BE 3rd year UG Math background sheet 1.5

Aldo Faisal
Dept. of Bioengineering, Imperial College London

DRAFT, NOT FOR GENERAL CIRCULATION. This is a summary of mathematical results with which you should be familiar from your first two years. These will be expected background knowledge for the 3rd year of your studies[1]. The document covers mostly just the basic results and occasionally gives reminders about their derivation.

The notation used is in general following the notation you will have initially encountered when the mathematical result was taught for the first time in one of your courses. This will help you recall the result, but you should be by now flexible to understand and handle changes of notation (e.g. variable names) in the various fields of study. The nomenclature (and notation) is adjusted, as you will have encountered several alternative versions in your courses. You will find "note"-pointers that hint at such alternatives, as well as point to issues when confronting different text books. In the final part of this document we have included a list of common and basic mathematical notation for logic and set theory, which will help you "read" mathematics (and by the way follows the engineering standard organisation ISO recommendations).

If anything below is unclear you should do some further reviewing and set yourself some exercises to get some practice and intuition. An excellent textbook to help you review is *Mathematical Methods of Physics and Engineering* by Riley, Hobson & Bence (Cambridge University Press). In addition some sections have pointers to freely available on-line material/books.

1 Linear Algebra

Scalars are individual numbers, vectors are columns of numbers, matrices are rectangular grids of numbers, eg:

$$x = 1, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}$$

In the above example x is 1×1, $\mathbf{x}$ is $n \times 1$ and A is $m \times n$. Dimensions

The transpose operator, $^\top$ ($'$ in Matlab), swaps the rows and columns: Transpose

$$x^\top = x, \quad \mathbf{x}^\top = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}, \quad \left(A^\top\right)_{ij} = A_{ji}$$

Quantities whose inner dimensions match may be "multiplied" by summing over this index. The outer dimensions give the dimensions of the answer. Multiplication

$$A\mathbf{x} \text{ has elements } (A\mathbf{x})_i = \sum_{j=1}^{n} A_{ij}\mathbf{x}_j \quad \text{and} \quad (AA^\top)_{ij} = \sum_{k=1}^{n} A_{ik}\left(A^\top\right)_{kj} = \sum_{k=1}^{n} A_{ik}A_{jk}$$

Check Dimensions

All the following are allowed (the dimensions of the answer are also shown):

$\mathbf{x}^\top\mathbf{x}$	$\mathbf{x}\mathbf{x}^\top$	$A\mathbf{x}$	$AA^\top$	$A^\top A$	$\mathbf{x}^\top A\mathbf{x}$
1×1	$n\times n$	$m\times 1$	$m\times m$	$n\times n$	1×1
scalar	matrix	vector	matrix	matrix	scalar

,

while $\mathbf{x}\mathbf{x}$, AA and $\mathbf{x}A$ *do not make sense* for $m \neq n \neq 1$ (Can you see why?).

Multiplication by scalar

An exception to the above rule is that we may write: αA. Every element of the matrix A is multiplied by the scalar α.

Summation convention

Care has to be taken, as in some contexts displaying the indices of matrices and vectors may imply a summation convention (which should be explicitly mentioned in its context). For example, the matrix-vector product is usually written as $x_i = \sum_j M_{ij}x_j$. However, under the "Siggers"[2] summation convention allows one to write the previous sum simply as $x_i = M_{ij}x_j$ where it is implied by the convention that the sum is taken over repeated identical indices. Other summation conventions (not encountered in your courses, but may be found in text books) include the Einstein summation convention (which involves upper and lower indices).

Scalar product/Dot product

The scalar product $\mathbf{x}^\top\mathbf{y}$ allows us to conveniently define the angle θ (the smaller of the two angles, between 0° and 180°) between the two vectors $\mathbf{x},\mathbf{y}$ as

$$\theta = \arccos\left(\frac{\mathbf{x}^\top\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}\right)$$

E.g. two orthogonal vectors (oriented 90° from each other) have always a scalar product of 0. Conveniently, the scalar product between a vector $\mathbf{x}$ and a unit vector $\hat{\mathbf{n}}$ (of length 1) yields the length l of the projection of $\mathbf{x}$ onto a line colinear with $\mathbf{n}$: $l = \mathbf{x}^\top\hat{\mathbf{n}}$.

Normal vector

The normal vector $\mathbf{n}$ or *normal* to a surface is a vector perpendicular to it. Often, the normal unit vector $\hat{\mathbf{n}}$ is desired, which is sometimes known as the *unit normal*. When normals are considered on closed surfaces, the inward-pointing normal vector (pointing towards the interior of the surface) and outward-pointing normal are usually distinguished. Defining b the distance of the closest point of the plane to the origin, we can conveniently test if a point $\mathbf{x}$ lies on this plane, as the following holds:

$$\mathbf{x}^\top\hat{\mathbf{n}} = b$$

(Can you see why?).

Cross product/Vector product

The cross product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\mathbf{x}\times\mathbf{y}$. The cross product is defined as a vector $\mathbf{z} = \mathbf{x}\times\mathbf{y}$ being perpendicular to both $\mathbf{x}$ and $\mathbf{y}$, with a direction given by the right-hand rule ($\mathbf{x}$ index finger, $\mathbf{y}$ middle finger, $\mathbf{z}$ thumb) and a magnitude equal to the area of the parallelogram that the vectors $\mathbf{x}$ and $\mathbf{y}$ span:

$$\mathbf{x}\times\mathbf{y} = ||\mathbf{x}||||\mathbf{y}||\sin(\theta)\mathbf{n}$$

with $\mathbf{n}$ the unit vector perpendicular to the plane on which $\mathbf{x}$ and $\mathbf{y}$ lie.

Basis vectors

A basis is a set of vectors $\mathbf{v_i}$ that as linear combination, can represent every vector in a given vector space, and such that no element of the set can be represented as a linear combination of the others. $\sum_{i=1..N} w_i\mathbf{v_i} = x$. We say that the $\mathbf{v_i}$ span the vector space. I.e. the basis vectors form a coordinate system we typically consider finite dimensional vector spaces $\mathbb{R}^N$, the N-dimensional space of real numbers), and the w_i are the coordinates of x. Note, that you also encounter orthonormal basis vectors, which are orthogonal to each other and their scalar product satisfies $\mathbf{v_n}^\top\mathbf{v_m} = 0, \forall n \neq m$, each orthonormal basis vector has normalised length 1. In everyday 3D space a typical (orthonormal) basis is e.g. $\mathbf{x_1} = (1,0,0)^\top, \mathbf{x_2} = (0,1,0)^\top, \mathbf{x_3} = (0,0,1)^\top$.

Basis functions

We can extend the notion of basis to the infinite dimensional space of functions, where an (infinite) set of linear basis functions $f_i(t)$ satisfy the inner product $f_n(t) \cdot f_m(t) := \int_a^b f_n(t) f_m(t) dt = 0, \forall n \neq m$ (also this notation is also often used $< f_n(t), f_m(t) >$).

Square Matrices

Diagonal matrices, the Identity

Now consider the square $n \times n$ matrix B. All off-diagonal elements of diagonal matrices are zero. The "Identity matrix", which leaves vectors and matrices unchanged on multiplication, is diagonal with each non-zero element equal to one.

$$\begin{array}{c} B_{ij} = 0 \text{ if } i \neq j \quad \Leftrightarrow \quad \text{“}B \text{ is diagonal”} \\ \mathbb{I}_{ij} = 0 \text{ if } i \neq j \text{ and } \mathbb{I}_{ii} = 1 \ \forall i \quad \Leftrightarrow \quad \text{“}\mathbb{I} \text{ is the identity matrix”} \\ \mathbb{I}\mathbf{x} = \mathbf{x} \quad \mathbb{I}B = B = B\mathbb{I} \quad \mathbf{x}^\top \mathbb{I} = \mathbf{x}^\top \end{array}$$

Easily proved results

Simple and valid manipulations:

$$(AB)C = A(BC) \quad A(B+C) = AB + AC \quad (A+B)^\top = A^\top + B^\top \quad (AB)^\top = B^\top A^\top$$

Note that $AB \neq BA$ in general.

Inverses

Some square matrices have inverses:

$$B^{-1}B = BB^{-1} = \mathbb{I} \quad \left(B^{-1}\right)^{-1} = B,$$

which have these properties:

$$(BC)^{-1} = C^{-1}B^{-1} \quad \left(B^{-1}\right)^\top = \left(B^\top\right)^{-1}$$

Pseudo Inverse

For a general matrix A (i.e. **not necessarily square**), we define the Moore-Penrose pseudo inverse $A^\dagger = (A^\top A)^{-1} A^\top$ which is no inverse $A^\dagger A \neq AA^\dagger \neq \mathbb{I}$, but satisfies $AA^\dagger A = A$ This generalises the concept of an Inverse (see below) for non-square matrices. You can solve over-constrained sets of linear equations (more known variables than unknowns) using $A^\dagger$ to find the least-square solution.

Solving Linear equations

Linear simultaneous equations could be solved (inefficiently) this way:

$$\text{if } B\mathbf{x} = \mathbf{y} \text{ then } \mathbf{x} = B^{-1}\mathbf{y}$$

Some other commonly used matrix definitions include:

Symmetry

$$B_{ij} = B_{ji} \Leftrightarrow \text{“}B \text{ is symmetric”}$$

Trace

$$\text{Trace}(B) = \text{Tr}(B) = \sum_{i=1}^{n} B_{ii} = \text{“sum of diagonal elements”}$$

A Trace Trick

Cyclic permutations are allowed inside trace. Trace of a scalar is a scalar:

$$\text{Tr}(BCD) = \text{Tr}(DBC) = \text{Tr}(CDB) \quad \mathbf{x}^\top B\mathbf{x} = \text{Tr}(\mathbf{x}^\top B\mathbf{x}) = \text{Tr}(\mathbf{x}\mathbf{x}^\top B)$$

Determinants

The determinant is written $\text{Det}(B)$ or $|B|$. It is a scalar regardless of n.

$$|BC| = |B||C|, \quad |x| = x, \quad |xB| = x^n |B|, \quad \left|B^{-1}\right| = \frac{1}{|B|}.$$

It *determines* if B can be inverted: $|B| = 0 \Rightarrow B^{-1}$ undefined. If the vector to every point of a shape is pre-multiplied by B then the shape's area or volume increases by a factor of $|B|$. For a diagonal matrix the volume scaling factor is simply the product of the diagonal elements. In general the determinant is the product of the eigenvalues.

$$B\mathbf{e}^{(i)} = \lambda^{(i)}\mathbf{e}^{(i)} \Leftrightarrow \text{“}\lambda^{(i)} \text{ is an eigenvalue of } B \text{ with eigenvector } \mathbf{e}^{(i)}\text{”}$$

$$|B| = \prod \text{eigenvalues} \qquad \text{Trace}(B) = \sum \text{eigenvalues}$$

Eigenvalues, Eigenvectors

If B is real and symmetric (eg a covariance matrix) the eigenvectors are orthogonal (perpendicular) and so form a basis (can be used as axes).

Tensor

”Tensors represent physical quantities and transform as such”. Tensors are objects that generalise the notion of scalar, vector, matrix to higher orders (0th order tensor are written like a scalar, 1st order tensor written like a vector, 2nd order tensor written like a matrix, etc.). E.g. a weather map with temperatures is a scalar-field (of temperatures) and is described by a rank-0 tensor. The multiplication of 2nd order tensors (the only Tensors you will have encountered are order 2 and rank 3, because of the 3 spatial dimensions of the vectors used in mechanics) by a vector follows the normal rules of matrix-vector multiplication. Thus, a stress tensor $\mathbf{T}$ takes a direction $\mathbf{v}$ as input and produces the stress $\mathbf{T}(\mathbf{v})$ on the surface normal to this vector as output and so expresses a relationship between these two vectors. It is possible to represent a tensor by examining what it does to a coordinate basis or frame of reference; the resulting quantity is then an organized multi-dimensional array of numerical values. Tensors express a relationship between vectors, tensors themselves are therefore independent of a particular choice of coordinate system. Note, unlike vector fields which assign vectorial quantities to positions in space, tensors formalise the idea that multiple vectorial quantities can exist at the same point in space.

2 Coordinate systems, complex numbers and transforms

Cylindrical coordinates

The cartesian coordinate system is the most commonly used, however in many practical 3-dimensional problems, 2 other coordinate systems are frequently more convenient to use. The cylindrical coordinate system is a 3-dimensional coordinate system that specifies point positions by the distance from a chosen reference axis ρ, the direction from the axis relative to a chosen reference direction ϕ, and the distance from a chosen reference plane perpendicular to the axis z (positive or negative depending on which side of the reference plane faces the point). One can convert between cartesian and cylindrical coordinates using:

$$\begin{aligned}
x &= \rho\cos(\phi) & \rho &= \sqrt{x^2+y^2} \\
y &= \rho\sin(\phi) & \phi &= 0, \text{if } x = y = 0 \\
& & \phi &= \arcsin\left(\frac{y}{\rho}\right), \text{if } x \geq 0 \\
& & \phi &= \pi - \arcsin\left(\frac{y}{\rho}\right), \text{if } x < 0 \\
z &= z & z &= z
\end{aligned}$$

Spherical coordinates

The spherical coordinate system defines the 3-D position of a point $\mathbf{x}$, as a radius r as Euclidean distance from the origin to $\mathbf{x}$, the inclination (or elevation or polar angle) θ is the angle between the zenith direction and line connecting the origin to $\mathbf{x}$, and the azimuth ϕ, is the signed angle, as measured from the azimuth direction to the orthogonal projection of the line from the origin to $\mathbf{x}$ onto the reference plane. One can convert between cartesian and spherical coordinates using:

$$\begin{aligned}
x &= r\sin(\theta)\cos(\phi) & r &= \sqrt{x^2+y^2+z^2} \\
y &= r\sin(\theta)\sin(\phi) & \theta &= \cos^{-1}\left(\frac{x}{y}\right) \\
z &= r\cos(\theta) & \phi &= \tan^{-1}\left(\frac{x}{y}\right)
\end{aligned}$$

[1]Acknowledgements: We thank Jennifer Siggers, Darryl Overby and Manos Drakakis for comments on versions of this manuscript.

[2]Jennifer to provide the proper name

Note 1: care has to be taken when taking derivatives, e.g. gradients (see below) in non-orthonormal coordinate systems, such as these two above, as derivatives are linked. Note 2: the coordinate variable names are recommended by an ISO standard (31-11).

Complex numbers and Fourier transform

Complex number

A complex number $z \in \mathcal{C}$ is a number consisting of a real part and an imaginary part.

$$z = a + bi$$

where a and b are real numbers ($a, b \in \mathbb{R}$) and i is a mathematical symbol which is called the *imaginary unit*. We define $i = \sqrt{-1}$. Note: sometimes the notation is reversed (ib vs bi) and the literature (especially in electrical engineering) uses the letter j for the imaginary unit. A basic relationships of complex numbers relates exponentials to trigonometric functions (which explains why when you encounter complex numbers as eigenvalues in mechanics or electrical circuits - oscillations are often the result).

$$e^{it} = \cos(t) + i\sin(t)$$

Fourier transform

Chief among the fundamental transformations of functions are the *Fourier* transform which transforms a (periodic) function f of variable t (e.g. time) into a function F with variable ω (e.g. frequency).

$$\text{forward transform: } F(\omega) = \int_{-\infty}^{\infty} f(t)e^{(i\ \omega\ t)}dt \quad \text{inverse transform: } f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{(i\ \omega\ t)}d\omega$$

There are extensive tables on the web that allow you to relate standard functions to a Fourier domain representation. Note, alternative definitions of the Fourier transform include a factor of $\frac{1}{2\pi}$, and may use oscillation frequency ν instead of angular frequency $\omega = 2\pi\nu$. Double check your assumptions based on the context.

Convolution

An important operation on functions is the *convolution*, a mathematical operation on two functions f and g that produces a third function h.

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau = h(t)$$

Convolution theorem

Convolution is a commutative, additive, distributive operation that is translation invariant to changes in the coordinate system. Importantly the *Convolution theorem* for the Fourier transform translates between convolution and multiplication of functions: Given the convolution of two confunctions

$$h(t) = (f \star g)(t)$$

then their Fourier transforms are simply products

$$H(\omega) = F(\omega)G(\omega)$$

On computers we typically use the Fast Fourier Transform (Matlab "fft") to compute these transforms efficiently.

Laplace transform

Laplace transform

Another important transform is the *Laplace* transform.

$$L(s) = \int f(t)e^{(-s\ t)}dt$$

We distinguish the uni-lateral form (where the integral goes from 0 to ∞) and which you will have normally encountered, as well as the bi-lateral form (where the integral goes from $-\infty$ to ∞. Caveat: The literature is divided or amiguous which of the two forms they denote by *the* Laplace transform, although you should have learned the uni-lateral form is the "correct" one. Importantly the Laplace transform allows us to transform operators (e.g. differentiation and integration) in a straightforward way:

$$f'(t) \to sF(s) - f(0) \quad \text{and} \quad \int_0^t f(\tau)d\tau \to \frac{1}{s}F(s)$$

Simple Laplace transforms

The properties of the unilateral Laplace transform allows simple and valid manipulations:

$$af(t)+bg(t) \quad \text{Linearity} \quad aF(s)+bG(s)$$

$$f(at) \quad \text{Scaling} \quad \frac{1}{a}F(\frac{s}{a})$$

$$t^n f(t) \quad n\text{-th differentiation } s \quad (-1)^n F^{(n)}(s)$$

$$f^n(t) \quad n\text{-th differentiation } t \quad s^n F(s) - s^{n-1} f(0) - \cdots - f^{(n-1)}(0)$$

$$\int_0^t f(\tau)d\tau \quad \text{Integration} \quad \frac{1}{s}F(s)$$

Linear-Time-Invariant

This makes the Laplace transform ideal method to solve linear differential equation with known initial conditions (in electrical engineering known as Linear-Time-Invariant (LTI) systems) and is therefore often encountered across engineering from signal processing (LRC circuits) to mechanics (damped oscillator).

Operators & Functional analysis

Link between Fourier and Laplace transform: the Fourier transform is obtained from the bi-lateral Laplace transform by setting the real part of the complex variable s to zero. In electrical engineering terms (where we use these transforms to characterize the simulus-response of a system/plant) this is equivalent to ignoring the impulse response of a system and dealing only with the steady-state (frequency) response. Fourier and Laplace transforms are functions that operate on functions, so called operators. Other such function operators include differentiation and integration. The mathematics of function of functions (although not explicitly covered in your course), so called , functional or linear analysis, shows how and why these operators are interlinked across very different areas of engineering.

3 Calculus

Any good A-level maths text book should cover basic calculus and have plenty of exercises. Undergraduate text books might cover it quickly in less than a few chapters.

Simple integral

For simple integrals, recall that these are the equivalent of sums for continuous variables. Eg: $\sum_{i=1}^{n} f(x_i)\Delta x$ becomes the integral $\int_a^b f(x)\mathrm{d}x$ in the limit $\Delta x \to 0$, $n \to \infty$, where $\Delta x = \frac{b-a}{n}$ and $x_i = a + i\Delta x$. The simple integral yields the area under a curve $f(x)$ inside the interval $[a,b]$. Common ways to solve such integrals are *Integration by substitution, Integration by parts, Changing integration order* and *Differentiating inside the integral*. Find an A-level text book with some diagrams if you want to refresh your intuition. For virtually all applications you will have learned and used the Riemann integral definition, you may however also encounter the Lebesgue interval. To compute the Riemann integral of f, one partitions the domain [a,b] into subintervals (think "vertical bar charts" that approximate the area under a function), while in the Lebesgue integral, one is in effect partitioning the range of f (think "horizontal bar charts"). In future courses, especially in control and probabilistic contexts you may also encounter the Ito integral, which deals with integrals over "random quantities".

Path integral

Recall that there is a difference between a "simple" integral and the path integral (sometimes also called line integral, contour integral, curve integral) where the function to be integrated is evaluated along a line or curve. The function to be integrated may be in fact a scalar field or a vector field. The value of the path integral is the sum of values of the field at all points on the curve, weighted by some scalar function on the curve. In case of a path integral taken on a vector field, we typically use the scalar product of the vector field with a differential vector in the curve. This idea of weighting distinguishes the path integral from simpler integrals defined on intervals or areas. Many simple relationships in engineering (for example, $W = Fx$) have natural continuous analogs in terms of line integrals ($W = \oint_C F dx$), e.g. the line integral finds the work done on an object moving through an gravitational or electric field.

Gradient

The gradient of a straight line $y = mx + c$ is a constant $y' = \frac{y(x+\Delta x)-y(x)}{\Delta x} = m$.

Differentiation

Many functions look like straight lines over a small enough range. The gradient of this line, the derivative, is not constant, but a new function:

$$y'(x) = \frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{y(x+\Delta x) - y(x)}{\Delta x}, \quad \text{which could be differentiated again:} \quad y'' = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \frac{\mathrm{d}y'}{\mathrm{d}x}$$

Standard derivatives

The following results are well known (c is a constant):

$f(x):$	c	cx	cx^n	$\log_e(x)$	$\exp(x)$
$f'(x):$	0	c	cnx^{n-1}	$1/x$	$\exp(x)$

.

Optimisation

At a *maximum* or *minimum* the function is rising on one side and falling on the other. In between the gradient must be zero. Therefore

$$\text{maxima and minima satisfy: } \frac{\mathrm{d}f(x)}{\mathrm{d}x} = 0 \quad \text{or} \quad \frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}} = \mathbf{0} \Leftrightarrow \frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}x_i} = 0 \ \ \forall i$$

If we can't solve this numerically we can evolve our variable x, or variables $\mathbf{x}$, on a computer using gradient information until we find a place where the gradient is zero. Note, that at *saddle points* the derivative can be zero, but these are not extrema (e.g. for the function $f(x) = x^3$, where the derivative is 0 at $x = 0$ but the function increases from left to right), therefore one should always check that the second order derivatives are positive (minimum) or negative (maximum) to ensure that the point considered is an extremum.

Taylor approximation

A function may be approximated "to first order" by a straight line about any point x_0.

$$f(x_0 + \Delta x) \approx f(x) + \Delta x f'(x_0), \qquad \text{eg: } \log(1+x) \approx \log(1+0) + x\frac{1}{1+0} = x$$

More generally any analytic function f can be represented as a Taylor series around a point x_0 to arbitrary precision:

$$f(x_0 + \Delta x) = \sum_{n=0}^{\infty} \frac{1}{n!} \Delta x \frac{d^n}{dx^n} f(x_0)$$

Linearity

The derivative operator is linear:

$$\frac{\mathrm{d}(f(x)+g(x))}{\mathrm{d}x} = \frac{\mathrm{d}f(x)}{\mathrm{d}x} + \frac{\mathrm{d}g(x)}{\mathrm{d}x}, \qquad \text{eg: } \frac{\mathrm{d}(x+\exp(x))}{\mathrm{d}x} = 1 + \exp(x).$$

Product Rule

Dealing with products is slightly more involved:

$$\frac{\mathrm{d}(u(x)v(x))}{\mathrm{d}x} = v\frac{\mathrm{d}u}{\mathrm{d}x} + u\frac{\mathrm{d}v}{\mathrm{d}x}, \qquad \text{eg: } \frac{\mathrm{d}(x \cdot \exp(x))}{\mathrm{d}x} = \exp(x) + x\exp(x).$$

Chain Rule

The "chain rule" $\frac{\mathrm{d}f(u)}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}f(u)}{\mathrm{d}u}$, allows derivatives of functions.

$$\text{For example: } \frac{\mathrm{d}\exp(ay^m)}{\mathrm{d}y} = \frac{\mathrm{d}(ay^m)}{\mathrm{d}y} \cdot \frac{\mathrm{d}\exp(ay^m)}{\mathrm{d}(ay^m)} \quad \text{"with } u = ay^m\text{"}$$
$$= amy^{m-1} \cdot \exp(ay^m)$$

Vector and Matrix differentiation

For our convenience using Matlab's vector and matrix notations, we can apply the normal rules of differentiation to vectors and matrices straightaway (instead of applying them by element as we usually would):

A vector differentiation operator on a function $f(\mathbf{x})$ can be defined as (see also below)

$$\frac{d}{d\mathbf{x}}f(\mathbf{x}) = [\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}]^\top = \text{grad}(f)$$

The following convenient properties hold:

$$\frac{d}{d\mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{d}{d\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$$

if $\mathbf{A}$ is symmetric: $\mathbf{A}^\top = \mathbf{A}$, otherwise use $\frac{d}{d\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$.

A great source of reference for these and other results is the Matrix Cookbook by K. B. Petersen and M. S. Pedersen available freely online at `http://matrixcookbook.com`.

4 Vector calculus

The elements in vector calculus are scalar fields (scalar-valued functions) and vector fields (vector-valued functions). These fields are transformed under various operators:

Gradient

The gradient measures the rate and direction of change in a scalar field. Maps scalar fields to vector fields. It always points in the direction of greatest increase of the scalar field f, and it has a magnitude equal to the maximum rate of increase at the point (i.e. the direction of and the rate of steepest ascent).

$$\text{grad}(f) = \nabla(f)$$

E.g. for a 3D field in cartesian space (x,y,z) $\nabla(f) = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} + \frac{\partial f}{\partial z}\hat{z}$, where the $\hat{x},\hat{y},\hat{z}$ are unit vectors pointing in the three spatial directions x,y,z. Caveat: Although for many puproses ∇ looks like a standard derivative it is not, it is an operator. Care has to be taken when operating with it, as e.g. ∇ does not commute.

Curl

The curl measures the tendency to rotate about a point in a vector field. Maps vector fields to a "pseudeo vector field"/scalar field. Note: This is the cross product of the gradient with the vector-valued function. Also, some literature may use the term "rot" instead of "curl".

$$\text{curl } \mathbf{v} = \nabla \times \mathbf{v}$$

Divergence

The divergence measures the magnitude of a source (converge) or sink (repel) at a given point in a vector field. Maps vector fields to a scalar fields.

$$\text{div } \mathbf{v} = \nabla \cdot \mathbf{v}$$

E.g. for a 3D field in cartesian space (x,y,z): $\text{div } \mathbf{v} = \frac{\partial \mathbf{v_x}}{\partial x} + \frac{\partial \mathbf{v_y}}{\partial y} + \frac{\partial \mathbf{v_z}}{\partial z}$ Note: This is the dot product of the gradient with the vector-valued function.

Laplacian

The laplacian is a composition of the divergence and gradient operations. Maps scalar fields to scalar fields or vector fields to vector fields.

$$\Delta f = \nabla^2 f = \nabla \cdot \nabla f$$

Divergence theorem

The Divergence or Gauss-Ostrogradsky theorem relates how the flow ("flux") of a vector field through a closed surface S to the behaviour of the (continuously differentiable) vector field $\mathbf{v}$ inside

the region R (with volume elements dR) bounded by the surface S (with surface element dS). On the boundary we have outward-pointing normal vectors $\mathbf{n}$.

$$\int\int\int_R (\nabla \cdot \mathbf{v}) dR = \int\int_S (\mathbf{F} \cdot n) dS$$

The left-hand side captures the source (inside the volume, hence the use of three integrals if in 3D space), the right-hand side captures the flux through the surface (hence the use of two integrals, which describe a 2D surface).

Stokes' theorem

The Kelvin-Stokes theorem allows us to evaluate vector fields $\mathbf{v}$ on an (open) surface S by evaluating the boundary curve C of the surface S and vice versa. The volume elements dS of the surface S are oriented (i.e. vectors), and the position elements of the curve C are positions (i.e. vectors). Note, that if S is a closed surface, than the equality is 0.

$$\int\int_C \text{curl } \mathbf{v} \cdot \mathbf{n} = \int_S \mathbf{v} \cdot \mathbf{dC}.$$

5 Probability Theory

Explained in more detail in Chapter 2, sections 2.1–2.3 of David MacKay's book (freely available online):`http://www.inference.phy.cam.ac.uk/mackay/itila/book.html`

The probability a discrete variable A takes value a is: $0 \leq P(A=a) \leq 1$

Alternatives

Probabilities of alternatives add: $P(A=a \text{ or } a') = P(A=a) + P(A=a')$

Normalisation

The probabilities of all outcomes must sum to one: $\sum_{\text{all possible } a} P(A=a) = 1$

Joint Probability

$P(A=a, B=b)$ is the joint probability that both $A=a$ and $B=b$ occur.

Marginalisation

Variables can be "summed out" of joint distributions:

$$P(A=a) = \sum_{\text{all possible } b} P(A=a, B=b)$$

Conditional Probability

$P(A=a|B=b)$ is the probability $A=a$ occurs given the knowledge $B=b$.

Product Rule

$$P(A=a, B=b) = P(A=a)\,P(B=b|A=a) = P(B=b)\,P(A=a|B=b)$$

Independence

The following hold, for all a and b, if and only if A and B are independent:

$$\begin{aligned} P(A=a|B=b) &= P(A=a) \\ P(B=b|A=a) &= P(B=b) \\ P(A=a, B=b) &= P(A=a)\,P(B=b)\,. \end{aligned}$$

Otherwise the product rule above *must* be used.

Bayes Rule

Bayes rule can be derived from the above:

$$P(A=a|B=b, \mathcal{H}) = \frac{P(B=b|A=a, \mathcal{H})\,P(A=a|\mathcal{H})}{P(B=b|\mathcal{H})} \propto P(A=a, B=b|\mathcal{H})$$

Note that here, as with any expression, we are free to condition the whole thing on any set of assumptions, $\mathcal{H}$, we like. Note $\sum_a P(A=a,B=b|\mathcal{H}) = P(B=b|\mathcal{H})$ gives the normalising constant of proportionality.

Bernoulli distribution

The Bernoulli distribution has probability mass function $B(k=1;q)=q$ and $B(k=0;q)=1-q$, e.g. captures the outcome of a coin-toss with outcomes $1,0$ and probability q for outcome 1.

Factorial

The factorial of an integer n is defined as $n! = 1\times 2\times 3\times\cdots\times n$.

n choose k

The binomial coefficient $\binom{n}{k}$ yields the number of ways to choose k elements from a set of n elements (i.e. it counts the number of possible *combinations*). It can be calcualted recursively as $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$ or directly using

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{n-k}$$

Useful example: there are $\binom{n+k-1}{k}$ ways to choose k elements from a set of n if repetitions are allowed.

Binomial distribution

The Binomial distribution has probability mass function

$$B(k;N,q) = \binom{N}{k} q^k (1-q)^{N-k}$$

It describes the probability that we observe k successes when having sampled n statistically independent trials, each with probability q to succeed (i.e. each being a Bernoulli experiment). It follows that if a random variable $x \sim B(k;q), k\in\{0,1\}$ is Bernoulli distributed then $y=\sum_{i=1}^{N} x$ is Binomial distributed with $y \sim B(k;N,q), k\in 0,\ldots,N$. It allows follows if N is large enough then that Binomial distribution is approximated by the Gaussian distribution (see below) $B(n;N,q) = \mathcal{N}(Np, Np(1-p))$.

Probability density function (PDF)

All the above theory basically still applies to continuous variables if sums are converted into integrals. The probability that X lies between x and $x+\mathrm{d}x$ is $p(x)\,\mathrm{d}x$, where $p(x)$ is a *probability density function* with range $[0,\infty]$.

$$P(x_1<X<x_2) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x, \quad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \text{ and } p(x) = \int_{-\infty}^{\infty} p(x,y)\,\mathrm{d}y.$$

Cumulative density function (PDF)

Correspondingly, the integral from $-\infty$ to the value x is known as the cumulative probability density function (CDF), i.e. $CDF(x) = \int_{-\infty}^{x} PDF(z)dz$, and captures that probability that the value of a random value is x or smaller.

Expectation/Mean

The expectation or mean under a probability distribution is:

$$\langle f(a)\rangle = \sum_a P(A=a)\,f(a) \quad \text{or} \quad \langle f(x)\rangle = \int_{-\infty}^{\infty} p(x)\,f(x)\mathrm{d}x$$

Moreover, if x and y are independent, then $\langle x+y\rangle = \langle x\rangle + \langle y\rangle$ and for a scalar constant α, $\langle \alpha x\rangle = \alpha\langle x\rangle$.

t-test

A simple collection of statistical significance tests involving the mean of sampled data are the t-test or Student's t-test methods (Matlab "ttest"). These can inform you (using strong statistical assumptions) if two sets of samples have the same mean (e.g. are men and women as groups equally tall ?) or if a single set of samples has a specific mean (e.g. are Imperial students on average 1.75m tall?).

Variance & standard deviation

The variance (under a probability distribution) is:

$$\mathrm{Var}(x) = \left\langle (x-\langle x\rangle)^2 \right\rangle = \left\langle x^2\right\rangle - \langle x\rangle^2$$

and the standard deviation being the square root of the variance. Moreover, if x and y are independent, then $\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y)$. For a scalar constant α, $\mathrm{Var}(\alpha x) = \alpha^2 \mathrm{Var}(x)$ (can you show why?).

ANOVA

A simple collection of statistical tests involving the variance of sampled data are the ANnalysis Of VAriance methods. In their simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to allow comparison of the means of more than two groups.

Gaussian distribution

A random scalar $\mathbf{x}$ is distributed as a Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma)$ with mean μ and standard deviation σ has probability density function :

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

If $\mu = 0, \sigma = 1$ then it is sometimes referred to as Normal distribution.

Error function

The error function is the integral of the Normal distribution.

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dz$$

Note, there is no closed form expression to calculate this, but *C* and Matlab have it as in-built function "erf". The complementary error function is defined as $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$. Note, you will encounter the same Gaussian and error functions when dealing with differential equations, although in a non-probabilistic context.

6 Differential equations

Homogenous ODE

Linear homogeneous differential equations of order n with constant coefficients A_k

$$y(x)^{(n)} + A_1 y(x)^{(n-1)} + \cdots + A_n = 0$$

, e.g. 2nd order $y'' + ay' + by = 0$ (e.g. damped harmonic oscillator), can be solved by using the *Laplace transform* method. Knowing that solutions have the form $\exp(zx) = e^{zx}$ (z possibly complex), by using the fact that the exponential function remain proportional after differentiation. Note: this is not true in the case of repeated roots (see below). Thus, for the sum of multiple derivatives of a function to sum up to zero, the derivatives must cancel each other out. The only way for them to do so is for the derivatives to have the same form as the initial function.

Laplace transform

Thus, to solve this type of differential equation we can set $y = e^{zx}$, resulting in $z^2 e^{xz} + aze^{xz} + be^{xz} = 0$. Dividing by e^{xz} gives a polynomial, the *characteristic equation*, $F(z) = z^2 + az + b = 0$. Solving the polynomial gives n values of z: $z_1, \ldots, z_n$. Substitution of any of these values for z intoe^{zx} gives a solution $e^{z_i x}$. Since homogeneous linear differential equations obey the *superposition principle*, any linear combination of these functions also solves the differential equation.

Non-homogenous ODE

Linear non-homogeneous (or inhomogeneous) equation with constant coefficients

$$zy(x)^{(n)} + A_1 y(x)^{(n_1)} + \cdots + A_n = f(x)$$

can be solved using the method of *variation of parameters* or the method of *undetermined coefficients* . In general the solution of this type differential equations is the sum of the solution to the homogenous equations (i.e. $f(x) \equiv 0$) and the *particular solution*.

Variable coefficient ODEs

Linear ODEs with variable coefficients

$$c_n(x) y^{(n)}(x) + c_{n-1}(x) y^{(n-1)}(x) + \cdots + c_0(x) y(x) = r(x)$$

are often encountered as a sub-step in the solution of Partial Differential Equations (PDEs), or in the solution of increasingly complex physical problems. There are three strategies for solution are

1. reduction of Order (to first order) if the original ODE can be reduced to a first order equation, 2. change of independent variable to create a constant coefficient ODE, and 3. solution by power series.

Euler's equation

A prominent type of variable coefficient differential equations is Euler's equation (or Euler-Cauchy or Cauchy-Euler equation)

$$a_n x^n y^{(n)}(x) + a_{(n-1)} x^{n-1} y^{(n-1)}(x) + \cdots + a_0 y(x) = 0$$

Because of its simple structure the equation can be replaced with an equivalent equation with constant coefficients which can then be solved explicitly. The substitution $x = e^u$ reduces this equation to a linear differential equation with constant coefficients (see above).

If this Euler equation has order 1, then the general form

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

Integrating factor

can be solved by using the *integrating factor method*: Multiply through with $\exp(\int f(x)dx)$ and then simplify using the product rule. Solving for y yields:

$$y = e^{a(x)} \left(\int g(x) e^{a(x)} dx + \text{const} \right) \quad \text{and} \quad a(x) = \int f(x)dx$$

E.g. $\frac{dy}{dx} + ay = 1$ (which has constant coefficients $f(x) = a, g(x) = 1$), as found in mass-damped systems and RC circuits, yields $y(x) = e^{-ax} \left(\frac{e^{bx}}{a} + \text{const} \right) = \frac{1}{b} + \text{const}\ e^{-ax}$

Partial differential equations

Diffusion equation

Partial differential equations (PDEs) involve functions and their partial derivatives. You will be familiar with two of them, the wave equation and the diffusion equation. The diffusion equation relates spatial 2nd order derivates to first order derivatives of time

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}$$

where D is a *diffusion* constant. Under *Dirichlet* boundary conditions (values of c are fixed at the boundary) this PDE has a solution $c(x,t) = c_0 \text{erfc}(\frac{x}{2\sqrt{Dt}})$. You can find this solution using the *Laplace transform* method or from a physical *similarity* argument that expresses variations in space and time as a single variable.

Wave equation

The wave equation is a second-order linear partial differential equation for the description of waves (it's general form is called hyperbolic partial differential equation)

$$\frac{\partial u^2}{\partial^2 t} = c^2 \nabla^2 u(x,t)$$

The constant c captures the speed of propagation of the wave. It's general solution in 1-dimension has the form $u(x,t) = F(x-ct) + G(x+ct)$, these are travelling solutions as the solution function shapes F and G remain the same and are simply offest overtime by the wave speed c.