Background material crib-sheet

Iain Murray <i.murray+ta@gatsby.ucl.ac.uk>, October 2003

Here are a summary of results with which you should be familiar. If anything here is unclear you should to do some further reading and exercises.

1 Probability Theory

Chapter 2, sections 2.1–2.3 of David MacKay's book covers this material:
`http://www.inference.phy.cam.ac.uk/mackay/itila/book.html`

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

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

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

$P(A\!=\!a, B\!=\!b)$ is the joint probability that both $A\!=\!a$ and $B\!=\!b$ occur. — Joint Probability

Variables can be "summed out" of joint distributions: — Marginalisation

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

$P(A\!=\!a|B\!=\!b)$ is the probability $A\!=\!a$ occurs given the knowledge $B\!=\!b$. — Conditional Probability

$P(A\!=\!a, B\!=\!b) = P(A\!=\!a)\,P(B\!=\!b|A\!=\!a) = P(B\!=\!b)\,P(A\!=\!a|B\!=\!b)$ — Product Rule

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

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

Otherwise the product rule above *must* be used.

Bayes rule can be derived from the above: — Bayes Rule

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

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.

All the above theory basically still applies to continuous variables if sums are converted into integrals[1]. 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 $[0,\infty]$.

Continuous variables

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

Continuous versions of some results

The expectation or mean under a probability distribution is:

Expectations

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

2 Linear Algebra

This is designed as a prequel to Sam Roweis's "matrix identities" sheet:
`http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf`

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

$$x = 3.4, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}$$

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

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

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

Multiplication

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

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

Check Dimensions

$\mathbf{x}^\top\mathbf{x}$	$\mathbf{x}\mathbf{x}^\top$	$A\mathbf{x}$	$AA^\top$	$A^\top A$	$\mathbf{x}^\top A\mathbf{x}$
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?

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

Multiplication by scalar

Simple and valid manipulations:

Easily proved results

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

Note that $AB \neq BA$ in general.

[1] Integrals 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 $x_i = a + i\Delta x$. Find an A-level text book with some diagrams if you have not seen this before.

2.1 Square Matrices

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.

Diagonal matrices, the Identity

$$\begin{aligned} B_{ij} = 0 \text{ if } i \neq j \quad &\Leftrightarrow \quad \text{"}B\text{ is diagonal"} \\ \mathbb{I}_{ij} = 0 \text{ if } i \neq j \text{ and } \mathbb{I}_{ii} = 1 \;\; \forall i \quad &\Leftrightarrow \quad \text{"}\mathbb{I}\text{ is the identity matrix"} \\ \mathbb{I}\mathbf{x} = \mathbf{x} \qquad \mathbb{I}B = B = B\mathbb{I} \qquad &\mathbf{x}^\top \mathbb{I} = \mathbf{x}^\top \end{aligned}$$

Some square matrices have inverses:

Inverses

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

which have these properties:

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

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

Solving Linear equations

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

Some other commonly used matrix definitions include:

Symmetry

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

Trace

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

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

A Trace Trick

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

The determinant[2] is written $\mathrm{Det}(B)$ or $|B|$. It is a scalar regardless of n.

Determinants

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

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 $|B|$. It also appears in the normalising constant of a Gaussian. 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.

Eigenvalues, Eigenvectors

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

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

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).

[2] This section is only intended to give you a flavour so you understand other references and Sam's crib sheet. More detailed history and overview is here: `http://www.wikipedia.org/wiki/Determinant`

3 Differentiation

Any good A-level maths text book should cover this material and have plenty of exercises. Undergraduate text books might cover it quickly in less than a chapter.

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

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: Differentiation

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

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

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

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 Optimisation

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

If we can't solve this 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.

A function may be approximated by a straight line[3] about any point a. Approximation

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

The derivative operator is linear: Linearity

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

Dealing with products is slightly more involved: Product Rule

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

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 results to be combined. Chain Rule

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

If you can't show the following you could do with some practice: *Exercise*

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{(b+cz)}\exp(az) + e\right] = \exp(az)\left(\frac{a}{b+cz} - \frac{c}{(b+cz)^2}\right)$$

Note that a, b, c and e are constants, that $\frac{1}{u} = u^{-1}$ and this is hard if you haven't done differentiation (for a long time). Again, get a text book.

[3] More accurate approximations can be made. Look up Taylor series.