Appendix A

Series and Complex Numbers

A.1 Power series

A function of a variable x can often be written in terms of a series of powers of x. For the sin function, for example, we have

$$\sin x = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... \tag{A.1}$$

We can find out what the appropriate coefficients are as follows. If we substitite $x = 0$ into the above equation we get $a_0 = 0$ since $sin 0 = 0$ and all the other terms disappear. If we now *differentiate* both sides of the equation and substitute $x = 0$ we get $a_1 = 1$ (because $\cos 0 = 1 = a_1$). Differentiating twice and setting $x = 0$ gives $a_2 = 0$. Continuing this process gives

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ... \tag{A.2}$$

Similarly, the series representations for $cos x$ and e^x can be found as

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + ... \tag{A.3}$$

and

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ... \tag{A.4}$$

More generally, for a function $f(x)$ we get the general result

$$f(x) = f(0) + x f'(0) + \frac{x^2}{2!} f''(0) + \frac{x^3}{3!} f'''(0) + ... \tag{A.5}$$

where $f'(0)$, $f''(0)$ and $f'''(0)$ are the first, second and third derivatives of $f(x)$ evaluated at $x = 0$. This expansion is called a *Maclaurin series*.

So far, to calculate the coefficients in the series we have differentiated and substituted $x = 0$. If, instead, we substitute $x = a$ we get

$$f(x) = f(a) + (x-a) f'(a) + \frac{(x-a)^2}{2!} f''(a) + \frac{(x-a)^3}{3!} f'''(a) + ... \tag{A.6}$$

which is called a *Taylor series.*

For a d-dimensional vector of parameters $\boldsymbol{x}$ the equivalent Taylor series is

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + (\boldsymbol{x} - \boldsymbol{a})^T \boldsymbol{g} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^T \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{a}) + ... \tag{A.7}$$

where

$$\boldsymbol{g} = [\partial f/\partial a_1, \partial f/\partial a_2, ..., \partial f/\partial a_d]^T \tag{A.8}$$

is the gradient vector and

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial f^2}{\partial a_1^2} & \frac{\partial f^2}{\partial a_1 \partial a_2} & .. & \frac{\partial f^2}{\partial a_1 \partial a_d} \\ \frac{\partial f^2}{\partial a_2 \partial a_1} & \frac{\partial f^2}{\partial a_2^2} & .. & \frac{\partial f^2}{\partial a_2 \partial a_d} \\ .. & .. & .. & .. \\ \frac{\partial f^2}{\partial a_d \partial a_1} & \frac{\partial f^2}{\partial a_d \partial a_2} & .. & \frac{\partial f^2}{\partial a_d^2} \end{bmatrix} \tag{A.9}$$

is the Hessian.

A.2 Complex numbers

Very often, when we try to find the roots of an equation [1], we may end up with our solution being the square root of a negative number. For example, the quadratic equation

$$ax^2 + bx + c = 0 \tag{A.10}$$

has solutions which may be found as follows. If we divide by a and *complete the square* [2] we get

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} = \frac{-c}{a} \tag{A.11}$$

Re-arranging gives the general solution

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{A.12}$$

Now, if $b^2 - 4ac < 0$ we are in trouble. What is the square root of a negative number ? To handle this problem, mathematicians have defined the number

$$i = \sqrt{-1} \tag{A.13}$$

allowing all square roots of negative numbers to be defined in terms of i, eg $\sqrt{-9} = \sqrt{9}\sqrt{-1} = 3i$. These numbers are called *imaginary numbers* to differentiate them from *real numbers.*

[1] We may wish to do this in a signal processing context in, for example, an autoregressive model, where, given a set of AR coefficients we wish to see what signals (ie. x) correspond to the AR model. See later in this chapter.

[2] This means re-arranging a term of the form $x^2 + kx$ into the form $(x + \frac{k}{2})^2 - \left(\frac{k}{2}\right)^2$ which is often convenient because x appears only once.

Finding the roots of equations, eg. the quadratic equation above, requires us to combine imaginary numbers and real numbers. These combinations are called *complex numbers*. For example, the equation

$$x^2 - 2x + 2 = 0 \tag{A.14}$$

has the solutions $x = 1 + i$ and $x = 1 - i$ which are complex numbers.

A complex number $z = a + bi$ has two components; a real part and an imaginary part which may be written

$$\begin{aligned} a &= Re\{z\} \\ b &= Im\{z\} \end{aligned} \tag{A.15}$$

The *absolute value* of a complex number is

$$R = Abs\{z\} = \sqrt{a^2 + b^2} \tag{A.16}$$

and the *argument* is

$$\theta = Arg\{z\} = \tan^{-1}\left(\frac{b}{a}\right) \tag{A.17}$$

The two numbers $z = a + bi$ and $z^* = a - bi$ are known as *complex conjugates*; one is the complex conjugate of the other. When multiplied together they form a real number. The roots of equations often come in complex conjugate pairs.

A.3 Complex exponentials

If we take the exponential function of an imaginary number and write it out as a series expansion, we get

$$e^{i\theta} = 1 + \frac{i\theta}{1!} + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + ... \tag{A.18}$$

By noting that $i^2 = -1$ and $i^3 = i^2 i = -i$ and similarly for higher powers of i we get

$$e^{i\theta} = \left[1 - \frac{\theta^2}{2!} + ...\right] + i\left[\frac{\theta}{1!} - \frac{\theta^3}{3!} + ...\right] \tag{A.19}$$

Comparing to the earlier expansions of $\cos\theta$ and $\sin\theta$ we can see that

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{A.20}$$

which is known as *Euler's formula*. Similar expansions for $e^{-i\theta}$ give the identity

$$e^{-i\theta} = \cos\theta - i\sin\theta \tag{A.21}$$

We can now express the sine and cosine functions in terms of complex exponentials

$$\begin{aligned} \cos\theta &= \frac{e^{i\theta} + e^{-i\theta}}{2} \\ \sin\theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i} \end{aligned} \tag{A.22}$$

A.4 DeMoivre's Theorem

By using the fact that

$$e^{i\theta}e^{i\theta} = e^{i\theta+i\theta} \tag{A.23}$$

(a property of the exponential function and exponents in general eg. $5^3 5^3 = 5^6$) or more generally

$$(e^{i\theta})^k = e^{ik\theta} \tag{A.24}$$

we can write

$$(cos\theta + i\sin\theta)^k = cosk\theta + isink\theta \tag{A.25}$$

which is known as *DeMoivre's theorem.*

A.5 Argand Diagrams

Any complex number can be represented as a complex exponential

$$a + bi = Re^{i\theta} = R(cos\theta + i\sin\theta) \tag{A.26}$$

and drawn on an *Argand diagram.*

Multiplication of complex numbers is equivalent to rotation in the complex plane (due to DeMoivre's Theorem).

$$(a+bi)^2 = R^2 e^{i2\theta} = R^2(cos2\theta + i\sin 2\theta) \tag{A.27}$$

Appendix B

Linear Regression

B.1 Univariate Linear Regression

We can find the slope a and offset b by minising the cost function

$$E = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \tag{B.1}$$

Differentiating with respect to a gives

$$\frac{\partial E}{\partial a} = -2 \sum_{i=1}^{N} x_i (y_i - ax_i - b) \tag{B.2}$$

Differentiating with respect to b gives

$$\frac{\partial E}{\partial b} = -2 \sum_{i=1}^{N} (y_i - ax_i - b) \tag{B.3}$$

By setting the above derivatives to zero we obtain the *normal equations* of the regression. Re-arranging the normal equations gives

$$a \sum_{i=1}^{N} x_i^2 + b \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \tag{B.4}$$

and

$$a \sum_{i=1}^{N} x_i + bN = \sum_{i=1}^{N} y_i \tag{B.5}$$

By substituting the mean observed values μ_x and μ_y into the last equation we get

$$b = \mu_y - a\mu_x \tag{B.6}$$

Now let

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{N} (x_i - \mu_x)^2 && \text{(B.7)} \\ &= \sum_{i=1}^{N} x_i^2 - N\mu_x^2 \end{aligned}$$

(B.8)

and

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) & \text{(B.9)} \\
&= \sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y
\end{aligned}
$$

$$\text{(B.10)}$$

Substiting for b into the first normal equation gives

$$a\sum_{i=1}^{N} x_i^2 + (\mu_y - a\mu_x)\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \tag{B.11}$$

Re-arranging gives

$$
\begin{aligned}
a &= \frac{\sum_{i=1}^{N} x_i y_i - \mu_y \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 + \mu_x \sum_{i=1}^{N} x_i} & \text{(B.12)} \\
&= \frac{\sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y}{\sum_{i=1}^{N} x_i^2 + N\mu_x^2} \\
&= \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \\
&= \frac{\sigma_{xy}}{\sigma_x^2}
\end{aligned}
$$

B.1.1 Variance of slope

The data points may be written as

$$
\begin{aligned}
y_i &= \hat{y}_i + e_i & \text{(B.13)} \\
&= ax_i + b + e_i
\end{aligned}
$$

where the noise, e_i has mean zero and variance σ_e^2. The mean and variance of each data point are

$$E(y_i) = ax_i + b \tag{B.14}$$

and

$$Var(y_i) = Var(e_i) = \sigma_e^2 \tag{B.15}$$

We now calculate the variance of the estimate a. From earlier we see that

$$a = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{B.16}$$

Let

$$c_i = \frac{(x_i - \mu_x)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{B.17}$$

We also note that $\sum_{i=1}^{N} c_i = 0$ and $\sum_{i=1}^{N} c_i x_i = 1$. Hence,

$$\begin{aligned} a &= \sum_{i=1}^{N} c_i (y_i - \mu_y) && \text{(B.18)} \\ &= \sum_{i=1}^{N} c_i y_i - \mu_y \sum_{i=1}^{N} c_i \\ & && \text{(B.19)} \end{aligned}$$

The mean estimate is therefore

$$\begin{aligned} E(a) &= \sum_{i=1}^{N} c_i E(y_i) - \mu_y \sum_{i=1}^{N} c_i && \text{(B.20)} \\ &= a \sum_{i=1}^{N} c_i x_i + b \sum_{i=1}^{N} c_i - \mu_y \sum_{i=1}^{N} c_i \\ &= a \\ & && \text{(B.21)} \end{aligned}$$

The variance is

$$Var(a) = Var(\sum_{i=1}^{N} c_i y_i - \mu_y \sum_{i=1}^{N} c_i) \tag{B.22}$$

The second term contains two fixed quantities so acts like a constant. From the later Appendix on Probability Distributions we see that

$$\begin{aligned} Var(a) &= Var(\sum_{i=1}^{N} c_i y_i) && \text{(B.23)} \\ &= \sum_{i=1}^{N} c_i^2 Var(y_i) \\ &= \sigma_e^2 \sum_{i=1}^{N} c_i^2 \\ &= \frac{\sigma_e^2}{\sum_{i=1}^{N} (x_i - \mu_x)^2} \\ &= \frac{\sigma_e^2}{(N-1)\sigma_x^2} \end{aligned}$$

B.2 Multivariate Linear Regression

B.2.1 Estimating the weight covariance matrix

Different instantiations of target noise will generate different estimated weight vectors according to the last equation. The corresponding weight covariance matrix is given by

$$Var(\hat{\boldsymbol{w}}) = Var((\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}) \tag{B.24}$$

Substituting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{e}$ gives

$$Var(\hat{\boldsymbol{w}}) = Var((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}) \tag{B.25}$$

This is in the form of equation B.28 in Appendix A with $\boldsymbol{d}$ being given by the first term which is constant, $\boldsymbol{C}$ being given by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{z}$ being given by $\boldsymbol{e}$. Hence,

$$\begin{aligned} Var(\hat{\boldsymbol{w}}) &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[Var(\boldsymbol{e})][(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned} \tag{B.26}$$

Re-arranging further gives

$$Var(\hat{\boldsymbol{w}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{B.27}$$

B.3 Functions of random vectors

For a vector of random variables, $\boldsymbol{z}$, and a matrix of constants, $\boldsymbol{C}$, and a vector of constants, $\boldsymbol{d}$, we have

$$Var(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{d}) = \boldsymbol{C}[Var(\boldsymbol{z})]\boldsymbol{C}^T \tag{B.28}$$

where, here, Var() denotes a covariance matrix. This is a generalisation of the result for scalar random variables $Var(cz) = c^2 Var(z)$.

The covariance between a pair of random vectors is given by

$$Var(\boldsymbol{C}_1\boldsymbol{z}, \boldsymbol{C}_2\boldsymbol{z}) = \boldsymbol{C}_1[Var(\boldsymbol{z})]\boldsymbol{C}_2^T \tag{B.29}$$

B.3.1 Estimating the weight covariance matrix

Different instantiations of target noise will generate different estimated weight vectors according to the equation 3.7. The corresponding weight covariance matrix is given by

$$\boldsymbol{\Sigma} = Var((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \tag{B.30}$$

Substituting $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{w}} + \boldsymbol{e}$ gives

$$\boldsymbol{\Sigma} = Var((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}) \tag{B.31}$$

This is in the form of $Var(\boldsymbol{C}\boldsymbol{z}+\boldsymbol{d})$ (see earlier) with $\boldsymbol{d}$ being given by the first term which is constant, $\boldsymbol{C}$ being given by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{z}$ being given by $\boldsymbol{e}$. Hence,

$$\begin{aligned}
\boldsymbol{\Sigma} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[Var(\boldsymbol{e})][(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \qquad \text{(B.32)} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma_e^2\boldsymbol{I})[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma_e^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

Re-arranging further gives

$$\boldsymbol{\Sigma} = \sigma_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{B.33}$$

B.3.2 Equivalence of t-test and F-test for feature selection

When adding a new variable x_p to a regression model we can test to see if the increase in the proportion of variance explained is *significant* by computing

$$F = \frac{(N-1)\sigma_y^2\,[r^2(y,\hat{y}_p) - r^2(y,\hat{y}_{p-1})]}{\sigma_e^2(p)} \tag{B.34}$$

where $r^2(y,\hat{y}_p)$ is the square of the correlation between y and the regression model with all p variables (ie. including x_p) and $r^2(y,\hat{y}_{p-1})$ is the square of the correlation between y and the regression model without x_p. The denominator is the noise variance from the model including x_p. This statistic is distributed according to the F-distribution with $v_1 = 1$ and $v_2 = N-p-2$ degrees of freedom.

This test is identical to the double sided t-test on the t-statistic computed from the regression coefficient a_p, described in this lecture (see also page 128 of [32]). This test is also equivalent to seeing if the partial correlation between x_p and y is significantly non-zero (see page 149 of [32]).

Appendix C

Matrix Identities

C.1 Multiplication

Matrix multiplication is associative

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \tag{C.1}$$

distributive

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC} \tag{C.2}$$

but not commutative

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{C.3}$$

C.2 Transposes

Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we have

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T \tag{C.4}$$

C.3 Inverses

Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we have

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1} \boldsymbol{A}^{-1} \tag{C.5}$$

The Matrix Inversion Lemma is

$$(\boldsymbol{XBX}^T + \boldsymbol{A})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}(\boldsymbol{B}^{-1} + \boldsymbol{X}^T\boldsymbol{A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^{-1} \tag{C.6}$$

The Sherman-Morrison-Woodury formula or Woodbury's identity is

$$(\boldsymbol{UV}^T + \boldsymbol{A})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1} \tag{C.7}$$

C.4 Eigendecomposition

$$\boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q} = \boldsymbol{\Lambda} \tag{C.8}$$

Pre-multiplying by $\boldsymbol{Q}$ and post-multiplying by $\boldsymbol{Q}^T$ gives

$$\boldsymbol{A} = \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^T \tag{C.9}$$

which is known as the *spectral theorem*. Any real, symmetric matrix can be represented as above where the columns of $\boldsymbol{Q}$ contain the eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues, λ_i. Equivalently,

$$\boldsymbol{A} = \sum_{k=1}^{d} \lambda_k \boldsymbol{q}_k \boldsymbol{q}_k^T \tag{C.10}$$

C.5 Determinants

If $\det(\boldsymbol{A}) = 0$ the matrix $\boldsymbol{A}$ is not invertible; it is *singular*. Conversely, if $\det(\boldsymbol{A}) \neq 0$ then $\boldsymbol{A}$ *is* invertible. Other properties of the determinant are

$$\begin{aligned} \det(\boldsymbol{A}^T) &= \det(\boldsymbol{A}) \\ \det(\boldsymbol{A}\boldsymbol{B}) &= \det(\boldsymbol{A}) \det(\boldsymbol{B}) \\ \det(\boldsymbol{A}^{-1}) &= 1/\det(\boldsymbol{A}) \\ \det(\boldsymbol{A}) &= \prod_k a_{kk} \det(\boldsymbol{A}) = \prod_k \lambda_k \end{aligned} \tag{C.11}$$

C.6 Traces

The *Trace* is the sum of the diagonal elements

$$Tr(\boldsymbol{A}) = \sum_k a_{kk} \tag{C.12}$$

and is also equal to the sum of the eigenvalues

$$Tr(\boldsymbol{A}) = \sum_k \lambda_k \tag{C.13}$$

Also

$$Tr(\boldsymbol{A} + \boldsymbol{B}) = Tr(\boldsymbol{A}) + Tr(\boldsymbol{B}) \tag{C.14}$$

C.7 Matrix Calculus

From [37] we know that the derivative of $\boldsymbol{c}^T \boldsymbol{B} \boldsymbol{c}$ with respect to $\boldsymbol{c}$ is $(\boldsymbol{B}^T + \boldsymbol{B})\boldsymbol{c}$.