University of Bristol
Bristol BS8 1UB
http://www.bristol.ac.uk/

Mathematical Basics

Ian T Nabney

Module Background November 29, 2019

1 Introduction

The purpose of this short section of lecture notes is to remind you of (hopefully, rather than introduce you to) the basic mathematical knowledge in real analysis, linear algebra, and multi-dimensional calculus you will need for the Data Science programme. The big area that is not covered is Probability and Statistics, but that will be addressed in some of the units during the programme.

2 Real Analysis

In this section we review, without proof, the main theorems in real analysis. This is concerned with the properties of 'smooth' functions on the real line.

2.1 Differentiation

The least amount of smoothness that a function can have is to be continuous:

Definition 2.1
A function f is continuous *at a if*

$$\lim_{x \to a} f(x) = f(a).$$

A function is continuous if the graph has no breaks, jumps or 'wild oscillations'. Continuous functions have many nice properties, but to make further progress we need stronger constraints on the smoothness of functions. To do this, we consider tangents to the graph of the function.

Definition 2.2
A function f is said to be differentiable *at a if*

$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists. We write $f'(a)$ for the derivative of f at a. A function is said to be differentiable if its derivative exists for all $a \in \mathbb{R}$.

The derivative can be interpreted as the gradient of a tangent to the curve at a. Here the tangent is defined as the limit of the chords $[a, a+h]$ as $h \to 0$.

We now list some of the simple properties of the derivative:

Theorem 2.3
If f and g are functions which are differentiable at a, then

1. *For $\lambda \in \mathbb{R}$ $\quad (f + \lambda g)'(a) = f'(a) + \lambda g'(a)$. That is, differentiation is a linear operator.*
2. *The product rule: $(f \cdot g)'(a) = f'(a)g(a) + f(a)g'(a)$.*
3. *The quotient rule. If $g'(a) \neq 0$ then*
$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{(g(a))^2}.$$

These properties allow us to differentiate many simple functions. However, to differentiate more interesting (or at any rate complicated) functions, we need to know how to calculate $(f \circ g)'$), that is the composition of f and g in terms of f' and g'. We can do this with the aid of the following theorem:

Theorem 2.4 (Chain Rule)
If g is differentiable at a and f is differentiable at $g(a)$ then $f \circ g$ is differentiable at a, and
$$(f \circ g)'(a) = f'(g(a)) \cdot g'(a).$$

This theorem will be generalised later to functions defined on higher dimensional spaces.

With the aid of these theorems, we can write down the derivatives of several functions.

Example 2.5
1. If $f(x) = c$, a constant, then $f'(x) = 0 \qquad \forall x \in \mathbb{R}$.
2. If $f(x) = x^n$, where $n \in \mathbb{Z}$, then $f'(x) = nx^{n-1}$.
3. If $f(x) = \sin(x)$, then $f'(x) = \cos(x)$. If $f(x) = \cos(x)$, then $f'(x) = -\sin(x)$.
4. If $f(x) = \log(x)$, then $f'(x) = 1/x$. If $f(x) = e^x$, then $f'(x) = e^x$.

2.2 Critical points

Although we have commented on how the derivative can be interpreted as the gradient of a tangent to the curve, it has much greater significance than that. For example, qualitative information about the graph of a function can be obtained by considering its derivative.

Definition 2.6
Let f be a function defined on an interval I. A point x in I is a local maximum [minimum] *for f in I if there is some $\delta > 0$ such that x is a maximum [minimum] for f on $I \cap (x - \delta, x + \delta)$.*

That is, x is a local maximum for f if the value of $f(x)$ is no less than the value of f in the whole of some (possibly small) interval around x. Note that we will be able to say quite a lot about local maxima and minima, but rather less about global 'optima'. This is because, unless there are fairly rigorous restrictions on the type of function that f is, its local behaviour puts no constraints on other parts of the function domain.

A key result for the practical use of derivatives is the following:

Theorem 2.7
If f is defined on (a, b) and f has a local maximum and is differentiable at x, then $f'(x) = 0$. The same is true if f has a local minimum at x.

Note carefully the direction of implications in this theorem. It is not necessarily the case that $f'(x) = 0$ implies that f has a local maximum or minimum at x. However, the points where $f'(x) = 0$ are important enough to warrant a special name. They are called the *critical points* of f.

There is one simple way in which we can determine the nature of critical points, and that is to consider the second derivative.

Theorem 2.8
Suppose that $f'(a) = 0$. If $f''(a) > 0$, then f has a local minimum at a. If $f''(a) < 0$, then f has a local maximum at a.

Later we will generalise this theorem to higher dimensional spaces, where the situation is complicated by geometrical considerations. Fortunately, the geometry of a line is not very interesting.

A useful technique for finding limits is l'Hôpital's rule.

Theorem 2.9
If $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = 0$ and

$$\lim_{x\to a} \frac{f'(x)}{g'(x)}$$

exists, then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Example 2.10
Work out $\lim_{r\to 0} r^2 \log r$.

First write the fraction in a form to which we can apply l'Hôpital's rule:

$$r^2 \log r = \frac{r^2}{1/\log r}.$$

Then we need to compute the derivative of $1/\log r$, which is $-1/r(\log r)^2$, and the derivative of r^2, which is $2r$. Now we can compute the required limit:

$$\lim_{r\to 0} \frac{r^2}{1/\log r} = \lim_{r\to 0} \frac{2r}{-1/r(\log r)^2} = -\lim_{r\to 0} 2(r\log r)^2 = 0.$$

2.3 Integration

The definition of integration is concerned with finding the area of a certain region: the set bounded by the x-axis, the graph of a function $f(x)$ and the vertical lines through $(a, 0)$ and $(b, 0)$.

We start by defining a partition of the interval $[a, b]$.

Definition 2.11
Let $a < b$. A partition of the interval $[a, b]$ is a finite collection of points in $[a, b]$ numbered in order of increasing magnitude such that the first point is a and the last point is b.

We can then define two sums such that the value of the integral lies between them.

Definition 2.12
Suppose that f is bounded on $[a, b]$ and $P = \{t_0, t_1, \ldots, t_n\}$ is a partition of $[a, b]$. Let

$$m_i = \inf\{f(x) : t_{i-1} \leq x \leq t_i\},$$
$$M_i = \sup\{f(x) : t_{i-1} \leq x \leq t_i\}.$$

The lower sum *of f for P is defined as*

$$L(f, P) = \sum_{i=1}^{n} m_i(t_i - t_{i-1}),$$

and the upper sum *of f for P is defined as*

$$U(f, P) = \sum_{i=1}^{n} M_i(t_i - t_{i-1}).$$

A bounded function f is integrable on $[a, b]$ if the lower sum and upper sum tend to a common value as the number of points in the partition tends to infinity. It is possible to make this more precise, but we have no space to do so here.

Although this defines the integral in a form that corresponds reasonably well with intuition, it leaves us with a problem. We do not know which functions are integrable, nor how to integrate those that are. Although progress can be made from the first principles definition, it is very tedious. Fortunately, we have some extremely powerful theorems that enable us to integrate a large number of functions that we meet. Even from our (incomplete) definition, it is clear that integration is a linear operator. That is:

$$\int_a^b f + \lambda g = \int f + \lambda \int g.$$

It is also reasonably easy to prove that the integral of a constant function is the product of the width of the interval $[a, b]$ and the value of the function.

To compute integrals, Newton's (or Leibniz's) theorem linking integration and differentiation is essential:

Theorem 2.13 (Fundamental Theorem of Calculus)

1. *If f is integrable and continuous on $[a, b]$ and*

$$F(x) = \int_a^x f$$

   then F is differentiable and $F'(c) = f(c)$.

2. *If f is integrable on $[a, b]$ and $f = g'$, then*

$$\int_a^b f = g(b) - g(a).$$

Two important techniques for more complicated integrals are integration by parts and substitution.

Theorem 2.14
Integration by parts *If f' and g' are continuous, then $\int fg' = fg - \int f'g$. More explicitly:*

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x.$$

Substitution *If f and g' are continuous, then*

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g)g'.$$

2.4 Taylor Series and Power Series

Many important functions are defined by integral equations. This is fine as a definition, but very difficult to compute with. For example, consider how you might find $\log(3.5)$ given that

$$\log(x) := \int_1^x \frac{1}{t}\,\mathrm{d}t.$$

The position for $\exp(x) = \log^{-1}(x)$ is even worse. The answer is to use a Taylor series expansion.

Define

$$a_k = \frac{f^k(a)}{k!}$$

and put

$$P_{n,a}(x) = a_0 + a_1(x-a) + \cdots + a_n(x-a)^n.$$

This is called the Taylor polynomial of degree n. Note that f and $P_{n,a}$ have the same first n derivatives at the point a.

For example, the first order Taylor polynomial is

$$f(x) = f(a) + (x-a)f'(a)$$

and the second order Taylor polynomial is

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(a).$$

Then the following two results are true:

1. $$\lim_{x\to a} \frac{f(x) - P_{n,a}(x)}{(x-a)^n} = 0.$$

   So we can write $f(x) = P_{n,a}(x) + R_{n,a}(x)$.

2. The remainder can be written in many forms. Two useful ones are

$$R_{n,a}(x) = \frac{f^{(n+1)}(t)(x-a)^{n+1}}{(n+1)!} \tag{2.1}$$

for some $t \in (a, x)$ or, if $f^{(n+1)}$ is integrable on $[a, x]$,

$$R_{n,a}(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n \, \mathrm{d}t \tag{2.2}$$

We can estimate this integral to give bounds on the error by taking the first n terms of the sum. We often expand 'smooth' functions as Taylor series and ignore 'higher order' terms.

If the function is very smooth, we get a *power series*. For example:

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\ e^x &= 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots \\ \log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad -1 < x \leq 1. \end{aligned}$$

3 Proof by Induction

This is an extremely useful method for proving results that breaks down a very difficult task into two much simpler ones.

If we want to prove that some statement $P(n)$ is true for all natural numbers n, then it is sufficient to prove that:

1. $P(0)$ is true;
2. *if* $P(k)$ is true, *then* $P(k+1)$ is also true.

Example 3.1
Let $P(n)$ be the statement

$$\sum_{i=0}^{n} i = \frac{n(n+1)}{2}.$$

Prove that this is true for $n \in \mathbb{N}$.

The proof breaks down into two parts:

1. Consider $P(0)$. The left-hand side (lhs) of the identity is equal to $\sum_{i=0}^{0} i = 0$ while the right-hand side (rhs) is $0(1)/2 = 0$. Hence $P(0)$ is true.
2. Suppose that $P(k)$ is true (this is known as the *inductive hypothesis*). This means that

$$\sum_{i=0}^{k} i = \frac{k(k+1)}{2}.$$

Now let us work out the lhs of the identity for $k+1$:

$$\begin{aligned}\sum_{i=0}^{k+1} i &= \sum_{i=0}^{k} +(k+1) \\ &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{k^2+k+2k+2}{2} \\ &= \frac{(k+1)(k+2)}{2}.\end{aligned}$$

4 Linear Algebra

4.1 Basic Definitions

Linear algebra is the study of linear systems. The basic space we work with is called a *vector space*. Vector spaces can be defined over any field (a type of mathematical algebraic structure), but for our purposes we will assume that they are defined over the reals $\mathbb{R}$ or complex numbers $\mathbb{C}$.

A vector space V consists of a set of vectors and scalars K (where K is $\mathbb{R}$ or $\mathbb{C}$) with the following properties.

1. Addition of vectors is well-defined. If $v, w \in V$ then $v + w \in V$, etc.
2. Scalar multiplication is well-defined. If $v \in V$ and $\lambda \in K$, then $\lambda v \in V$.
3. There is linear structure. If $v, w \in V$ and $\lambda \in K$, then

$$\lambda(v + w) = \lambda v + \lambda w.$$

There are many examples of vector spaces ("I've been speaking prose all my life"[1]) including standard Euclidean spaces (with the usual coordinate systems), functions on an interval, polynomials etc.

The advantage of such an abstract approach is that many results can be applied more widely than you might think. There is an approach to statistics called *information geometry* that starts with the vector space view of probability density functions.

A *vector subspace* $W \subset V$ is a subset that is also a vector space. This means that it is linear and contains a zero vector. Examples are straight lines, planes, or the polynomials inside function space.

A set S of vectors is said to be *linearly dependent* if it contains a finite subset $v_1, \ldots, v_n$ and scalars $\lambda_1, \ldots \lambda_n$, not all equal to zero, such that

$$\sum_{i=1}^{n} \lambda_i v_i = 0.$$

Of course, a set is *linearly independent* if it is not linearly dependent.

[1] "Il y a plus de quarante ans que je dis de la prose sans que j'en susse rien."

A set S of vectors is said to *span* a vector space V if every vector $v \in V$ can be written as a linear combination of vectors in S:

$$v \in V \Rightarrow \exists w_1, \ldots w_n \in S, \lambda_1, \ldots \lambda_n \in K \quad \text{such that} \quad v = \sum_{i=1}^{n} \lambda_i w_i.$$

A set S of vectors is a *basis* for a vector space V if it is a linearly independent spanning set. If a basis is finite, then all bases are finite and all have the same size, which is known as the *dimension* of the vector space. If a vector space has a basis then every vector can be written as a unique linear combination of basis vectors (i.e. we have a proper coordinate system).

Some examples of bases:

1. The *standard basis* $e_1, \ldots, e_n$ for $\mathbb{R}^n$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the ith coordinate.
2. The set $\{e_1, e_2 + e_1, e_3 + e_1, \ldots, e_n + e_1\}$ is also a basis for $\mathbb{R}^n$.
3. The set of powers of x, $1, x, x^2, \ldots$ forms a basis for the vector space of polynomials $\mathbb{R}[x]$. Note that a vector space need not be finite dimensional.

4.2 Linear Transformations

Vector spaces by themselves can't do much; we need to consider how they can be mapped. A mapping $A : V \to W$ is said to be a *linear transformation* if it preserves linearity:

$$A(\lambda v + w) = \lambda A v + A w \qquad \forall v, w \in V.$$

This means that linear subspaces are mapped to linear subspaces. Another consequence is that the origin (the zero vector) is mapped to the origin.

For a finite dimensional vector space, every linear transformation can be written as a matrix, once a basis is chosen. If the standard basis is used, then the jth column of the matrix consists of the coordinates of Ae_j. The entries in the matrix depend on the basis. But we can also define linear transformations on infinite dimensional vector spaces: differentiation and integration are two examples.

The *kernel* of a linear transformation A is the set of vectors that it maps to zero. It can be shown that this is a subspace; the dimension of this subspace is called the *nullity* of A. The kernel of the derivative map is the set of constant functions.

The *range* of a linear transformation A is the image AV. This is also a linear subspace; its dimension is the *rank* of A. The range of the derivative map in the vector space V of polynomials is the whole of V.

There is a useful relationship between the rank and nullity of a linear transformation $A : V \to W$. If W has finite dimension n, then the sum of the rank and nullity is equal to n: $r(A) + n(A) = n$.

A linear transformation $A : V \to W$ is *invertible* if its range is equal to W and its kernel is just the zero vector. It is not a simple matter to compute the inverse of a linear transformation in matrix

form, and it should be avoided if at all possible! The Fundamental Theorem of Calculus essentially says that differentiation is the inverse linear operator to integration.

Mapping from one basis to another is always an invertible linear transformation (and vice versa).

Theorem 4.1 (Change of Basis for Coordinates)
Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ be two (ordered) bases for V, and suppose that B is the linear transformation that maps $B(x_i) = y_i$ with matrix (β_{ij}). If

$$v = \sum_{i=1}^{n} \xi_i x_i = \sum_{i=1}^{n} \eta_i y_i,$$

then we know that

$$\xi_i = \sum_{j=1}^{n} \beta_{ij} \eta_j.$$

Hence the coordinates of v with respect to $(y_1, \ldots, y_n)$ are given by $B^{-1} v_x$.

We said earlier that the matrix of a linear transformation depends on the chosen basis. We can now write down how to convert the matrix from one basis to another.

Theorem 4.2 (Change of Basis for Linear Transformation)
Let $A : V \to V$ be a linear transformation. Suppose that $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are (ordered) bases for V, and that B is the (invertible) linear transformation that maps $B(x_i) = y_i$. Let the matrix of A with respect to $(x_1, \ldots, x_n)$ be $\alpha = (\alpha_{ij})$, so that

$$Ax_i = \sum_{j=1}^{n} \alpha_{ji} x_j.$$

Let the matrix of B with respect to $(x_1, \ldots, x_n)$ be $\beta = (\beta_{ij})$. Then the matrix of A with respect to $(y_1, \ldots, y_n)$ is given by $\beta^{-1}\alpha\beta$. This map is called a similarity.

4.3 Solution of Linear Equations

Forget whatever else you have been taught about solving simultaneous linear equations (determinants, inverses, Cramer's rule). Apart from theoretical proofs, by far the best way of solving such systems is to use row reduction to echelon form. This is best illustrated by an example.

Let us solve the equations

$$\begin{bmatrix} 2 & 1 & -2 \\ 2 & -3 & 2 \\ -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 9 \\ -3.5 \end{bmatrix} \tag{4.1}$$

We work with the augmented matrix

$$\begin{bmatrix} 2 & 1 & -2 & -1 \\ 2 & -3 & 2 & 9 \\ -1 & 1 & -1 & 3.5 \end{bmatrix} \tag{4.2}$$

Stage 1 We use the first row to eliminate the elements in the first column below a_{11}.

1. We divide the first row by a_{11}. Note that if a_{11} were zero, we would swap row 1 with a row without a zero entry in the first column. This gives

$$\begin{bmatrix} 1 & 0.5 & -1 & -0.5 \\ 2 & -3 & 2 & 9 \\ -1 & 1 & -1 & -3.5 \end{bmatrix}$$

2. Take Row $2 - 2 \times$ (Row 1) and Row $3 +$ Row 1.

$$\begin{bmatrix} 1 & 0.5 & -1 & -0.5 \\ 0 & -4 & 0 & 10 \\ 0 & 1.5 & -2 & -4 \end{bmatrix}$$

Stage 2 In stage 2 we work on the second column. We can't use the first row to zero entries in the second column because that would mess up the first column. Instead we use the second row.

1. We divide the second row by a_{22}. Again, if this entry were zero, we would swap row 2 with a later row which had a non-zero entry in the second column.

$$\begin{bmatrix} 1 & 0.5 & -1 & -0.5 \\ 0 & 1 & -1 & -2.5 \\ 0 & 1.5 & -2 & -4 \end{bmatrix}$$

2. Eliminate the entries below a_{22} in the second column. Row $3 - 1.5 \times$ (Row 2).

$$\begin{bmatrix} 1 & 0.5 & -1 & -0.5 \\ 0 & 1 & -1 & -2.5 \\ 0 & 0 & -0.5 & -0.25 \end{bmatrix}$$

Stage 3 Divide the third row by a_{32}.

$$\begin{bmatrix} 1 & 0.5 & -1 & -0.5 \\ 0 & 1 & -1 & -2.5 \\ 0 & 0 & 1 & 0.5 \end{bmatrix}$$

The matrix is now in echelon form.

Stage 4 We can now solve the equations easily by back-substitution.

1. From the third row, we see that $z = 0.5$.
2. Now use the second row to find y:

$$y - z = -2.5 \Longrightarrow y = -2.5 + 0.5 = -2.$$

Note how we *substitute* the value of z into the equation to solve it.

3. Now use the first row to find x:

$$x + 0.5 \times y - z = -0.5 \Longrightarrow x = -0.5 + 1 + 0.5 = 1.$$

Check We can check the solution by substituting the vector $[1, -2, 0.5]$ back into the original equations.

4.4 Determinant

Determinants are calculated by hand using the matrix of cofactors approach that you may have seen before. We won't need to do this (not, at least for general matrices), and the right way to do it on computer is quite different (as you will see later). We will actually need determinants for change of variables in integration, not solving linear equations, but it is still a useful quantity to work with.

The formal definition is that the determinant of a linear transformation A on a finite dimensional vector space with matrix (a_{ij}) is given by

$$\det A = \sum_{\sigma \in S_n} (-1)^{(sgn(\sigma))} a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

It takes a little bit of work to show that this is a property that is inherent in the linear transformation and not just in its matrix. This involves showing that the definition is invariant under a change of basis. Unlikely as it might seem, this is rather easier for the horrible looking definition here than the usual one.

The important properties of the determinant are:

1. $\det AB = \det A \det B$. (This is relatively easy to prove from the definition, and gives the invariance property straightaway.)
2. $\det A^{-1} = 1/\det(A)$, so $\det A \neq 0$ if and only if A is invertible. A matrix with zero determinant is said to be *singular*.

4.5 Eigenvalues and Eigenvectors

A scalar λ is an *eigenvalue* and a non-zero vector v is an *eigenvector* of a linear transformation A if

$$Av = \lambda v.$$

The set of eigenvectors V_λ together with the zero vector is called the *eigenspace* for λ and is a vector subspace. An eigenvalue can have multiple linearly independent eigenvectors. For example, for the identity matrix I_n, every non-zero vector is an eigenvector with eigenvalue 1. In function spaces, we can have eigenvectors as well. For example, for the linear operator of differentiation, the function $f(x) = e^{\lambda x}$ is an eigenvector with eigenvalue λ since $f'(x) = \lambda e^{\lambda x} = \lambda f(x)$.

Eigenspaces are invariant under similarity (i.e. a change of basis) so they are a *geometric* property of the linear transformation. The dimension of the corresponding eigenspace is called the *geometric multiplicity* of the eigenvalue. We note that if $p(x)$ is a polynomial, then

$$p(A)v = p(\lambda)v,$$

when v is an eigenvector of A with eigenvalue λ.

This is all very well, but we need to be able to calculate the eigensystem for a linear transformation. We can do this by noting that λ is an eigenvalue of A if and only if $A - \lambda I$ is singular, which occurs if and only if $\det(A - \lambda I) = 0$. From this we derive the *characteristic equation*

$$c(t) := \det(A - tI) = 0,$$

which is a polynomial of degree $n = \dim V$. The solutions to this equation are precisely the eigenvalues of A. Note that these may be complex, even if A is defined on a real vector space. Over $\mathbb{C}$, we can write the characteristic polynomial in the form

$$\prod_{j=1}^{l}(t-\lambda_j)^{m_j},$$

where the powers m_j are the *algebraic multiplicities* of the corresponding eigenvalues. The Cayley-Hamilton theorem states that $c(A) = 0$.

The ideal form for a linear transformation is to be diagonal. This is equivalent to finding a basis of V that consists entirely of eigenvectors of A. Since the geometric multiplicity is no greater than the algebraic multiplicity, and $\sum_{j=1}^{l} m_j = n$ it follows that this can happen if and only if the algebraic and geometric multiplicities are the same for all eigenvalues.

Unfortunately, we cannot always achieve the ideal: consider the matrix

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

4.6 Inner Products

We have come a long way with describing linear structure, but have not yet introduced the concepts of distance and angle. That is precisely what an inner product allows us to define.

An *inner product* is a map $(,) : V \times V \to K$ with the following properties:

1. It is *bilinear:*

$$(u, \lambda v + w) = (u, v) + \overline{\lambda}(u, w)$$
$$(u + \lambda v, w) = (u, w) + \lambda(v, w).$$

2. It is positive definite:

$$(v, v) \geq 0$$

   with equality iff $v = 0$.

3. It is conjugate-symmetric:

$$(u, v) = \overline{(v, u)}.$$

The *norm* defined by the inner product is given by

$$\|v\| = \sqrt{(v, v)}.$$

The norm defines a distance measure on the vector space, and the inner product defines an angle by

$$\cos(\angle(v, w)) = \frac{(v, w)}{\|v\|\|w\|}.$$

An inner product space over $\mathbb{R}$ is said to be *Euclidean*, and one over $\mathbb{C}$ is said to be *unitary*.

Some examples:

1. The usual dot product in $\mathbb{R}^n$ is an inner product:
$$(u, v) = \sum_{i=1}^{n} u_i \overline{v_i}.$$
2. Integration defines an inner product on the vector space of real-valued functions:
$$(f, g) = \int_0^1 f(t)\overline{g(t)}\, dt.$$

Note that we say '*an* inner product': many different inner products can be defined on the same vector space.

Two useful Euclidean inequalities are true more generally:

Schwarz inequality
$$|(u, v)| \leq \|u\|\|v\|.$$

Triangle inequality
$$\|u + v\| \leq \|u\| + \|v\|.$$

4.7 Inner Products and Geometry

What are the geometrical implications of defining an inner product?

A pair of vectors u, v are *orthogonal* if their inner product is zero. A set of vectors is orthogonal if every distinct pair of members is orthogonal. u and v are *orthonormal* if they are orthogonal and norm 1 (i.e. $\|u\| = \|v\| = 1$). It is fairly easy to show that an orthogonal set of vectors (that doesn't contain the zero vector) must be linearly independent.

An orthonormal basis is a basis that has a nice additional property related to the geometry of the inner product. It turns out that it is easy to generate orthonormal bases.

Theorem 4.3 (Gram-Schmidt Process)
If S is a set of orthonormal vectors in a finite dimensional vector space V, then S can always be extended to an orthonormal basis of V.

Proof: We will show how to extend S by a single vector, and then the result is proved by induction.

If S is empty, we just need to pick any non-zero $v \in V$ and then set $u = v/\|v\|$, which certainly has norm 1 and forms an orthonormal set with one element.

Suppose that $S = \{u_1, \ldots, u_m\}$ is an orthonormal set of vectors. If it isn't already a basis for V, then we can find a vector $v \in V$ not in the span of S. Now consider
$$w = v - \sum_{i=1}^{m} (v, u_i) u_i.$$
It is easy to show that $(w, u_i) = 0$ for $i = 1, \ldots, m$ (i.e. that w is orthogonal to all the elements of S). It is also clear that $w \neq 0$, since otherwise v would be in the span of S. Hence if we set $u_{m+1} = w/\|w\|$, the set $S \cup u_{m+1}$ is orthonormal. □

4.8 The Adjoint Transformation

It is now time to let the geometry and transformations interact. If A is a linear transformation defined on an inner product space, then its adjoint A^* is defined by

$$(Au, v) = (u, A^*v).$$

Replacing u and v by appropriate basis vectors, it is easy to show that if A is given by the matrix (a_{ij}) with respect to a basis, then A^* is $(\overline{a_{ji}})$, i.e. is the complex conjugate transpose.

A linear transformation is *self-adjoint* if $A = A^*$; over $\mathbb{R}$ this is a *symmetric* matrix, and over $\mathbb{C}$ this is a *Hermitian* matrix. A is Hermitian iff $(Av, v) \in \mathbb{R}$ for all $v \in V$. This implies that A is strictly positive definite and is invertible.

Using the adjoint we can decide which transformations are *isometries*, i.e. leave the geometry intact. (These are also known as *orthogonal* transformations). A transformation U is an isometry iff

$$\begin{aligned} \|Uv\| &= \|v\| \qquad \forall v \in V \\ \Leftrightarrow (Uv, Uw) &= (v, w) \qquad \forall u, v \in V \\ \Leftrightarrow U^*U &= I. \end{aligned}$$

We also note that changing from one orthonormal basis of V to another uses an isometry.

4.9 Diagonalisation

What does all this tell us about eigenvalues and eigenspaces?

- If A is self-adjoint, then every eigenvalue is real.
- If A is an isometry, then every eigenvalue has modulus 1.
- In both cases, eigenspaces for different eigenvalues are orthogonal, which implies that both types of matrix are diagonalisable with an orthogonal change of basis $P^{-1}AP$ for P orthogonal.

  Although we won't use it, this last point is true for any *normal* transformation, that is one with the property that $A^*A = AA^*$.

4.10 Symmetric Bilinear and Quadratic Forms

A map $b : V \times V \to \mathbb{R}$ is a *symmetric bilinear form* if

1. it is symmetric: $b(u, v) = b(v, u)$;
2. and bilinear: $b(\lambda u + v, w) = \lambda b(u, w) + b(v, w)$.

A map $q : V \to \mathbb{R}$ is a *quadratic form* if it can be written as $q(v) = b(v, v)$ for a symmetric bilinear form b.

A bilinear form is *positive definite* if $b(v, v) > 0$ for all $v \in V$; it is *positive semi-definite* if $b(v, v) \geq 0$.

By choosing a basis for V, we can represent the bilinear form b by a matrix B so that

$$b(v, w) = v^T Bw.$$

A change of basis by an invertible transformation A means that the matrix B is mapped to $A^T BA$.

Can we find a particularly nice matrix for a bilinear transformation? And will it take a lot of work? With a bit of cunning, we can do very nicely with the theorems we already have.

The bilinear transformation b is symmetric, and hence the matrix B is symmetric. By the results of Section 4.9 we can find a transformation A so that $A^{-1}BA$ is diagonal. Unfortunately, that doesn't match the change of basis formula for bilinear forms, which is $A^T BA$. However, we are guaranteed that A can be chosen to be orthogonal, which means that $A^{-1} = A^* = A^T$ since A is real. Hence we can transform B to a diagonal matrix where the entries λ_i are the eigenvalues of B. By pre- and post-multiplying by the matrix

$$S = \text{diag}(1/\sqrt{|\lambda_i|}),$$

(where the value 0 is used if $\lambda_i = 0$) we obtain a diagonal matrix which has entries 1 (if $\lambda_i > 0$), -1 (if $\lambda_i < 0$) and zero.

5 Multivariate Calculus

In this section we will consider functions of many variables of the form

$$f : \mathbb{R}^n \to \mathbb{R}.$$

5.1 Definition of Derivatives

The *partial derivatives* are defined in terms of one-dimensional derivatives. As an example, consider $f : \mathbb{R}^2 \to \mathbb{R}$. Then the partial derivative with respect to x is defined as:

$$\left.\frac{\partial f}{\partial x}\right]_{(a,b)} = f_x(a, b) = \lim_{h\to 0} \frac{f(a + h, b) - f(a, b)}{h}.$$

The partial derivative with respect to y is defined in a similar way:

$$\left.\frac{\partial f}{\partial y}\right]_{(a,b)} = f_y(a, b) = \lim_{k\to 0} \frac{f(a, b) - f(a, b + k)}{k}.$$

It is also possible to define partial derivatives of higher order. It is very generally true (certainly for all the functions that you are likely to meet) that

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

We define the *Hessian matrix* to be the matrix of second order partial derivatives. If $f : \mathbb{R}^2 \to \mathbb{R}$, then

$$\nabla\nabla f = H = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

5.2 Chain Rule

There is an generalisation of the one-dimensional chain rule to the multivariate case, but it is more complicated. Consider a function $f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$. If each coordinate x_j is a function of m variables $u_1, \ldots, u_m$, i.e.

$$x_j = x_j(u_1, \ldots, u_m),$$

then what is $\partial f / \partial u_\alpha$?

Consider the small change in f caused by small changes in all the inputs:

$$\delta f = f(x_1 + \delta x_1, \ldots, x_n + \delta x_n) - f(x_1, \ldots, x_n) \tag{5.1}$$

$$= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \delta x_j + \text{higher order terms.} \tag{5.2}$$

We can apply this equation also to a small change δx_j to give

$$\delta x_j = \sum_{\alpha=1}^{m} \frac{\partial x_j}{\partial u_\alpha} \delta u_\alpha + \text{higher order terms.} \tag{5.3}$$

Combining equations (5.2) and (5.3), we get

$$\frac{\partial f}{\partial u_\alpha} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial u_\alpha}. \tag{5.4}$$

We can rewrite this in vector notation as a dot product:

$$\frac{\partial}{\partial u_\alpha} f(x(u)) = (\nabla_x f(x))^T \left(\frac{\partial x(u)}{\partial u_\alpha} \right). \tag{5.5}$$

5.3 Directional Derivatives

So far we have only seen how to define derivatives in directions that are parallel to the axes. Now let us consider how to compute a *directional derivative* of a function f in a direction $\mathbf{v}$ evaluated at a point $\mathbf{c}$. Let $\mathbf{x}$ denote a point on the line through $\mathbf{c}$ in the direction of $\mathbf{v}$, so $\mathbf{x} = \mathbf{c} + t\mathbf{v}$. Then we can write each coordinate of $\mathbf{x}$ as $x_j = c_j + tv_j$.

The directional derivative of f at $\mathbf{c}$ is given by

$$(D_\mathbf{v} f)(c) := \frac{\mathrm{d}f}{\mathrm{d}t}(\mathbf{c} + t\mathbf{v}) \rfloor_{t=0}.$$

Using the chain rule (5.4) and the fact that $\mathrm{d}x_j / \mathrm{d}t = v_j$, we have that

$$\frac{\mathrm{d}f}{\mathrm{d}t}(\mathbf{c} + t\mathbf{v}) = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(\mathbf{c} + t\mathbf{v}).$$

So

$$(D_\mathbf{v} f)(c) = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(c) = \nabla f^T . \mathbf{v} \tag{5.6}$$

This means that ∇f is perpendicular to surfaces where f is constant, as all $\mathbf{v}$ on such surfaces cause no change to f.

5.4 Taylor's Theorem

Armed with a definition of a directional derivative, we can generalise Taylor's Theorem to the multivariate case as well. We want to relate $f(\mathbf{c})$ and $f(\mathbf{c}+\mathbf{v})$ to the derivatives of f at $\mathbf{c}$. The easiest way to do this is to consider the one-dimensional function $F(t) = f(\mathbf{c}+t\mathbf{v})$ where $t \in [0,1]$. By the first form of Taylor's theorem (2.1), we can write

$$f(\mathbf{c}+\mathbf{v}) = F(1) = \sum_{i=0}^{m} \frac{1}{i!} F^{(i)}(0) + \frac{1}{m!} F^{(m)}(\theta), \tag{5.7}$$

where $\theta \in (0,1)$. We now just need to relate derivatives of F to derivatives of f.

$$F'(0) = \frac{\mathrm{d}}{\mathrm{d}t}(f(\mathbf{c}+t\mathbf{v}))\rfloor_{t=0} = (D_{\mathbf{v}} f)(c).$$

In a similar way (using induction),

$$F^{(i)} = (D_v^i f)(\mathbf{c}).$$

We shall frequently use a second-order Taylor series expansion, which has the form

$$f(\mathbf{c}+\mathbf{v}) = f(\mathbf{c}) + \nabla f^T \mathbf{v} + \frac{1}{2}\mathbf{v}^T \mathbf{H} \mathbf{v} + \text{higher order}, \tag{5.8}$$

where $\mathbf{H}$ is the Hessian matrix of f evaluated at $\mathbf{c}$. Note that $\mathbf{v}^T\mathbf{H}\mathbf{v}$ is a quadratic form (see Section 4.10).

5.5 Critical Points

When all the first order partial derivatives are zero (i.e. $\nabla f = 0$), a point $\mathbf{c}$ is said to be *stationary* or *critical*, and $f(\mathbf{c})$ is the stationary (or critical) *value*. Let us work out the relationship between stationary points and local maxima.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has a local maximum at $\mathbf{c} \in \mathbb{R}^n$. This implies that $f(\mathbf{c}+\mathbf{v}) \leq f(\mathbf{c})$ for all 'small' vectors $\mathbf{v}$. The first-order Taylor series expansion of f near $\mathbf{c}$ is

$$f(\mathbf{c}+\mathbf{v}) = f(\mathbf{c}) + (D_{\mathbf{v}} f)(\mathbf{c}) + \text{higher order terms}.$$

Expanding this, and using the fact that f has a maximum at $\mathbf{c}$,

$$\sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(\mathbf{c}) + \text{higher order terms} \leq 0. \tag{5.9}$$

Now $\mathbf{v}$ is a small vector; write it as $\mathbf{v} = \epsilon\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a fixed vector and $\epsilon \in \mathbb{R}$ is small. Substituting this into (5.9) gives

$$\epsilon \sum_{j=1}^{n} \beta_j \frac{\partial f}{\partial x_j}(\mathbf{c}) + O(\epsilon^2) \leq 0 \tag{5.10}$$

$$\Rightarrow \sum_{j=1}^{n} \beta_j \frac{\partial f}{\partial x_j}(\mathbf{c}) \leq 0, \tag{5.11}$$

for all possible vectors $\boldsymbol{\beta}$. Now choose $\boldsymbol{\beta}$ so that

$$\beta_j = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases} \tag{5.12}$$

Then $\partial f/\partial x_k(\mathbf{c}) \leq 0$. Choosing

$$\beta_j = \begin{cases} 0 & j \neq k \\ -1 & j = k \end{cases}$$

implies that $\partial f/\partial x_k(\mathbf{c}) \geq 0$. Hence

$$\frac{\partial f}{\partial x_k}(\mathbf{c}) = 0 \qquad \forall k. \tag{5.13}$$

We can conclude that

$$f(\mathbf{c}+\mathbf{v}) = f(\mathbf{c}) + \frac{1}{2}\mathbf{v}^T\mathbf{H}\mathbf{v} + \text{higher order terms.} \tag{5.14}$$

5.6 Constrained Stationary Points

In this section we discuss how to find stationary points subject to a set of constraints (e.g. subject to lying on a sphere). Suppose that f and g are continuously differentiable functions $\mathbb{R}^n \to \mathbb{R}$ and that $\nabla g \neq 0$. Then if f has a stationary value subject to $g(\mathbf{x}) = 0$ at a point bfc, then $\exists\lambda$ such that, at the point $\mathbf{c}$, the following equation holds:

$$\frac{\partial f}{\partial x_j} + \lambda\frac{\partial g}{\partial x_j} = 0 \qquad \text{for } j = 1, \ldots, n. \tag{5.15}$$

The parameter λ is known as a *Lagrange multiplier*.

The argument used for (unconstrained) stationary points remains valid up until equation (5.11). This still holds, but it is no longer necessarily true that we can choose the vector $\boldsymbol{\beta}$ as in (5.12) since the point $\mathbf{c}+\epsilon\boldsymbol{\beta}$ must satisfy the constraint $g(\mathbf{c}+\epsilon\boldsymbol{\beta}) = 0$. To make progress we have to incorporate this constraint into our analysis. We know that $g(\mathbf{c}+\epsilon\boldsymbol{\beta}) = g(\mathbf{c}) = 0$. Hence, by a first-order Taylor series expansion, it follows that

$$0 = g(\mathbf{c}+\epsilon\boldsymbol{\beta}) - g(\mathbf{c}) = \epsilon\sum_{j=1}^{n}\beta_j\frac{\partial g}{\partial x_j} + O(\epsilon^2). \tag{5.16}$$

Thus, to a first order approximation, for any $\lambda \in \mathbb{R}$,

$$\sum_{j=1}^{n}\beta_j\frac{\partial f}{\partial x_j} = \sum_{j=1}^{n}\beta_j\left(\frac{\partial f}{\partial x_j} + \lambda\frac{\partial g}{\partial x_j}\right). \tag{5.17}$$

Now $\nabla g \neq 0$, so at least one $\partial g/\partial x_j$ is non-zero. Without loss of generality, suppose that $\partial g/\partial x_n \neq 0$. Then, from equation (5.16), it follows that

$$\beta_n = -\sum_{j=1}^{n-1}\beta_j\left(\frac{\partial g}{\partial x_j}\Big/\frac{\partial g}{\partial x_n}\right). \tag{5.18}$$

Thus we can choose $\beta_1, \beta_2, \ldots, \beta_{n-1}$ arbitrarily, provided that we choose β_n to satisfy equation (5.18).

Now let

$$\lambda = -\left(\frac{\partial f}{\partial x_n}\Big/\frac{\partial g}{\partial x_n}\right). \tag{5.19}$$

Substituting this into equation (5.17), the term in β_n cancels out and we are left with the equation

$$0 = \sum_{j=1}^{n}\beta_j\frac{\partial f}{\partial x_j} = \sum_{j=1}^{n-1}\beta_j\left(\frac{\partial f}{\partial x_j} + \lambda\frac{\partial g}{\partial x_j}\right). \tag{5.20}$$

The difference from equation (5.17) is that we can now incorporated the constraint g into the choice of $\boldsymbol{\beta}$. We can now apply the argument of Section 5.5 to the choice of β_1, β_2, $\ldots \beta_{n-1}$ to show that each term on the right-hand side is equal to zero, i.e.

$$\frac{\partial f}{\partial x_j} + \lambda \frac{\partial g}{\partial x_j} = 0 \qquad \text{for } j = 1, \ldots, n-1. \tag{5.21}$$

This is also true for $j = n$ by the choice of λ.

5.7 Substitution in Multivariate Integrals

Multivariate integrals are to be avoided if you can, as they are even more difficult to evaluate than one-dimensional integrals. There is one technique that it is important to know about, and that is substitution.

A *coordinate transformation* $\rho : U \to V \subseteq \mathbb{R}^n$, also called a *diffeomorphism*, is a function $\rho : (x_1, \ldots, x_n) \in U \to (u_1, \ldots, u_n) \in V$ with the following properties:

1. The functions $u_1(\mathbf{x}), \ldots, u_n(\mathbf{x})$ have continuous partial derivatives with respect to $x_1, \ldots, x_n$.
2. ρ is a bijection (i.e. it is one-to-one and onto; it can be inverted).
3. The *Jacobian*
$$\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)} = \begin{vmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \cdots & \frac{\partial u_1}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial u_n}{\partial x_1} & \frac{\partial u_n}{\partial x_2} & \cdots & \frac{\partial u_n}{\partial x_n} \end{vmatrix} \neq 0 \qquad \text{on } U. \tag{5.22}$$

We can then prove the following facts about coordinate transformations and integrals.

- If $\rho : U \to V$ is a diffeomorphism, then so is $\rho^{-1} : V \to U$, and
$$\frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} = \frac{1}{\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)}}.$$
To prove this, multiply out the matrices for the two Jacobians; using the chain rule, one obtains a matrix A where $a_{ij} = \partial u_i / \partial u_j$ and this is the identity.
- If D and its boundary are contained in U and $D' = \rho(D) \subseteq V$, then for any integrable function f,
$$\int_D \cdots \int f(\mathbf{x}) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n = \int_{D'} \cdots \int f(\mathbf{x}(\mathbf{u})) \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} \right| \mathrm{d}u_1 \ldots \mathrm{d}u_n.$$

As an example, consider the integral

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} \, \mathrm{d}x \, \mathrm{d}y.$$

To compute this integral, we make a substitution using radial coordinates (r, θ) so that

$$x = x(r, \theta) = r \cos\theta$$
$$y = y(r, \theta) = r \sin\theta$$

We need to compute the Jacobian

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r.$$

We also note that $x^2 + y^2 = r^2\cos^2\theta + r^2\sin^2\theta = r^2$. Hence

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y &= \int_0^{\infty}\int_0^{2\pi} re^{-r^2}\,\mathrm{d}\theta\,\mathrm{d}r \\ &= 2\pi\int_0^{\infty} re^{-r^2}\,\mathrm{d}r \\ &= 2\pi\left[\frac{-e^{-r^2}}{2}\right]_0^{\infty} \\ &= \pi. \end{aligned}$$

You will use this result often when studying the Gaussian probability density function.

6 Reading List for Mathematical Background

- G. Strang, *Introduction to Linear Algebra*. 5th ed. Wellesley-Cambridge Press, 2016.
- S. Lipchutz, *Linear Algebra*, Schaum's outline series, McGraw-Hill, NY.
- F. Ayres, *Differential and Integral Calculus*, Schaum's outline series, McGraw-Hill, NY.
- K. A. Stroud, *Further Engineering Mathematics* 3rd edition, Macmillan Press, London, 1996.
- P. V. O'Neil, *Advanced Engineering Mathematics* 4th edition, PWS Publishing, London, 1995.
- E. Kreyzig, *Advanced Engineering Mathematics* 6th edition, Wiley, NY, 1988.

You may also find the following free online courses helpful, though I haven't tried them out in detail myself.

- `https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/` (this goes go a lot deeper than is needed to prepare for the Data Science programme)
- `https://www.coursera.org/specializations/mathematics-machine-learning`
- `https://www.khanacademy.org/math/multivariable-calculus` first two sections
- `https://courses.edx.org/courses/course-v1:UCSanDiegoX+DSE210x+3T2017/course/`