\boldsymbol{y}\right)=\sum\left\{\text{sgn}\left(\det\left(D\boldsymbol{g}\left(\boldsymbol{x}\right)\right)\right):\boldsymbol{x}\in\boldsymbol{g}^{-1}\left(\boldsymbol{y}\right)\right\}=\int_{\Omega}\phi_{\varepsilon}\left(\boldsymbol{h}\left(\boldsymbol{x},1\right)\right)\det D_{1}\boldsymbol{h}\left(\boldsymbol{x},1\right)dx$$

$$=\int_{\Omega}\phi_{\varepsilon}\left(\boldsymbol{h}\left(\boldsymbol{x},0\right)\right)\det D_{1}\boldsymbol{h}\left(\boldsymbol{x},0\right)dx=\sum\left\{\text{sgn}\left(\det\left(D\hat{\boldsymbol{g}}\left(\boldsymbol{x}\right)\right)\right):\boldsymbol{x}\in\hat{\boldsymbol{g}}^{-1}\left(\boldsymbol{y}\right)\right\}\;\blacksquare$$

13.2 Properties of the Degree

Now that the degree for a continuous function has been defined, it is time to consider properties of the degree. In particular, it is desirable to prove a theorem about homotopy invariance which depends only on continuity considerations.

Theorem 13.2.1 *If $\boldsymbol{h}$ is in $C\left(\overline{\Omega}\times[a,b],\mathbb{R}^p\right)$, and $\mathbf{0}\notin \boldsymbol{h}(\partial\Omega\times[a,b])$ for each t, then $t\to d(\boldsymbol{h}(\cdot,t),\Omega,\mathbf{0})$ is constant for $t\in[a,b]$.*

Proof: Let $0<\delta=\min|\boldsymbol{h}(\partial\Omega\times[a,b])|$. By Corollary 13.1.1, there exists $\boldsymbol{h}_m(\cdot,t)=\sum_{k=0}^m p_k(t)\,\boldsymbol{h}(\cdot,t_k)$ for $p_k(t)$ a polynomial in t of degree m such that $p_0(a)=1$ but $p_k(a)=0$ if $k\neq 0$ and $p_m(b)=1$ but $p_k(b)=0$ if $k\neq m$ and

$$\max_{t\in[a,b]}\|\boldsymbol{h}_m(\cdot,t)-\boldsymbol{h}(\cdot,t)\|_{\infty,\overline{\Omega}}<\delta, t_0=a, t_m=b \tag{13.5}$$

Now replace $\boldsymbol{h}(\cdot,t_k)$ with $\boldsymbol{g}_k^m(\cdot)\in C^\infty\left(\overline{\Omega},\mathbb{R}^p\right)$ and $\mathbf{0}$ is a regular value of $\boldsymbol{g}_k^m$ and let $\boldsymbol{g}_m(\cdot,t)\equiv\sum_{k=0}^m p_k(t)\,\boldsymbol{g}_k^m(\cdot)$ where the functions $\boldsymbol{g}_k^m$ are close enough to $\boldsymbol{h}(\cdot,t_k)$ that

$$\max_{t\in[a,b]}\|\boldsymbol{g}_m(\cdot,t)-\boldsymbol{h}(\cdot,t)\|_{\infty,\overline{\Omega}}<\delta. \tag{13.6}$$

$\boldsymbol{g}_m\in C^\infty\left(\overline{\Omega}\times[a,b];\mathbb{R}^p\right)$ because all partial derivatives with respect to either t or $\boldsymbol{x}$ are continuous. Thus $\boldsymbol{g}_0^m(\cdot)=\boldsymbol{g}_m(\cdot,a)$, $\boldsymbol{g}_m^m(\cdot)=\boldsymbol{g}_m(\cdot,b)$. Also, from the definition of the degree and Lemma 13.1.13, for small enough ε,

$$d(\boldsymbol{h}(\cdot,a),\Omega,\mathbf{0})=d(\boldsymbol{g}_0^m(\cdot),\Omega,\mathbf{0})=\int_\Omega\phi_\varepsilon(\boldsymbol{g}_m(\boldsymbol{x},a))\det D_1\boldsymbol{g}_m(\boldsymbol{x},a)\,dx$$

$$=\int_\Omega\phi_\varepsilon(\boldsymbol{g}_m(\boldsymbol{x},b))\det D_1\boldsymbol{g}_m(\boldsymbol{x},b)\,dx=d(\boldsymbol{g}_m^m(\cdot),\Omega,\mathbf{0})=d(\boldsymbol{h}(\cdot,b),\Omega,\mathbf{0})$$

Since a,b are arbitrary, this proves the theorem. ■

Now the following theorem is a summary of the main result on properties of the degree.

Theorem 13.2.2 *Definition 13.1.6 is well defined and the degree satisfies the following properties.*

1. *(homotopy invariance) If $\boldsymbol{h}\in C\left(\overline{\Omega}\times[0,1],\mathbb{R}^p\right)$ and $\boldsymbol{y}(t)\notin\boldsymbol{h}(\partial\Omega,t)$ for all $t\in[0,1]$ where $\boldsymbol{y}$ is continuous, then*

$$t\to d(\boldsymbol{h}(\cdot,t),\Omega,\boldsymbol{y}(t))$$

is constant for $t\in[0,1]$.

2. *If $\Omega\supseteq\Omega_1\cup\Omega_2$ where $\Omega_1\cap\Omega_2=\emptyset$, for Ω_i an open set, then if*

$$\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\setminus(\Omega_1\cup\Omega_2)\right),$$

then

$$d(\boldsymbol{f},\Omega_1,\boldsymbol{y})+d(\boldsymbol{f},\Omega_2,\boldsymbol{y})=d(\boldsymbol{f},\Omega,\boldsymbol{y})$$

3. *$d(I,\Omega,\boldsymbol{y})=1$ if $\boldsymbol{y}\in\Omega$.*

4. *$d(\boldsymbol{f},\Omega,\cdot)$ is continuous and constant on every connected component of $\mathbb{R}^p\setminus\boldsymbol{f}(\partial\Omega)$.*

5. *$d(\boldsymbol{g},\Omega,\boldsymbol{y})=d(\boldsymbol{f},\Omega,\boldsymbol{y})$ if $\boldsymbol{g}|_{\partial\Omega}=\boldsymbol{f}|_{\partial\Omega}$.*

6. *If $\boldsymbol{y}\notin\boldsymbol{f}(\partial\Omega)$, and if $d(\boldsymbol{f},\Omega,\boldsymbol{y})\neq 0$, then there exists $\boldsymbol{x}\in\Omega$ such that $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{y}$.*

Proof: That the degree is well defined follows from Lemma 13.1.13.

Consider 1., the first property about homotopy. This follows from Theorem 13.2.1 applied to $H(\boldsymbol{x},t)\equiv\boldsymbol{h}(\boldsymbol{x},t)-\boldsymbol{y}(t)$.

Consider 2. where $\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\setminus(\Omega_1\cup\Omega_2)\right)$. Note that

$$\operatorname{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\overline{\Omega}\setminus(\Omega_1\cup\Omega_2)\right)\right)\leq\operatorname{dist}(\boldsymbol{y},\boldsymbol{f}(\partial\Omega))$$

Then let $\boldsymbol{g}$ be in $C\left(\overline{\Omega};\mathbb{R}^p\right)$ and

$$\begin{aligned}\|\boldsymbol{g}-\boldsymbol{f}\|_\infty &< \operatorname{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\overline{\Omega}\setminus(\Omega_1\cup\Omega_2)\right)\right)\\ &\leq \min\left(\operatorname{dist}(\boldsymbol{y},\boldsymbol{f}(\partial\Omega_1)),\operatorname{dist}(\boldsymbol{y},\boldsymbol{f}(\partial\Omega_2)),\operatorname{dist}(\boldsymbol{y},\boldsymbol{f}(\partial\Omega))\right)\end{aligned}$$

where $\boldsymbol{y}$ is a regular value of $\boldsymbol{g}$. Then by definition,

$$d(\boldsymbol{f},\Omega,\boldsymbol{y})\equiv\sum\left\{\det(D\boldsymbol{g}(\boldsymbol{x})):\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$$

$$\begin{aligned}&= \sum\left\{\det(D\boldsymbol{g}(\boldsymbol{x})):\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y}),\boldsymbol{x}\in\Omega_1\right\}\\ &\quad+\sum\left\{\det(D\boldsymbol{g}(\boldsymbol{x})):\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y}),\boldsymbol{x}\in\Omega_2\right\}\\ &\equiv d(\boldsymbol{f},\Omega_1,\boldsymbol{y})+d(\boldsymbol{f},\Omega_2,\boldsymbol{y})\end{aligned}$$

It is of course obvious that this can be extended by induction to any finite number of disjoint open sets Ω_i.

Note that 3. is obvious because $I(\boldsymbol{x})=\boldsymbol{x}$ and so if $\boldsymbol{y}\in\Omega$, then $I^{-1}(\boldsymbol{y})=\boldsymbol{y}$ and $DI(\boldsymbol{x})=I$ for any $\boldsymbol{x}$ so the definition gives 3.

Now consider 4. Let U be a connected component of $\mathbb{R}^p\setminus\boldsymbol{f}(\partial\Omega)$. This is open as well as connected and arc wise connected by Theorem 3.11.12. Hence, if $\boldsymbol{u},\boldsymbol{v}\in U$, there is a continuous function $\boldsymbol{y}(t)$ which is in U such that $\boldsymbol{y}(0)=\boldsymbol{u}$ and $\boldsymbol{y}(1)=\boldsymbol{v}$. By homotopy invariance, it follows $d(\boldsymbol{f},\Omega,\boldsymbol{y}(t))$ is constant. Thus $d(\boldsymbol{f},\Omega,\boldsymbol{u})=d(\boldsymbol{f},\Omega,\boldsymbol{v})$.

Next consider 5. When $\boldsymbol{f}=\boldsymbol{g}$ on $\partial\Omega$, it follows that if $\boldsymbol{y}\notin\boldsymbol{f}(\partial\Omega)$, then $\boldsymbol{y}\notin\boldsymbol{f}(\boldsymbol{x})+t(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x}))$ for $t\in[0,1]$ and $\boldsymbol{x}\in\partial\Omega$ so $d(\boldsymbol{f}+t(\boldsymbol{g}-\boldsymbol{f}),\Omega,\boldsymbol{y})$ is constant for $t\in[0,1]$ by homotopy invariance in part 1. Therefore, let $t=0$ and then $t=1$ to obtain 5.

Claim 6. follows from Lemma 13.1.13 which says that if $\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\right)$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y})=0$. ■

From the above, there is an easy corollary which gives related properties of the degree.

Corollary 13.2.3 *The following additional properties of the degree are also valid.*

1. *If $\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\setminus\Omega_1\right)$ and Ω_1 is an open subset of Ω, then $d(\boldsymbol{f},\Omega,\boldsymbol{y})=d(\boldsymbol{f},\Omega_1,\boldsymbol{y})$.*

2. *$d(\cdot,\Omega,\boldsymbol{y})$ is defined and constant on*

$$\left\{\boldsymbol{g}\in C\left(\overline{\Omega};\mathbb{R}^p\right):\|\boldsymbol{g}-\boldsymbol{f}\|_\infty<r\right\}$$

where $r=\operatorname{dist}(\boldsymbol{y},\boldsymbol{f}(\partial\Omega))$.

3. *If* $\boldsymbol{y} \in \boldsymbol{f}(\Omega)$, $\operatorname{dist}(\boldsymbol{y}, \boldsymbol{f}(\partial\Omega)) \geq \delta$ *and* $|\boldsymbol{z}-\boldsymbol{y}| < \delta$, *then* $d(\boldsymbol{f}, \Omega, \boldsymbol{y}) = d(\boldsymbol{f}, \Omega, \boldsymbol{z})$.

Proof: Consider 1. You can take $\Omega_2 = \emptyset$ in 2 of Theorem 13.2.2 or you can modify the proof of 2 slightly. Consider 2. To verify, let $\boldsymbol{h}(\boldsymbol{x}, t) = \boldsymbol{f}(\boldsymbol{x}) + t(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))$. Then note that $\boldsymbol{y} \notin \boldsymbol{h}(\partial\Omega, t)$ and use Property 1 of Theorem 13.2.2.

Finally, consider 3. Let $\boldsymbol{y}(t) \equiv (1-t)\boldsymbol{y} + t\boldsymbol{z}$. Then for $\boldsymbol{x} \in \partial\Omega$

$$\begin{aligned} |(1-t)\boldsymbol{y} + t\boldsymbol{z} - \boldsymbol{f}(\boldsymbol{x})| &= |\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}) + t(\boldsymbol{z}-\boldsymbol{y})| \\ &\geq \delta - t|\boldsymbol{z}-\boldsymbol{y}| > \delta - \delta = 0 \end{aligned}$$

Then by 1 of Theorem 13.2.2, $d(\boldsymbol{f}, \Omega, (1-t)\boldsymbol{y} + t\boldsymbol{z})$ is constant. When $t = 0$ you get $d(\boldsymbol{f}, \Omega, \boldsymbol{y})$ and when $t = 1$ you get $d(\boldsymbol{f}, \Omega, \boldsymbol{z})$. ■

Corollary 13.2.4 *Let* $\boldsymbol{h} \in C^{\infty}\left(\overline{\Omega}, \mathbb{R}^n\right)$ *where* Ω *is a bounded open set in* $\mathbb{R}^n$*and let* $\boldsymbol{y} \notin \boldsymbol{h}(\partial\Omega)$. *Then* $d(\boldsymbol{h}, \Omega, \boldsymbol{y}) = \lim_{\varepsilon \to 0} \int_{\Omega} \phi_{\varepsilon}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{h}(\boldsymbol{x})\, dx$.

Proof: Let $\left\|\tilde{\boldsymbol{h}} - \boldsymbol{h}\right\|_{\infty, \overline{\Omega}} < \delta$ where $0 < \delta < \operatorname{dist}(\boldsymbol{y}, \boldsymbol{h}(\partial\Omega))$ and $\boldsymbol{y}$ is a regular value for $\tilde{\boldsymbol{h}}$, and $D\tilde{\boldsymbol{h}}(\boldsymbol{x}) = D\boldsymbol{h}(\boldsymbol{x})$. Then

$$\begin{aligned} d(\boldsymbol{h}, \Omega, \boldsymbol{y}) &= d\left(\tilde{\boldsymbol{h}}, \Omega, \boldsymbol{y}\right) = \lim_{\varepsilon \to 0} \int_{\Omega} \phi_{\varepsilon}\left(\tilde{\boldsymbol{h}}(\boldsymbol{x}) - \boldsymbol{y}\right) \det D\tilde{\boldsymbol{h}}(\boldsymbol{x})\, dx \\ &= \lim_{\varepsilon \to 0} \int_{\Omega} \phi_{\varepsilon}\left(\tilde{\boldsymbol{h}}(\boldsymbol{x}) - \boldsymbol{y}\right) \det D\tilde{\boldsymbol{h}}(\boldsymbol{x})\, dx \\ &= \lim_{\varepsilon \to 0} \int_{\Omega} \phi_{\varepsilon}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{h}(\boldsymbol{x})\, dx \end{aligned}$$

because for $\boldsymbol{h}(\boldsymbol{x}, t) = t(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y}) + (1-t)\left(\tilde{\boldsymbol{h}}(\boldsymbol{x}) - \boldsymbol{y}\right)$,

$$t \to \int_{\Omega} \phi_{\varepsilon}(\boldsymbol{h}(\boldsymbol{x}, t)) \det D_1 \boldsymbol{h}(\boldsymbol{x}, t)\, dx$$

is constant for $t \in [0,1]$. ■

13.3 Brouwer Fixed Point Theorem

The degree makes it possible to give a very simple proof of the Brouwer fixed point theorem.

Theorem 13.3.1 *(Brouwer fixed point) Let* $B = \overline{B(\mathbf{0}, r)} \subseteq \mathbb{R}^p$ *and let* $\boldsymbol{f} : B \to B$ *be continuous. Then there exists a point* $\boldsymbol{x} \in B$, *such that* $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.

Proof: Assume there is no fixed point. Consider $\boldsymbol{h}(\boldsymbol{x}, t) \equiv \boldsymbol{x} - t\boldsymbol{f}(\boldsymbol{x})$ for $t \in [0,1]$. Then for $\|\boldsymbol{x}\| = r$, $\mathbf{0} \notin \boldsymbol{x} - t\boldsymbol{f}(\boldsymbol{x})$, $t \in [0,1]$. By homotopy invariance, $t \to d(I - t\boldsymbol{f}, B, \mathbf{0})$ is constant. But when $t = 0$, this is $d(I, B, \mathbf{0}) = 1 \neq 0$. This is a contradiction so there must be a fixed point after all. ■

You can use standard stuff from Hilbert space to get this the fixed point theorem for a compact convex set. Let K be a closed bounded convex set and let $\boldsymbol{f} : K \to K$ be continuous. Let P be the projection map onto K as in Problem 10 on Page 138. Then P is continuous because $|P\boldsymbol{x}-P\boldsymbol{y}|\leq|\boldsymbol{x}-\boldsymbol{y}|$. Recall why this is. From the characterization of the projection map P, $(\boldsymbol{x}-P\boldsymbol{x},\boldsymbol{y}-P\boldsymbol{x})\leq 0$ for all $\boldsymbol{y}\in K$. Therefore,

$$(\boldsymbol{x}-P\boldsymbol{x},P\boldsymbol{y}-P\boldsymbol{x})\leq 0,\ (\boldsymbol{y}-P\boldsymbol{y},P\boldsymbol{x}-P\boldsymbol{y})\leq 0 \text{ so } (\boldsymbol{y}-P\boldsymbol{y},P\boldsymbol{y}-P\boldsymbol{x})\geq 0$$

Hence, subtracting the first from the last,

$$(\boldsymbol{y}-P\boldsymbol{y}-(\boldsymbol{x}-P\boldsymbol{x}),P\boldsymbol{y}-P\boldsymbol{x})\geq 0$$

consequently,

$$|\boldsymbol{x}-\boldsymbol{y}|\,|P\boldsymbol{y}-P\boldsymbol{x}|\geq(\boldsymbol{y}-\boldsymbol{x},P\boldsymbol{y}-P\boldsymbol{x})\geq|P\boldsymbol{y}-P\boldsymbol{x}|^2$$

and so $|P\boldsymbol{y}-P\boldsymbol{x}|\leq|\boldsymbol{y}-\boldsymbol{x}|$ as claimed.

Now let r be so large that $K\subseteq B(\mathbf{0},r)$. Then consider $\boldsymbol{f}\circ P$. This map takes $\overline{B(\mathbf{0},r)}\to\overline{B(\mathbf{0},r)}$. In fact it maps $\overline{B(\mathbf{0},r)}$ to K. Therefore, being the composition of continuous functions, it is continuous and so has a fixed point in $\overline{B(\mathbf{0},r)}$ denoted as $\boldsymbol{x}$. Hence $\boldsymbol{f}(P(\boldsymbol{x}))=\boldsymbol{x}$. Now, since $\boldsymbol{f}$ maps into K, it follows that $\boldsymbol{x}\in K$. Hence $P\boldsymbol{x}=\boldsymbol{x}$ and so $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{x}$. This has proved the following general Brouwer fixed point theorem.

Theorem 13.3.2 *Let $\boldsymbol{f}:K\to K$ be continuous where K is compact and convex and nonempty, $K\subseteq\mathbb{R}^p$. Then $\boldsymbol{f}$ has a fixed point.*

Definition 13.3.3 *$\boldsymbol{f}$ is a retract of $\overline{B(\mathbf{0},r)}$ onto $\partial B(0,r)$ if $\boldsymbol{f}$ is continuous,*

$$\boldsymbol{f}\left(\overline{B(\mathbf{0},r)}\right)\subseteq\partial B(\mathbf{0},r)$$

and $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{x}$ for all $\boldsymbol{x}\in\partial B(\mathbf{0},r)$.

Theorem 13.3.4 *There does not exist a retract of $\overline{B(\mathbf{0},r)}$ onto $\partial B(\mathbf{0},r)$, its boundary.*

Proof: Suppose $\boldsymbol{f}$ were such a retract. Then for all $\boldsymbol{x}\in\partial B(\mathbf{0},r)$, $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{x}$ and so from the properties of the degree, the one which says if two functions agree on $\partial\Omega$, then they have the same degree, $1=d(I,B(\mathbf{0},r),\mathbf{0})=d(\boldsymbol{f},B(\mathbf{0},r),\mathbf{0})$ which is clearly impossible because $\boldsymbol{f}^{-1}(\mathbf{0})=\emptyset$ which implies $d(\boldsymbol{f},B(\mathbf{0},r),\mathbf{0})=0$. ■

You should now use this theorem to give another proof of the Brouwer fixed point theorem.

13.4 Borsuk's Theorem

In this section is an important theorem which can be used to verify that $d(\boldsymbol{f},\Omega,\boldsymbol{y})\neq 0$. This is significant because when this is known, it follows from Theorem 13.2.2 that $\boldsymbol{f}^{-1}(\boldsymbol{y})\neq\emptyset$. In other words there exists $\boldsymbol{x}\in\Omega$ such that $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{y}$.

Definition 13.4.1 *A bounded open set, Ω is symmetric if $-\Omega=\Omega$. A continuous function $\boldsymbol{f}:\overline{\Omega}\to\mathbb{R}^p$ is odd if $\boldsymbol{f}(-\boldsymbol{x})=-\boldsymbol{f}(\boldsymbol{x})$.*

Suppose Ω is symmetric and $\boldsymbol{g}\in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ is an odd map for which $\mathbf{0}$ is a regular value. Then the chain rule implies $D\boldsymbol{g}(-\boldsymbol{x})=D\boldsymbol{g}(\boldsymbol{x})$ and so $d(\boldsymbol{g},\Omega,\mathbf{0})$ must equal an odd integer because if $\boldsymbol{x}\in\boldsymbol{g}^{-1}(\mathbf{0})$, it follows that $-\boldsymbol{x}\in\boldsymbol{g}^{-1}(\mathbf{0})$ also and since $D\boldsymbol{g}(-\boldsymbol{x})=$

$D\boldsymbol{g}(\boldsymbol{x})$, it follows the overall contribution to the degree from $\boldsymbol{x}$ and $-\boldsymbol{x}$ must be an even integer. Also $\mathbf{0} \in \boldsymbol{g}^{-1}(\mathbf{0})$ and so the degree equals an even integer added to $\text{sgn}(\det D\boldsymbol{g}(\mathbf{0}))$, an odd integer, either -1 or 1. It seems reasonable to expect that something like this would hold for an arbitrary continuous odd function defined on symmetric Ω. In fact this is the case and this is next. The following lemma is the key result used. This approach is due to Gromes [21]. See also Deimling [11] which is where I found this argument. I think it is one of the cleverest calculus manipulations I have seen.

To get an idea consider the case of $p = 1$. Then Ω is bounded and symmetric and h is odd and in $C^\infty\left(\overline{\Omega}\right)$. Suppose that $h'(0) \neq 0$. I want to find arbitrarily small ε such that $\hat{h}(x) \equiv h(x) - \varepsilon x^3$ has 0 as a regular value for $x \neq 0$. Let ε be a regular value for $\frac{h(x)}{x^3} \equiv f(x)$ for $x \neq 0$. By Sard's lemma the singular values of $f(x)$ contain no balls so we can take ε as small as desired and have ε a regular value of f. Then at a point where $\hat{h}(x) = 0, f(x) = \varepsilon$ and so $\hat{h}(x) + \varepsilon x^3 = x^3 f(x)$. Now differentiate this. $\hat{h}'(x) + 3\varepsilon x^2 = 3x^2 f(x) + x^3 f'(x) = 3x^2\varepsilon + x^3 f'(x)$ so $\hat{h}'(x) = x^3 f'(x) \neq 0$. This is the motivation for the following process.

The idea is to start with a smooth odd map and approximate it with a smooth odd map which also has $\mathbf{0}$ a regular value. Note that $\mathbf{0}$ is a value because $\boldsymbol{g}(\mathbf{0}) = -\boldsymbol{g}(\mathbf{0})$.

Process: Suppose $\boldsymbol{h}_0 \in C^\infty\left(\overline{\Omega}, \mathbb{R}^p\right)$ is odd and $\det(D\boldsymbol{h}_0(\mathbf{0})) \neq 0$. Let Ω_k be those points of Ω where $x_k \neq 0$. Here $\boldsymbol{x} \equiv (x_1, \ldots, x_p)$. Then $\boldsymbol{x} \to \frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3}$ is a smooth map defined on Ω_k so by Sard's lemma, its singular values do not contain $B(\mathbf{0}, \eta)$. Therefore, there is $\boldsymbol{y}^k$ with $\boldsymbol{y}^k$ a regular value and $\left\|\boldsymbol{y}^k\right\| < \eta$ where $\eta > 0$ is given. Then consider $\hat{\boldsymbol{h}}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - x_k^3 \boldsymbol{y}^k$. I want to argue that $\mathbf{0}$ is a regular value of $\hat{\boldsymbol{h}}$ on Ω_k. Note that $\frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3} = \boldsymbol{y}^k$ if and only if $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \mathbf{0}$.

Letting $\boldsymbol{f}(\boldsymbol{x}) \equiv \frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3} = \frac{\hat{\boldsymbol{h}}(\boldsymbol{x}) + x_k^3 \boldsymbol{y}^k}{x_k^3}$, then $\hat{\boldsymbol{h}}(\boldsymbol{x}) = x_k^3\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^k\right)$ and $D\boldsymbol{f}(\boldsymbol{x})$ is invertible at the $\boldsymbol{x}$ of interest, one where $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^k = \mathbf{0}$. Then

$$D\hat{\boldsymbol{h}}(\boldsymbol{x})(\boldsymbol{u}) = 3x_k^2 \left(\overset{=\mathbf{0}}{\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^k} \right)(\boldsymbol{u}) + x_k^3 D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}). \tag{13.7}$$

At the point of interest, the first term on the right is $\mathbf{0}$ and so

$$\det\left(D\hat{\boldsymbol{h}}(\boldsymbol{x})\right) = x_k^3 \det(D\boldsymbol{f}(\boldsymbol{x})) \neq 0.$$

If $\mathbf{0}$ is a regular value for $\boldsymbol{h}_0$ on $\mathscr{U} \subseteq \Omega$, will $\mathbf{0}$ be a regular value for $\hat{\boldsymbol{h}}$ on $\mathscr{U}$ where $\hat{\boldsymbol{h}}$ is described above? The only points of concern are those $\boldsymbol{x} \in \mathscr{U}$ for which $x_k = 0$ because if $x_k \neq 0$ then $\boldsymbol{x} \in \Omega_k$. But for these points where $x_k = 0$, $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \boldsymbol{h}_0(\boldsymbol{x})$ and $D\hat{\boldsymbol{h}}(\boldsymbol{x}) = D\boldsymbol{h}_0(\boldsymbol{x})$ because $3x_k^2 = 0$ when $x_k = 0$. Thus the new function $\hat{\boldsymbol{h}}$ has $\mathbf{0}$ a regular value for all $\boldsymbol{x} \in \mathscr{U} \cup \Omega_k$. This **Process** is the basis for the following lemma.

Lemma 13.4.2 *Let $\boldsymbol{h}_0 \in C^\infty\left(\overline{\Omega}, \mathbb{R}^p\right)$ is odd and $\det(D\boldsymbol{h}_0(\mathbf{0})) \neq 0$ for Ω a symmetric open set and let $\eta > 0$. Then there are vectors $\boldsymbol{y}^k$ each with $\left\|\boldsymbol{y}^k\right\| < \eta$ such that $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - \sum_{k=1}^p x_k^3 \boldsymbol{y}^k$ has $\mathbf{0}$ as a regular value.*

Proof: Use the above process leading to 13.7 repeatedly. Start with $\boldsymbol{h}_0$ which has $\mathbf{0}$ a regular value on $\{\mathbf{0}\}$. Then use the process to get $\boldsymbol{h}_1(\boldsymbol{x}) = \boldsymbol{h}_0(\boldsymbol{x}) - \boldsymbol{y}^1 x_1^3$ which has $\mathbf{0}$ as a regular value on $\{\mathbf{0}\} \cup \Omega_1$. Then repeat the process to get $\boldsymbol{h}_2(\boldsymbol{x}) = \boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{y}^2 x_2^3$ which has $\mathbf{0}$ as a regular value on $\{\mathbf{0}\} \cup \Omega_1 \cup \Omega_2$. Continue this way and let $\boldsymbol{h} = \boldsymbol{h}_p$ which has $\mathbf{0}$ a regular value on $\{\mathbf{0}\} \cup \Omega_1 \cup \cdots \cup \Omega_p = \Omega$. ■

Lemma 13.4.3 *Let $\boldsymbol{g} \in C^{\infty}\left(\overline{\Omega};\mathbb{R}^p\right)$ be an odd map. Then for every $\varepsilon > 0$, there exists $\boldsymbol{h} \in C^{\infty}\left(\overline{\Omega};\mathbb{R}^p\right)$ such that $\boldsymbol{h}$ is also an odd map, $\|\boldsymbol{h}-\boldsymbol{g}\|_{\infty} < \varepsilon$, and $\mathbf{0}$ is a regular value of $\boldsymbol{h}, \mathbf{0} \notin \boldsymbol{g}(\partial\Omega)$. Here Ω is a symmetric bounded open set. In addition, $d(\boldsymbol{g},\Omega,\mathbf{0})$ is an odd integer.*

Proof: In this argument $\eta > 0$ will be a small positive number. Let $\boldsymbol{h}_0(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) + \eta\boldsymbol{x}$ where η is sufficiently small but nonzero that $\det D\boldsymbol{h}_0(\mathbf{0}) \neq 0$. See Lemma 9.14.1. Note that $\boldsymbol{h}_0$ is odd and $\mathbf{0}$ is a value of $\boldsymbol{h}_0$ thanks to $\boldsymbol{h}_0(\mathbf{0}) = \mathbf{0}$. This has taken care of $\mathbf{0}$. However, it is not known whether $\mathbf{0}$ is a regular value of $\boldsymbol{h}_0$ because there may be other $\boldsymbol{x}$ where $\boldsymbol{h}_0(\boldsymbol{x}) = \mathbf{0}$. By Lemma 13.4.2, there are vectors $\boldsymbol{y}^j$ with $\left\|\boldsymbol{y}^k\right\| \leq \eta$ and $\mathbf{0}$ is a regular value of $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - \sum_{j=1}^{p} \boldsymbol{y}^j x_j^3$. Then

$$\begin{aligned}\|\boldsymbol{h}-\boldsymbol{g}\|_{\infty,\,\overline{\Omega}} &\leq \max_{\boldsymbol{x}\in\overline{\Omega}}\left\{\|\eta\boldsymbol{x}\| + \sum_{k=1}^{p}\left\|\boldsymbol{y}^k\right\|\|\boldsymbol{x}\|\right\} \\ &\leq \eta\left((p+1)\operatorname{diam}(\Omega)\right) < \varepsilon < \operatorname{dist}(\boldsymbol{g}(\partial\Omega),\mathbf{0})\end{aligned}$$

provided η was chosen sufficiently small to begin with.

So what is $d(\boldsymbol{h},\Omega,\mathbf{0})$? Since $\mathbf{0}$ is a regular value and $\boldsymbol{h}$ is odd,

$$\boldsymbol{h}^{-1}(\mathbf{0}) = \{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_r,-\boldsymbol{x}_1,\cdots,-\boldsymbol{x}_r,\mathbf{0}\}.$$

So consider $D\boldsymbol{h}(\boldsymbol{x})$ and $D\boldsymbol{h}(-\boldsymbol{x})$.

$$\begin{gathered}D\boldsymbol{h}(-\boldsymbol{x})\boldsymbol{u} + \boldsymbol{o}(\boldsymbol{u}) = \boldsymbol{h}(-\boldsymbol{x}+\boldsymbol{u}) - \boldsymbol{h}(-\boldsymbol{x}) = -\boldsymbol{h}(\boldsymbol{x}+(-\boldsymbol{u})) + \boldsymbol{h}(\boldsymbol{x}) \\ = -\left(D\boldsymbol{h}(\boldsymbol{x})(-\boldsymbol{u})\right) + \boldsymbol{o}(-\boldsymbol{u}) = D\boldsymbol{h}(\boldsymbol{x})(\boldsymbol{u}) + \boldsymbol{o}(\boldsymbol{u})\end{gathered}$$

Hence $D\boldsymbol{h}(\boldsymbol{x}) = D\boldsymbol{h}(-\boldsymbol{x})$ and so the determinants of these two are the same. It follows from the definition that $d(\boldsymbol{g},\Omega,\mathbf{0}) = d(\boldsymbol{h},\Omega,\mathbf{0})$

$$\begin{aligned}&= \sum_{i=1}^{r}\operatorname{sgn}(\det(D\boldsymbol{h}(\boldsymbol{x}_i))) + \sum_{i=1}^{r}\operatorname{sgn}(\det(D\boldsymbol{h}(-\boldsymbol{x}_i) + \operatorname{sgn}(\det(D\boldsymbol{h}(\mathbf{0}))))) \\ &= 2m \pm 1 \text{ some integer } m \blacksquare\end{aligned}$$

Theorem 13.4.4 *(Borsuk) Let $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$ be odd and let Ω be symmetric with $\mathbf{0} \notin \boldsymbol{f}(\partial\Omega)$. Then $d(\boldsymbol{f},\Omega,\mathbf{0})$ equals an odd integer.*

Proof: Let ψ_n be a mollifier which is symmetric, $\psi(-\boldsymbol{x}) = \psi(\boldsymbol{x})$. Also recall that $\boldsymbol{f}$ is the restriction to $\overline{\Omega}$ of a continuous function, still denoted as $\boldsymbol{f}$ which is defined on all of $\mathbb{R}^p$. Let $\boldsymbol{g}$ be the odd part of this function. That is,

$$\boldsymbol{g}(\boldsymbol{x}) \equiv \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(-\boldsymbol{x})\right) = \boldsymbol{f}(\boldsymbol{x}) \text{ on } \overline{\Omega}$$

Thus $d(\boldsymbol{f},\Omega,\mathbf{0}) = d(\boldsymbol{g},\Omega,\mathbf{0})$. Then

$$\begin{gathered}\boldsymbol{g}_n(-\boldsymbol{x}) \equiv \boldsymbol{g} * \psi_n(-\boldsymbol{x}) = \int_{\Omega}\boldsymbol{g}(-\boldsymbol{x}-\boldsymbol{y})\psi_n(\boldsymbol{y})\,dy \\ = -\int_{\Omega}\boldsymbol{g}(\boldsymbol{x}+\boldsymbol{y})\psi_n(\boldsymbol{y})\,dy = -\int_{\Omega}\boldsymbol{g}(\boldsymbol{x}-(-\boldsymbol{y}))\psi_n(-\boldsymbol{y})\,dy = -\boldsymbol{g}_n(\boldsymbol{x})\end{gathered}$$

Thus $\boldsymbol{g}_n$ is odd and is infinitely differentiable. Let n be large enough that

$$\|\boldsymbol{g}_n - \boldsymbol{g}\|_{\infty,\overline{\Omega}} < \delta < \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\mathbf{0}) = \operatorname{dist}(\boldsymbol{g}(\partial\Omega),\mathbf{0})$$

Then by definition of the degree, $d(\boldsymbol{f},\Omega,\mathbf{0}) = d(\boldsymbol{g},\Omega,\mathbf{0}) = d(\boldsymbol{g}_n,\Omega,\mathbf{0})$ and by Lemma 13.4.3 this is an odd integer. ■

13.5 Some Applications

With Borsuk's theorem it is possible to give relatively easy proofs of some very important and difficult theorems.

Lemma 13.5.1 *Let $\boldsymbol{g} : \overline{B(\mathbf{0},r)} \subseteq \mathbb{R}^p \to \mathbb{R}^p$ be one to one and continuous. Then there exists $\delta > 0$ such that $B(\boldsymbol{g}(\mathbf{0}),\delta) \subseteq \boldsymbol{g}(B(\mathbf{0},r))$.*

Proof: For $t \in [0,1]$, let $\boldsymbol{h}(\boldsymbol{x},t) \equiv \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(-t\boldsymbol{x})$. Then for $\boldsymbol{x} \in \partial B(\mathbf{0},r)$, $\boldsymbol{h}(\boldsymbol{x},t) \neq \mathbf{0}$ because if this were so, the fact $\boldsymbol{g}$ is one to one implies $\boldsymbol{x} = -t\boldsymbol{x}$ and this requires $\boldsymbol{x} = \mathbf{0}$, not the case since $\|\boldsymbol{x}\| = r$. Then $d(\boldsymbol{h}(\cdot,t), B(\mathbf{0},r), \mathbf{0})$ is constant by Theorem 13.2.1, homotopy invariance. Hence it is an odd integer for all t thanks to Borsuk's theorem, because $\boldsymbol{h}(\cdot,1)$ is odd. Now let $B(\mathbf{0},\delta)$ be such that $B(\mathbf{0},\delta) \cap \boldsymbol{h}(\partial\Omega,0) = \emptyset$. Then $0 \neq d(\boldsymbol{h}(\cdot,0), B(\mathbf{0},r), \mathbf{0}) = d(\boldsymbol{h}(\cdot,0), B(\mathbf{0},r), \boldsymbol{z})$ for $\boldsymbol{z} \in B(\mathbf{0},\delta)$ because the degree is constant on connected components of $\mathbb{R}^p \setminus \boldsymbol{h}(\partial\Omega,0)$ by Theorem 13.2.2. Hence $\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x},0) = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\mathbf{0})$ for some $\boldsymbol{x} \in B(0,r)$. Thus

$$\boldsymbol{g}(B(\mathbf{0},r)) \supseteq \boldsymbol{g}(\mathbf{0}) + B(\mathbf{0},\delta) = B(\boldsymbol{g}(\mathbf{0}),\delta). \blacksquare$$

Theorem 13.5.2 *(invariance of domain)Let Ω be any open subset of $\mathbb{R}^p$ and let $\boldsymbol{f} : \Omega \to \mathbb{R}^p$ be continuous and one to one. Then $\boldsymbol{f}$ maps open subsets of Ω to open sets in $\mathbb{R}^p$.*

Proof: Let $\overline{B(\boldsymbol{x}_0,r)} \subseteq \Omega$ where $\boldsymbol{f}$ is one to one on $\overline{B(\boldsymbol{x}_0,r)}$. Let $\boldsymbol{g}$ be defined on $\overline{B(\mathbf{0},r)}$ given by

$$\boldsymbol{g}(\boldsymbol{x}) \equiv \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{x}_0)$$

Then $\boldsymbol{g}$ satisfies the conditions of Lemma 13.5.1, being one to one and continuous. It follows from that lemma that there exists $\delta > 0$ such that

$$\begin{aligned} \boldsymbol{f}(\Omega) &\supseteq \boldsymbol{f}(B(\boldsymbol{x}_0,r)) = \boldsymbol{f}(\boldsymbol{x}_0 + B(\mathbf{0},r)) \\ &= \boldsymbol{g}(B(\mathbf{0},r)) \supseteq \boldsymbol{g}(\mathbf{0}) + B(\mathbf{0},\delta) \\ &= \boldsymbol{f}(\boldsymbol{x}_0) + B(\mathbf{0},\delta) = B(\boldsymbol{f}(\boldsymbol{x}_0),\delta) \end{aligned}$$

This shows that for any $\boldsymbol{x}_0 \in \Omega, \boldsymbol{f}(\boldsymbol{x}_0)$ is an interior point of $\boldsymbol{f}(\Omega)$ which shows $\boldsymbol{f}(\Omega)$ is open. ■

Definition 13.5.3 *If $\boldsymbol{f} : U \subseteq \mathbb{R}^p \to \mathbb{R}^p$ where U is an open set. Then $\boldsymbol{f}$ is locally one to one if for every $\boldsymbol{x} \in U$, there exists $\delta > 0$ such that $\boldsymbol{f}$ is one to one on $B(\boldsymbol{x},\delta)$.*

Then an examination of the proof of the above theorem shows the following corollary.

Corollary 13.5.4 *In Theorem 13.5.2 it suffices to assume $\boldsymbol{f}$ is locally one to one.*

With the above, one gets easily the following amazing result. It is something which is clear for linear maps but this is a statement about continuous maps.

Corollary 13.5.5 *If $p > m$ there does not exist a continuous one to one map from $\mathbb{R}^p$ to $\mathbb{R}^m$.*

Proof: Suppose not and let $\boldsymbol{f}$ be such a continuous map, $\boldsymbol{f}(\boldsymbol{x}) \equiv (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}))^T$. Then let $\boldsymbol{g}(\boldsymbol{x}) \equiv (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}), 0, \cdots, 0)^T$ where there are $p-m$ zeros added in. Then $\boldsymbol{g}$ is a one to one continuous map from $\mathbb{R}^p$ to $\mathbb{R}^p$ and so $\boldsymbol{g}(\mathbb{R}^p)$ would have to be open from the invariance of domain theorem and this is not the case. ■

Corollary 13.5.6 *Let $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ and $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$ where $\boldsymbol{f}$ is locally one to one and continuous. Then $\boldsymbol{f}$ maps $\mathbb{R}^p$ onto $\mathbb{R}^p$.*

Proof: By the invariance of domain theorem, $\boldsymbol{f}(\mathbb{R}^p)$ is an open set. It is also true that $\boldsymbol{f}(\mathbb{R}^p)$ is a closed set. Here is why. If $\boldsymbol{f}(\boldsymbol{x}_k) \to \boldsymbol{y}$, the growth condition ensures that $\{\boldsymbol{x}_k\}$ is a bounded sequence. Taking a subsequence which converges to $\boldsymbol{x} \in \mathbb{R}^p$ and using the continuity of $\boldsymbol{f}$, it follows $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$. Thus $\boldsymbol{f}(\mathbb{R}^p)$ is both open and closed which implies $\boldsymbol{f}$ must be an onto map since otherwise, $\mathbb{R}^p$ would not be connected. ■

The proofs of the next two theorems make use of the Tietze extension theorem, Theorem 5.7.5.

Theorem 13.5.7 *Let Ω be a symmetric open set in $\mathbb{R}^p$ such that $\mathbf{0} \in \Omega$ and let $\boldsymbol{f} : \partial\Omega \to V$ be continuous where V is an m dimensional subspace of $\mathbb{R}^p, m < p$. Then $\boldsymbol{f}(-\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})$ for some $\boldsymbol{x} \in \partial\Omega$.*

Proof: You could reduce to the case where $V = \mathbb{R}^m$ if desired. Suppose not. Using the Tietze extension theorem on components of the function, extend $\boldsymbol{f}$ to all of $\mathbb{R}^p$, $\boldsymbol{f}\left(\overline{\Omega}\right) \subseteq V$. (Here the extended function is also denoted by $\boldsymbol{f}$.) Let $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(-\boldsymbol{x})$. Then $\mathbf{0} \notin \boldsymbol{g}(\partial\Omega)$ and so for some $r > 0$, $B(\mathbf{0},r) \subseteq \mathbb{R}^p \setminus \boldsymbol{g}(\partial\Omega)$. For $\boldsymbol{z} \in B(\mathbf{0},r)$, $d(\boldsymbol{g},\Omega,\boldsymbol{z}) = d(\boldsymbol{g},\Omega,\mathbf{0}) \neq 0$ because $B(\mathbf{0},r)$ is contained in a component of $\mathbb{R}^p \setminus \boldsymbol{g}(\partial\Omega)$ and Borsuk's theorem implies that $d(\boldsymbol{g},\Omega,\mathbf{0}) \neq 0$ since $\boldsymbol{g}$ is odd. Hence $V \supseteq \boldsymbol{g}(\Omega) \supseteq B(\mathbf{0},r)$ and this is a contradiction because V is m dimensional. ■

This theorem is called the Borsuk Ulam theorem. Note that it implies there exist two points on opposite sides of the surface of the earth which have the same atmospheric pressure and temperature, assuming the earth is symmetric and that pressure and temperature are continuous functions. The next theorem is an amusing result which is like combing hair. It gives the existence of a "cowlick".

Theorem 13.5.8 *Let p be odd and let Ω be an open bounded set in $\mathbb{R}^p$ with $\mathbf{0} \in \Omega$. Suppose $\boldsymbol{f} : \partial\Omega \to \mathbb{R}^p \setminus \{\mathbf{0}\}$ is continuous. Then for some $\boldsymbol{x} \in \partial\Omega$ and $\lambda \neq 0$, $\boldsymbol{f}(\boldsymbol{x}) = \lambda\boldsymbol{x}$.*

Proof: Using the Tietze extension theorem, extend $\boldsymbol{f}$ to all of $\mathbb{R}^p$. Also denote the extended function by $\boldsymbol{f}$. Suppose for all $\boldsymbol{x} \in \partial\Omega$, $\boldsymbol{f}(\boldsymbol{x}) \neq \lambda\boldsymbol{x}$ for all $\lambda \in \mathbb{R}$. Then

$$\mathbf{0} \notin t\boldsymbol{f}(\boldsymbol{x}) + (1-t)\boldsymbol{x}, \ \ (\boldsymbol{x},t) \in \partial\Omega \times [0,1].$$

$$\mathbf{0} \notin t\boldsymbol{f}(\boldsymbol{x}) - (1-t)\boldsymbol{x}, \ \ (\boldsymbol{x},t) \in \partial\Omega \times [0,1].$$

Thus there exists a homotopy of $\boldsymbol{f}$ and I and a homotopy of $\boldsymbol{f}$ and $-I$. Then by the homotopy invariance of degree,

$$d(\boldsymbol{f},\Omega,\mathbf{0}) = d(I,\Omega,\mathbf{0}), \ d(\boldsymbol{f},\Omega,\mathbf{0}) = d(-I,\Omega,\mathbf{0}).$$

But this is impossible because $d(I,\Omega,\mathbf{0}) = 1$ but $d(-I,\Omega,\mathbf{0}) = (-1)^n = -1$. ■

13.6 Product Formula, Separation Theorem

This section is on the product formula for the degree which is used to prove the Jordan separation theorem. To begin with is a significant observation which is used without comment below. Recall that the connected components of an open set are open. The formula is all about the composition of continuous functions.

$$\Omega \xrightarrow{\boldsymbol{f}} \boldsymbol{f}(\Omega) \subseteq \mathbb{R}^p \xrightarrow{\boldsymbol{g}} \mathbb{R}^p$$

Lemma 13.6.1 *Let $\{K_i\}_{i=1}^N, N \leq \infty$ be the connected components of $\mathbb{R}^p \setminus C$ 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. See Theorem 3.11.12. Thus ∂K_i consists of all limit points of K_i which are not in K_i. Let $\boldsymbol{p}$ 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 $\boldsymbol{x}$ is a point in U it cannot be a limit point of V for V disjoint from U. ■

Definition 13.6.2 *Let the connected components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ be denoted by K_i. From the properties of the degree listed in Theorem 13.2.2, $d(\boldsymbol{f},\Omega,\cdot)$ is constant on each of these components. Denote by $d(\boldsymbol{f},\Omega,K_i)$ the constant value on the component K_i.*

The following is the product formula. Note that if K is an unbounded component of $\boldsymbol{f}(\partial\Omega)^C$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in K$ by homotopy invariance and the fact that for large enough $\|\boldsymbol{y}\|$, $\boldsymbol{f}^{-1}(\boldsymbol{y}) = \emptyset$ since $\boldsymbol{f}\left(\overline{\Omega}\right)$ is compact.

Theorem 13.6.3 *(product formula)Let $\{K_i\}_{i=1}^\infty$ be the bounded components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ for $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$, let $\boldsymbol{g} \in C(\mathbb{R}^p,\mathbb{R}^p)$, and suppose that $\boldsymbol{y} \notin \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$ or in other words, $\boldsymbol{g}^{-1}(\boldsymbol{y}) \cap \boldsymbol{f}(\partial\Omega) = \emptyset$. Then*

$$d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{i=1}^\infty d(\boldsymbol{f},\Omega,K_i)\, d(\boldsymbol{g},K_i,\boldsymbol{y}). \tag{13.8}$$

All but finitely many terms in the sum are zero. If there are no bounded components of $\boldsymbol{f}(\partial\Omega)^C$, then $d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) = 0$.

Proof: The compact set $\boldsymbol{f}\left(\overline{\Omega}\right) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$ is contained in $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ so $\boldsymbol{f}\left(\overline{\Omega}\right) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$ is covered by finitely many of the components K_j one of which may be the unbounded component. Since these components are disjoint, the other components fail to intersect $\boldsymbol{f}\left(\overline{\Omega}\right) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$. Thus, if K_i is one of these others, either it fails to intersect $\boldsymbol{g}^{-1}(\boldsymbol{y})$ or K_i fails to intersect $\boldsymbol{f}\left(\overline{\Omega}\right)$. Thus either $d(\boldsymbol{f},\Omega,K_i) = 0$ because K_i fails to intersect $\boldsymbol{f}\left(\overline{\Omega}\right)$ or $d(\boldsymbol{g},K_i,\boldsymbol{y}) = 0$ if K_i fails to intersect $\boldsymbol{g}^{-1}(\boldsymbol{y})$. Thus the sum is always a finite sum. I am using Theorem 13.2.2, the part which says that if $\boldsymbol{y} \notin \boldsymbol{h}\left(\overline{\Omega}\right)$, then $d(\boldsymbol{h},\Omega,\boldsymbol{y}) = 0$. Note that by Lemma 13.6.1 $\partial K_i \subseteq \boldsymbol{f}(\partial\Omega)$ so $\boldsymbol{g}(\partial K_i) \subseteq \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$ and so $\boldsymbol{y} \notin \boldsymbol{g}(\partial K_i)$ because it is assumed that $\boldsymbol{y} \notin \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$.

Let $\tilde{\boldsymbol{g}}$ be in $C^\infty(\mathbb{R}^p,\mathbb{R}^p)$ and let $\|\boldsymbol{g}-\tilde{\boldsymbol{g}}\|_{\infty,\boldsymbol{f}\left(\overline{\Omega}\right)} < \operatorname{dist}(\boldsymbol{y},\boldsymbol{g}(\boldsymbol{f}(\partial\Omega)))$. Thus, for each of the finitely many K_i intersecting $\boldsymbol{f}\left(\overline{\Omega}\right) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$,

$$\begin{aligned} d(\boldsymbol{g},K_i,\boldsymbol{y}) &= d(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y}) \text{ and} \\ d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) &= d(\tilde{\boldsymbol{g}}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) \end{aligned} \tag{13.9}$$

By Lemma 13.1.5, there exists $\tilde{\boldsymbol{g}}$ such that $\boldsymbol{y}$ is a regular value of $\tilde{\boldsymbol{g}}$ in addition to 13.9 and $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap \boldsymbol{f}(\partial\Omega) = \emptyset$. Then $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ is contained in the union of the K_i along with the unbounded component(s) and by Lemma 13.1.5 $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ is countable. As discussed there, $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap K_i$ is finite if K_i is bounded. Let $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap K_i = \left\{\boldsymbol{x}_j^i\right\}_{j=1}^{m_i}, m_i \leq \infty$. m_i could only be ∞ on the unbounded component.

Now use Lemma 13.1.5 again to get $\tilde{\boldsymbol{f}}$ in $C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ such that each $\boldsymbol{x}_j^i$ is a regular value of $\tilde{\boldsymbol{f}}$ on Ω and also $\left\|\tilde{\boldsymbol{f}} - \boldsymbol{f}\right\|_\infty$ is very small, so small that

$$d\left(\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}, \Omega, \boldsymbol{y}\right) = d\left(\tilde{\boldsymbol{g}} \circ \boldsymbol{f}, \Omega, \boldsymbol{y}\right) = d\left(\boldsymbol{g} \circ \boldsymbol{f}, \Omega, \boldsymbol{y}\right)$$

and $d\left(\tilde{\boldsymbol{f}}, \Omega, \boldsymbol{x}_j^i\right) = d\left(\boldsymbol{f}, \Omega, \boldsymbol{x}_j^i\right)$ for each i, j.

Thus, from the above,

$$\begin{aligned} d\left(\boldsymbol{g} \circ \boldsymbol{f}, \Omega, \boldsymbol{y}\right) &= d\left(\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}, \Omega, \boldsymbol{y}\right), \\ d\left(\tilde{\boldsymbol{f}}, \Omega, \boldsymbol{x}_j^i\right) &= d\left(\boldsymbol{f}, \Omega, \boldsymbol{x}_j^i\right) = d\left(\boldsymbol{f}, \Omega, K_i\right) \\ d\left(\tilde{\boldsymbol{g}}, K_i, \boldsymbol{y}\right) &= d\left(\boldsymbol{g}, K_i, \boldsymbol{y}\right) \end{aligned}$$

Is $\boldsymbol{y}$ a regular value for $\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}$ on Ω? Suppose $\boldsymbol{z} \in \Omega$ and $\boldsymbol{y} = \tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}(\boldsymbol{z})$ so $\tilde{\boldsymbol{f}}(\boldsymbol{z}) \in \tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$. Then $\tilde{\boldsymbol{f}}(\boldsymbol{z}) = \boldsymbol{x}_j^i$ for some i, j and $D\tilde{\boldsymbol{f}}(\boldsymbol{z})^{-1}$ exists. Hence

$$D\left(\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}\right)(\boldsymbol{z}) = D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right) D\tilde{\boldsymbol{f}}(\boldsymbol{z}),$$

both linear transformations invertible. Thus $\boldsymbol{y}$ is a regular value of $\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}$ on Ω.

What of $\boldsymbol{x}_j^i$ in K_i where K_i is unbounded? As observed, the sum of sgn $\left(\det D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)$ for $\boldsymbol{z} \in \tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)$ is $d\left(\tilde{\boldsymbol{f}}, \Omega, \boldsymbol{x}_j^i\right)$ and is 0 because the degree is constant on K_i which is unbounded.

From the definition of the degree, the left side of 13.8 $d\left(\boldsymbol{g} \circ \boldsymbol{f}, \Omega, \boldsymbol{y}\right)$ equals

$$\sum\left\{\operatorname{sgn}\left(\det D\tilde{\boldsymbol{g}}\left(\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right) \operatorname{sgn}\left(\det D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right) : \boldsymbol{z} \in \tilde{\boldsymbol{f}}^{-1}\left(\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})\right)\right\}$$

The $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ are the $\boldsymbol{x}_j^i$. Thus the above is of the form

$$= \sum_i \sum_j \sum_{\boldsymbol{z} \in \tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)} \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right) \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)$$

As mentioned, if $\boldsymbol{x}_j^i \in K_i$ an unbounded component, then

$$\sum_{\boldsymbol{z} \in \tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)} \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right) \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right) = 0$$

and so, it suffices to only consider bounded components in what follows and the sum makes sense because there are finitely many $\boldsymbol{x}_j^i$ in bounded K_i. This also shows that if there are no bounded components of $\boldsymbol{f}(\partial\Omega)^C$, then $d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y})=0$. Thus $d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y})$ equals

$$\begin{aligned}
&= \sum_i\sum_j \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right)\sum_{\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)} \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right) \\
&= \sum_i d(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y})\, d\left(\tilde{\boldsymbol{f}},\Omega,K_i\right)
\end{aligned}$$

To explain the last step,

$$\sum_{\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)} \operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)\equiv d\left(\tilde{\boldsymbol{f}},\Omega,\boldsymbol{x}_j^i\right)=d\left(\tilde{\boldsymbol{f}},\Omega,K_i\right).$$

This proves the product formula because $\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{f}}$ were chosen close enough to $\boldsymbol{f},\boldsymbol{g}$ respectively that

$$\sum_i d\left(\tilde{\boldsymbol{f}},\Omega,K_i\right)d(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y})=\sum_i d(\boldsymbol{f},\Omega,K_i)\,d(\boldsymbol{g},K_i,\boldsymbol{y})\ \blacksquare$$

Before the general Jordan separation theorem, I want to first consider the examples of most interest.

Recall that if a function $\boldsymbol{f}$ is continuous and one to one on a compact set K, then $\boldsymbol{f}$ is a homeomorphism of K and $\boldsymbol{f}(K)$. Also recall that if U is a nonempty open set, the boundary of U, denoted as ∂U and meaning those points $\boldsymbol{x}$ with the property that for all $r>0$ $B(\boldsymbol{x},r)$ intersects both U and U^C, is $\overline{U}\setminus U$.

Proposition 13.6.4 *Let C be a compact set and let $\boldsymbol{f}:C\to D\subseteq\mathbb{R}^p, p\geq 2$ be one to one and continuous so that C and $\boldsymbol{f}(C)\equiv D$ are homeomorphic. Suppose C^C has only one connected component so C^C is connected. Then D^C also has only one component.*

Proof: Extend $\boldsymbol{f}$, using the Tietze extension theorem on its entries to all of $\mathbb{R}^p$ and let $\boldsymbol{g}$ be an extension of $\boldsymbol{f}^{-1}$ to all of $\mathbb{R}^p$. Suppose D^C has a bounded component K. Then from Lemma 13.6.1,$\partial K\subseteq D, \boldsymbol{g}(\partial K)\subseteq\boldsymbol{g}(D)=C$. It follows that $d(\boldsymbol{f}\circ\boldsymbol{g},K,\boldsymbol{z})=1$ where $\boldsymbol{z}\in K$ because on ∂K, $\boldsymbol{f}\circ\boldsymbol{g}=id$.

If $\boldsymbol{z}\in K$, then $\boldsymbol{z}\neq\boldsymbol{f}\circ\boldsymbol{g}(\boldsymbol{k})$ for any $\boldsymbol{k}\in\partial K$ because $\boldsymbol{f}\circ\boldsymbol{g}=id$ on $\partial K\subseteq C$, this by Lemma 13.6.1. Then $\boldsymbol{g}(\partial K)^C\supseteq C^C$. If Q is a bounded component of $\boldsymbol{g}(\partial K)^C$ then if Q contains a point of C^C it follows that C^C is connected, has no points of C and hence no points of $\boldsymbol{g}(\partial K)$ so $Q\supseteq C^C$ and Q is not bounded after all. Thus $\boldsymbol{g}(\partial K)^C$ has no bounded components. Then from the product formula Theorem 13.6.3, $d(\boldsymbol{f}\circ\boldsymbol{g},K,\boldsymbol{z})=0$ which is a contradiction. Thus there is no bounded component of D^C. ■

This says that if a compact set H fails to separate $\mathbb{R}^p$ and if $\boldsymbol{f}$ is continuous and one to one, then also $\boldsymbol{f}(H)$ fails to separate $\mathbb{R}^p$.

It is obvious that the unit sphere S^{p-1} divides $\mathbb{R}^p$ into two disjoint open sets, the inside and the outside. The following shows that this also holds for any homeomorphic image of S^{p-1}.

Proposition 13.6.5 *Let B be the ball $B(\mathbf{0},1)$ with S^{p-1} its boundary, $p\geq 2$. Suppose $\boldsymbol{f}:S^{p-1}\to C\equiv\boldsymbol{f}\left(S^{p-1}\right)\subseteq\mathbb{R}^p$ is a homeomorphism. Then C^C also has exactly two components, one bounded and one unbounded.*

Proof: By Proposition 13.6.4 there is at least one component of $\boldsymbol{f}(\partial B)^C$ called K since it is clear that $\left(S^{p-1}\right)^C$ is not connected. Let $\boldsymbol{f}$ denote the extension of $\boldsymbol{f}$ to all of $\mathbb{R}^p$ and let $\boldsymbol{g} = \boldsymbol{f}^{-1}$ on $\boldsymbol{f}(\partial B)$ where $\boldsymbol{g}$ is also extended using the Tietze extension theorem to all of $\mathbb{R}^p$. Let H be the unbounded component of $\mathbb{R}^p \setminus S^{p-1}$.

From Lemma 13.6.1, $\partial K \subseteq \boldsymbol{f}(\partial B)$ so $\boldsymbol{g}(\partial K) \subseteq \partial B$. Also,

$$\boldsymbol{f}\circ\boldsymbol{g}(\partial K) \subseteq \boldsymbol{f}\circ\boldsymbol{g}(\boldsymbol{f}(\partial B)) = \boldsymbol{f}(\partial B).$$

Recall that K has no points in $\boldsymbol{f}(\partial B)$ so if $\boldsymbol{p} \in K$, then $\boldsymbol{p}$ cannot be in $\boldsymbol{f}(\partial B)$ and consequently $\boldsymbol{p}$ cannot be in $\boldsymbol{f}\circ\boldsymbol{g}(\partial K)$ either. Summarizing this,

$$\partial K \subseteq \boldsymbol{f}(\partial B),\ \boldsymbol{g}(\partial K) \subseteq \partial B,\ \boldsymbol{f}\circ\boldsymbol{g}(\partial K)\cap K = \emptyset$$

Then picking $\boldsymbol{p} \in K$, by the product rule,

$$1 = d(id,K,\boldsymbol{p}) = d(\boldsymbol{f}\circ\boldsymbol{g},K,\boldsymbol{p}) = \sum_i d(\boldsymbol{g},K,Q_i)\,d(\boldsymbol{f},Q_i,\boldsymbol{p})$$

where here the Q_i are the bounded components of $(\boldsymbol{g}(\partial K))^C$. These are maximal open connected sets in $\mathbb{R}^p$. Recall $\boldsymbol{g}(\partial K) \subseteq \partial B$. If Q_i has a point of H, then H would be connected and contain no points of $\boldsymbol{g}(\partial K)$ 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 $\boldsymbol{g}(\partial K)$ because it contains no points of ∂B which is a larger set than $\boldsymbol{g}(\partial K)$ and so in fact $Q_i = B$ and there is only one term in the above sum. Thus, from properties of the degree,

$$\begin{aligned} 1 &= d(id,K,\boldsymbol{p}) = d(\boldsymbol{f}\circ\boldsymbol{g},K,\boldsymbol{p}) = d(\boldsymbol{g},K,B)\,d(\boldsymbol{f},B,\boldsymbol{p}) \\ &= d(\boldsymbol{g},K,\mathbf{0})\,d(\boldsymbol{f},B,K) = d(\boldsymbol{g}\circ\boldsymbol{f},B,\mathbf{0}) \end{aligned}$$

so by the product rule there is no more than one bounded component of $\boldsymbol{f}(\partial B)^C$ the K just mentioned. To emphasize this, if you had bounded components K_i of $\boldsymbol{f}(\partial B)^C$, $i \leq m \leq \infty$ Then $1 = d(\boldsymbol{g},K_i,\mathbf{0})\,d(\boldsymbol{f},B,K_i) = d(\boldsymbol{g}\circ\boldsymbol{f},B,\mathbf{0})$, but then, by the product rule, you would have for $K \equiv K_0$, $1 = d(\boldsymbol{g}\circ\boldsymbol{f},B,\mathbf{0}) = \sum_{k=0}^m \overset{=1}{d(\boldsymbol{g},K_i,\mathbf{0})\,d(\boldsymbol{f},B,K_i)} = m+1$. Thus there is exactly one bounded component of $\boldsymbol{f}(\partial B)^C$. ■

A repeat of the above proof yields the following corollary. Replace B with Ω.

Corollary 13.6.6 *Let $\Omega \subseteq \mathbb{R}^p$, $p \geq 2$be a bounded open connected set such that $\partial\Omega^C$ has two components, a bounded and an unbounded component. Suppose $\boldsymbol{f} : \partial\Omega \to C \equiv \boldsymbol{f}(\partial\Omega) \subseteq \mathbb{R}^p$ is a homeomorphism. Then C^C also has exactly two components, one bounded and one unbounded.*

As an application, here is a very interesting little result. It has to do with $d(\boldsymbol{f},\Omega,\boldsymbol{f}(\boldsymbol{x}))$ in the case where $\boldsymbol{f}$ is one to one and Ω is open and connected. You might imagine this should equal 1 or -1 based on one dimensional analogies. Recall a one to one map defined on an interval is either increasing or decreasing. It either preserves or reverses orientation. It is similar in n dimensions and it is a nice application of the Jordan separation theorem and the product formula.

Proposition 13.6.7 *Let Ω be an open connected bounded set in $\mathbb{R}^p, p \geq 2$ such that $\mathbb{R}^p \setminus \partial\Omega$ consists of two connected components. Let $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$ be continuous and one to one. Then $\boldsymbol{f}(\Omega)$ is the bounded component of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ and for $\boldsymbol{y} \in \boldsymbol{f}(\Omega)$, $d(\boldsymbol{f},\Omega,\boldsymbol{y})$ either equals 1 or -1.*

Proof: By the Jordan separation theorem, Corollary 13.6.6, $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ consists of two components, a bounded component B and an unbounded component U. Using the Tietze extention theorem, there exists $\boldsymbol{g}$ defined on $\mathbb{R}^p$ such that $\boldsymbol{g} = \boldsymbol{f}^{-1}$ on $\boldsymbol{f}\left(\overline{\Omega}\right)$. Thus on $\partial\Omega, \boldsymbol{g} \circ \boldsymbol{f} = \text{id}$. It follows from this and the product formula that

$$1 = d(\text{id}, \Omega, \boldsymbol{g}(\boldsymbol{y})) = d(\boldsymbol{g} \circ \boldsymbol{f}, \Omega, \boldsymbol{g}(\boldsymbol{y})) = d(\boldsymbol{g}, B, \boldsymbol{g}(\boldsymbol{y}))\, d(\boldsymbol{f}, \Omega, B)$$

Therefore, $d(\boldsymbol{f}, \Omega, B) \neq 0$ and so for every $\boldsymbol{z} \in B$, it follows $\boldsymbol{z} \in \boldsymbol{f}(\Omega)$. Thus $B \subseteq \boldsymbol{f}(\Omega)$. On the other hand, $\boldsymbol{f}(\Omega)$ cannot have points in both U and B because it is a connected set. Therefore $\boldsymbol{f}(\Omega) \subseteq B$ and this shows $B = \boldsymbol{f}(\Omega)$. Thus $d(\boldsymbol{f}, \Omega, B) = d(\boldsymbol{f}, \Omega, \boldsymbol{y})$ for each $\boldsymbol{y} \in B$ and the above formula shows this equals either 1 or -1 because the degree is an integer. ■

The one dimensional case also fits into this although it is easier to do by more elementary means. In the case where $n = 1$, the argument is essentially the same. There is one and only one bounded component for $\mathbb{R} \setminus f(\{a, b\})$. 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 $\boldsymbol{f}\left(S^{p-1}\right)$ one wants to verify that this is the is the boundary of both components, the bounded one and the unbounded one.

Theorem 13.6.8 *Let S^{p-1} be the unit sphere in $\mathbb{R}^p, p \geq 2$. Suppose $\gamma : S^{p-1} \to \Gamma \subseteq \mathbb{R}^p$ is one to one onto and continuous. Then $\mathbb{R}^p \setminus \Gamma$ 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 $\mathbb{R}^p \setminus \Gamma$ is Γ and Γ has empty interior.*

Proof: γ^{-1} is continuous since S^{p-1} is compact and γ is one to one. By the Jordan separation theorem, $\mathbb{R}^p \setminus \Gamma = U_o \cup U_i$ where these on the right are the connected components of the set on the left, both open sets. Only U_i is bounded. Thus $\Gamma \cup U_i \cup U_o = \mathbb{R}^p$. Since both U_i, U_o are open, $\partial U \equiv \overline{U} \setminus U$ for U either U_o or U_i. If $\boldsymbol{x} \in \Gamma$, and is not a limit point of U_i, then there is $B(\boldsymbol{x}, r)$ which contains no points of U_i. Let S be those points $\boldsymbol{x}$ of Γ for which, $B(\boldsymbol{x}, r)$ contains no points of U_i for some $r > 0$. This S is open in Γ. Let $\hat{\Gamma}$ be $\Gamma \setminus S$. Then if $\hat{C} = \gamma^{-1}\left(\hat{\Gamma}\right)$, it follows that $\hat{C}$ is a closed set in $S^{p-1}$and is a proper subset of S^{p-1}. It is obvious that taking a relatively open set from S^{p-1} results in a compact set whose complement in $\mathbb{R}^p$ is an open connected set. By Proposition 13.6.4, $\mathbb{R}^p \setminus \hat{\Gamma}$ is also an open connected set. Start with $\boldsymbol{x} \in U_i$ and consider a continuous curve which goes from $\boldsymbol{x}$ to $\boldsymbol{y} \in U_o$ which is contained in $\mathbb{R}^p \setminus \hat{\Gamma}$. Thus the curve contains no points of $\hat{\Gamma}$. However, it must contain points of Γ which can only be in S. The first point of Γ 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 Γ is in $\overline{U_i}$. Similarly, every point of Γ is in $\overline{U_o}$. Thus $\Gamma \subseteq \overline{U_i} \setminus U_i$ and $\Gamma \subseteq \overline{U_o} \setminus U_o$. However, if $\boldsymbol{x} \in \overline{U_i} \setminus U_i$, then $\boldsymbol{x} \notin U_o$ because it is a limit point of U_i and so $\boldsymbol{x} \in \Gamma$. It is similar with U_o. Thus $\Gamma = \overline{U_i} \setminus U_i$ and $\Gamma = \overline{U_o} \setminus U_o$. This could not happen if Γ had an interior point. Such a point would be in Γ but would fail to be in either ∂U_i or ∂U_o. ■

When $p = 2$, this theorem is called the Jordan curve theorem.

What if γ maps $\bar{B}$ to $\mathbb{R}^p$ instead of γ only being defined on S^{p-1}? Obviously, one should be able to say a little more.

Corollary 13.6.9 *Let B be an open ball and let $\gamma : \bar{B} \to \mathbb{R}^p$ be one to one and continuous. Let U_i, U_o be as in the above theorem, the bounded and unbounded components of $\gamma(\partial B)^C$. Then $U_i = \gamma(B)$.*

Proof: This follows from Proposition 13.6.7.

Note how this yields the invariance of domain theorem. If $\boldsymbol{f}$ is one to one on U an open set, you could consider $\bar{B} \subseteq U$ and then $\boldsymbol{f}(B)$ is the bounded component of $\boldsymbol{f}(\partial B)^C$. You can do this for each ball contained in U. Thus $\boldsymbol{f}(U)$ is open.

13.7 General Jordan Separation Theorem

What follows is the general Jordan separation theorem. First note that if C,D are compact sets and $\boldsymbol{f} : C \to D$ is a homeomorphism, continuous, one to one and onto, then if C,D are both in $\mathbb{R}$ and if C^C, has no bounded components, then C would be a closed interval and so would D. Thus C^C, D^C have the same number of bounded components. In general for $\mathbb{R}^p$, Proposition 13.6.4 says C^C, D^C both have no bounded components together. The Jordan Separation Theorem shows that C^C, D^C have the same number of bounded components in general.

Lemma 13.7.1 *Let Ω be a bounded open set in $\mathbb{R}^p$, $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$, and suppose the sequence $\{\Omega_i\}_{i=1}^\infty$ are disjoint open sets contained in Ω such that*

$$\boldsymbol{y} \notin \boldsymbol{f}\left(\overline{\Omega} \setminus \cup_{j=1}^\infty \Omega_j\right)$$

Then $d(\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{j=1}^\infty d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$ where the sum has only finitely many terms equal to 0.

Proof: By assumption, the compact set $\boldsymbol{f}^{-1}(\boldsymbol{y}) \equiv \left\{\boldsymbol{x} \in \overline{\Omega} : \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}\right\}$ has empty intersection with $\overline{\Omega} \setminus \cup_{j=1}^\infty \Omega_j$ and so this compact set is covered by finitely many of the Ω_j, say $\{\Omega_1, \cdots, \Omega_{n-1}\}$ and $\boldsymbol{y} \notin \boldsymbol{f}\left(\cup_{j=n}^\infty \Omega_j\right)$. By Theorem 13.2.2 and letting $O = \cup_{j=n}^\infty \Omega_j$,

$$d(\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{j=1}^{n-1} d(\boldsymbol{f},\Omega_j,\boldsymbol{y}) + d(\boldsymbol{f},O,\boldsymbol{y}) = \sum_{j=1}^{\infty} d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$$

because $d(\boldsymbol{f},O,\boldsymbol{y}) = 0$ as is $d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$ for every $j \geq n$. ■

Theorem 13.7.2 *(Jordan separation theorem) Let $\boldsymbol{f}$ be a homeomorphism of C and $\boldsymbol{f}(C) \equiv D$ where C is a compact set in $\mathbb{R}^p$. Then $\mathbb{R}^p \setminus C$ and $\mathbb{R}^p \setminus D$ have the same number of connected components.*

Proof: If either C or D has no bounded components, then so does the other, this from Proposition 13.6.4. Let $\boldsymbol{f}$ denote a Tietze extension of $\boldsymbol{f}$ to all of $\mathbb{R}^p$ and let $\boldsymbol{g}$ be a Tietze extension of $\boldsymbol{f}^{-1}$ to all of $\mathbb{R}^p$. Let the bounded components of C^C be $\{J_r\}_{r=1}^n \equiv \mathscr{J}$ and let the bounded components of D^C be $\{K_s\}_{s=1}^m \equiv \mathscr{K}, n,m \leq \infty$. If both are ∞ then we consider the theorem proved. Assume one of n,m is less than ∞. Pick $\boldsymbol{x}_r \in J_r$ and $\boldsymbol{y}_s \in K_s$. By Lemma 13.6.1, $\partial K_s \subseteq D$ and so $\boldsymbol{g}(\partial K_s) \subseteq \boldsymbol{g}(D) = C$. $\boldsymbol{f} \circ \boldsymbol{g}(\partial K_s) \subseteq \boldsymbol{f}(C) = D$ and K_s is a component of D^C and so $\boldsymbol{y}_s \notin \boldsymbol{f} \circ \boldsymbol{g}(\partial K_s)$. Then from the definition of the degree and its properties along with the product formula,

$$1 = d(\boldsymbol{f} \circ \boldsymbol{g}, K_s, \boldsymbol{y}_s) = \sum_j d(\boldsymbol{g}, K_s, Q_j)\, d(\boldsymbol{f}, Q_j, \boldsymbol{y}_s) \tag{13.10}$$

where the Q_j are the bounded components of $\boldsymbol{g}(\partial K_s)^C$. If the unbounded component of C^C is U, then considering Q_j, it can't have any point of U. This is because U has no points of $\boldsymbol{g}(\partial K_s)$ a smaller set than C and so $Q_j \cup U$ would be connected, open, and contained in $\boldsymbol{g}(\partial K_s)^C$ so it would equal Q_j resulting in Q_j not being bounded after all. Could Q_j intersect some J_r? If it does, then $J_r \subseteq Q_j$ because J_r is connected and does not intersect $\boldsymbol{g}(\partial K_s)^C$. Consider $\boldsymbol{f}\left(\bar{Q}_j \diagdown \cup \mathscr{J}_j\right)$ where $\mathscr{J}_j$ are the components J_r contained in Q_j. Is $\boldsymbol{y}_s \in \boldsymbol{f}\left(\bar{Q}_j \diagdown \cup \mathscr{J}_j\right)$? From Lemma 13.6.1, $\partial Q_j \subseteq \boldsymbol{g}(\partial K_s) \subseteq C$ so $\boldsymbol{f}(\partial Q_j) \subseteq \partial K_s$ and so $\boldsymbol{y}_s \notin \boldsymbol{f}(\partial Q_j)$. Suppose $\boldsymbol{y}_s = \boldsymbol{f}(\boldsymbol{z})$ where $\boldsymbol{z} \in Q_j$. If $\boldsymbol{z}$ is not in any of the J_r but is in $\bar{Q}_j$ then $\boldsymbol{z} \in C$ so $\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{y}_s \in D$. But $\boldsymbol{y}_s$ is in K_s a component of D^C so this is impossible. Hence $\boldsymbol{z}$ is in one of the J_r and so this J_r is in $\mathscr{J}_j$. Therefore, $\boldsymbol{y}_s \notin \boldsymbol{f}\left(\bar{Q}_j \diagdown \cup \mathscr{J}_j\right)$ and so we can apply Lemma 13.7.1 in 13.10. First note that if $J_r \in \mathscr{J}_j$ then $d(\boldsymbol{g},K_s,Q_j) = d(\boldsymbol{g},K_s,J_r)$

$$1 = d(\boldsymbol{f}\circ\boldsymbol{g},K_s,\boldsymbol{y}_s) = \sum_j d(\boldsymbol{g},K_s,Q_j)\, d(\boldsymbol{f},Q_j,\boldsymbol{y}_s) = \sum_j \sum_{J\in\mathscr{J}_j} d(\boldsymbol{g},K_s,J)\, d(\boldsymbol{f},J,\boldsymbol{y}_s)$$

Since the Q_j cover at least C^C, it follows that each J intersects some Q_j and from the above is contained in Q_j. Thus the $\mathscr{J}_j$ cover $\cup\mathscr{J}$. Therefore, the above equals

$$= \sum_{J\in\mathscr{J}} d(\boldsymbol{g},K_s,J)\, d(\boldsymbol{f},J,\boldsymbol{y}_s) = \sum_{r=1}^{n} d(\boldsymbol{g},K_s,J_r)\, d(\boldsymbol{f},J_r,K_s)$$

where $\mathscr{J}$ is the set of components of C^C. Recall that in the product formula the sums are finite. Then adding over s, it follows

$$m = \sum_{s=1}^{m}\sum_{r=1}^{n} d(\boldsymbol{g},K_s,J_r)\, d(\boldsymbol{f},J_r,K_s)$$

However, we could do the same thing in the other order starting with components in C^C and obtain

$$1 = \sum_{s=1}^{m} d(\boldsymbol{g},K_s,J_r)\, d(\boldsymbol{f},J_r,K_s)$$

and then summing over r,

$$n = \sum_{r=1}^{n}\sum_{s=1}^{m} d(\boldsymbol{g},K_s,J_r)\, d(\boldsymbol{f},J_r,K_s) = \sum_{s=1}^{m}\sum_{r=1}^{n} d(\boldsymbol{g},K_s,J_r)\, d(\boldsymbol{f},J_r,K_s) = m. \blacksquare$$

13.8 Uniqueness of the Degree

I am mainly interested in the topological theorems which can be proved using the above topological degree. To me this justifies its importance. Nevertheless, there are other methods for finding the degree which are based more directly on topological considerations and algebra. These other methods are older than the presentation given here. Nevertheless if the degree satisfies the properties of the degree given in Theorem 13.2.2 along with the following condition, then this is sufficient to determine the degree.

Condition 13.8.1 *Let* $\boldsymbol{f} : \overline{B(\boldsymbol{w},R)} \to \mathbb{R}^p$ *be such that* $\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{w})) = \{\boldsymbol{w}\}$ *and suppose* $D\boldsymbol{f}(\boldsymbol{w})$ *is invertible. Then* $d(\boldsymbol{f},B(\boldsymbol{w},R),\boldsymbol{f}(\boldsymbol{w})) = \operatorname{sgn}(\det(D\boldsymbol{f}(\boldsymbol{w})))$.

This follows from a repeat of the arguments which led to the degree in the above. Homotopy invariance and the properties of Theorem 13.2.2 can be used to get the same definition of the degree for continuous functions given in the above. From this all the rest followed. In an appendix to my book "Linear Algebra and Analysis" such an approach to the degree based on algebra is given and it verifies the above condition. Thus this other approach based on homology gives the same degree function. Also, the above condition will end up following from Theorem 13.2.2 and by insisting that if $\boldsymbol{s}(\boldsymbol{x}) = \hat{\boldsymbol{x}}$ where $\hat{\boldsymbol{x}}$ has two components switched so it corresponds to that elementary matrix then the degree is -1, this will suffice with the other properties to show the above condition. This process is followed in that other approach to the degree. That something more is required follows because the degree also keeps track of orientation.

13.9 Exercises

1. Show that if $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_r$ in $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$, then if $\tilde{\boldsymbol{f}}$ has the property that

   $$\left\| \tilde{\boldsymbol{f}} - \boldsymbol{f} \right\|_\infty < \min_{i \leq r} \operatorname{dist}(\boldsymbol{y}_i, \boldsymbol{f}(\partial\Omega)),$$

   then $d(\boldsymbol{f}, \Omega, \boldsymbol{y}_i) = d\left(\tilde{\boldsymbol{f}}, \Omega, \boldsymbol{y}_i\right)$ for each $\boldsymbol{y}_i$. **Hint:** Consider for

   $$t \in [0,1], \boldsymbol{f}(\boldsymbol{x}) + t\left(\tilde{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\right) - \boldsymbol{y}_i$$

   and homotopy invariance.

2. Show the Brouwer fixed point theorem is equivalent to the nonexistence of a continuous retraction onto the boundary of $B(\mathbf{0}, r)$.

3. Give a version of Proposition 13.6.7 which is valid for the case where $n = 1$.

4. It was shown that if $\lim_{|\boldsymbol{x}| \to \infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty, \boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is locally one to one and continuous, then $\boldsymbol{f}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$. Suppose you have $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$ where $\boldsymbol{f}$ is one to one, continuous, and $\lim_{|\boldsymbol{x}| \to \infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$, $m < n$. Show that $\boldsymbol{f}$ cannot be onto.

5. Can there exist a one to one onto continuous map, $\boldsymbol{f}$ which takes the unit interval to the unit disk?

6. Let $m < n$ and let $B_m(\mathbf{0}{,}r)$ be the ball in $\mathbb{R}^m$ and $B_n(\mathbf{0}, r)$ be the ball in $\mathbb{R}^n$. Show that there is no one to one continuous map from $\overline{B_m(\mathbf{0}{,}r)}$ to $\overline{B_n(\mathbf{0}{,}r)}$. **Hint:** It is like the above problem.

7. Consider the unit disk, $\left\{(x,y) : x^2 + y^2 \leq 1\right\} \equiv D$ and the annulus

   $$\left\{(x,y) : \frac{1}{2} \leq x^2 + y^2 \leq 1\right\} \equiv A.$$

   Is it possible there exists a one to one onto continuous map $\boldsymbol{f}$ such that $\boldsymbol{f}(D) = A$? Thus D has no holes and A is really like D but with one hole punched out. Can you generalize to different numbers of holes? **Hint:** Consider the invariance of domain theorem. The interior of D would need to be mapped to the interior of A. Where do the points of the boundary of A come from? Consider Theorem 3.11.3.

8. Suppose C is a compact set in $\mathbb{R}^n$ which has empty interior and $\boldsymbol{f} : C \to \Gamma \subseteq \mathbb{R}^n$ is one to one onto and continuous with continuous inverse. Could Γ have nonempty interior? Show also that if $\boldsymbol{f}$ is one to one and onto Γ then if it is continuous, so is $\boldsymbol{f}^{-1}$.

9. Let K be a nonempty closed and convex subset of $\mathbb{R}^p$. Recall K is convex means that if $\boldsymbol{x}, \boldsymbol{y} \in K$, then for all $t \in [0,1]$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in K$. Show that if $\boldsymbol{x} \in \mathbb{R}^p$ there exists a unique $\boldsymbol{z} \in K$ such that $|\boldsymbol{x} - \boldsymbol{z}| = \min\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{y} \in K\}$. This $\boldsymbol{z}$ will be denoted as $P\boldsymbol{x}$. **Hint:** First note you do not know K is compact. Establish the parallelogram identity if you have not already done so,

$$|\boldsymbol{u} - \boldsymbol{v}|^2 + |\boldsymbol{u} + \boldsymbol{v}|^2 = 2|\boldsymbol{u}|^2 + 2|\boldsymbol{v}|^2.$$

Then let $\{\boldsymbol{z}_k\}$ be a minimizing sequence,

$$\lim_{k\to\infty} |\boldsymbol{z}_k - \boldsymbol{x}|^2 = \inf\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{y} \in K\} \equiv \lambda.$$

Using convexity, explain why

$$\left|\frac{\boldsymbol{z}_k - \boldsymbol{z}_m}{2}\right|^2 + \left|\boldsymbol{x} - \frac{\boldsymbol{z}_k + \boldsymbol{z}_m}{2}\right|^2 = 2\left|\frac{\boldsymbol{x} - \boldsymbol{z}_k}{2}\right|^2 + 2\left|\frac{\boldsymbol{x} - \boldsymbol{z}_m}{2}\right|^2$$

and then use this to argue $\{\boldsymbol{z}_k\}$ is a Cauchy sequence. Then if $\boldsymbol{z}_i$ works for $i = 1,2$, consider $(\boldsymbol{z}_1 + \boldsymbol{z}_2)/2$ to get a contradiction.

10. In Problem 9 show that $P\boldsymbol{x}$ satisfies the following variational inequality. $(\boldsymbol{x} - P\boldsymbol{x}) \cdot (\boldsymbol{y} - P\boldsymbol{x}) \leq 0$ for all $\boldsymbol{y} \in K$. Then show that $|P\boldsymbol{x}_1 - P\boldsymbol{x}_2| \leq |\boldsymbol{x}_1 - \boldsymbol{x}_2|$. **Hint:** For the first part note that if $\boldsymbol{y} \in K$, the function

$$t \to |\boldsymbol{x} - (P\boldsymbol{x} + t(\boldsymbol{y} - P\boldsymbol{x}))|^2$$

achieves its minimum on $[0,1]$ at $t = 0$. For the second part,

$$(\boldsymbol{x}_1 - P\boldsymbol{x}_1) \cdot (P\boldsymbol{x}_2 - P\boldsymbol{x}_1) \leq 0,\ (\boldsymbol{x}_2 - P\boldsymbol{x}_2) \cdot (P\boldsymbol{x}_1 - P\boldsymbol{x}_2) \leq 0.$$

Explain why

$$(\boldsymbol{x}_2 - P\boldsymbol{x}_2 - (\boldsymbol{x}_1 - P\boldsymbol{x}_1)) \cdot (P\boldsymbol{x}_2 - P\boldsymbol{x}_1) \geq 0$$

and then use a some manipulations and the Cauchy Schwarz inequality to get the desired inequality.

11. Suppose D is a set which is homeomorphic to $\overline{B(\mathbf{0},1)}$. This means there exists a continuous one to one map, $\boldsymbol{h}$ such that $\boldsymbol{h}\left(\overline{B(\mathbf{0},1)}\right) = D$ such that $\boldsymbol{h}^{-1}$ is also one to one. Show that if $\boldsymbol{f}$ is a continuous function which maps D to D then $\boldsymbol{f}$ has a fixed point. Now show that it suffices to say that $\boldsymbol{h}$ is one to one and continuous. In this case the continuity of $\boldsymbol{h}^{-1}$ is automatic. Sets which have the property that continuous functions taking the set to itself have at least one fixed point are said to have the fixed point property. Work Problem 7 using this notion of fixed point property. What about a solid ball and a donut? Could these be homeomorphic?

12. Using the definition of the derivative and the Vitali covering theorem, show that if $\boldsymbol{f} \in C^1(\overline{U}, \mathbb{R}^n)$ and ∂U has n dimensional measure zero then $\boldsymbol{f}(\partial U)$ also has measure zero. (This problem has little to do with this chapter. It is a review.)

13. Suppose Ω is any open bounded subset of $\mathbb{R}^n$ which contains $\mathbf{0}$ and that $\boldsymbol{f}:\overline{\Omega}\to\mathbb{R}^n$ is continuous with the property that $\boldsymbol{f}(\boldsymbol{x})\cdot\boldsymbol{x}\geq 0$ for all $\boldsymbol{x}\in\partial\Omega$. Show that then there exists $\boldsymbol{x}\in\Omega$ such that $\boldsymbol{f}(\boldsymbol{x})=\mathbf{0}$. Give a similar result in the case where the above inequality is replaced with $\leq$. **Hint:** You might consider the function $\boldsymbol{h}(t,\boldsymbol{x})\equiv t\boldsymbol{f}(\boldsymbol{x})+(1-t)\boldsymbol{x}$.

14. Suppose Ω is an open set in $\mathbb{R}^n$ containing $\mathbf{0}$ and suppose that $\boldsymbol{f}:\overline{\Omega}\to\mathbb{R}^n$ is continuous and $|\boldsymbol{f}(\boldsymbol{x})|\leq|\boldsymbol{x}|$ for all $\boldsymbol{x}\in\partial\Omega$. Show $\boldsymbol{f}$ has a fixed point in $\overline{\Omega}$. **Hint:** Consider $\boldsymbol{h}(t,\boldsymbol{x})\equiv t(\boldsymbol{x}-\boldsymbol{f}(\boldsymbol{x}))+(1-t)\boldsymbol{x}$ for $t\in[0,1]$. If $t=1$ and some $\boldsymbol{x}\in\partial\Omega$ is sent to $\mathbf{0}$, then you are done. Suppose therefore, that no fixed point exists on $\partial\Omega$. Consider $t<1$ and use the given inequality.

15. Let Ω be an open bounded subset of $\mathbb{R}^n$ and let $\boldsymbol{f},\boldsymbol{g}:\overline{\Omega}\to\mathbb{R}^n$ both be continuous, $\mathbf{0}\notin\boldsymbol{f}(\partial\Omega)$, such that $|\boldsymbol{f}(\boldsymbol{x})|-|\boldsymbol{g}(\boldsymbol{x})|>0$ for all $\boldsymbol{x}\in\partial\Omega$. Show that then $d(\boldsymbol{f}-\boldsymbol{g},\Omega,\mathbf{0})=d(\boldsymbol{f},\Omega,\mathbf{0})$. Show that if there exists $\boldsymbol{x}\in\boldsymbol{f}^{-1}(\mathbf{0})$, then there exists $\boldsymbol{x}\in(\boldsymbol{f}-\boldsymbol{g})^{-1}(\mathbf{0})$. **Hint:** Consider $\boldsymbol{h}(t,\boldsymbol{x})\equiv(1-t)\boldsymbol{f}(\boldsymbol{x})+t(\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}))$ and argue $\mathbf{0}\notin\boldsymbol{h}(t,\partial\Omega)$ for $t\in[0,1]$.

16. Let $f:\mathbb{C}\to\mathbb{C}$ where $\mathbb{C}$ is the field of complex numbers. Thus f has a real and imaginary part. Letting $z=x+iy, f(z)=u(x,y)+iv(x,y)$. Recall that the norm in $\mathbb{C}$ is given by $|x+iy|=\sqrt{x^2+y^2}$ and this is the usual norm in $\mathbb{R}^2$ for the ordered pair (x,y). Thus complex valued functions defined on $\mathbb{C}$ can be considered as $\mathbb{R}^2$ valued functions defined on some subset of $\mathbb{R}^2$. Such a complex function is said to be analytic if the usual definition holds. That is $f'(z)=\lim_{h\to 0}\frac{f(z+h)-f(z)}{h}$. In other words,
$$f(z+h)=f(z)+f'(z)h+o(h) \tag{13.11}$$
at a point z where the derivative exists. Let $f(z)=z^n$ where n is a positive integer. Thus $z^n=p(x,y)+iq(x,y)$ for p,q suitable polynomials in x and y. Show this function is analytic. Next show that for an analytic function and u and v the real and imaginary parts, the Cauchy Riemann equations hold, $u_x=v_y,\ u_y=-v_x$. In terms of mappings show 13.11 has the form
$$\begin{pmatrix} u(x+h_1,y+h_2) \\ v(x+h_1,y+h_2)\end{pmatrix}$$
$$\begin{aligned} &= \begin{pmatrix} u(x,y) \\ v(x,y)\end{pmatrix}+\begin{pmatrix} u_x(x,y) & u_y(x,y) \\ v_x(x,y) & v_y(x,y)\end{pmatrix}\begin{pmatrix} h_1 \\ h_2\end{pmatrix}+\boldsymbol{o}(\boldsymbol{h}) \\ &= \begin{pmatrix} u(x,y) \\ v(x,y)\end{pmatrix}+\begin{pmatrix} u_x(x,y) & -v_x(x,y) \\ v_x(x,y) & u_x(x,y)\end{pmatrix}\begin{pmatrix} h_1 \\ h_2\end{pmatrix}+\boldsymbol{o}(\boldsymbol{h})\end{aligned}$$
where $\boldsymbol{h}=(h_1,h_2)^T$ and h is given by h_1+ih_2. Thus the determinant of the above matrix is always nonnegative. Letting B_r denote the ball $B(0,r)=B((0,0),r)$ show $d(f,B_r,\mathbf{0})=n$ where $f(z)=z^n$. As a mapping on $\mathbb{R}^2$, $\boldsymbol{f}(x,y)=\begin{pmatrix} u(x,y) \\ v(x,y)\end{pmatrix}$. Thus show $d(\boldsymbol{f},B_r,\mathbf{0})=n$. **Hint:** You might consider $g(z)\equiv\prod_{j=1}^{n}(z-a_j)$ where the a_j are small real distinct numbers and argue that both this function and f are analytic but that $\mathbf{0}$ is a regular value for $\boldsymbol{g}$ although it is not so for $\boldsymbol{f}$. However, for each a_j small but distinct $d(\boldsymbol{f},B_r,\mathbf{0})=d(\boldsymbol{g},B_r,\mathbf{0})$.

17. Using Problem 16, prove the fundamental theorem of algebra as follows. Let $p(z)$ be a nonconstant polynomial of degree n, $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots$ Show that for large enough r, $|p(z)| > |p(z) - a_n z^n|$ for all $z \in \partial B(0,r)$. Now from Problem 15 you can conclude $d(p, B_r, 0) = d(f, B_r, 0) = n$ where $f(z) = a_n z^n$.

18. Suppose $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ satisfies $|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})| \geq \alpha |\boldsymbol{x} - \boldsymbol{y}|$, $\alpha > 0$. Show that $\boldsymbol{f}$ must map $\mathbb{R}^p$ onto $\mathbb{R}^p$. **Hint:** First show $\boldsymbol{f}$ is one to one. Then use invariance of domain. Next show, using the inequality, that the points not in $\boldsymbol{f}(\mathbb{R}^p)$ must form an open set because if $\boldsymbol{y}$ is such a point, then there can be no sequence $\{\boldsymbol{f}(\boldsymbol{x}_n)\}$ converging to it. Finally recall that $\mathbb{R}^p$ is connected.

19. Suppose D is a nonempty bounded open set in $\mathbb{R}^p$ and suppose $\boldsymbol{f} : D \to \partial D$ is continuous with $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ for $\boldsymbol{x} \in \partial D$. Show this cannot happen. **Hint:** Let $\boldsymbol{y} \in D$ and note that id and $\boldsymbol{f}$ agree on ∂D. Therefore, from properties of the degree, $d(\boldsymbol{f}, D, \boldsymbol{y}) = d(\mathrm{id}, D, \boldsymbol{y})$. Explain why this cannot occur.

20. Assume D is a closed ball in $\mathbb{R}^p$ and suppose $\boldsymbol{f} : D \to D$ is continuous. Use the above problem to conclude $\boldsymbol{f}$ has a fixed point. **Hint:** If no fixed point, let $\boldsymbol{g}(\boldsymbol{x})$ be the point on ∂D which results from extending the ray starting at $\boldsymbol{f}(\boldsymbol{x})$ to $\boldsymbol{x}$. This would be a continuous map from D to ∂D which does not move any point on ∂D. Draw a picture. This may be the easiest proof of the Brouwer fixed point theorem but note how dependent it is on the properties of the degree.

Appendix A

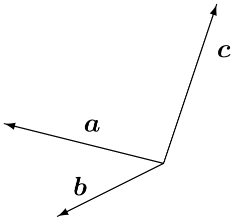
Basic Vector Analysis

A.1 The Cross Product

The cross product is the other way of multiplying two vectors in $\mathbb{R}^3$. It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product.

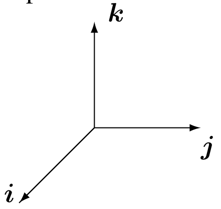
Definition A.1.1 *Three vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ form a right handed system if when you extend the fingers of your right hand along the vector $\boldsymbol{a}$ and close them in the direction of $\boldsymbol{b}$, the thumb points roughly in the direction of $\boldsymbol{c}$.*

For an example of a right handed system of vectors, see the following picture.



In this picture the vector $\boldsymbol{c}$ points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector $\boldsymbol{c}$ would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ will always form a right handed system. To repeat, if you extend the fingers of our right hand along $\boldsymbol{i}$ and close them in the direction $\boldsymbol{j}$, the thumb points in the direction of $\boldsymbol{k}$.



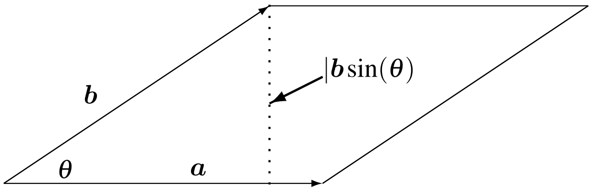
The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

Definition A.1.2 *Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^3$. Then $\boldsymbol{a} \times \boldsymbol{b}$ is defined by the following two rules.*

1. $|\boldsymbol{a} \times \boldsymbol{b}| = |\boldsymbol{a}| |\boldsymbol{b}| \sin \theta$ *where θ is the included angle.*

2. $\boldsymbol{a} \times \boldsymbol{b} \cdot \boldsymbol{a} = 0$, $\boldsymbol{a} \times \boldsymbol{b} \cdot \boldsymbol{b} = 0$, *and $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{a} \times \boldsymbol{b}$ forms a right hand system.*

Note that $|\boldsymbol{a} \times \boldsymbol{b}|$ is the area of the parallelogram spanned by $\boldsymbol{a}$ and $\boldsymbol{b}$.



The cross product satisfies the following properties.

$$\boldsymbol{a}\times\boldsymbol{b} = -(\boldsymbol{b}\times\boldsymbol{a})\ ,\ \boldsymbol{a}\times\boldsymbol{a} = \boldsymbol{0}, \tag{1.1}$$

For α a scalar,

$$(\alpha\boldsymbol{a})\times\boldsymbol{b} = \alpha(\boldsymbol{a}\times\boldsymbol{b}) = \boldsymbol{a}\times(\alpha\boldsymbol{b}), \tag{1.2}$$

For $\boldsymbol{a},\boldsymbol{b},$ and $\boldsymbol{c}$ vectors, one obtains the distributive laws,

$$\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c}) = \boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c}, \tag{1.3}$$

$$(\boldsymbol{b}+\boldsymbol{c})\times\boldsymbol{a} = \boldsymbol{b}\times\boldsymbol{a}+\boldsymbol{c}\times\boldsymbol{a}. \tag{1.4}$$

Formula 1.1 follows immediately from the definition. The vectors $\boldsymbol{a}\times\boldsymbol{b}$ and $\boldsymbol{b}\times\boldsymbol{a}$ have the same magnitude, $|\boldsymbol{a}|\,|\boldsymbol{b}|\sin\theta$, and an application of the right hand rule shows they have opposite direction. Formula 1.2 is also fairly clear. If α is a nonnegative scalar, the direction of $(\alpha\boldsymbol{a})\times\boldsymbol{b}$ is the same as the direction of $\boldsymbol{a}\times\boldsymbol{b}$,$\alpha(\boldsymbol{a}\times\boldsymbol{b})$ and $\boldsymbol{a}\times(\alpha\boldsymbol{b})$ while the magnitude is just α times the magnitude of $\boldsymbol{a}\times\boldsymbol{b}$ which is the same as the magnitude of $\alpha(\boldsymbol{a}\times\boldsymbol{b})$ and $\boldsymbol{a}\times(\alpha\boldsymbol{b})$. Using this yields equality in 1.2. In the case where $\alpha<0$, everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by $|\alpha|$ when comparing their magnitudes. The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using 1.1,

$$(\boldsymbol{b}+\boldsymbol{c})\times\boldsymbol{a} = -\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c}) = -(\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c}) = \boldsymbol{b}\times\boldsymbol{a}+\boldsymbol{c}\times\boldsymbol{a}.$$

A proof of the distributive law is given later.

Now from the definition of the cross product,

$$\begin{array}{ll}\boldsymbol{i}\times\boldsymbol{j}=\boldsymbol{k}, & \boldsymbol{j}\times\boldsymbol{i}=-\boldsymbol{k}\\ \boldsymbol{k}\times\boldsymbol{i}=\boldsymbol{j}, & \boldsymbol{i}\times\boldsymbol{k}=-\boldsymbol{j}\\ \boldsymbol{j}\times\boldsymbol{k}=\boldsymbol{i}, & \boldsymbol{k}\times\boldsymbol{j}=-\boldsymbol{i}\end{array}$$

With this information, the following gives the coordinate description of the cross product.

Proposition A.1.3 *Let* $\boldsymbol{a}=a_1\boldsymbol{i}+a_2\boldsymbol{j}+a_3\boldsymbol{k}$ *and* $\boldsymbol{b}=b_1\boldsymbol{i}+b_2\boldsymbol{j}+b_3\boldsymbol{k}$ *be two vectors. Then*

$$\boldsymbol{a}\times\boldsymbol{b} = (a_2b_3-a_3b_2)\boldsymbol{i}+(a_3b_1-a_1b_3)\boldsymbol{j}+(a_1b_2-a_2b_1)\boldsymbol{k}. \tag{1.5}$$

Proof: From the above table and the properties of the cross product listed,

$$(a_1\boldsymbol{i}+a_2\boldsymbol{j}+a_3\boldsymbol{k})\times(b_1\boldsymbol{i}+b_2\boldsymbol{j}+b_3\boldsymbol{k}) =$$

$$a_1b_2\boldsymbol{i}\times\boldsymbol{j}+a_1b_3\boldsymbol{i}\times\boldsymbol{k}+a_2b_1\boldsymbol{j}\times\boldsymbol{i}+a_2b_3\boldsymbol{j}\times\boldsymbol{k}+a_3b_1\boldsymbol{k}\times\boldsymbol{i}+a_3b_2\boldsymbol{k}\times\boldsymbol{j}$$

$$= a_1 b_2 \boldsymbol{k} - a_1 b_3 \boldsymbol{j} - a_2 b_1 \boldsymbol{k} + a_2 b_3 \boldsymbol{i} + a_3 b_1 \boldsymbol{j} - a_3 b_2 \boldsymbol{i}$$
$$= (a_2 b_3 - a_3 b_2)\,\boldsymbol{i} + (a_3 b_1 - a_1 b_3)\,\boldsymbol{j} + (a_1 b_2 - a_2 b_1)\,\boldsymbol{k} \qquad (1.6)$$

■

It is probably impossible for most people to remember 1.5. Fortunately, there is a somewhat easier way to remember it.

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \qquad (1.7)$$

where you formally expand the determinant along the top row. For those who have not seen determinants, here is a short description. All you need here is how to evaluate 2 × 2 and 3 × 3 determinants.

$$\begin{vmatrix} x & y \\ z & w \end{vmatrix} = xw - yz$$

and

$$\begin{vmatrix} a & b & c \\ x & y & z \\ u & v & w \end{vmatrix} = a \begin{vmatrix} y & z \\ v & w \end{vmatrix} - b \begin{vmatrix} x & z \\ u & w \end{vmatrix} + c \begin{vmatrix} x & y \\ u & v \end{vmatrix}.$$

Here is the rule: You look at an entry in the top row and cross out the row and column which contain that entry. If the entry is in the i^{th} column, you multiply $(-1)^{1+i}$ times the determinant of the 2 × 2 which remains. This is the cofactor. You take the element in the top row times this cofactor and add all such terms. The rectangular array enclosed by the vertical lines is called a **matrix** and a lot more can be said about these, but this is enough for our purposes here.

Example A.1.4 *Find* $(\boldsymbol{i} - \boldsymbol{j} + 2\boldsymbol{k}) \times (3\boldsymbol{i} - 2\boldsymbol{j} + \boldsymbol{k})$.

Use 1.7 to compute this.

$$\begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ 1 & -1 & 2 \\ 3 & -2 & 1 \end{vmatrix} = \begin{vmatrix} -1 & 2 \\ -2 & 1 \end{vmatrix} \boldsymbol{i} - \begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix} \boldsymbol{j} + \begin{vmatrix} 1 & -1 \\ 3 & -2 \end{vmatrix} \boldsymbol{k} = 3\boldsymbol{i} + 5\boldsymbol{j} + \boldsymbol{k}.$$

Example A.1.5 *Find the area of the parallelogram determined by the vectors*

$$(\boldsymbol{i} - \boldsymbol{j} + 2\boldsymbol{k}), \ (3\boldsymbol{i} - 2\boldsymbol{j} + \boldsymbol{k}).$$

These are the same two vectors in Example A.1.4.

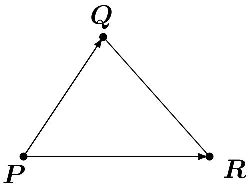
From Example A.1.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example A.1.4. Thus the area is $\sqrt{9+25+1} = \sqrt{35}$.

Example A.1.6 *Find the area of the triangle with verticies* $(1,2,3), (0,2,5)$, *and* $(5,1,2)$.

This triangle is obtained by connecting the three points with lines. Picking $(1,2,3)$ as a starting point, there are two displacement vectors $(-1,0,2)$ and $(4,-1,-1)$ such that the given vector added to these displacement vectors gives the other two vectors. The area of the triangle is half the area of the parallelogram determined by $(-1,0,2)$ and $(4,-1,-1)$. Thus $(-1,0,2) \times (4,-1,-1) = (2,7,1)$ and so the area of the triangle is $\frac{1}{2}\sqrt{4+49+1} = \frac{3}{2}\sqrt{6}$.

Observation A.1.7 *In general, if you have three points in* $\mathbb{R}^3, \boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{R}$ *the area of the triangle is given by*

$$\frac{1}{2}\left|(\boldsymbol{Q}-\boldsymbol{P})\times(\boldsymbol{R}-\boldsymbol{P})\right|.$$



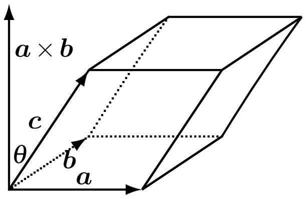
A.2 The Box Product

Definition A.2.1 *A parallelepiped determined by the three vectors* $\boldsymbol{a}, \boldsymbol{b},$ *and* $\boldsymbol{c}$ *consists of*

$$\{r\boldsymbol{a} + s\boldsymbol{b} + t\boldsymbol{c} : r,s,t \in [0,1]\}.$$

That is, if you pick three numbers, r, s, and t each in $[0,1]$ *and form* $r\boldsymbol{a} + s\boldsymbol{b} + t\boldsymbol{c}$, *then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.*

The following is a picture of such a thing.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ has area equal to $|\boldsymbol{a}\times\boldsymbol{b}|$ while the altitude of the parallelepiped is $|\boldsymbol{c}|\cos\theta$ where θ is the angle shown in the picture between $\boldsymbol{c}$ and $\boldsymbol{a}\times\boldsymbol{b}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

$$|\boldsymbol{a}\times\boldsymbol{b}|\,|\boldsymbol{c}|\cos\theta = \boldsymbol{a}\times\boldsymbol{b}\cdot\boldsymbol{c}.$$

This expression is known as the box product and is sometimes written as $[\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}]$. You should consider what happens if you interchange the $\boldsymbol{b}$ with the $\boldsymbol{c}$ or the $\boldsymbol{a}$ with the $\boldsymbol{c}$. You can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume.

Example A.2.2 *Find the volume of the parallelepiped determined by the vectors* $\boldsymbol{i}+2\boldsymbol{j}-5\boldsymbol{k}, \boldsymbol{i}+3\boldsymbol{j}-6\boldsymbol{k}, 3\boldsymbol{i}+2\boldsymbol{j}+3\boldsymbol{k}$.

According to the above discussion, pick any two of these, take the cross product and then take the dot product of this with the third of these vectors. The result will be either the desired volume or minus the desired volume.

$$(\boldsymbol{i}+2\boldsymbol{j}-5\boldsymbol{k})\times(\boldsymbol{i}+3\boldsymbol{j}-6\boldsymbol{k}) = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} = 3\boldsymbol{i}+\boldsymbol{j}+\boldsymbol{k}$$

Now take the dot product of this vector with the third which yields

$$(3\boldsymbol{i}+\boldsymbol{j}+\boldsymbol{k})\cdot(3\boldsymbol{i}+2\boldsymbol{j}+3\boldsymbol{k})=9+2+3=14.$$

This shows the volume of this parallelepiped is 14 cubic units.

There is a fundamental observation which comes directly from the geometric definitions of the cross product and the dot product.

Lemma A.2.3 *Let $\boldsymbol{a},\boldsymbol{b},$ and $\boldsymbol{c}$ be vectors. Then $(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{c}=\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c})$.*

Proof: This follows from observing that either $(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{c}$ and $\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c})$ both give the volume of the parallelepiped or they both give -1 times the volume. ■

A.3 Proof of the Distributive Law

Let $\boldsymbol{x}$ be a vector. From the above observation,

$$\begin{aligned}\boldsymbol{x}\cdot\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c}) &= (\boldsymbol{x}\times\boldsymbol{a})\cdot(\boldsymbol{b}+\boldsymbol{c})=(\boldsymbol{x}\times\boldsymbol{a})\cdot\boldsymbol{b}+(\boldsymbol{x}\times\boldsymbol{a})\cdot\boldsymbol{c}\\ &= \boldsymbol{x}\cdot\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{x}\cdot\boldsymbol{a}\times\boldsymbol{c}=\boldsymbol{x}\cdot(\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c}).\end{aligned}$$

Therefore,

$$\boldsymbol{x}\cdot[\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c})-(\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c})]=0$$

for all $\boldsymbol{x}$. In particular, this holds for $\boldsymbol{x}=\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c})-(\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c})$ showing that

$$\boldsymbol{a}\times(\boldsymbol{b}+\boldsymbol{c})=\boldsymbol{a}\times\boldsymbol{b}+\boldsymbol{a}\times\boldsymbol{c}$$

and this proves the distributive law for the cross product.

Observation A.3.1 *Suppose you have three vectors, $\boldsymbol{u}=(a,b,c)$, $\boldsymbol{v}=(d,e,f)$, and $\boldsymbol{w}=(g,h,i)$. Then $\boldsymbol{u}\cdot\boldsymbol{v}\times\boldsymbol{w}$ is given by the following.*

$$\begin{aligned}\boldsymbol{u}\cdot\boldsymbol{v}\times\boldsymbol{w} &= (a,b,c)\cdot\begin{vmatrix}\boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k}\\ d & e & f\\ g & h & i\end{vmatrix}=a\begin{vmatrix}e & f\\ h & i\end{vmatrix}-b\begin{vmatrix}d & f\\ g & i\end{vmatrix}+c\begin{vmatrix}d & e\\ g & h\end{vmatrix}\\ &= \det\begin{pmatrix}a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}.\end{aligned}$$

The message is that to take the box product, you can simply take the determinant of the matrix which results by letting the rows be the rectangular components of the given vectors in the order in which they occur in the box product.

A.4 Vector Identities and Notation

To begin with consider $\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w})$ and it is desired to simplify this expression. It turns out this expression comes up in many different contexts. Let $\boldsymbol{u}=(u_1,u_2,u_3)$ and let $\boldsymbol{v}$ and $\boldsymbol{w}$ be defined similarly.

$$\boldsymbol{v}\times\boldsymbol{w}=\begin{vmatrix}\boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k}\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix}=(v_2w_3-v_3w_2)\boldsymbol{i}+(w_1v_3-v_1w_3)\boldsymbol{j}+(v_1w_2-v_2w_1)\boldsymbol{k}$$

Next consider $\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w})$ which is given by

$$\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w}) = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ u_1 & u_2 & u_3 \\ (v_2w_3 - v_3w_2) & (w_1v_3 - v_1w_3) & (v_1w_2 - v_2w_1) \end{vmatrix}.$$

When you multiply this out, you get

$$\boldsymbol{i}\,(v_1u_2w_2 + u_3v_1w_3 - w_1u_2v_2 - u_3w_1v_3) + \boldsymbol{j}\,(v_2u_1w_1 + v_2w_3u_3 - w_2u_1v_1 - u_3w_2v_3)$$

$$+\boldsymbol{k}\,(u_1w_1v_3 + v_3w_2u_2 - u_1v_1w_3 - v_2w_3u_2)$$

and if you are clever, you see right away that

$$(\boldsymbol{i}v_1 + \boldsymbol{j}v_2 + \boldsymbol{k}v_3)(u_1w_1 + u_2w_2 + u_3w_3) - (\boldsymbol{i}w_1 + \boldsymbol{j}w_2 + \boldsymbol{k}w_3)(u_1v_1 + u_2v_2 + u_3v_3).$$

Thus

$$\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w}) = \boldsymbol{v}\,(\boldsymbol{u}\cdot\boldsymbol{w}) - \boldsymbol{w}\,(\boldsymbol{u}\cdot\boldsymbol{v}). \tag{1.8}$$

A related formula is

$$\begin{aligned}(\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w} &= -[\boldsymbol{w}\times(\boldsymbol{u}\times\boldsymbol{v})] = -[\boldsymbol{u}\,(\boldsymbol{w}\cdot\boldsymbol{v}) - \boldsymbol{v}\,(\boldsymbol{w}\cdot\boldsymbol{u})] \\ &= \boldsymbol{v}\,(\boldsymbol{w}\cdot\boldsymbol{u}) - \boldsymbol{u}\,(\boldsymbol{w}\cdot\boldsymbol{v}). \end{aligned} \tag{1.9}$$

This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, 1.8 and 1.9 are sufficient for most applications if you are creative in using them, but there is another way. This other way allows you to discover such vector identities as the above without any creativity or any cleverness. Therefore, it is far superior to the above nasty and tedious computation. It is a vector identity discovering machine and it is this which is the main topic in what follows. I cannot understand why it is not routinely presented in calculus texts. The engineers I have known seem to know all about it.

There are two special symbols, δ_{ij} and ε_{ijk} which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

Definition A.4.1 *The symbol δ_{ij}, called the Kroneker delta symbol is defined as follows.*

$$\delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}.$$

With the Kroneker symbol i and j can equal any integer in $\{1,2,\cdots,n\}$ for any $n \in \mathbb{N}$.

Definition A.4.2 *For i, j, and k integers in the set, $\{1,2,3\}$, ε_{ijk} is defined as follows.*

$$\varepsilon_{ijk} \equiv \begin{cases} 1 \text{ if } (i,j,k) = (1,2,3),(2,3,1), \text{ or } (3,1,2) \\ -1 \text{ if } (i,j,k) = (2,1,3),(1,3,2), \text{ or } (3,2,1) \\ 0 \text{ if there are any repeated integers} \end{cases}.$$

The subscripts ijk and ij in the above are called indices. A single one is called an index. This symbol ε_{ijk} is also called the permutation symbol.

The way to think of ε_{ijk} is that $\varepsilon_{123} = 1$ and if you switch any two of the numbers in the list i, j, k, it changes the sign. Thus $\varepsilon_{ijk} = -\varepsilon_{jik}$ and $\varepsilon_{ijk} = -\varepsilon_{kji}$ etc. You should check that this rule reduces to the above definition. For example, it immediately implies that if there is a repeated index, the answer is zero. This follows because $\varepsilon_{iij} = -\varepsilon_{iij}$ and so $\varepsilon_{iij} = 0$.

It is useful to use the Einstein summation convention when dealing with these symbols. Simply stated, the convention is that you sum over the repeated index. Thus $a_i b_i$ means $\sum_i a_i b_i$. Also, $\delta_{ij} x_j$ means $\sum_j \delta_{ij} x_j = x_i$. When you use this convention, there is one very important thing to never forget. It is this: Never have an index be repeated more than once. Thus $a_i b_i$ is all right but $a_{ii} b_i$ is not. The reason for this is that you end up getting confused about what is meant. If you want to write $\sum_i a_i b_i c_i$ it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

Lemma A.4.3 *The following holds.*

$$\varepsilon_{ijk}\varepsilon_{irs} = \left(\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}\right).$$

Proof: If $\{j,k\} \neq \{r,s\}$ then every term in the sum on the left must have either ε_{ijk} or ε_{irs} contains a repeated index. Therefore, the left side equals zero. The right side also equals zero in this case. To see this, note that if the two sets of indices are not equal, then there is one of the indices in one of the sets which is not in the other set. For example, it could be that j is not equal to either r or s. Then the right side equals zero.

Therefore, it can be assumed $\{j,k\} = \{r,s\}$. If $i = r$ and $j = s$ for $s \neq r$, then there is exactly one term in the sum on the left and it equals 1. The right also reduces to 1 in this case. If $i = s$ and $j = r$, there is exactly one term in the sum on the left which is nonzero and it must equal -1. The right side also reduces to -1 in this case. If there is a repeated index in $\{j,k\}$, then every term in the sum on the left equals zero. The right also reduces to zero in this case because then $j = k = r = s$ and so the right side becomes $(1)(1) - (-1)(-1) = 0$. ■

Proposition A.4.4 *Let $\boldsymbol{u}, \boldsymbol{v}$ be vectors in $\mathbb{R}^p$ where the Cartesian coordinates of $\boldsymbol{u}$ are $(u_1, \cdots, u_p)$ and the Cartesian coordinates of $\boldsymbol{v}$ are $(v_1, \cdots, v_p)$. Then $\boldsymbol{u} \cdot \boldsymbol{v} = u_i v_i$. If $\boldsymbol{u}, \boldsymbol{v}$ are vectors in $\mathbb{R}^3$, then*

$$(\boldsymbol{u} \times \boldsymbol{v})_i = \varepsilon_{ijk} u_j v_k.$$

Also, $\delta_{ik} a_k = a_i$.

Proof: The first claim is obvious from the definition of the dot product. The second is verified by simply checking it works. For example,

$$\boldsymbol{u} \times \boldsymbol{v} \equiv \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

and so

$$(\boldsymbol{u} \times \boldsymbol{v})_1 = (u_2 v_3 - u_3 v_2).$$

From the above formula in the proposition,

$$\varepsilon_{1jk} u_j v_k \equiv u_2 v_3 - u_3 v_2,$$

the same thing. The cases for $(\boldsymbol{u} \times \boldsymbol{v})_2$ and $(\boldsymbol{u} \times \boldsymbol{v})_3$ are verified similarly. The last claim follows directly from the definition. ■

With this notation, you can easily discover vector identities and simplify expressions which involve the cross product.

Example A.4.5 *Discover a formula which simplifies* $(\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w}$*.*

From the above reduction formula,

$$\begin{aligned}((\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w})_i &= \varepsilon_{ijk}(\boldsymbol{u}\times\boldsymbol{v})_j w_k = \varepsilon_{ijk}\varepsilon_{jrs}u_r v_s w_k \\ &= -\varepsilon_{jik}\varepsilon_{jrs}u_r v_s w_k = -(\delta_{ir}\delta_{ks}-\delta_{is}\delta_{kr})u_r v_s w_k \\ &= -(u_i v_k w_k - u_k v_i w_k) = \boldsymbol{u}\cdot\boldsymbol{w}v_i - \boldsymbol{v}\cdot\boldsymbol{w}u_i \\ &= ((\boldsymbol{u}\cdot\boldsymbol{w})\boldsymbol{v}-(\boldsymbol{v}\cdot\boldsymbol{w})\boldsymbol{u})_i .\end{aligned}$$

Since this holds for all i, it follows that

$$(\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w} = (\boldsymbol{u}\cdot\boldsymbol{w})\boldsymbol{v}-(\boldsymbol{v}\cdot\boldsymbol{w})\boldsymbol{u}.$$

A.5 Divergence and Curl of a Vector Field

Here the important concepts of divergence and curl are defined in terms of rectangular coordinates.

Definition A.5.1 *Let* $\boldsymbol{f}:U\to\mathbb{R}^p$ *for* $U\subseteq\mathbb{R}^p$ *denote a vector field. A scalar valued function is called a* ***scalar field****. The function* $\boldsymbol{f}$ *is called a* C^k ***vector field*** *if the function* $\boldsymbol{f}$ *is a* C^k *function. For a* C^1 *vector field, as just described* $\nabla\cdot\boldsymbol{f}(\boldsymbol{x})\equiv\operatorname{div}\boldsymbol{f}(\boldsymbol{x})$ *known as the* ***divergence****, is defined as*

$$\nabla\cdot\boldsymbol{f}(\boldsymbol{x})\equiv\operatorname{div}\boldsymbol{f}(\boldsymbol{x})\equiv\sum_{i=1}^{p}\frac{\partial f_i}{\partial x_i}(\boldsymbol{x}).$$

Using the repeated summation convention, this is often written as

$$f_{i,i}(\boldsymbol{x})\equiv\partial_i f_i(\boldsymbol{x})$$

where the comma indicates a partial derivative is being taken with respect to the i^{th} *variable and* ∂_i *denotes differentiation with respect to the* i^{th} *variable. In words, the divergence is the sum of the* i^{th} *derivative of the* i^{th} *component function of* $\boldsymbol{f}$ *for all values of i. If* $p=3$*, the* ***curl*** *of the vector field yields another vector field and it is defined as follows.*

$$(\operatorname{curl}(\boldsymbol{f})(\boldsymbol{x}))_i\equiv(\nabla\times\boldsymbol{f}(\boldsymbol{x}))_i\equiv\varepsilon_{ijk}\partial_j f_k(\boldsymbol{x})$$

where here ∂_j *means the partial derivative with respect to* x_j *and the subscript of i in* $(\operatorname{curl}(\boldsymbol{f})(\boldsymbol{x}))_i$ *means the* i^{th} *Cartesian component of the vector* $\operatorname{curl}(\boldsymbol{f})(\boldsymbol{x})$*. Thus the curl is evaluated by expanding the following determinant along the top row.*

$$\begin{vmatrix}\boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k}\\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z}\\ f_1(x,y,z) & f_2(x,y,z) & f_3(x,y,z)\end{vmatrix}.$$

Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.) Also

$$\nabla^2 f\equiv\nabla\cdot(\nabla f).$$

This last symbol is important enough that it is given a name, the ***Laplacian****.It is also denoted by* Δ*. Thus* $\nabla^2 f = \Delta f$*. In addition for* $\boldsymbol{f}$ *a vector field, the symbol* $\boldsymbol{f}\cdot\nabla$ *is defined as a "differential operator" in the following way.*

$$\boldsymbol{f}\cdot\nabla(\boldsymbol{g}) \equiv f_1(\boldsymbol{x})\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial x_1} + f_2(\boldsymbol{x})\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial x_2} + \cdots + f_p(\boldsymbol{x})\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial x_p}.$$

Thus $\boldsymbol{f}\cdot\nabla$ *takes vector fields and makes them into new vector fields.*

This definition is in terms of a given rectangular coordinate system but later coordinate free definitions of the curl and div are presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence and curl have profound physical significance and this will be discussed later. For now it is important to understand their definition in terms of coordinates. Be sure you understand that for $\boldsymbol{f}$ a vector field, $\operatorname{div}\boldsymbol{f}$ is a scalar field meaning it is a scalar valued function of three variables. For a scalar field f, ∇f is a vector field described earlier. For $\boldsymbol{f}$ a vector field having values in $\mathbb{R}^3$, $\operatorname{curl}\boldsymbol{f}$ is another vector field.

Example A.5.2 *Let* $\boldsymbol{f}(\boldsymbol{x}) = xy\boldsymbol{i} + (z-y)\boldsymbol{j} + (\sin(x)+z)\boldsymbol{k}$*. Find* $\operatorname{div}\boldsymbol{f}$ *and* $\operatorname{curl}\boldsymbol{f}$*.*

First the divergence of $\boldsymbol{f}$ is

$$\frac{\partial(xy)}{\partial x} + \frac{\partial(z-y)}{\partial y} + \frac{\partial(\sin(x)+z)}{\partial z} = y + (-1) + 1 = y.$$

Now $\operatorname{curl}\boldsymbol{f}$ is obtained by evaluating

$$\begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z-y & \sin(x)+z \end{vmatrix} =$$

$$\boldsymbol{i}\left(\frac{\partial}{\partial y}(\sin(x)+z) - \frac{\partial}{\partial z}(z-y)\right) - \boldsymbol{j}\left(\frac{\partial}{\partial x}(\sin(x)+z) - \frac{\partial}{\partial z}(xy)\right) +$$

$$\boldsymbol{k}\left(\frac{\partial}{\partial x}(z-y) - \frac{\partial}{\partial y}(xy)\right) = -\boldsymbol{i} - \cos(x)\boldsymbol{j} - x\boldsymbol{k}.$$

A.6 Vector Identities

There are many interesting identities which relate the gradient, divergence and curl.

Theorem A.6.1 *Assuming* $\boldsymbol{f},\boldsymbol{g}$ *are a* C^2 *vector fields whenever necessary, the following identities are valid.*

1. $\nabla\cdot(\nabla\times\boldsymbol{f}) = 0$
2. $\nabla\times\nabla\phi = \mathbf{0}$
3. $\nabla\times(\nabla\times\boldsymbol{f}) = \nabla(\nabla\cdot\boldsymbol{f}) - \nabla^2\boldsymbol{f}$ *where* $\nabla^2\boldsymbol{f}$ *is a vector field whose* i^{th} *component is* $\nabla^2 f_i$*.*
4. $\nabla\cdot(\boldsymbol{f}\times\boldsymbol{g}) = \boldsymbol{g}\cdot(\nabla\times\boldsymbol{f}) - \boldsymbol{f}\cdot(\nabla\times\boldsymbol{g})$

5. $\nabla\times(\boldsymbol{f}\times\boldsymbol{g}) = (\nabla\cdot\boldsymbol{g})\,\boldsymbol{f} - (\nabla\cdot\boldsymbol{f})\,\boldsymbol{g} + (\boldsymbol{g}\cdot\nabla)\,\boldsymbol{f} - (\boldsymbol{f}\cdot\nabla)\,\boldsymbol{g}$

Proof: These are all easy to establish if you use the repeated index summation convention and the reduction identities.

$$\begin{aligned}
\nabla\cdot(\nabla\times\boldsymbol{f}) &= \partial_i(\nabla\times\boldsymbol{f})_i = \partial_i\left(\varepsilon_{ijk}\partial_j f_k\right) = \varepsilon_{ijk}\partial_i(\partial_j f_k) \\
&= \varepsilon_{jik}\partial_j(\partial_i f_k) = -\varepsilon_{ijk}\partial_j(\partial_i f_k) = -\varepsilon_{ijk}\partial_i(\partial_j f_k) \\
&= -\nabla\cdot(\nabla\times\boldsymbol{f})\,.
\end{aligned}$$

This establishes the first formula. The second formula is done similarly. Now consider the third.

$$\begin{aligned}
(\nabla\times(\nabla\times\boldsymbol{f}))_i &= \varepsilon_{ijk}\partial_j(\nabla\times\boldsymbol{f})_k = \varepsilon_{ijk}\partial_j(\varepsilon_{krs}\partial_r f_s) \\
&= \overbrace{\varepsilon_{kij}}^{=\varepsilon_{ijk}}\varepsilon_{krs}\partial_j(\partial_r f_s) = (\delta_{ir}\delta_{js}-\delta_{is}\delta_{jr})\,\partial_j(\partial_r f_s) \\
&= \partial_j(\partial_i f_j) - \partial_j(\partial_j f_i) = \partial_i(\partial_j f_j) - \partial_j(\partial_j f_i) \\
&= \Big(\nabla(\nabla\cdot\boldsymbol{f}) - \nabla^2\boldsymbol{f}\Big)_i
\end{aligned}$$

This establishes the third identity.

Consider the fourth identity.

$$\begin{aligned}
\nabla\cdot(\boldsymbol{f}\times\boldsymbol{g}) &= \partial_i(\boldsymbol{f}\times\boldsymbol{g})_i = \partial_i\varepsilon_{ijk}f_j g_k \\
&= \varepsilon_{ijk}(\partial_i f_j)\,g_k + \varepsilon_{ijk}f_j(\partial_i g_k) \\
&= \left(\varepsilon_{kij}\partial_i f_j\right)g_k - \left(\varepsilon_{jik}\partial_i g_k\right)f_k \\
&= \nabla\times\boldsymbol{f}\cdot\boldsymbol{g} - \nabla\times\boldsymbol{g}\cdot\boldsymbol{f}.
\end{aligned}$$

This proves the fourth identity.

Consider the fifth.

$$\begin{aligned}
(\nabla\times(\boldsymbol{f}\times\boldsymbol{g}))_i &= \varepsilon_{ijk}\partial_j(\boldsymbol{f}\times\boldsymbol{g})_k = \varepsilon_{ijk}\partial_j\varepsilon_{krs}f_r g_s \\
&= \varepsilon_{kij}\varepsilon_{krs}\partial_j(f_r g_s) = (\delta_{ir}\delta_{js}-\delta_{is}\delta_{jr})\,\partial_j(f_r g_s) \\
&= \partial_j(f_i g_j) - \partial_j(f_j g_i) \\
&= (\partial_j g_j)\,f_i + g_j\partial_j f_i - (\partial_j f_j)\,g_i - f_j(\partial_j g_i) \\
&= ((\nabla\cdot\boldsymbol{g})\,\boldsymbol{f} + (\boldsymbol{g}\cdot\nabla)(\boldsymbol{f}) - (\nabla\cdot\boldsymbol{f})\,\boldsymbol{g} - (\boldsymbol{f}\cdot\nabla)(\boldsymbol{g}))_i
\end{aligned}$$

and this establishes the fifth identity. ■

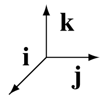
Appendix B

Curvilinear Coordinates

B.1 Basis Vectors

In this chapter, I will use the repeated index summation convention unless stated otherwise. Thus, **a repeated index indicates a sum.** Also, it is helpful in order to keep things straight to always have the two repeated indices be on different levels. That is, I will write $a_i^j b_j$ and not $a_{ij} b_j$. The reason for this will become clear as the exposition proceeds.

The usual basis vectors are denoted by $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ and are as the following picture describes.

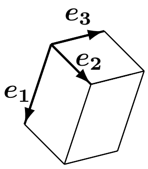


The vectors, $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$, are fixed. If $\boldsymbol{v}$ is a vector, there are unique scalars called components such that $\boldsymbol{v} = v^1\, \boldsymbol{i} + v^2\, \boldsymbol{j} + v^3 \boldsymbol{k}$. This is what it means that $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ is a basis.

Now suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are three vectors which satisfy

$$\boldsymbol{e}_1 \times \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 \neq 0.$$

Recall this means the volume of the box spanned by the three vectors is not zero.



Suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are as just described. Does it follow that they form a basis? In other words, for any vector $\boldsymbol{v}$, there are unique scalars v^i such that $\boldsymbol{v} = v^i \boldsymbol{e}_i$. Of course this is the case because the box product is really the determinant of the matrix which has $\boldsymbol{e}_i$ as the i^{th} row (column). This is the content of the following theorem.

Theorem B.1.1 *If $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are three vectors, then they form a basis if and only if*

$$\boldsymbol{e}_1 \times \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 \neq 0.$$

This gives a simple geometric condition which determines whether a list of three vectors forms a basis in $\mathbb{R}^3$. One simply takes the box product. If the box product is not equal to zero, then the vectors form a basis. If not, the list of three vectors does not form a basis. This condition generalizes to $\mathbb{R}^p$ as follows. If $\boldsymbol{e}_i = a_i^j \boldsymbol{i}_j$, then $\{\boldsymbol{e}_i\}_{i=1}^p$ forms a basis if and only if $\det\left(a_i^j\right) \neq 0$.

These vectors may or may not be orthonormal. In any case, it is convenient to define something called the dual basis.

Definition B.1.2 *Let $\{\boldsymbol{e}_i\}_{i=1}^p$ form a basis for $\mathbb{R}^p$. Then $\left\{\boldsymbol{e}^i\right\}_{i=1}^p$ is called the dual basis if*

$$\boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta_j^i \equiv \left\{ \begin{array}{l} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{array} \right. . \qquad (2.1)$$

Theorem B.1.3 *If $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis then $\left\{\boldsymbol{e}^i\right\}_{i=1}^p$ is also a basis provided 2.1 holds.*

Proof: Suppose

$$\boldsymbol{v} = v_i \boldsymbol{e}^i. \qquad (2.2)$$

Then taking the dot product of both sides of 2.2 with $\boldsymbol{e}_j$,yields

$$v_j = \boldsymbol{v} \cdot \boldsymbol{e}_j. \qquad (2.3)$$

Thus there is at most one choice of scalars v_j such that $\boldsymbol{v} = v_j \boldsymbol{e}^j$ and it is given by 2.3.

$$\left(\boldsymbol{v} - \boldsymbol{v}\cdot\boldsymbol{e}_j\boldsymbol{e}^j\right)\cdot\boldsymbol{e}_k = 0$$

and so, since $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis,

$$\left(\boldsymbol{v} - \boldsymbol{v}\cdot\boldsymbol{e}_j\boldsymbol{e}^j\right)\cdot\boldsymbol{w} = 0$$

for all vectors $\boldsymbol{w}$. It follows $\boldsymbol{v} - \boldsymbol{v}\cdot\boldsymbol{e}_j\boldsymbol{e}^j = \boldsymbol{0}$ and this shows $\left\{\boldsymbol{e}^i\right\}_{i=1}^p$ is a basis. ■

In the above argument are obtained formulas for the components of a vector $\boldsymbol{v}$, v_i, with respect to the dual basis, found to be $v_j = \boldsymbol{v}\cdot\boldsymbol{e}_j$. In the same way, one can find the components of a vector with respect to the basis $\{\boldsymbol{e}_i\}_{i=1}^p$. Let $\boldsymbol{v}$ be any vector and let

$$\boldsymbol{v} = v^j\boldsymbol{e}_j. \tag{2.4}$$

Then taking the dot product of both sides of 2.4 with $\boldsymbol{e}^i$ we see $v^i = \boldsymbol{e}^i\cdot\boldsymbol{v}$.

Does there exist a dual basis and is it uniquely determined?

Theorem B.1.4 *If $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis for $\mathbb{R}^p$, then there exists a unique dual basis, $\left\{\boldsymbol{e}^j\right\}_{j=1}^p$ satisfying*

$$\boldsymbol{e}^j\cdot\boldsymbol{e}_i = \delta_i^j.$$

Proof: First I show the dual basis is unique. Suppose $\left\{\boldsymbol{f}^j\right\}_{j=1}^p$ is another set of vectors which satisfies $\boldsymbol{f}^j\cdot\boldsymbol{e}_i = \delta_i^j$. Then

$$\boldsymbol{f}^j = \boldsymbol{f}^j\cdot\boldsymbol{e}_i\boldsymbol{e}^i = \delta_i^j\boldsymbol{e}^i = \boldsymbol{e}^j.$$

Note that from the definition, the dual basis to $\left\{\boldsymbol{i}_j\right\}_{j=1}^p$ is just $\boldsymbol{i}^j = \boldsymbol{i}_j$. It remains to verify the existence of the dual basis. Consider the matrix $g_{ij} \equiv \boldsymbol{e}_i\cdot\boldsymbol{e}_j$. This is called the **metric tensor.** If the resulting matrix is denoted as G, does it follow that G^{-1} exists? Suppose you have $\boldsymbol{e}_i\cdot\boldsymbol{e}_j x^j = 0$. Then, since i is arbitrary, this implies $\boldsymbol{e}_j x^j = \boldsymbol{0}$ and since $\left\{\boldsymbol{e}_j\right\}$ is a basis, this requires each x^j to be zero. Thus G is invertible. Denote by g^{ij} the ij^{th} entry of this inverse matrix. Consider $\boldsymbol{e}^j \equiv g^{jk}\boldsymbol{e}_k$. Is this the dual basis as the notation implies?

$$\boldsymbol{e}^j\cdot\boldsymbol{e}_i = g^{jk}\boldsymbol{e}_k\cdot\boldsymbol{e}_i = g^{jk}g_{ki} = \delta_i^j$$

so yes, it is indeed the dual basis. This has shown both existence and uniqueness of the dual basis. ■

From this is a useful observation.

Proposition B.1.5 $\{\boldsymbol{e}_i\}_{i=1}^p$ *is a basis for $\mathbb{R}^p$ if and only if when $\boldsymbol{e}_i = a_i^j\boldsymbol{i}_j$,* $\det\left(a_i^j\right) \neq 0$.

Proof: First suppose $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis for $\mathbb{R}^p$. Letting $A_{ij} \equiv a_i^j$, we need to show that $\det(A) \neq 0$. This is equivalent to showing that A or A^T is one to one. But

$$a_i^j x^i = 0 \Rightarrow a_i^j x^i\boldsymbol{i}_j = 0 \Rightarrow \boldsymbol{e}_i x^i = 0 \Rightarrow x^i = 0$$

so A^T is one to one if and only if $\det(A) = \det\left(A^T\right) \neq 0$.

Conversely, suppose A has nonzero determinant. Why are the $\boldsymbol{e}_k$ a basis? Suppose $x^k\boldsymbol{e}_k = \mathbf{0}$. Is each $x^k = 0$? Then $x^k a_k^j \boldsymbol{i}_j = \mathbf{0}$ and so for each j, $a_k^j x^k = 0$ and since A has nonzero determinant, $x^k = 0$. ■

Summarizing what has been shown so far, we know that $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis for $\mathbb{R}^p$ if and only if when $\boldsymbol{e}_i = a_i^j \boldsymbol{i}_j$,

$$\det\left(a_i^j\right) \neq 0. \tag{2.5}$$

If $\{\boldsymbol{e}_i\}_{i=1}^p$ is a basis, then there exists a unique dual basis, $\left\{\boldsymbol{e}^j\right\}_{j=1}^p$ satisfying

$$\boldsymbol{e}^j \cdot \boldsymbol{e}_i = \delta_i^j, \tag{2.6}$$

and that if $\boldsymbol{v}$ is any vector,

$$\boldsymbol{v} = v_j \boldsymbol{e}^j,\ \boldsymbol{v} = v^j \boldsymbol{e}_j. \tag{2.7}$$

The components of $\boldsymbol{v}$ which have the index on the top are called the contravariant components of the vector while the components which have the index on the bottom are called the covariant components. In general $v_i \neq v^j$! We also have formulae for these components in terms of the dot product.

$$v_j = \boldsymbol{v} \cdot \boldsymbol{e}_j,\ v^j = \boldsymbol{v} \cdot \boldsymbol{e}^j. \tag{2.8}$$

As indicated above, define $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j$ and $g^{ij} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}^j$. The next theorem describes the process of raising or lowering an index.

Theorem B.1.6 *The following hold.*

$$g^{ij}\boldsymbol{e}_j = \boldsymbol{e}^i,\ g_{ij}\boldsymbol{e}^j = \boldsymbol{e}_i, \tag{2.9}$$

$$g^{ij}v_j = v^i,\ g_{ij}v^j = v_i, \tag{2.10}$$

$$g^{ij}g_{jk} = \delta_k^i, \tag{2.11}$$

$$\det\left(g_{ij}\right) > 0,\ \det\left(g^{ij}\right) > 0. \tag{2.12}$$

Proof: First,

$$\boldsymbol{e}^i = \boldsymbol{e}^i \cdot \boldsymbol{e}^j \boldsymbol{e}_j = g^{ij}\boldsymbol{e}_j$$

by 2.7 and 2.8. Similarly, by 2.7 and 2.8,

$$\boldsymbol{e}_i = \boldsymbol{e}_i \cdot \boldsymbol{e}_j \boldsymbol{e}^j = g_{ij}\boldsymbol{e}^j.$$

This verifies 2.9. To verify 2.10,

$$v^i = \boldsymbol{e}^i \cdot \boldsymbol{v} = g^{ij}\boldsymbol{e}_j \cdot \boldsymbol{v} = g^{ij}v_j.$$

The proof of the remaining formula in 2.10 is similar.

To verify 2.11,

$$g^{ij}g_{jk} = \boldsymbol{e}^i \cdot \boldsymbol{e}^j \boldsymbol{e}_j \cdot \boldsymbol{e}_k = \left(\left(\boldsymbol{e}^i \cdot \boldsymbol{e}^j\right)\boldsymbol{e}_j\right) \cdot \boldsymbol{e}_k = \boldsymbol{e}^i \cdot \boldsymbol{e}_k = \delta_k^i.$$

This shows the two determinants in 2.12 are non zero because the two matrices are inverses of each other. It only remains to verify that one of these is greater than zero. Letting $\boldsymbol{e}_i = a_i^j \boldsymbol{i}_j = b_j^i \boldsymbol{i}^j$, we see that since $\boldsymbol{i}_j = \boldsymbol{i}^j, a_i^j = b_j^i$. Therefore,

$$\boldsymbol{e}_i \cdot \boldsymbol{e}_j = a_i^r \boldsymbol{i}_r \cdot b_k^j \boldsymbol{i}^k = a_i^r b_k^j \delta_r^k = a_i^k b_k^j = a_i^k a_j^k.$$

It follows that for G the matrix whose ij^{th} entry is $\boldsymbol{e}_i \cdot \boldsymbol{e}_j$, $G = AA^T$ where the ik^{th} entry of A is a_i^k. Therefore, $\det(G) = \det(A)\det(A^T) = \det(A)^2 > 0$. It follows from 2.11 that if H is the matrix whose ij^{th} entry is g^{ij}, then $GH = I$ and so $H = G^{-1}$ and

$$\det(G)\det(G^{-1}) = \det(g^{ij})\det(G) = 1.$$

Therefore, $\det(G^{-1}) > 0$ also. ■

Note that $\det(AA^T) \geq 0$ always, because the eigenvalues are nonegative.

As noted above, we have the following definition.

Definition B.1.7 *The matrix $(g_{ij}) = G$ is called the metric tensor.*

B.2 Exercises

1. Let $\boldsymbol{e}_1 = \boldsymbol{i}+\boldsymbol{j}, \boldsymbol{e}_2 = \boldsymbol{i}-\boldsymbol{j}, \boldsymbol{e}_3 = \boldsymbol{j}+\boldsymbol{k}.$ Find $\boldsymbol{e}^1, \boldsymbol{e}^2, \boldsymbol{e}^3$, (g_{ij}), (g^{ij}). If $\boldsymbol{v} = \boldsymbol{i}+2\boldsymbol{j}+\boldsymbol{k},$ find v^i and v_j, the contravariant and covariant components of the vector.

2. Let $\boldsymbol{e}^1 = 2\boldsymbol{i}+\boldsymbol{j}, \boldsymbol{e}^2 = \boldsymbol{i}-\boldsymbol{2j}, \boldsymbol{e}^3 = \boldsymbol{k}.$ Find $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, (g_{ij}), (g^{ij}). If $\boldsymbol{v} = 2\,\boldsymbol{i}-2\boldsymbol{j}+\boldsymbol{k},$ find v^i and v_j, the contravariant and covariant components of the vector.

3. Suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ have the property that $\boldsymbol{e}_i \cdot \boldsymbol{e}_j = 0$ whenever $i \neq j$. Show the same is true of the dual basis.

4. Let $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_3$ be a basis for $\mathbb{R}^n$ and let $\boldsymbol{v} = v^i\boldsymbol{e}_i = v_i\boldsymbol{e}^i, \boldsymbol{w} = w^j\boldsymbol{e}_j = w_j\boldsymbol{e}^j$ be two vectors. Show
$$\boldsymbol{v}\cdot\boldsymbol{w} = g_{ij}v^iw^j = g^{ij}v_iw_j.$$

5. Show if $\{\boldsymbol{e}_i\}_{i=1}^3$ is a basis in $\mathbb{R}^3$
$$\boldsymbol{e}^1 = \frac{\boldsymbol{e}_2\times\boldsymbol{e}_3}{\boldsymbol{e}_2\times\boldsymbol{e}_3\cdot\boldsymbol{e}_1},\ \boldsymbol{e}^2 = \frac{\boldsymbol{e}_1\times\boldsymbol{e}_3}{\boldsymbol{e}_1\times\boldsymbol{e}_3\cdot\boldsymbol{e}_2},\ \boldsymbol{e}^3 = \frac{\boldsymbol{e}_1\times\boldsymbol{e}_2}{\boldsymbol{e}_1\times\boldsymbol{e}_2\cdot\boldsymbol{e}_3}.$$

6. Let $\{\boldsymbol{e}_i\}_{i=1}^n$ be a basis and define
$$\boldsymbol{e}_i^* \equiv \frac{\boldsymbol{e}_i}{|\boldsymbol{e}_i|},\ \boldsymbol{e}^{*i} \equiv \boldsymbol{e}^i\,|\boldsymbol{e}_i|.$$
Show $\boldsymbol{e}^{*i}\cdot\boldsymbol{e}_j^* = \delta_j^i$.

7. If $\boldsymbol{v}$ is a vector, v_i^* and v^{*i}, are defined by
$$\boldsymbol{v} \equiv v_i^*\boldsymbol{e}^{*i} \equiv v^{*i}\boldsymbol{e}_i^*.$$
These are called the physical components of $\boldsymbol{v}.$ Show
$$v_i^* = \frac{v_i}{|\boldsymbol{e}_i|},\ v^{*i} = v^i\,|\boldsymbol{e}_i| \text{ (No summation on } i \text{).}$$

B.3 Curvilinear Coordinates

There are many ways to identify a point in n dimensional space with an ordered list of real numbers. Some of these are spherical coordinates, cylindrical coordinates and rectangular coordinates and these particular examples are discussed earlier. I will denote by $\boldsymbol{y}$ the rectangular coordinates of a point in n dimensional space which I will go on writing as $\mathbb{R}^n$. Thus $\boldsymbol{y} = \begin{pmatrix} y^1 & \cdots & y^n \end{pmatrix}$. It follows there are equations which relate the rectangular coordinates to some other coordinates $\begin{pmatrix} x^1 & \cdots & x^n \end{pmatrix}$. In spherical coordinates, these were ρ, ϕ, θ where the geometric meaning of these were described earlier. However, completely general systems are to be considered here, with certain stipulations. The idea is

$$y^k = y^k\left(x^1,...,x^n\right),\ \boldsymbol{y} = \boldsymbol{y}\left(x^1,...,x^n\right)$$

Let $\begin{pmatrix} x^1 & \cdots & x^n \end{pmatrix} \in D \subseteq \mathbb{R}^n$ be an open set and let

$$\boldsymbol{x} \to \boldsymbol{y}\left(x^1,...,x^n\right) \equiv \boldsymbol{M}\left(x^1,...,x^n\right)$$

satisfy

$$\boldsymbol{M} \text{ is } C^2, \tag{2.13}$$

$$\boldsymbol{M} \text{ is one to one.} \tag{2.14}$$

Letting $\boldsymbol{x} \in D$, we can write

$$\boldsymbol{M}(\boldsymbol{x}) = M^k(\boldsymbol{x})\, \boldsymbol{i}_k$$

where, as usual, $\boldsymbol{i}_k$ are the standard basis vectors for $\mathbb{R}^n$, $\boldsymbol{i}_k$ being the vector in $\mathbb{R}^n$ which has a one in the k^{th} coordinate and a 0 in every other spot. Thus $y^k = M^k(\boldsymbol{x})$ where this y^k refers to the k^{th} rectangular coordinate of the point $\boldsymbol{y}$ as just described.

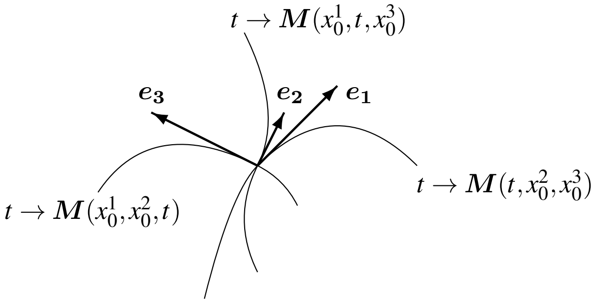
For a fixed $\boldsymbol{x} \in D$, we can consider the space curves,

$$t \to \boldsymbol{M}(\boldsymbol{x} + t\boldsymbol{i}_k) \equiv \boldsymbol{y}(\boldsymbol{x} + t\boldsymbol{i}_k)$$

for $t \in I$, some open interval containing 0. Then for the point $\boldsymbol{x}$,we let

$$\boldsymbol{e}_k \equiv \frac{\partial \boldsymbol{M}}{\partial x^k}(\boldsymbol{x}) \equiv \frac{d}{dt}\left(\boldsymbol{M}(\boldsymbol{x} + t\boldsymbol{i}_k)\right)|_{t=0} \equiv \frac{\partial \boldsymbol{y}}{\partial x^k}(\boldsymbol{x})$$

Denote this vector as $\boldsymbol{e}_k(\boldsymbol{x})$ to emphasize its dependence on $\boldsymbol{x}$. The following picture illustrates the situation in $\mathbb{R}^3$.


$t \to \boldsymbol{M}(x_0^1, t, x_0^3)$
$\boldsymbol{e_3}$
$\boldsymbol{e_2}$
$\boldsymbol{e_1}$
$t \to \boldsymbol{M}(t, x_0^2, x_0^3)$
$t \to \boldsymbol{M}(x_0^1, x_0^2, t)$

I want $\{\boldsymbol{e}_k\}_{k=1}^n$ to be a basis. Thus, from Proposition B.1.5,

$$\det\left(\frac{\partial M^i}{\partial x^k}\right) \equiv \det\left(D\boldsymbol{y}\left(\boldsymbol{x}\right)\right) \equiv \det\left(D\left(\boldsymbol{M}\right)\left(\boldsymbol{x}\right)\right) \neq 0. \tag{2.15}$$

Let

$$y^i = M^i\left(\boldsymbol{x}\right)\ i = 1,\cdots,n \tag{2.16}$$

so that the y^i are the usual rectangular coordinates with respect to the usual basis vectors $\{\boldsymbol{i}_k\}_{k=1}^n$ of the point $\boldsymbol{y} = \boldsymbol{M}\left(\boldsymbol{x}\right)$. Letting $\boldsymbol{x} \equiv \left(x^1,\cdots,x^n\right)$, it follows from the inverse function theorem (See Chapter 7) that $\boldsymbol{M}\left(D\right)$ is open, and that 2.15, 2.13, and 2.14 imply the equations 2.16 define each x^i as a C^2 function of $\boldsymbol{y} \equiv \left(y^1,\cdots,y^n\right)$. Thus, abusing notation slightly, the equations 2.16 are equivalent to

$$x^i = x^i\left(y^1,...,y^n\right),\ i = 1,\cdots,n$$

where x^i is a C^2 function of the rectangular coordinates of a point $\boldsymbol{y}$. It follows from the material on the gradient described earlier,

$$\nabla x^k\left(\boldsymbol{y}\right) = \frac{\partial x^k\left(\boldsymbol{y}\right)}{\partial y^j}\boldsymbol{i}^j.$$

Then

$$\nabla x^k\left(\boldsymbol{y}\right)\cdot\boldsymbol{e}_j = \frac{\partial x^k}{\partial y^s}\boldsymbol{i}^s\cdot\frac{\partial y^r}{\partial x^j}\boldsymbol{i}_r = \frac{\partial x^k}{\partial y^s}\frac{\partial y^s}{\partial x^j} = \delta_j^k$$

by the chain rule. Therefore, the dual basis is given by

$$\boldsymbol{e}^k\left(\boldsymbol{x}\right) = \nabla x^k\left(\boldsymbol{y}\left(\boldsymbol{x}\right)\right). \tag{2.17}$$

Notice that it might be hard or even impossible to solve algebraically for x^i in terms of the y^j. Thus the straight forward approach to finding $\boldsymbol{e}^k$ by 2.17 might be impossible. Also, this approach leads to an expression in terms of the $\boldsymbol{y}$ coordinates rather than the desired $\boldsymbol{x}$ coordinates. Therefore, it is expedient to use another method to obtain these vectors in terms of $\boldsymbol{x}$. Indeed, this is the main idea in this chapter, doing everything in terms of $\boldsymbol{x}$ rather than $\boldsymbol{y}$. The vectors, $\boldsymbol{e}^k\left(\boldsymbol{x}\right)$ may always be found by using formula 2.9 and the result is in terms of the curvilinear coordinates $\boldsymbol{x}$. Here is a familiar example.

Example B.3.1 *$D \equiv (0,\infty)\times(0,\pi)\times(0,2\pi)$ and*

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1\sin\left(x^2\right)\cos\left(x^3\right) \\ x^1\sin\left(x^2\right)\sin\left(x^3\right) \\ x^1\cos\left(x^2\right) \end{pmatrix}$$

(We usually write this as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho\sin\left(\phi\right)\cos\left(\theta\right) \\ \rho\sin\left(\phi\right)\sin\left(\theta\right) \\ \rho\cos\left(\phi\right) \end{pmatrix}$$

where (ρ,ϕ,θ) are the spherical coordinates. We are calling them x^1, x^2, and x^3 to preserve the notation just discussed.) Thus

$$\boldsymbol{e}_1\left(\boldsymbol{x}\right) = \sin\left(x^2\right)\cos\left(x^3\right)\boldsymbol{i}_1 + \sin\left(x^2\right)\sin\left(x^3\right)\boldsymbol{i}_2 + \cos\left(x^2\right)\boldsymbol{i}_3,$$

$$
\begin{gathered}
\boldsymbol{e}_2(\boldsymbol{x}) = x^1 \cos\left(x^2\right)\cos\left(x^3\right)\boldsymbol{i}_1 \\
+x^1\cos\left(x^2\right)\sin\left(x^3\right)\boldsymbol{i}_2 - x^1\sin\left(x^2\right)\boldsymbol{i}_3, \\
\boldsymbol{e}_3(\boldsymbol{x}) = -x^1\sin\left(x^2\right)\sin\left(x^3\right)\boldsymbol{i}_1 + x^1\sin\left(x^2\right)\cos\left(x^3\right)\boldsymbol{i}_2 + 0\boldsymbol{i}_3.
\end{gathered}
$$

It follows the metric tensor is

$$
G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^2 & 0 \\ 0 & 0 & \left(x^1\right)^2\sin^2\left(x^2\right) \end{pmatrix} = \left(g_{ij}\right) = \left(\boldsymbol{e}_i \cdot \boldsymbol{e}_j\right). \tag{2.18}
$$

Therefore, by Theorem B.1.6

$$
\begin{gathered}
G^{-1} = \left(g^{ij}\right) \\
= \left(\boldsymbol{e}^i, \boldsymbol{e}^j\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^{-2} & 0 \\ 0 & 0 & \left(x^1\right)^{-2}\sin^{-2}\left(x^2\right) \end{pmatrix}.
\end{gathered}
$$

To obtain the dual basis, use Theorem B.1.6 to write

$$
\begin{gathered}
\boldsymbol{e}^1(\boldsymbol{x}) = g^{1j}\boldsymbol{e}_j(\boldsymbol{x}) = \boldsymbol{e}_1(\boldsymbol{x}) \\
\boldsymbol{e}^2(\boldsymbol{x}) = g^{2j}\boldsymbol{e}_j(\boldsymbol{x}) = \left(x^1\right)^{-2}\boldsymbol{e}_2(\boldsymbol{x}) \\
\boldsymbol{e}^3(\boldsymbol{x}) = g^{3j}\boldsymbol{e}_j(\boldsymbol{x}) = \left(x^1\right)^{-2}\sin^{-2}\left(x^2\right)\boldsymbol{e}_3(\boldsymbol{x}).
\end{gathered}
$$

Note that $\frac{\partial \boldsymbol{y}}{\partial y^k} \equiv \boldsymbol{e}_k(\boldsymbol{y}) = \boldsymbol{i}^k = \boldsymbol{i}_k$ where, as described, $\begin{pmatrix} y^1 & \cdots & y^n \end{pmatrix}$ are the rectangular coordinates of the point in $\mathbb{R}^n$.

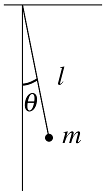
B.4 Exercises

1. Let

$$
\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^2 + x^3 \\ x^1 - 2x^2 \end{pmatrix}
$$

where the y^i are the rectangular coordinates of the point. Find $\boldsymbol{e}^i, \boldsymbol{e}_i, i = 1,2,3$, and find $\left(g_{ij}\right)(\boldsymbol{x})$ and $\left(g^{ij}(\boldsymbol{x})\right)$.

2. Let $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x},t)$ where t signifies time and $\boldsymbol{x} \in U \subseteq \mathbb{R}^m$ for U an open set, while $\boldsymbol{y} \in \mathbb{R}^n$ and suppose $\boldsymbol{x}$ is a function of t. Physically, this corresponds to an object moving over a surface in $\mathbb{R}^n$ which may be changing as a function of t. The point $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}(t), t)$ is the point in $\mathbb{R}^n$ corresponding to t. For example, consider the pendulum



in which $n = 2, l$ is fixed and $y^1 = l\sin\theta, y^2 = l - l\cos\theta$. Thus, in this simple example, $m = 1$. If l were changing in a known way with respect to t, then this would be of the form $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, t)$. In general, the kinetic energy is defined as

$$T \equiv \frac{1}{2} m \dot{\boldsymbol{y}} \cdot \dot{\boldsymbol{y}} \qquad (*)$$

where the dot on the top signifies differentiation with respect to t. Show

$$\frac{\partial T}{\partial \dot{x}^k} = m \dot{\boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial x^k}.$$

Hint: First show

$$\dot{\boldsymbol{y}} = \frac{\partial \boldsymbol{y}}{\partial x^j} \dot{x}^j + \frac{\partial \boldsymbol{y}}{\partial t} \qquad (**)$$

and so

$$\frac{\partial \dot{\boldsymbol{y}}}{\partial \dot{x}^j} = \frac{\partial \boldsymbol{y}}{\partial x^j}.$$

3. ↑ Show

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) = m \ddot{\boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial x^k} + m \dot{\boldsymbol{y}} \cdot \frac{\partial^2 \boldsymbol{y}}{\partial x^k \partial x^r} \dot{x}^r + m \dot{\boldsymbol{y}} \cdot \frac{\partial^2 \boldsymbol{y}}{\partial t \partial x^k}.$$

4. ↑ Show

$$\frac{\partial T}{\partial x^k} = m \dot{\boldsymbol{y}} \cdot \left(\frac{\partial^2 \boldsymbol{y}}{\partial x^r \partial x^k} \dot{x}^r + \frac{\partial^2 \boldsymbol{y}}{\partial t \partial x^k}\right).$$

Hint: Use $*$ and $**$.

5. ↑ Now show from Newton's second law (mass times acceleration equals force) that for $\boldsymbol{F}$ the force,

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) - \frac{\partial T}{\partial x^k} = m \ddot{\boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial x^k} = \boldsymbol{F} \cdot \frac{\partial \boldsymbol{y}}{\partial x^k}. \qquad (***)$$

6. ↑ In the example of the simple pendulum above,

$$\boldsymbol{y} = \begin{pmatrix} l\sin\theta \\ l - l\cos\theta \end{pmatrix} = l\sin\theta\, \boldsymbol{i} + (l - l\cos\theta)\, \boldsymbol{j}.$$

Use $***$ to find a differential equation which describes the vibrations of the pendulum in terms of θ. First write the kinetic energy and then consider the force acting on the mass which is $-mg\boldsymbol{j}$.

7. Of course, the idea is to write equations of motion in terms of the variables x^k, instead of the rectangular variables y^k. Suppose $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ and $\boldsymbol{x}$ is a function of t. Letting G denote the metric tensor, show that the kinetic energy is of the form $\frac{1}{2} m \dot{\boldsymbol{x}}^T G \boldsymbol{x}$ where m is a point mass with m its mass.

8. The pendulum problem is fairly easy to do without the formalism developed. Now consider the case where $\boldsymbol{x} = (\rho, \theta, \phi)$, spherical coordinates, and write differential equations for ρ, θ, and ϕ to describe the motion of an object in terms of these coordinates given a force, $\boldsymbol{F}$.

9. Suppose the pendulum is not assumed to vibrate in a plane. Let it be suspended at the origin and let ϕ be the angle between the negative z axis and the positive x axis while θ is the angle between the projection of the position vector onto the xy plane and the positive x axis in the usual way. Thus

$$x = \rho \sin\phi \cos\theta, y = \rho \sin\phi \sin\theta, z = -\rho \cos\phi$$

10. If there are many masses, $m_\alpha, \alpha = 1, \cdots, R$, the kinetic energy is the sum of the kinetic energies of the individual masses. Thus,

$$T \equiv \frac{1}{2} \sum_{\alpha=1}^{R} m_\alpha \left|\dot{\boldsymbol{y}}_\alpha\right|^2.$$

Generalize the above problems to show that, assuming

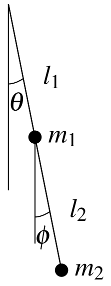
$$\boldsymbol{y}_\alpha = \boldsymbol{y}_\alpha(\boldsymbol{x}, t),$$

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha=1}^{R} \boldsymbol{F}_\alpha \cdot \frac{\partial \boldsymbol{y}_\alpha}{\partial x^k}$$

where $\boldsymbol{F}_\alpha$ is the force acting on m_α.

11. Discuss the equivalence of these formulae with Newton's second law, force equals mass times acceleration. What is gained from the above so called Lagrangian formalism?

12. The double pendulum has two masses instead of only one.



Write differential equations for θ and ϕ to describe the motion of the double pendulum.

B.5 Transformation of Coordinates.

How do we write $\boldsymbol{e}^k(\boldsymbol{x})$ in terms of the vectors, $\boldsymbol{e}^j(\boldsymbol{z})$ where $\boldsymbol{z}$ is some other type of curvilinear coordinates? This is next.

Consider the following picture in which U is an open set in $\mathbb{R}^n$, D and $\widehat{D}$ are open sets in $\mathbb{R}^n$, and $\boldsymbol{M}, \boldsymbol{N}$ are C^2 mappings which are one to one from D and $\widehat{D}$ respectively. The only reason for this is to ensure that the mixed partial derivatives are equal. We will suppose that a point in U is identified by the curvilinear coordinates $\boldsymbol{x}$ in D and $\boldsymbol{z}$ in $\widehat{D}$.

$$\begin{tikzcd}
& U & \\
\underset{(x^1,x^2,x^3)}{D} \arrow[ur, bend left, "\boldsymbol{M}"] & & \underset{(z^1,z^2,z^3)}{\widehat{D}} \arrow[ul, bend right, "\boldsymbol{N}"']
\end{tikzcd}$$

Thus $\boldsymbol{M}(\boldsymbol{x}) = \boldsymbol{N}(\boldsymbol{z})$ and so $\boldsymbol{z} = \boldsymbol{N}^{-1}(\boldsymbol{M}(\boldsymbol{x}))$. The point in U will be denoted in rectangular coordinates as $\boldsymbol{y}$ and we have $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{y}(\boldsymbol{z})$ Now by the chain rule,

$$\boldsymbol{e}_i(\boldsymbol{z}) = \frac{\partial \boldsymbol{y}}{\partial z^i} = \frac{\partial \boldsymbol{y}}{\partial x^j}\frac{\partial x^j}{\partial z_i} = \frac{\partial x^j}{\partial z^i}\boldsymbol{e}_j(\boldsymbol{x}) \tag{2.19}$$

Define the covariant and contravariant coordinates for the various curvilinear coordinates in the obvious way. Thus,

$$\boldsymbol{v} = v_i(\boldsymbol{x})\boldsymbol{e}^i(\boldsymbol{x}) = v^i(\boldsymbol{x})\boldsymbol{e}_i(\boldsymbol{x}) = v_j(\boldsymbol{z})\boldsymbol{e}^j(\boldsymbol{z}) = v^j(\boldsymbol{z})\boldsymbol{e}_j(\boldsymbol{z}).$$

Then the following theorem tells how to transform the vectors and coordinates.

Theorem B.5.1 *The following transformation rules hold for pairs of curvilinear coordinates.*

$$v_i(\boldsymbol{z}) = \frac{\partial x^j}{\partial z_i}v_j(\boldsymbol{x}),\ v^i(\boldsymbol{z}) = \frac{\partial z^i}{\partial x^j}v^j(\boldsymbol{x}), \tag{2.20}$$

$$\boldsymbol{e}_i(\boldsymbol{z}) = \frac{\partial x^j}{\partial z_i}\boldsymbol{e}_j(\boldsymbol{x}),\ \boldsymbol{e}^i(\boldsymbol{z}) = \frac{\partial z^i}{\partial x^j}\boldsymbol{e}^j(\boldsymbol{x}), \tag{2.21}$$

$$g_{ij}(\boldsymbol{z}) = \frac{\partial x^r}{\partial z^i}\frac{\partial x^s}{\partial z^j}g_{rs}(\boldsymbol{x}),\ g^{ij}(\boldsymbol{z}) = \frac{\partial z^i}{\partial x^r}\frac{\partial z^j}{\partial x^s}g^{rs}(\boldsymbol{x}). \tag{2.22}$$

Proof: We already have shown the first part of 2.21 in 2.19. Then, from 2.19,

$$\begin{aligned}
\boldsymbol{e}^i(\boldsymbol{z}) &= \boldsymbol{e}^i(\boldsymbol{z})\cdot\boldsymbol{e}_j(\boldsymbol{x})\boldsymbol{e}^j(\boldsymbol{x}) = \boldsymbol{e}^i(\boldsymbol{z})\cdot\frac{\partial z^k}{\partial x^j}\boldsymbol{e}_k(\boldsymbol{z})\boldsymbol{e}^j(\boldsymbol{x}) \\
&= \delta^i_k\frac{\partial z^k}{\partial x^j}\boldsymbol{e}^j(\boldsymbol{x}) = \frac{\partial z^i}{\partial x^j}\boldsymbol{e}^j(\boldsymbol{x})
\end{aligned}$$

and this proves the second part of 2.21. Now to show 2.20,

$$v_i(\boldsymbol{z}) = \boldsymbol{v}\cdot\boldsymbol{e}_i(\boldsymbol{z}) = \boldsymbol{v}\cdot\frac{\partial x^j}{\partial z_i}\boldsymbol{e}_j(\boldsymbol{x}) = \frac{\partial x^j}{\partial z_i}\boldsymbol{v}\cdot\boldsymbol{e}_j(\boldsymbol{x}) = \frac{\partial x^j}{\partial z_i}v_j(\boldsymbol{x})$$

and

$$v^i(\boldsymbol{z}) = \boldsymbol{v}\cdot\boldsymbol{e}^i(\boldsymbol{z}) = \boldsymbol{v}\cdot\frac{\partial z^i}{\partial x^j}\boldsymbol{e}^j(\boldsymbol{x}) = \frac{\partial z^i}{\partial x^j}\boldsymbol{v}\cdot\boldsymbol{e}^j(\boldsymbol{x}) = \frac{\partial z^i}{\partial x^j}v^j(\boldsymbol{x}).$$

To verify 2.22,

$$g_{ij}(\boldsymbol{z}) = \boldsymbol{e}_i(\boldsymbol{z})\cdot\boldsymbol{e}_j(\boldsymbol{z}) = \boldsymbol{e}_r(\boldsymbol{x})\frac{\partial x^r}{\partial z^i}\cdot\boldsymbol{e}_s(\boldsymbol{x})\frac{\partial x^s}{\partial z^j} = g_{rs}(\boldsymbol{x})\frac{\partial x^r}{\partial z^i}\frac{\partial x^s}{\partial z^j}.\ \blacksquare$$

B.6 Differentiation and Christoffel Symbols

Let $\boldsymbol{F} : U \to \mathbb{R}^n$ be differentiable. We call $\boldsymbol{F}$ a vector field and it is used to model force, velocity, acceleration, or any other vector quantity which may change from point to point in U. Then $\frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^j}$ is a vector and so there exist scalars, $F^i_{,j}(\boldsymbol{x})$ and $F_{i,j}(\boldsymbol{x})$ such that

$$\frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^j} = F^i_{,j}(\boldsymbol{x})\,\boldsymbol{e}_i(\boldsymbol{x}), \; \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^j} = F_{i,j}(\boldsymbol{x})\,\boldsymbol{e}^i(\boldsymbol{x}) \tag{2.23}$$

We will see how these scalars transform when the coordinates are changed.

Theorem B.6.1 *If $\boldsymbol{x}$ and $\boldsymbol{z}$ are curvilinear coordinates,*

$$F^r_{,s}(\boldsymbol{x}) = F^i_{,j}(\boldsymbol{z})\frac{\partial x^r}{\partial z^i}\frac{\partial z^j}{\partial x^s},\; F_{r,s}(\boldsymbol{x})\frac{\partial x^r}{\partial z^i}\frac{\partial x^s}{\partial z^j} = F_{i,j}(\boldsymbol{z}). \tag{2.24}$$

Proof:

$$F^r_{,s}(\boldsymbol{x})\,\boldsymbol{e}_r(\boldsymbol{x}) \equiv \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^s} = \frac{\partial \boldsymbol{F}(\boldsymbol{z})}{\partial z^j}\frac{\partial z^j}{\partial x^s} \equiv$$

$$F^i_{,j}(\boldsymbol{z})\,\boldsymbol{e}_i(\boldsymbol{z})\frac{\partial z^j}{\partial x^s} = F^i_{,j}(\boldsymbol{z})\frac{\partial z^j}{\partial x^s}\frac{\partial x^r}{\partial z^i}\boldsymbol{e}_r(\boldsymbol{x})$$

which shows the first formula of 2.23. To show the other formula,

$$F_{i,j}(\boldsymbol{z})\,\boldsymbol{e}^i(\boldsymbol{z}) \equiv \frac{\partial \boldsymbol{F}(\boldsymbol{z})}{\partial z^j} = \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^s}\frac{\partial x^s}{\partial z^j} \equiv$$

$$F_{r,s}(\boldsymbol{x})\,\boldsymbol{e}^r(\boldsymbol{x})\frac{\partial x^s}{\partial z^j} = F_{r,s}(\boldsymbol{x})\frac{\partial x^s}{\partial z^j}\frac{\partial x^r}{\partial z^i}\boldsymbol{e}^i(\boldsymbol{z}),$$

and this shows the second formula for transforming these scalars. ■

Now $\boldsymbol{F}(\boldsymbol{x}) = F^i(\boldsymbol{x})\,\boldsymbol{e}_i(\boldsymbol{x})$ and so by the product rule,

$$\frac{\partial \boldsymbol{F}}{\partial x^j} = \frac{\partial F^i}{\partial x^j}\boldsymbol{e}_i(\boldsymbol{x}) + F^i(\boldsymbol{x})\frac{\partial \boldsymbol{e}_i(\boldsymbol{x})}{\partial x^j}. \tag{2.25}$$

Now $\frac{\partial \boldsymbol{e}_i(\boldsymbol{x})}{\partial x^j}$ is a vector and so there exist scalars, $\left\{ \begin{array}{c} k \\ ij \end{array} \right\}$ such that

$$\frac{\partial \boldsymbol{e}_i(\boldsymbol{x})}{\partial x^j} = \left\{ \begin{array}{c} k \\ ij \end{array} \right\} \boldsymbol{e}_k(\boldsymbol{x}).$$

Thus

$$\left\{ \begin{array}{c} k \\ ij \end{array} \right\} \boldsymbol{e}_k(\boldsymbol{x}) = \frac{\partial^2 \boldsymbol{y}}{\partial x^j \partial x^i}$$

and so

$$\left\{ \begin{array}{c} k \\ ij \end{array} \right\} \boldsymbol{e}_k(\boldsymbol{x}) \cdot \boldsymbol{e}^r(\boldsymbol{x}) = \left\{ \begin{array}{c} k \\ ij \end{array} \right\} \delta^r_k = \left\{ \begin{array}{c} r \\ ij \end{array} \right\} = \frac{\partial^2 \boldsymbol{y}}{\partial x^j \partial x^i} \cdot \boldsymbol{e}^r(\boldsymbol{x}) \tag{2.26}$$

Therefore, from 2.25, $\frac{\partial \boldsymbol{F}}{\partial x^j} = \frac{\partial F^k}{\partial x^j}\boldsymbol{e}_k(\boldsymbol{x}) + F^i(\boldsymbol{x}) \left\{ \begin{array}{c} r \\ ij \end{array} \right\} \boldsymbol{e}_k(\boldsymbol{x})$ which shows

$$F^k_{,j}(\boldsymbol{x}) = \frac{\partial F^k}{\partial x^j} + F^i(\boldsymbol{x}) \left\{ \begin{array}{c} k \\ ij \end{array} \right\}. \tag{2.27}$$

This is sometimes called the covariant derivative.

Theorem B.6.2 *The Christoffel symbols of the second kind satisfy the following*

$$\frac{\partial \boldsymbol{e}_i(\boldsymbol{x})}{\partial x^j} = \left\{ \begin{array}{c} k \\ ij \end{array} \right\} \boldsymbol{e}_k(\boldsymbol{x}), \tag{2.28}$$

$$\frac{\partial \boldsymbol{e}^i(\boldsymbol{x})}{\partial x^j} = -\left\{ \begin{array}{c} i \\ kj \end{array} \right\} \boldsymbol{e}^k(\boldsymbol{x}), \tag{2.29}$$

$$\left\{ \begin{array}{c} k \\ ij \end{array} \right\} = \left\{ \begin{array}{c} k \\ ji \end{array} \right\}, \tag{2.30}$$

$$\left\{ \begin{array}{c} m \\ ik \end{array} \right\} = \frac{g^{jm}}{2} \left[\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right]. \tag{2.31}$$

Proof: Formula 2.28 is the definition of the Christoffel symbols. We verify 2.29 next. To do so, note

$$\boldsymbol{e}^i(\boldsymbol{x}) \cdot \boldsymbol{e}_k(\boldsymbol{x}) = \delta^i_k.$$

Then from the product rule,

$$\frac{\partial \boldsymbol{e}^i(\boldsymbol{x})}{\partial x^j} \cdot \boldsymbol{e}_k(\boldsymbol{x}) + \boldsymbol{e}^i(\boldsymbol{x}) \cdot \frac{\partial \boldsymbol{e}_k(\boldsymbol{x})}{\partial x^j} = 0.$$

Now from the definition,

$$\frac{\partial \boldsymbol{e}^i(\boldsymbol{x})}{\partial x^j} \cdot \boldsymbol{e}_k(\boldsymbol{x}) = -\boldsymbol{e}^i(\boldsymbol{x}) \cdot \left\{ \begin{array}{c} r \\ kj \end{array} \right\} \boldsymbol{e}_r(\boldsymbol{x}) = -\left\{ \begin{array}{c} r \\ kj \end{array} \right\} \delta^i_r = -\left\{ \begin{array}{c} i \\ kj \end{array} \right\}.$$

But also, using the above,

$$\frac{\partial \boldsymbol{e}^i(\boldsymbol{x})}{\partial x^j} = \frac{\partial \boldsymbol{e}^i(\boldsymbol{x})}{\partial x^j} \cdot \boldsymbol{e}_k(\boldsymbol{x}) \boldsymbol{e}^k(\boldsymbol{x}) = -\left\{ \begin{array}{c} i \\ kj \end{array} \right\} \boldsymbol{e}^k(\boldsymbol{x}).$$

This verifies 2.29. Formula 2.30 follows from 2.26 and equality of mixed partial derivatives.

It remains to show 2.31.

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{\partial \boldsymbol{e}_i}{\partial x^k} \cdot \boldsymbol{e}_j + \boldsymbol{e}_i \cdot \frac{\partial \boldsymbol{e}_j}{\partial x^k} = \left\{ \begin{array}{c} r \\ ik \end{array} \right\} \boldsymbol{e}_r \cdot \boldsymbol{e}_j + \boldsymbol{e}_i \cdot \boldsymbol{e}_r \left\{ \begin{array}{c} r \\ jk \end{array} \right\}.$$

Therefore,

$$\frac{\partial g_{ij}}{\partial x^k} = \left\{ \begin{array}{c} r \\ ik \end{array} \right\} g_{rj} + \left\{ \begin{array}{c} r \\ jk \end{array} \right\} g_{ri}. \tag{2.32}$$

Switching i and k while remembering 2.30 yields

$$\frac{\partial g_{kj}}{\partial x^i} = \left\{ \begin{array}{c} r \\ ik \end{array} \right\} g_{rj} + \left\{ \begin{array}{c} r \\ ji \end{array} \right\} g_{rk}. \tag{2.33}$$

Now switching j and k in 2.32,

$$\frac{\partial g_{ik}}{\partial x^j} = \left\{ \begin{array}{c} r \\ ij \end{array} \right\} g_{rk} + \left\{ \begin{array}{c} r \\ jk \end{array} \right\} g_{ri}. \tag{2.34}$$

Adding 2.32 to 2.33 and subtracting 2.34 yields

$$\frac{\partial g_{ij}}{\partial x^k}+\frac{\partial g_{kj}}{\partial x^i}-\frac{\partial g_{ik}}{\partial x^j}=2\left\{\begin{array}{c} r \\ ik \end{array}\right\}g_{rj}.$$

Now multiplying both sides by g^{jm} and using the fact shown earlier in Theorem B.1.6 that $g_{rj}g^{jm}=\delta_r^m$, it follows

$$2\left\{\begin{array}{c} m \\ ik \end{array}\right\}=g^{jm}\left(\frac{\partial g_{ij}}{\partial x^k}+\frac{\partial g_{kj}}{\partial x^i}-\frac{\partial g_{ik}}{\partial x^j}\right)$$

which proves 2.31. ■

This is a very interesting formula because it shows the Christoffel symbols are completely determined by the metric tensor and its partial derivatives which illustrates the fundamental nature of the metric tensor. Note that the inner product is determined by this metric tensor.

B.7 Gradients and Divergence

The purpose of this section is to express the gradient and the divergence of a vector field in general curvilinear coordinates. As before, $\left(y^1,...,y^n\right)$ will denote the standard coordinates with respect to the usual basis vectors. Thus

$$\boldsymbol{y}\equiv y^k\boldsymbol{i}_k,\ \boldsymbol{e}_k(\boldsymbol{y})=\boldsymbol{i}_k=\boldsymbol{e}^k(\boldsymbol{y}).$$

Let $\phi:U\to\mathbb{R}$ be a differentiable scalar function, sometimes called a “scalar field” in this subject. Write $\phi(\boldsymbol{x})$ to denote the value of ϕ at the point whose coordinates are $\boldsymbol{x}$**.** The same convention is used for a vector field. Thus $\boldsymbol{F}(\boldsymbol{x})$ is the value of a vector field at the point of U determined by the coordinates $\boldsymbol{x}$**.** In the standard rectangular coordinates, the gradient is well understood from earlier.

$$\nabla\phi(\boldsymbol{y})=\frac{\partial\phi(\boldsymbol{y})}{\partial y^k}\boldsymbol{e}^k(\boldsymbol{y})=\frac{\partial\phi(\boldsymbol{y})}{\partial y^k}\boldsymbol{i}^k.$$

However, the idea is to express the gradient in arbitrary coordinates. Therefore, using the chain rule, if the coordinates of the point of U are given as $\boldsymbol{x}$**,**

$$\nabla\phi(\boldsymbol{x})=\nabla\phi(\boldsymbol{y})=\frac{\partial\phi(\boldsymbol{x})}{\partial x^r}\frac{\partial x^r}{\partial y^k}\boldsymbol{e}^k(\boldsymbol{y})=$$

$$\frac{\partial\phi(\boldsymbol{x})}{\partial x^r}\frac{\partial x^r}{\partial y^k}\frac{\partial y^k}{\partial x^s}\boldsymbol{e}^s(\boldsymbol{x})=\frac{\partial\phi(\boldsymbol{x})}{\partial x^r}\delta_s^r\boldsymbol{e}^s(\boldsymbol{x})=\frac{\partial\phi(\boldsymbol{x})}{\partial x^r}\boldsymbol{e}^r(\boldsymbol{x}).$$

This shows the covariant components of $\nabla\phi(\boldsymbol{x})$ are

$$\left(\nabla\phi(\boldsymbol{x})\right)_r=\frac{\partial\phi(\boldsymbol{x})}{\partial x^r}, \tag{2.35}$$

Formally the same as in rectangular coordinates. To find the contravariant components, “raise the index” in the usual way. Thus

$$\left(\nabla\phi(\boldsymbol{x})\right)^r=g^{rk}(\boldsymbol{x})\left(\nabla\phi(\boldsymbol{x})\right)_k=g^{rk}(\boldsymbol{x})\frac{\partial\phi(\boldsymbol{x})}{\partial x^k}. \tag{2.36}$$

What about the divergence of a vector field? The divergence of a vector field $\boldsymbol{F}$ defined on U is a scalar field, $\operatorname{div}(\boldsymbol{F})$ which from calculus is

$$\frac{\partial F^k}{\partial y^k}(\boldsymbol{y}) = F^k_{,k}(\boldsymbol{y})$$

in terms of the usual rectangular coordinates $\boldsymbol{y}$**.** The reason the above equation holds in this case is that $\boldsymbol{e}_k(\boldsymbol{y})$ is a constant and so the Christoffel symbols are zero. We want an expression for the divergence in arbitrary coordinates. From Theorem B.6.1,

$$F^i_{,j}(\boldsymbol{y}) = F^r_{,s}(\boldsymbol{x}) \frac{\partial x^s}{\partial y^j}\frac{\partial y^i}{\partial x^r}$$

From 2.27,

$$= \left(\frac{\partial F^r(\boldsymbol{x})}{\partial x^s} + F^k(\boldsymbol{x})\left\{\begin{array}{c} r \\ ks \end{array}\right\}(\boldsymbol{x})\right)\frac{\partial x^s}{\partial y^j}\frac{\partial y^i}{\partial x^r}.$$

Letting $j = i$ yields

$$\begin{aligned}
\operatorname{div}(\boldsymbol{F}) &= \left(\frac{\partial F^r(\boldsymbol{x})}{\partial x^s} + F^k(\boldsymbol{x})\left\{\begin{array}{c} r \\ ks \end{array}\right\}(\boldsymbol{x})\right)\frac{\partial x^s}{\partial y^i}\frac{\partial y^i}{\partial x^r} \\
&= \left(\frac{\partial F^r(\boldsymbol{x})}{\partial x^s} + F^k(\boldsymbol{x})\left\{\begin{array}{c} r \\ ks \end{array}\right\}(\boldsymbol{x})\right)\delta^s_r \\
&= \left(\frac{\partial F^r(\boldsymbol{x})}{\partial x^r} + F^k(\boldsymbol{x})\left\{\begin{array}{c} r \\ kr \end{array}\right\}(\boldsymbol{x})\right). \qquad (2.37)
\end{aligned}$$

$\left\{\begin{array}{c} r \\ kr \end{array}\right\}$ is simplified using the description of it in Theorem B.6.2. Thus, from this theorem,

$$\left\{\begin{array}{c} r \\ rk \end{array}\right\} = \frac{g^{jr}}{2}\left[\frac{\partial g_{rj}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^r} - \frac{\partial g_{rk}}{\partial x^j}\right]$$

Now consider $\frac{g^{jr}}{2}$ times the last two terms in $[\cdot]$. Relabeling the indices r and j in the second term implies

$$\frac{g^{jr}}{2}\frac{\partial g_{kj}}{\partial x^r} - \frac{g^{jr}}{2}\frac{\partial g_{rk}}{\partial x^j} = \frac{g^{jr}}{2}\frac{\partial g_{kj}}{\partial x^r} - \frac{g^{rj}}{2}\frac{\partial g_{jk}}{\partial x^r} = 0.$$

Therefore,

$$\left\{\begin{array}{c} r \\ rk \end{array}\right\} = \frac{g^{jr}}{2}\frac{\partial g_{rj}}{\partial x^k}. \qquad (2.38)$$

Now recall $g \equiv \det(g_{ij}) = \det(G) > 0$ from Theorem B.1.6. Also from the formula for the inverse of a matrix and this theorem,

$$g^{jr} = A^{rj}(\det G)^{-1} = A^{jr}(\det G)^{-1}$$

where A^{rj} is the rj^{th} cofactor of the matrix (g_{ij}). Also recall that

$$g = \sum_{r=1}^{n} g_{rj}A^{rj} \text{ no sum on } j.$$

Therefore, g is a function of the variables $\left\{g_{rj}\right\}$ and $\frac{\partial g}{\partial g_{rj}} = A^{rj}$. From 2.38,

$$\left\{ \begin{array}{c} r \\ rk \end{array} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g_{rj}}{\partial x^k} A^{jr} = \frac{1}{2g} \frac{\partial g}{\partial g_{rj}} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g}{\partial x^k}$$

and so from 2.37,

$$\operatorname{div}(\boldsymbol{F}) = \frac{\partial F^k(\boldsymbol{x})}{\partial x^k} +$$

$$+F^k(\boldsymbol{x}) \frac{1}{2g(\boldsymbol{x})} \frac{\partial g(\boldsymbol{x})}{\partial x^k} = \frac{1}{\sqrt{g(\boldsymbol{x})}} \frac{\partial}{\partial x^i} \left(F^i(\boldsymbol{x}) \sqrt{g(\boldsymbol{x})} \right). \tag{2.39}$$

This is the formula for the divergence of a vector field in general curvilinear coordinates. Note that it uses the contravariant components of $\boldsymbol{F}$.

The Laplacian of a scalar field is nothing more than the divergence of the gradient. In symbols, $\Delta\phi \equiv \nabla \cdot \nabla\phi$. From 2.39 and 2.36 it follows

$$\Delta\phi(\boldsymbol{x}) = \frac{1}{\sqrt{g(\boldsymbol{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\boldsymbol{x}) \frac{\partial \phi(\boldsymbol{x})}{\partial x^k} \sqrt{g(\boldsymbol{x})} \right). \tag{2.40}$$

We summarize the conclusions of this section in the following theorem.

Theorem B.7.1 *The following hold for gradient, divergence, and Laplacian in general curvilinear coordinates.*

$$(\nabla\phi(\boldsymbol{x}))_r = \frac{\partial \phi(\boldsymbol{x})}{\partial x^r}, \tag{2.41}$$

$$(\nabla\phi(\boldsymbol{x}))^r = g^{rk}(\boldsymbol{x}) \frac{\partial \phi(\boldsymbol{x})}{\partial x^k}, \tag{2.42}$$

$$\operatorname{div}(\boldsymbol{F}) = \frac{1}{\sqrt{g(\boldsymbol{x})}} \frac{\partial}{\partial x^i} \left(F^i(\boldsymbol{x}) \sqrt{g(\boldsymbol{x})} \right), \tag{2.43}$$

$$\Delta\phi(\boldsymbol{x}) = \frac{1}{\sqrt{g(\boldsymbol{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\boldsymbol{x}) \frac{\partial \phi(\boldsymbol{x})}{\partial x^k} \sqrt{g(\boldsymbol{x})} \right). \tag{2.44}$$

Example B.7.2 *Define curvilinear coordinates as follows*

$$x = r\cos\theta, y = r\sin\theta$$

Find $\nabla^2 f(r,\theta)$. That is, find the Laplacian in terms of these new variables r, θ.

First find the metric tensor. From the definition, this is

$$G = \left(\begin{array}{cc} 1 & 0 \\ 0 & r^2 \end{array} \right), G^{-1} = \left(\begin{array}{cc} 1 & 0 \\ 0 & r^{-2} \end{array} \right)$$

The contravariant components of the gradient are

$$\left(\begin{array}{cc} 1 & 0 \\ 0 & r^{-2} \end{array} \right) \left(\begin{array}{c} f_r \\ f_\theta \end{array} \right) = \left(\begin{array}{c} f_r \\ \frac{1}{r^2} f_\theta \end{array} \right)$$

Then also $\sqrt{g} = r$. Therefore, using the formula,

$$\nabla^2 f(u,v) = \frac{1}{r} \left[(rf_r)_r + \left(r \frac{1}{r^2} f_\theta \right)_\theta \right] = \frac{1}{r}(rf_r)_r + \frac{1}{r^2} f_{\theta\theta}$$

Notice how easy this is. It is anything but easy if you try to do it by brute force with none of the machinery developed here.

B.8 Exercises

1. Let $y^1 = x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let
$$\boldsymbol{F}(\boldsymbol{x}) = x^1\boldsymbol{e}_1(\boldsymbol{x}) + x^2\boldsymbol{e}_2(\boldsymbol{x}) + \left(x^3\right)^2\boldsymbol{e}(\boldsymbol{x}).$$
Find $\operatorname{div}(\boldsymbol{F})(\boldsymbol{x})$.

2. For the coordinates of the preceding problem, and ϕ a scalar field, find
$$(\nabla\phi(\boldsymbol{x}))^3$$
in terms of the partial derivatives of ϕ taken with respect to the variables x^i.

3. Let $y^1 = 7x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let ϕ be a scalar field. Find $\nabla^2\phi(\boldsymbol{x})$.

4. Derive $\nabla^2 u$ in cylindrical coordinates, r, θ, z, where u is a scalar field on $\mathbb{R}^3$.
$$x = r\cos\theta,\ y = r\sin\theta,\ z = z.$$

5. ↑ Find all solutions to $\nabla^2 u = 0$ which depend only on r where $r \equiv \sqrt{x^2 + y^2}$.

6. Derive $\nabla^2 u$ in spherical coordinates.

7. ↑Let u be a scalar field on $\mathbb{R}^3$. Find all solutions to $\nabla^2 u = 0$ which depend only on
$$\rho \equiv \sqrt{x^2 + y^2 + z^2}.$$

8. The temperature, u, in a solid satisfies $\nabla^2 u = 0$ after a long time. Suppose in a long pipe of inner radius 9 and outer radius 10 the exterior surface is held at 100° while the inner surface is held at 200° find the temperature in the solid part of the pipe.

9. Show velocity can be expressed as $\boldsymbol{v} = v_i(\boldsymbol{x})\boldsymbol{e}^i(\boldsymbol{x})$, where
$$v_i(\boldsymbol{x}) = \frac{\partial r_i}{\partial x^j}\frac{dx^j}{dt} - r_p(\boldsymbol{x})\left\{\begin{array}{c} p \\ ik \end{array}\right\}\frac{dx^k}{dt}$$
and $r_i(\boldsymbol{x})$ are the covariant components of the displacement vector,
$$\boldsymbol{r} = r_i(\boldsymbol{x})\boldsymbol{e}^i(\boldsymbol{x}).$$

10. Find the covariant components of velocity in spherical coordinates. **Hint:** $\boldsymbol{v} = \frac{d\boldsymbol{y}}{dt}$. Now use chain rule and identify the contravariant components. Then use the technique of lowering or raising index.

11. Show that $\boldsymbol{v}\cdot\boldsymbol{w} = g_{ij}(\boldsymbol{x})v^i(\boldsymbol{x})v^j(\boldsymbol{x}) = g^{ij}(\boldsymbol{x})v_i(\boldsymbol{x})v_j(\boldsymbol{x})$.

Bibliography

[1] **Apostol T. M.,** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Apostol, T. M.,** *Mathematical Analysis,* Addison Wesley Publishing Co., 1974.

[3] **Ash, Robert,** *Complex Variables*, Academic Press, 1971.

[4] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[5] **Balakrishnan A.V.,** *Applied Functional Analysis, Springer Verlag 1976.*

[6] **Brézis, H.** *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert,* Math Studies, 5, North Holland, 1973.

[7] **Buck, R. C.** *Advanced Calculus* 2 edition. McGraw-Hill, 1965.

[8] **Cheney, E. W.,** *Introduction To Approximation Theory,* McGraw Hill 1966.

[9] **Chow S.N. and Hale J.K.,** *Methods of bifurcation Theory ,* Springer Verlag, New York 1982.

[10] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[11] **Deimling K.** *Nonlinear Functional Analysis,* Springer-Verlag, 1985.

[12] **Dontchev A.L.** The Graves theorem Revisited, *Journal of Convex Analysis,* Vol. 3, 1996, No.1, 45-53.

[13] **Donal O'Regan, Yeol Je Cho, and Yu-Qing Chen,** Topological Degree Theory and Applications, Chapman and Hall/CRC 2006.

[14] **Dunford N.** and **Schwartz J.T.** *Linear Operators,* Interscience Publishers, a division of John Wiley and Sons, New York, part 1 1958, part 2 1963, part 3 1971.

[15] **Evans L.C. and Gariepy,** *Measure Theory and Fine Properties of Functions,* CRC Press, 1992.

[16] **Evans L.C.** *Partial Differential Equations,* Berkeley Mathematics Lecture Notes. 1993.

[17] **Fitzpatrick P. M.,** *Advanced Calculus a course in Mathematical Analysis,* PWS Publishing Company 1996.

[18] **Fonesca I. and Gangbo W.** *Degree theory in analysis and applications* Clarendon Press 1995.

[19] **Gasinski L. and Papageorgiou N.,** *Nonlinear Analysis,* Volume 9, Chapman and Hall, 2006.

[20] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[21] **Gromes W.** Ein einfacher Beweis des Satzes von Borsuk. *Math. Z.* 178, pp. 399 -400 (1981).

[22] **Hardy, G.H., Littlewood, J.E. and Polya, G.,** Inequalities, Cambridge University Press 1964.

[23] **Hewitt E.** and **Stromberg K.** *Real and Abstract Analysis,* Springer-Verlag, New York, 1965.

[24] **Heinz, E.**An elementary analytic theory of the degree of mapping in n dimensional space. *J. Math. Mech. 8, 231-247* 1959

[25] **Hobson E.W.,** The Theory of functions of a Real Variable and the Theory of Fourier's Series V. 1, Dover 1957.

[26] **Hocking J. and Young G.,** Topology, Addison-Wesley Series in Mathematics, 1961.

[27] **Horn R. and Johnson C.,** *matrix Analysis,* Cambridge University Press, 1985.

[28] **Kato T.** Perturbation Theory for Linear Operators, Springer, 1966.

[29] **Kreyszig E.** *Introductory Functional Analysis With applications,* Wiley 1978.

[30] **Kuratowski K.** and **Ryll-Nardzewski C.** A general theorem on selectors, *Bull. Acad. Pol. Sc.,* **13,** 397-403.

[31] **Kuttler K. L.**, *Linear Algebra and Analysis*, web page Web Page

[32] **Marsden J. E. and Hoffman J. M.,** *Elementary Classical Analysis,* Freeman, 1993.

[33] **McShane E. J.** *Integration,* Princeton University Press, Princeton, N.J. 1944.

[34] **Munkres, James R.,***Topology A First Course*, Prentice Hall, Englewood Cliffs, New Jersey 1975

[35] **Natanson I. P.,** *Theory Of Functions Of A Real Variable,* Fredrick Ungar Publishing Co. 1955.

[36] **Naylor A. and Sell R.,** *Linear Operator Theory in Engineering and Science*, Holt Rinehart and Winston, 1971.

[37] **Ray W.O.** *Real Analysis,* Prentice-Hall, 1988.

[38] **Rudin W.,** *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[39] **Rudin W.** *Real and Complex Analysis,* third edition, McGraw-Hill, 1987.

[40] **Rudin W.** *Functional Analysis,* second edition, McGraw-Hill, 1991.

[41] **Spivak M.**, *Calculus On Manifolds*, Benjamin 1965.

[42] **Taylor A. E.** *General Theory of Functions and Integration,* Blaisdell Publishing, 1965.

[43] **Widder, D**. *Advanced Calculus,* second edition, Prentice Hall 1961.

[44] **Yosida K.** *Functional Analysis,* Springer-Verlag, New York, 1978.

Index