Appendix - Mathematical Methods

11.1 Introduction

The book assumes a familiarity with the methods of basic linear algebra, differential equations and probability theory, as covered by many standard texts in mathematical methods for physicists. This chapter discusses the slightly more exotic mathematical methods used and defines some of the notation employed in the text. Explanations and theory are also sketchy, and completeness and elegance are sacrificed for convenience.

11.2 Linear Algebra

The definitions and styles of the basic quantities in linear algebra are:

vectors are written as $\mathbf{a}$ with components $[\mathbf{c}]_k = c_k$ and are usually treated *column-wise.*

matrices are written as $\mathbf{K}$ with elements $\mathbf{K}_{kl}$. Some important quantities for square matrices include:

Quantity	Notation	Definition
identity matrix	$\mathbb{I}$	δ_{kl}
trace	$\mathrm{tr}(\mathbf{K})$	$\sum_i \mathbf{K}_{kk}$
determinant	$\det(\mathbf{K})$	see references
inverse	$\mathbf{K}^{-1}$	$\mathbf{K}\mathbf{K}^{-1} = \mathbb{I}$

where

$$\delta_{kl} = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases} \tag{11.1}$$

The determinant is sometimes written $|\mathbf{K}| = |\det(\mathbf{K})|$.

products of vectors with each other and with matrices are written without transpose signs unless this would be confusing, so:

Quantity	Definition			Aliases
dot product	$\mathbf{c}\cdot\mathbf{d}$	$=$	$\sum_k c_k d_k$	$\mathbf{c}^T\mathbf{d}$
norm	$\lvert\mathbf{c}\rvert^2$	$=$	$\mathbf{c}\cdot\mathbf{c} = \sum_k c_k^2$	$\lVert\mathbf{c}\rVert^2$
vector-matrix product	$[\mathbf{K}\cdot\mathbf{c}]_k$	$=$	$\sum_l \mathbf{K}_{kl} c_l$	$\mathbf{Kc}$
vector-matrix product	$[\mathbf{c}\cdot\mathbf{K}]_l$	$=$	$\sum_k c_k \mathbf{K}_{kl}$	$\mathbf{c}^T\mathbf{K}$
quadratic form	$\mathbf{c}\cdot\mathbf{K}\cdot\mathbf{d}$	$=$	$\sum_{kl} c_k \mathbf{K}_{kl} d_l$	$\mathbf{c}^T\mathbf{Kd}$
outer product	$[\mathbf{cd}]_{kl}$	$=$	$c_k d_l$	$\mathbf{cd}^T$
matrix-matrix product	$[\mathbf{KL}]_{kl}$	$=$	$\sum_m \mathbf{K}_{km}\mathbf{L}_{ml}$	

An *eigenvector* of a square matrix $\mathbf{K}$ is a non-zero vector $\mathbf{e}$ such that

$$\mathbf{K}\cdot\mathbf{e} = \lambda\mathbf{e}. \tag{11.2}$$

Factor λ is called the *eigenvalue* of $\mathbf{e}$ and satisfies the polynomial equation:

$$\det(\mathbf{K} - \lambda\mathbf{I}) = 0 \tag{11.3}$$

Even if $\mathbf{K}$ has only real-valued entries, the eigenvectors and eigenvalues may still come in complex conjugate pairs.

If $\mathbf{e}$ is an eigenvector, then $\epsilon\mathbf{e}$ is also an eigenvector for any $0 \neq \epsilon \in \Re$. It is often convenient to use normalised eigenvectors, for which $\lvert\mathbf{e}\rvert = 1$. If two eigenvectors $\mathbf{e}^1, \mathbf{e}^2$ have the same eigenvalue λ, then $\alpha\mathbf{e}^1 + \beta\mathbf{e}^2$ is also an eigenvector with the same eigenvalue, for any $\alpha, \beta \in \Re$. Apart from such degeneracies, a square matrix $\mathbf{K}$ operating over $\Re^n$ can have at most n eigenvectors, although some matrices have fewer. If $\mathbf{K}$ has exactly n non-degenerate eigenvectors, then they form a basis for $\Re^n$, so that any vector $\mathbf{c}$ can be written as

$$\mathbf{c} = \sum_a \alpha_a \mathbf{e}^a \tag{11.4}$$

for some unique set of real coefficients α^a. Furthermore, the representation of $\mathbf{K}$ in the eigenvector basis is given by

$$\Gamma^{-1}\mathbf{K}\Gamma = \Gamma^{-1}\left(\lambda^1\mathbf{e}^1\ \lambda^2\mathbf{e}^2\ \ldots\ \lambda^n\mathbf{e}^n\right) = \begin{pmatrix} \lambda^1 & 0 & \ldots & 0 \\ 0 & \lambda^2 & \ldots & 0 \\ & & \vdots & \\ 0 & 0 & \ldots & \lambda^n \end{pmatrix}. \tag{11.5}$$

where the columns of $\Gamma = \left(\mathbf{e}^1\ \mathbf{e}^2\ \ldots\ \mathbf{e}^n\right)$ consist of the eigenvectors. Write $\mathrm{diag}(\lambda^1, \ldots, \lambda^n)$ for the diagonal matrix. Then a diagonalised form of $\mathbf{K}$ is:

$$\mathbf{K} = \Gamma\ \mathrm{diag}(\lambda^1, \ldots, \lambda^n)\ \Gamma^{-1}, \tag{11.6}$$

and, if none of the eigenvalues of $\mathbf{K}$ is 0, the diagonalised form of $\mathbf{K}^{-1}$ as:

$$\mathbf{K} = \Gamma\ \mathrm{diag}(1/\lambda^1, \ldots, 1/\lambda^n)\ \Gamma^{-1}. \tag{11.7}$$

If 0 is an eigenvalue of $\mathbf{K}$, then it is not invertible. Matrices $\mathbf{K}$ and $\mathbf{K}^T$ have the same eigenvalues but, in general, have different eigenvectors. Not all matrices have a full complement of eigenvectors and therefore not all matrices are diagonalisable.

Symmetric matrices (for which $\mathbf{K}^T = \mathbf{K}$) have particularly convenient eigendecompositions. First, their eigenvalues and vectors are always real-valued. Second, an *orthonormal* set of eigenvectors $\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^n$ for $\mathbf{K}$ can be chosen, meaning that $\Gamma^{-1} = \Gamma^T$. In this case, we can represent arbitrary vectors $\mathbf{c}$ in this basis as:

$$\mathbf{c} = \sum_{a=1}^{n} \left(\mathbf{c} \cdot \mathbf{e}^a\right) \mathbf{e}^a \tag{11.8}$$

making it easy to calculate the components in the direction of each eigenvector. Finally, we can calculate

$$\mathbf{Kc} = \mathbf{K}\left(\sum_{a=1}^{n} \left(\mathbf{c} \cdot \mathbf{e}^a\right) \mathbf{e}^a\right) = \sum_{a=1}^{n} \left(\mathbf{c} \cdot \mathbf{e}^a\right) \lambda^a \mathbf{e}^a, \tag{11.9}$$

and so the effect of $\mathbf{K}$ is to boost components in the direction of eigenvectors with large eigenvalues. A symmetric matrix is positive definite if and only if all its eigenvalues are greater than 0. It is positive semi-definite if and only if all its eigenvalues are greater than or equal to 0.

Töplitz matrices are another class for which the eigendecomposition is particularly simple. Matrix $\mathbf{K}$ is Töplitz if $\mathbf{K}_{kl}$ is a function only of $(k-l) \bmod n$. To avoid the modulus, we will write $\mathbf{K}_{kl} = \mathcal{K}(k-l)$ where $\mathcal{K}$ satisfies

$$\mathcal{K}(-k) = \mathcal{K}(n-k), \ \forall a. \tag{11.10}$$

For instance, a $4d$ Töplitz matrix has entries:

$$\mathbf{K} = \begin{pmatrix} \mathcal{K}(0) & \mathcal{K}(-1) & \mathcal{K}(-2) & \mathcal{K}(-3) \\ \mathcal{K}(1) & \mathcal{K}(0) & \mathcal{K}(-1) & \mathcal{K}(-2) \\ \mathcal{K}(2) & \mathcal{K}(1) & \mathcal{K}(0) & \mathcal{K}(-1) \\ \mathcal{K}(3) & \mathcal{K}(2) & \mathcal{K}(1) & \mathcal{K}(0) \end{pmatrix} = \begin{pmatrix} \mathcal{K}(0) & \mathcal{K}(3) & \mathcal{K}(2) & \mathcal{K}(1) \\ \mathcal{K}(1) & \mathcal{K}(0) & \mathcal{K}(3) & \mathcal{K}(2) \\ \mathcal{K}(2) & \mathcal{K}(1) & \mathcal{K}(0) & \mathcal{K}(3) \\ \mathcal{K}(3) & \mathcal{K}(2) & \mathcal{K}(1) & \mathcal{K}(0) \end{pmatrix} \tag{11.11}$$

If $\mathcal{K}(k) = \mathcal{K}(n-k)$ for all k, then the Töplitz matrix is also symmetric. Töplitz matrices are important class since they are routinely used to describe spatially and temporally stationary linear filters.

The eigenvectors of general Töplitz matrices can be written as complex exponentials:

$$[\mathbf{e}^a]_k = e^{\imath\omega^a k} = \cos\left(\omega^a k\right) + \imath \sin\left(\omega^a k\right) \tag{11.12}$$

where $\omega^a = 2\pi\frac{a}{n}$ for $a = 1 \ldots n$ and $\imath = \sqrt{-1}$. We can show this directly, for then:

$$[\mathbf{K}\cdot\mathbf{e}^a]_k = \sum_{l=1}^{n} \mathbf{K}_{kl} e^{\imath\omega^a l} = e^{\omega^a k} \sum_{l=1}^{n} \mathcal{K}(k-l) e^{\imath\omega^a (l-k)}$$
$$= e^{\omega^a k} \sum_{p=k-1}^{k-n} \mathcal{K}(p) e^{-\imath\omega^a p}. \qquad (11.13)$$

The sum is independent of k if $e^{-\imath\omega^a p}$ satisfies the same Töplitz condition as $\mathcal{K}$ in equation 11.10, namely that

$$e^{-\imath\omega^a p} = e^{-\imath\omega^k (p-n)}. \qquad (11.14)$$

This requires that $e^{\imath\omega^a n} = 1$, making $\omega^a = 2\pi\frac{a}{n}$ for $a = 1 \ldots n$ an n^{th} root of unity.

By inspection of equation 11.13, the eigenvalues of Töplitz $\mathbf{K}$ are closely related to the components of the discrete-time Fourier transform of the function $\mathcal{K}$ that generates $\mathbf{K}$. This is a very useful result that occurs in general settings.

In a well defined limit as $n \to \infty$, one abandons the toroidal boundary conditions, and writes $\mathbf{K}(k,l) = \mathcal{K}(k-l)$ as a function of continuous rather than discrete indices. For this, the equivalent of the eigenvectors are now eigenfunctions $\mathbf{e}^\omega(k) = e^{\imath\omega k}$ where ω is continuous rather than discrete, and the eigenspectrum is exactly the standard continuous Fourier transform of $\mathcal{K}(k)$.

11.3 Differential Equations

A first order homogeneous autonomous linear differential equation in n variables $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots x_n(t))^T$ can be written as

$$\dot{\mathbf{x}}(t) = \mathbf{K}\cdot\mathbf{x}(t) \qquad (11.15)$$

for a matrix $\mathbf{K}$. If $\mathbf{K}$ is real and has n linearly independent eigenvectors $\mathbf{e}^1, \ldots, \mathbf{e}^n$ with different eigenvalues $\lambda^1, \ldots, \lambda^n$, then the time evolution of $\mathbf{x}(t)$ is best understood by using them as a basis. If $\mathbf{x}(0) = \sum_{a=1}^n \alpha_a \mathbf{e}^a$, then

$$\mathbf{x}(t) = \sum_{a=1}^{n} \alpha_a e^{\lambda^a t} \mathbf{e}^a \qquad (11.16)$$

satisfies equation 11.15 by direct substitution. The different $\mathbf{e}^a$ are called *modes* (or sometimes *eigenmodes*) of the system and their growth is plainly decoupled from each other. Note that this does not require the eigenvectors to be orthogonal. If the eigenvalues and eigenvectors are complex,

$\mathbf{x}(t)$ will nevertheless remain real-valued if $\mathbf{x}(0)$ is real-valued, since the conjugate pairs cancel appropriately. If some of the eigenvalues are equal, then this expression is no longer true – consult the references for the solution in this case.

Given this form of the solution, the qualitative properties of the dynamics given by equation 11.15 can be deduced by studying the eigenvalues of $\mathbf{K}$. If $\lambda^a = \mu^a + \imath\nu^a$, where $\imath = \sqrt{-1}$, then

$$e^{\lambda^a t} = e^{\mu^a t}\left(\cos(\nu^a t) + \imath \sin(\nu^a t)\right) \tag{11.17}$$

Therefore the real part of λ^a controls the growth of the amplitude of the mode; and the complex part the nature of any oscillations. If the μ^a are ordered so that $\mu^1 > \mu^2 > \ldots > \mu^n$ (ignoring equality, for convenience), then component $\mathbf{e}^1$ will come to dominate relative to the other $\mathbf{e}^i$. Furthermore, if $\mu^a > 0$ for any i, then the system is unstable, since $|\mathbf{x}(t)| \to \infty$ as $t \to \infty$ except possibly for very particular starting conditions. If $\mu^a < 0$ for all a, then $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$. If $\nu^a \neq 0$ then the system will oscillate about $\mathbf{0}$ in mode $\mathbf{e}^a$. If $\mathbf{K}$ is symmetric, then we are guaranteed that all its eigenvalues are real ($\nu^a = 0, \forall a$), and therefore that there will be no oscillations.

If $\mathbf{K}$ is symmetric and negative definite, so $\mathbf{x} \cdot \mathbf{K} \cdot \mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$, then we can define the function

$$E(\mathbf{x}(t)) = -\frac{1}{2}\mathbf{x}(t) \cdot \mathbf{K} \cdot \mathbf{x}(t). \tag{11.18}$$

The dynamics of equation 11.15 are such that

$$\dot{\mathbf{x}}(t) = -\nabla_{\mathbf{x}(t)} E(\mathbf{x}(t)) \tag{11.19}$$

and so

$$\dot{E}(\mathbf{x}(t)) = \dot{\mathbf{x}}(t) \cdot \nabla_{\mathbf{x}(t)} E(\mathbf{x}(t)) = -|\nabla_{\mathbf{x}(t)} E(\mathbf{x}(t))|^2 \leq 0. \tag{11.20}$$

This means that $E(\mathbf{x}(t))$ strictly decreases, since $\nabla_{\mathbf{x}(t)} E(\mathbf{x}(t))$ is non-zero if $\mathbf{x}(t)$ is non-zero. Since $E(\mathbf{x}(t))$ is positive, bounded below (by 0) and decreases with the dynamics, it is called a *Lyapunov* function. Lyapunov functions are used to analyse the stability of differential equations, showing, for instance, that we can make $|\mathbf{x}(t)| < \epsilon$ for all t provided that we start with $|\mathbf{x}(0)| < \delta$ for some particular $\delta > 0$. Since matrix $\mathbf{K}$ is symmetric, all its eigenvalues are real. Since, further, it is negative definite, all the eigenvalues must be negative. This is just the condition we saw above as making $\mathbf{x}(t) \to \mathbf{0}$. Although $E(\mathbf{x}(t))$ was chosen so that the dynamics perform *gradient descent* on it (as seen in equation 11.19), it is possible to define Lyapunov functions that are not just direct integrals of the underlying dynamics.

Linear analysis tools can also be used to understand part of the behavior of non-linear differential equations too. Consider the equation

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) \tag{11.21}$$

for some non-linear function $\mathbf{f}(\mathbf{x}(t)) : \Re^n \rightarrow \Re^n$. The *equilibrium* or *stationary* points of equation 11.21 are those points $\mathbf{s}$ for which $f(\mathbf{s}) = \mathbf{0}$, since if $\mathbf{x}(0) = \mathbf{s}$, then $\mathbf{x}(t) = \mathbf{s}$ for all t. In the linear equation above, the only equilibrium point is $\mathbf{s} = \mathbf{0}$, but non-linear functions can have multiple equilibria.

It is standard to analyse the behavior of equation 11.21 *near to* the equilibrium points, since they often collectively characterise well the behavior of the whole system. Consider writing

$$\mathbf{x}(t) = \mathbf{s} + \boldsymbol{\epsilon}(t) \tag{11.22}$$

where $\boldsymbol{\epsilon}(t)$ is small. Up to terms in first order,

$$\dot{\boldsymbol{\epsilon}}(t) = \mathbf{J}(\mathbf{s}) \cdot \boldsymbol{\epsilon}(t) \tag{11.23}$$

where $\mathbf{J}(\mathbf{x})$ is the *Jacobian* matrix of function $\mathbf{f}(\mathbf{x})$ with elements

$$\mathbf{J}(\mathbf{x})_{kl} = \frac{\partial f_k(\mathbf{x})}{\partial x_l}. \tag{11.24}$$

Therefore, we can understand the behavior of equation 11.21 about $\mathbf{s}$ by considering the eigenvalues of the Jacobian at $\mathbf{s}$ just as above. If they all have negative real parts, then $\boldsymbol{\epsilon}(t) \rightarrow \mathbf{0}$, and therefore $\mathbf{x}(t) \rightarrow \mathbf{s}$. If, in addition their complex parts are non-zero, then $\mathbf{x}(t)$ will tend to spiral in to $\mathbf{s}$, otherwise it will move radially. If any of the eigenvalues has a positive real part, then the equilibrium point is unstable. In non-linear systems, this does not necessarily imply that $\mathbf{x}(t) \rightarrow \infty$, since $\mathbf{x}(t)$ could, for instance, get 'trapped' by another equilibrium point or saturate.

We have therefore derived a range of behaviors near to the equilibrium points. Stable equilibrium points act as attractors, sucking nearby trajectories into themselves. The global behavior of the system, showing the way that the local dynamics is stitched together, can be much more complicated, and there is also the possibility of phenomena such as deterministic chaos.

We have so far considered the case in which there is no external input to the system. The solution to the more general one-dimensional differential equation

$$\dot{x}(t) = \lambda x(t) + f(t) \tag{11.25}$$

is

$$x(t) = e^{\lambda t} \int d\tau \, e^{-\lambda \tau} f(\tau) + \gamma e^{\lambda \tau} \tag{11.26}$$

for a constant γ chosen to match the initial conditions. When $\mathbf{K}$ is symmetric, and the driving function $\mathbf{f}(t)$ is a vector, the different modes decouple as in equation 11.16 to give a set of equations like those in equation 11.25, in which eigenmode a is driven by $\mathbf{e}^a \cdot \mathbf{f}(t)$.

In some cases, we will consider discrete rather than continuous dynamics. For example, this is a natural description in cases for which patterns are presented to a network sequentially and weights adapt on the basis of each pattern. A natural model for this is the difference equation:

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \mathbf{K}\mathbf{x}(n) + \mathbf{f}(n) \tag{11.27}$$

using n as an index rather than t.

The strategy for solving such difference equations is very similar to that for solving differential equations. We will only consider the simplest case of this without a driving function, so $\mathbf{f}(n) = \mathbf{0}$. We will also assume that $\mathbf{K}$ has a complete eigensystem with linearly independent eigenvectors $\mathbf{e}^1, \ldots, \mathbf{e}^n$ and (different) eigenvalues $\lambda^1, \ldots, \lambda^n$. Again, the eigenmodes separate, and the general solution is

$$\mathbf{x}(n) = \sum_{a=1}^{n} \alpha_a (1+\lambda^a)^T \mathbf{e}^a \tag{11.28}$$

where the multiplicative growth rate of mode a is $(1+\lambda^a)$ and $\mathbf{x}(0) = \sum_{a=1}^{n} \alpha_a \mathbf{e}^a$. This has some similar characteristics to the solution of the continuous differential equation in expression 11.16. If $\lambda^a = \mu^a + \imath\nu^a$, then, again, the system will be oscillatory in mode a if $\nu^a \neq 0$. However, the stability of the system is now governed by the magnitude of:

$$|1+\lambda^a|^2 = \left(1+\mu^a\right)^2 + \left(\nu^a\right)^2 . \tag{11.29}$$

If this is greater than 1, then $|\mathbf{x}(n)| \to \infty$ as $n \to \infty$.

Linearisation about equilibrium points can also be used for non-linear versions of equation 11.27 and similar classes of behavior about equilibrium points can be observed.

11.4 Probability Theory

Probabilities are based on sample spaces of possible outcomes (*eg* $\Omega = \{\omega\}$) over which there are one or more measures $p(\omega)$. Subsets of Ω are called *events*. If Ω is countable, the measures must satisfy $0 \leq p(\omega \leq 1$ and $\sum_\omega p(\omega) = 1$. We are mostly interested in *random variables*, which are functions $X(\omega)$ from Ω to some other space. We write

$$P[X = x] = \sum_{\substack{\omega \in \Omega: \\ X(\omega) = x}} p(\omega) \tag{11.30}$$

as the *distribution function* for X and often use $P[x]$ as a shorthand for $P[X = x]$. Some key statistics include:

Quantity	Definition	Alias
mean	$\langle X \rangle = \sum_x P[x]x$	$\mathcal{E}[X]$
variance	$\mathrm{var}(X) = \langle X^2 \rangle - \langle X \rangle^2$	$\mathcal{V}[X]$
covariance	$\mathrm{cov}(X_1, X_2) = \langle X_1 X_2 \rangle - \langle X_1 \rangle \langle X_2 \rangle$	
entropy	$\mathcal{H}[X] = -\sum P[x] \log P[x]$	

where X_1 and X_2 are two random variables defined over the same sample space, and, in the form for the entropy, substituting $P[x] \log P[x] = 0$ for $P[x] = 0$. Two random variables are *independent* if $P[X_1 = x_1, X_2 = x_2] = P[X_1 = x_1]P[X_2 = x_2]$ for all x_1 and x_2. The *conditional probability* that $X_1 = x_1$ given that $X_2 = x_2$, is the proportional measure of the subset of $\{\omega : X_2(\omega) = x_2\}$ for which $X_1(\omega) = x_1$ and is

$$P[X_1 = x_1 | X_2 = x_2] = \frac{P[X_1 = x_1, X_2 = x_2]}{P[X_2 = x_2]} \tag{11.31}$$

for $P[X_2 = x_2] \neq 0$, where $P[X_1 = x_1, X_2 = x_2]$ is the probability that $X_1 = x_1$ *and* $X_2 = x_2$. By straightforward algebra, conditional probabilities satisfy Bayes theorem:

$$P[X_1 = x_1 | X_2 = x_2] = P[X_2 = x_2 | X_1 = x_1] \frac{P[X_1 = x_1]}{P[X_2 = x_2]} \,. \tag{11.32}$$

The *Kullback-Leibler (*KL) *divergence* is a measure of the difference between two different distributions $P[X = x]$ and $Q[X = x]$ for the same random variable:

$$KL[Q[X], P[X]] = \sum_x Q[X = x] \log \left[\frac{Q[X = x]}{P[X = x]} \right] . \tag{11.33}$$

The KL divergence is always non-negative (see chapter 4), although it can be infinite if $P[X = x] = 0$ when $Q[X = x] \neq 0$. We will often drop the random variables and write $KL[Q, P]$. For discrete distributions, $KL[Q, P] = 0$ if and only if $Q[X = x] = P[X = x]$ for all x. Crucially (and almost by inspection), the KL divergence is not symmetric: $KL[Q, P] \neq KL[P, Q]$. This means that the KL divergence is not a *metric* on the space of distributions, even though it is often referred to as the KL distance.

The *mutual information* between two random variables X_1 and X_2 is defined as:

$$I(X_1, X_2) = \mathcal{H}[X_1] - \mathcal{H}[X_1 | X_2] = \mathcal{H}[X_2] - \mathcal{H}[X_2 | X_1] \tag{11.34}$$

and is a measure of their non-independence. Here $\mathcal{H}[X_1 | X_2]$ is defined as the entropy of the conditional distribution $P[X_1 | X_2 = x_2]$ *averaged* over the possible values of X_2:

$$\begin{aligned} &\mathcal{H}[X_1 | X_2] = \\ &-\sum_{x_2} P[X_2 = x_2] \sum_{x_1} P[X_1 = x_1 | X_2 = x_2] \log(P[X_1 = x_1 | X_2 = x_2]) . \end{aligned} \tag{11.35}$$

Knowing the value of X_2 can never increase the uncertainty about X_1, therefore $\mathcal{H}[X_2|X_1] \leq \mathcal{H}[X_1]$, with equality if and only if X_1 and X_2 are statistically independent. This means that $I(X_1, X_2) \geq 0$ with equality under similar conditions. The mutual information is also symmetric, unlike the KL divergence. One can also write the mutual information as

$$I(X_1, X_2) = KL[P[X_1, X_2], P[X_1]P[X_2]]. \tag{11.36}$$

indicating that the mutual information quantifies the divergence between the actual joint distribution $P[X_1, X_2]$ and the joint distribution they would have if they were independent $P[X_1]P[X_2]$.

Sample spaces can also be uncountable. In this case there are technical conditions on what subsets of Ω count as events and also on the measure $p(\omega)$ that weighs elements of the sample space. All the sums in the expressions for discrete sample spaces turn into integrations and the density function $p(x)$ defined by $P[X \in [x, x+\delta x]] \simeq p(x)\delta x$ replaces the distribution function. Chapter 4 discusses the problems of defining the entropy in this case.

Some common discrete and continuous distributions include:

Name	Range	Probability	Mean	Variance
Bernoulli	$x \in \{0, 1\}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$
Poisson	$x \in \{0, 1, \ldots\}$	$e^{-\lambda}\lambda^x/x$	λ	λ
Exponential	$x > 0$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
Gaussian	$\mathbf{x} \in \Re^n$	$\frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\cdot\Sigma^{-1}\cdot(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{n/2}\lvert\Sigma\rvert^{1/2}}$	$\boldsymbol{\mu}$	Σ
Cauchy	$x \in \Re$	$\frac{\beta/\pi}{(x-\alpha)^2+\beta^2}$	α	$\cdot$

where Σ is actually the whole covariance matrix $\langle(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})\rangle$ for the Gaussian, and the Cauchy distribution has such heavy tails that the integral defining its variance does not converge. Parameter β still controls the dispersion for the Cauchy distribution – the larger β, the heavier the tails.

The Gaussian distribution is particularly important because of the *central limit theorem.* Consider a sequence $X_1, X_2, X_3, \ldots$ of random variables that are independent, have identical distributions and have finite mean μ and variance σ^2, and define $S_m = \frac{1}{m}\sum_{i=1}^{m} X_m$. The central limit theorem states that for every x,

$$\lim_{m\to\infty} P\left(\sqrt{m}\frac{S_m - \mu}{\sigma} \leq x\right) = \frac{1}{2\pi}\int_{u=-\infty}^{x} du\, e^{-u^2/2} \tag{11.37}$$

has the probability distribution of a Gaussian random variable $\mathcal{N}[0, 1]$. This means, for instance, that sources of noise that consist of the combination of many identical disturbances often have Gaussian distributions.

11.5 Fourier Transforms

Fourier transforms re-express data that vary over time or space as integrals (or sums) of sine and cosine functions of various temporal or spatial frequencies. They thus allow us to determine 'how much' of a given frequency of oscillation is contained in a particular dataset. They are also useful for calculations that involve convolutions of two functions. Typically, Fourier transforms are useful in cases where there is a translational symmetry with respect to the argument of the functions being considered.

The Fourier transform of an *aperiodic* function $f(t)$ is a complex function of a real argument ω given by

$$\tilde{f}(\omega) = \int dt\, f(t) e^{\imath\omega t} . \tag{11.38}$$

The complex exponential function can be broken into its real and imaginary parts using the identity

$$\exp(\imath\omega t) = \cos(\omega t) + \imath \sin(\omega t) \tag{11.39}$$

where $\imath = \sqrt{-1}$. The Fourier transform $\tilde{f}(\omega)$ provides a complete description of the original function $f(t)$ because the Fourier transform can be inverted,

$$f(t) = \frac{1}{2\pi} \int d\omega\, \tilde{f}(\omega) e^{-\imath\omega t} . \tag{11.40}$$

This works because one can show that

$$\delta(t) = \frac{1}{2\pi} \int d\omega\, e^{\imath\omega t} . \tag{11.41}$$

Function $f(t)$ has to be continuous and to satisfy regularity conditions called the Dirichlet conditions for the inversion of the Fourier transform to be exact.

The *convolution* of two functions f and g is the integral

$$h(\tau) = f(t) * g(t) = \int dt\, f(t) g(\tau - t) . \tag{11.42}$$

Convolutions arise frequently in calculating the response of a spatially or temporally stationary linear filter to input, and Fourier transforms are useful for dealing with them because the Fourier transform of h can be written as a simple product

$$\tilde{h}(\omega) = \tilde{f}(\omega) \tilde{g}(\omega) . \tag{11.43}$$

We can show this directly, since

$$\begin{aligned} \tilde{h}(\omega) &= \int d\tau dt f(t) g(\tau - t) e^{\imath\omega\tau} \\ &= \int d\tau dt f(t) e^{\imath\omega t} g(\tau - t) e^{\imath\omega(\tau - t)} \\ &= \int du dt f(t) e^{\imath\omega t} g(u) e^{\imath\omega u} \qquad u = \tau - t \\ &= \tilde{f}(\omega) \tilde{g}(\omega). \end{aligned}$$

Parseval's theorem states that

$$\int d\omega\, |\tilde{f}(\omega)|^2 = \frac{1}{2\pi}\int d(t) f(t)^2. \tag{11.44}$$

We will also be interested in taking Fourier transforms of functions of more than one dimension, for instance the two dimensions of projected visual space. In this case

$$\tilde{f}(\omega_x, \omega_y) = \int dxdy f(x,y) e^{\imath(\omega_x x + \omega_y y)}. \tag{11.45}$$

The properties are similar to those of one-dimensional transforms.

If $f(t)$ is *periodic*, with period T (this means that $f(t+T) = f(t)$ for all t, then we can represent $f(t)$ as a Fourier *series* rather than a Fourier integral. This means that

$$f(t) = \sum_{i=-\infty}^{\infty} \tilde{f}_i e^{-2\pi\imath i \frac{t}{T}} \tag{11.46}$$

where $\tilde{f}_i$ is given by:

$$\tilde{f}_i = \frac{1}{T}\int_{t=0}^{T} dt\, f(t) e^{2\pi\imath i \frac{t}{T}}. \tag{11.47}$$

Equation 11.46 is the equivalent of the inverse Fourier transform in equation 11.40. Again, regularity conditions have to hold for the series to converge and for equation 11.46 to hold. Fourier series satisfy very similar properties to Fourier transforms, including a convolution theorem and a version of Parseval's theorem.

Often, rather than having a function $f(t)$ for all t, we will only have a *finite* number of *samples* of the function, $f_k = f(t_k)$ for $t_k = k\delta$ and $k = 0, 1, \ldots n-1$. The *discrete Fourier transform* of these samples estimates $\tilde{f}_a = \tilde{f}(\omega^a)$ for $\omega^a = 2\pi\frac{a}{n\delta}$ where $a = -n/2, \ldots, n/2$ (assuming n is even, for convenience) by

$$\tilde{f}_a = \sum_{k=0}^{n-1} f_k e^{\imath\omega^a k}. \tag{11.48}$$

This effectively represents a change of basis in the underlying finite dimensional space in which $\mathbf{f} = \{f_k\}$ resides. Note that $\tilde{f}_{-n/2} = \tilde{f}_{n/2}$. The inverse discrete Fourier transform changes back to the canonical basis, using

$$f_k = \frac{1}{n}\sum_{a=0}^{n-1} \tilde{f}_a e^{-\imath\omega^a k} \tag{11.49}$$

This equation implies a periodic continuation of f outside the range $0 \leq k < n$, so that $f_{k+n} = f_k$ for all k. Consult the references for an analysis of the properties of the discrete Fourier transform and the quality of its approximation to the continuous Fourier transform. Note in particular that there is a difference between the *discrete-time* Fourier transform, which is the Fourier transform of a signal that is inherently discrete (*ie* is only defined at discrete points) and the *discrete* Fourier transform, given above, which is based on a finite number of samples of an underlying continuous signal.

If function $f(t)$ is *band-limited*, having $\tilde{f}(\omega)=0$ for $|\omega| > \frac{2\pi}{2\delta}$, then the Nyquist theorem states that $f(t)$ is completely determined by regular samples spaced at $1/\delta$.

11.6 The δ Function

Despite its name, the Dirac δ function is not a properly defined function at all, but rather the limit of a sequence of functions. In this limit, the δ function approaches zero everywhere except where its argument is zero, and there it grows without bound. The infinite height and infinitesimal width of this function are matched so that its integral is one. The sequence of functions used to construct the δ function as a limit is not unique. Basically, any function that integrates to one and has a single peak that gets continually narrower as the limit is taken can be used. For example, the δ function can be expressed as the limit of a square pulse

$$\delta(t) = \lim_{\Delta t \to 0} \begin{cases} 1/\Delta t & \text{if } \; -\Delta t/2 < t < \Delta t/2 \\ 0 & \text{otherwise} \end{cases} \tag{11.50}$$

or a Gaussian function

$$\delta(t) = \lim_{\Delta t \to 0} \frac{1}{\sqrt{2\pi}\Delta t} \exp\left[-\frac{1}{2} \left(\frac{t}{\Delta t} \right)^2 \right] . \tag{11.51}$$

The δ function integrates to one,

$$\int dt\, \delta(t) = 1 \tag{11.52}$$

provided only that the limits of integration surround the point $t=0$ (otherwise the integral is zero). The integral of the product of a δ function with any smooth function g is

$$\int d\tau\, \delta(t-\tau) g(\tau) = g(t) \tag{11.53}$$

for any value of t contained within the integration interval (if t is not within this interval, the integral is zero). These two identities normally provide enough information to use the δ function in calculations despite its unwieldy definition.

One can also write the delta function as

$$\delta(t) = \int_{\omega=-\infty}^{\infty} d\omega\, e^{\imath\omega t}. \tag{11.54}$$

This underlies the inverse Fourier transform.

11.7 Lagrange Multipliers

When first encountered, the method of Lagrange multipliers for finding maxima and minima of functions with constraints appears to be an incomprehensible magic trick. There is a bit of magic to the method but hopefully it can become comprehensible. The basic problem is simple; find an extreme value of a function of several variables, $F(x_1, x_2, \ldots, x_N)$ subject to a constraint $G(x_1, x_2, \ldots, x_N)=$constant. The condition that characterizes an extreme value of the function is that small changes Δx_k in the values of the variables x_k should not change the value of the function. This results in the condition

$$\sum_{k=1}^{N} F_k \Delta x_k = 0 \tag{11.55}$$

where we use the notation

$$F_k = \frac{\partial F}{\partial x_k} \tag{11.56}$$

to make the equations more compact. Equation 11.55 expresses the fact that the change in the function caused by the small changes in its variables should be zero. Without a constraint, this condition must be satisfied for all possible values of the different Δx_k provided that they are small. This can only occur if each separate term in 11.55 vanishes separately so we get the usual condition for an extremum

$$F_k = \frac{\partial F}{\partial x_k} = 0 \tag{11.57}$$

for all k. However, when a constraint like $G(x_1, x_2, \ldots, x_N)=$constant, equation 11.55 does not have to hold for all possible values of the different Δx's, only for those that satisfy the constraint. The condition on the Δx's imposed by the constraint is that they cannot change the value of G therefore we have the condition

$$\sum_{k=1}^{N} G_k \Delta x_k = 0 \tag{11.58}$$

with the same notation for the derivative used for G as for F.

The most obvious way to deal with the constraint equation 11.58 is to solve for one of the Δx's, say Δx_l, writing

$$\Delta x_l = \frac{1}{G_k} \sum_{k \neq l} G_k \Delta x_k \,. \tag{11.59}$$

Then, we substitute this expression into 11.55 to obtain

$$\sum_{k \neq l} F_k \Delta x_k - \frac{F_l}{G_l} \sum_{k \neq l} G_k \Delta x_k = 0 \,. \tag{11.60}$$

Since we have eliminated the constraint, this equation must be satisfied for all values of the remaining Δx's, those with $k \neq l$, and thus, as in the case of no constraint, we find

$$F_k - \frac{F_l}{G_l} G_k = 0 \tag{11.61}$$

for all $k \neq l$. The terms involving derivatives of F and G in this set of equations depend are functions of the x's so these equations can finally be solved to determine where the extremum point is located.

In the above treatment we have singled variable l for special treatment. We have no way of knowing until we get to the final step whether the particular l we chose will lead to a simple or a complex set of final equations. The clever idea of the Lagrange multiplier is to notice that the whole problem is symmetrical with respect to the different x variables. Choosing one l value as we did above breaks this symmetry and often complicates the algebra. To introduce the Lagrange multiplier we simply define it as

$$\lambda = -\frac{F_l}{G_l} \tag{11.62}$$

using the j value chosen above. Using this notation, or final set of equations can be written as

$$F_k + \lambda G_k = 0 \,. \tag{11.63}$$

Before we had to say that these equations only held for $k \neq l$ because l was treated differently. Now, however, notice that the above equation when k is set to l is algebraically equivalent to the definition 11.62. Thus, we can say that equation 11.63 apply for all k and they thus provide a symmetric formulation of the problem of finding an extremum that often results in simpler algebra.

The final realization is that the set of equation 11.63 are exactly what we would have derived if we had set out in the first place to find an extremum of the function $F + \lambda G$ and forgot about the constraint entirely. Of course this lunch is not completely free. From equation 11.63, we derive a set of extremum points parameterized by the undetermined variable λ. To fix λ we must substitute this family of solutions back into G and find the value of λ that satisfies the constraint G=constant. This is then the solution of the constrained problem and it explains the mysterious procedure followed by people solving constrained problems of this sort: introduce an undetermined variable (λ), change the function being minimized or maximized (from F to $F + \lambda G$) and proceed as if no constraint existed.

11.8 Further Reading

Further Reading

Boas, ML (1966) *Mathematical Methods in the Physical Sciences.* New York, NY:Wiley.

Cover, TM & Thomas, JA (1991) *Elements of Information Theory.* New York, NY:Wiley.

Feller, W (1968) *An Introduction to Probability Theory and its Application, 3rd edition.* New York, NT:Wiley.

Gershenfeld, N (1999) *The Nature of Mathematical Modeling.* Cambridge, UK:CUP.

Jordan, DW & Smith, P (1977) *Nonlinear Ordinary Differential Equations.* Oxford, England:Clarendon Press.

Lindgren, BW (1993) *Statistical Theory, 4th edition.* New York, NY:Chapman & Hall.

Oppenheim, AV & Willsky, AS with Nawab, H (1997) *Signals and Systems, 2nd edition.* Upper Saddle River, NJ:Prentice Hall.

Press, WH, Teukolsky, SA, Vetterling, WT & Flannery, BP (1992) *Numerical recipes in C*. Cambridge, UK:Cambridge University Press.

Strang, G (1976) *Linear Algebra and its Applications.* New York, NY:Academic Press.