

MATHEMATICS FOR MULTIDISCIPLINARY MASTERS

A course through exercises

Hugo van den Berg

WARWICK INTERDISCIPLINARY SCIENCE
DOCTORAL TRAINING CENTRES

This is a series of exercises for students following a multidisciplinary master's degree in the natural sciences. Such students typically constitute a diverse crowd, drawn from various disciplines (life sciences, physics, chemistry, mathematics, statistics, engineering...) and drawn toward the interface between the natural sciences. Such students require a crash course in those relevant subjects that their first degree was *not* in; this book gives an introduction to mathematical skills for non-mathematicians at master's level.

The guiding principle is that many concepts are best grasped if one works one's way toward them through a series of exercises that gradually admit more complexity. Also, since I've found that non-mathematicians shut down their brains pre-emptively in a defensive response to certain terminology, I introduce new terminology at the very last moment, after the student has (I hope) discovered the concept for him/herself in the exercises. I would not introduce terminology at all were it not for the fact that the student also has to gain access to the literature.

Part I reviews basic skills. Part II deals with the core techniques needed to apply mathematics in the natural sciences. Part III deals with various topics which frequently arise useful in multidisciplinary research. Chapter 1 is a prerequisite for chapters 3, 4, 7, and 8. Chapter 2 is a prerequisite for chapters 4, 5, and 9. Chapter 3 is a prerequisite for chapter 6. Chapter 4 is a prerequisite for parts of chapter 7.

I am very grateful to students who have pointed out errors in the exercises: Jonathan Armond, Nigel Dyer, Robert Gardner, Jouko Koecher, Steven Kiddle, and Hayley Morley.

How to use this book Working through the following exercises you will quickly glean an overview or review of some of the essential mathematical skills that are needed for a multidisciplinary master's degree in the natural sciences. Remember: when you are stuck is when you learn. Harder exercises start with the words "Can you...". Successful completion of these harder exercises is not necessary when you are first working through the material. However, the you must still take notice of any results stated in the exercise, since what follows will generally depend on this material.

Occasionally, you are asked to review or revise a topic. You then have to research the material using a text-book or an on-line resource. While the choice of these resources is entirely up to you, here is a list of textbooks that I particularly recommend:

Linear algebra David C. Lay: Linear Algebra and Its Applications (Addison-Wesley)

Differential equations James C. Robinson: An Introduction to Ordinary Differential Equations (Cambridge Texts in Applied Mathematics)

Probability G. R. Grimmett, D. R. Stirzaker: Probability and Random Processes (Oxford Science Publications)

Statistics Lee J. Bain, Max Engelhardt: Introduction to Probability and Mathematical Statistics (Duxbury Classic Series)

Numerical analysis Richard Burden, J. Douglas Faires: Numerical Analysis (Brooks Cole)

Contents

I	The basics	1
1	Differentiation & Integration	3
1.1	The fundamental definition	3
1.2	Powers	3
1.3	Differentiation rules	3
1.4	More differentiation rules	3
1.5	Fractional powers	4
1.6	Antiderivatives	4
1.7	A special area operator	5
1.8	An area function	5
1.9	Differentiating the area function	5
1.10	The relationship between derivatives and areas	5
1.11	Areas and antiderivatives	6
1.12	The fundamental theorem of calculus	6
1.13	A problem with powers	6
1.14	A special antiderivative	6
1.15	L is a logarithm	7
1.16	The natural logarithm	7
1.17	The exponential function	7
1.18	The number e	7
1.19	Some interesting (anti)derivatives	7
1.20	An approximation	8
1.21	Higher derivatives	8
1.22	Approximating a high derivative	8
1.23	Many antidifferentiation steps	8
1.24	A useful formula	9
1.25	Two very special functions	9
1.26	The link with trigonometry	10
2	Vectors and matrices	11
2.1	A simple start	11
2.2	Matrices	11
2.3	Vector length	11
2.4	Vector equality	11
2.5	A vector times a number	12
2.6	Adding vectors	12
2.7	Vector equations	12
2.8	More on vector equations	13
2.9	A shorthand notation	13
2.10	Multiplying matrices	14
2.11	The order of matrix multiplication matters	14
2.12	Columns and rows	15

2.13	Inverse matrices	15
2.14	Matrix powers	15
2.15	Transposition	15
2.16	The eigenvalue problem	16
2.17	Eigenvalues	16
2.18	Eigenvectors	16
3	Probability	19
3.1	Intuitive probability	19
3.2	Events	19
3.3	Conditional probability	20
3.4	Random variables	20
3.5	Assigning probabilities	21
3.6	The Bernoulli variate	21
3.7	The binomial variate	21
3.8	The distribution function	21
3.9	Discrete and continuous random variables	22
3.10	A useful property	22
3.11	Probability density	22
3.12	Examples of probability density functions	22
3.13	Important probability density functions	23
3.14	Transformations of random variables	23
3.15	Probability transforms	24
3.16	Expectation	24
3.17	Expectation of a transformed random variable	25
3.18	Variance	25
3.19	Expectation of a sum	26
3.20	Independence	26
3.21	The chi-square variate	26
3.22	The Central Limit Theorem	27
3.23	Student's t	27
3.24	Covariance	28
3.25	Variance of a sum	28
3.26	Some commonly used statistics	28
3.27	Conditioning on a rare event	28
3.28	A tricky question	29
3.29	An improved approximation for rare events	29
3.30	Sampling complicated probabilistic structures	29
3.31	Markov Chain Monte Carlo	30
3.32	Choosing the next step	30
II	Core skills	33
4	Dynamics	35
4.1	State-transition functions	35
4.2	Continuous time	35
4.3	Bernoulli's method	36
4.4	The ubiquitous use of differential equations	36
4.5	Finding the state-transition function	37
4.6	Autonomy	38
4.7	Finding solutions	38
4.8	Violation of the ' L ' condition	39
4.9	Arrowheads for qualitative analysis	39

4.10	The phase plane	40
4.11	The phase flow	40
4.12	The phase portrait	41
4.13	Finding whence & whither	41
4.14	A linear system	42
4.15	The characteristic equation	43
4.16	Local linear approximation	43
4.17	Discrete time	44
4.18	Iterate maps	44
4.19	Fixpoints	44
4.20	Stochastic dynamics	46
4.21	The persistence function	47
4.22	The hazard rate	47
4.23	The no-memory property	48
4.24	Accounting for an extra influx	48
4.25	Taking the ensemble limit	49
4.26	The jump chain	49
4.27	The Markov chain	49
4.28	Detailed balance	50
4.29	Ergodicity	50
4.30	Controlled dynamics	50
4.31	Control	50
4.32	Control as a state	50
4.33	Optimized control	51
4.34	Optimal choice	51
4.35	Dynamic programming	52
4.36	The dynamic programming table	52
5	Modelling Methodology	53
5.1	What is a model?	53
5.2	Assumptions	53
5.3	Hidden assumptions	53
5.4	Strong and weak assumptions	54
5.5	Consistency	54
5.6	Model complexity	54
5.7	Simulation	55
5.8	Dimensional analysis	55
5.9	Empirical dimension	55
5.10	Equality of dimensions	56
5.11	Dimensionless quantities	56
5.12	A basis of empirical dimensions	56
5.13	Dimension formulæ	57
5.14	Alternate bases	57
5.15	Natural units	57
5.16	Units	58
5.17	Choosing model-derived units	58
5.18	How to choose units	58
5.19	Buckingham's theorem	59
5.20	Forming a basis	59
5.21	Dimensionless quantities	60
5.22	The proof concluded	60
5.23	Measurement strength	60
5.24	The nominal scale of measurement	60
5.25	The ordinal scale of measurement	60

5.26	The interval scale of measurement	61
5.27	The ratio scale of measurement	61
5.28	The absolute scales of measurement	61
5.29	Data	62
5.30	Testing predictions	62
5.31	Parameter estimation	62
5.32	Goodness of fit	62
5.33	Best-fit values	63
5.34	Simultaneous fitting	63
5.35	A justification for least-squares	63
5.36	Multiple data sets	64
5.37	Scaling by variance	64
5.38	A fundamental trade-off	64
5.39	Estimating confidence	65
5.40	Fisher information	65
5.41	Out-of-sample prediction	65
6	Inferential statistics	67
6.1	Populations	67
6.2	Observations	67
6.3	Means	67
6.4	The population distribution function	67
6.5	The statistical population	68
6.6	Repeated observations	68
6.7	Sample variance	68
6.8	An unbiased estimate of sample variance	69
6.9	Hypotheses	69
6.10	The critical region	70
6.11	Choosing alpha	70
6.12	Alpha or P-value?	70
6.13	A useful result	70
6.14	Another type of error	70
6.15	Power	71
6.16	Estimation	71
6.17	Likelihood	71
6.18	Multiple observations and likelihood	72
6.19	Logistic regression	72
6.20	A standard test problem	73
6.21	Characterizing the samples	73
6.22	A test statistic	73
6.23	The null hypothesis	73
6.24	Determining the critical region	74
6.25	Using an estimate for the variance	74
6.26	The choice of a critical region, in general	75
6.27	The Neyman-Pearson lemma	75
6.28	Proving the Neyman-Pearson lemma	76
6.29	Generalized likelihood ratio principle	76
6.30	The likelihood ratio	77
6.31	Analysis of variance	77
6.32	Chi-square tests	78
6.33	Further chi-square tests	79
6.34	Testing for independence	79
6.35	Goodness-of-fit tests	80
6.36	Screening	80

6.37	Sensitivity and specificity	81
6.38	Sharing alpha	81
6.39	Conservation of statistical power	81
6.40	A key observation	82
6.41	The next step	82
6.42	The step-down procedure	82
III Advanced topics		83
7	Optimization	85
7.1	Simple extrema	85
7.2	Bounds	85
7.3	Dealing with constraints	85
7.4	Maximizing the H-function	86
7.5	Lagrange multipliers	86
7.6	Optimal process control	86
7.7	Respecting the state transitions	87
7.8	Finding an optimal solution	87
7.9	The switching function	87
7.10	Towards continuous time	88
7.11	Singular control	88
7.12	Control in feedback form	89
7.13	The calculus of variations	89
7.14	A planar objective function	89
7.15	Internal point	90
7.16	Convex domains	91
7.17	A systematic procedure	91
7.18	Extending the technique to higher dimensions	92
7.19	Vertex moves	92
8	Fourier series, Fourier transform, & Sampling	93
8.1	Functions that repeat	93
8.2	Periodic functions	93
8.3	The fundamental period	93
8.4	Finding the coefficients	93
8.5	The Fourier series representation	94
8.6	The amplitude spectrum	94
8.7	The phase spectrum	94
8.8	Accommodating aperiodic functions	94
8.9	Toward the Fourier transform	95
8.10	The Fourier transform and its spectra	96
8.11	Signals	96
8.12	Sampling	96
8.13	The Fourier transform of the sampled process	96
8.14	Band-limited signals	97
8.15	The Nyquist frequency	97
9	Principal Component Analysis and Clustering	99
9.1	A data set with a peculiar property	99
9.2	The mean-deviation form	99
9.3	Almost-linear data sets	99
9.4	Mean-deviation in general	100
9.5	Information-rich and information-poor coordinates	100

9.6	A useful property	100
9.7	Generalizing to higher dimensions	101
9.8	Diagonalization	101
9.9	Principal components	101
9.10	The scree plot	102
9.11	Maximum Likelihood clustering	103
9.12	Obtaining ML estimators	103
9.13	Assigning data points to clusters	103
9.14	k -Means clustering	104
9.15	Hierarchical clustering	104
9.16	Distance-based clustering	104
9.17	Amalgamation	104
9.18	Distances	105
9.19	The amalgamation rule	105
9.20	Validation	105
9.21	Maximum Likelihood clustering	106

|

The basics

Differentiation & Integration

1.1 The fundamental definition

If f is some function, the function *derived from* f , often denoted f' and called the *derivative of* f , satisfies

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (1.1)$$

for all x for which this limit exists.

Exercise 1 Let $f(x) = \alpha x$. Verify that $f(x+h) = \alpha(x+h) = \alpha x + \alpha h$ and thus

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \alpha.$$

Hence, $f'(x) = \alpha$ (a “constant function”).

Exercise 2 Let $f(x) = \alpha x + \beta$. Verify that $\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \alpha$.

Exercise 3 Let $f(x) = k$ and determine $f'(x)$, using equation (1.1). Sketch graphs of $f(x)$ and $f'(x)$.

Exercise 4 Let $f(x) = \alpha x^2$. Determine $f'(x)$, using equation (1.1).

1.2 Powers

If $f(x) = x^n$, where $n = 1, n = 2, n = 3$, etc., then $f'(x) = nx^{n-1}$ (rule: “exponent in front, exponent minus one”).

Exercise 5 Can you see why this is so? (Hint: the definition, equation (1.1); start with $n = 1$, and work your way up.)

Exercise 6 Let $f(x) = \alpha x + \beta x^2 + \gamma x^3 + \delta x^4$. Determine $f'(x)$.

1.3 Differentiation rules

In what follows, f, g , and u are functions; f and g have derivatives f' and g' .

Exercise 7 (The sum rule) Let $u(x) = f(x) + g(x)$ (“ f plus g ”). Use equation (1.1) to show that $u'(x) = f'(x) + g'(x)$.

Exercise 8 (The product rule) Let $u(x) = f(x)g(x)$ (“ f times g ”). Can you show that $u'(x) = f'(x)g(x) + f(x)g'(x)$? (Hint: write $f(x+h)$ as $f(x) + h(f(x+h) - f(x))/h$ and $g(x+h)$ as $g(x) + h(g(x+h) - g(x))/h$. Substitute these longer expressions for $f(x+h)$ and $g(x+h)$ in the definition, $u'(x) = \lim_{h \rightarrow 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h}$.)

1.4 More differentiation rules

We can use the product rule to find the derivative of $f(x) = \sqrt{x}$. Observe that $x = x^1 = x^{\frac{1}{2} + \frac{1}{2}} = x^{\frac{1}{2}}x^{\frac{1}{2}}$. The derivative of this is clearly 1 (since $g'(x) = 1$ when $g(x) = x$). But the product rule says

$$(x^{\frac{1}{2}}x^{\frac{1}{2}})' = f'(x)x^{\frac{1}{2}} + x^{\frac{1}{2}}f'(x) = 2f'(x)x^{\frac{1}{2}}$$

and equating this to 1, we obtain $2f'(x)x^{\frac{1}{2}} = 1$.

Exercise 9 Hence, work out that $f'(x) = 1/(2\sqrt{x})$.

Exercise 10 (The quotient rule) Let $u(x) = f(x)/g(x)$ (“ f divided by g ”). Can you show that

$$u'(x) = \frac{f'(x)}{g(x)} - \frac{g'(x)}{g(x)}u(x) ?$$

(Note: you may have encountered a different formula in school; verify that they come to the same thing; the form given here is easier to remember!)

Exercise 11 Let

$$y(x) = \frac{\alpha + \beta x}{\gamma + \delta x}$$

where $x \geq 0$ and $\alpha, \beta, \gamma,$ and δ are all positive parameters. Show that $y'(x) \geq 0$ iff¹ $y(x) \leq \beta/\delta$. Sketch a graph of the function $y(x)$ (i) when $\alpha\delta < \beta\gamma$ and (ii) when $\alpha\delta > \beta\gamma$.

Exercise 12 (“ g of f of x ”) Let $f(x) = x + 3$ and $g(x) = x^2$. Determine $g(f(x))$. (Hint: substitute the right member of the definition of f into the place of x in the right member of the definition of g .)

Exercise 13 (The chain rule) Let $u(x) = g(f(x))$ (“ g of f of x ” i.e. “first f , then g ”). Can you show that $u'(x) = g'(f(x))f'(x)$? (Note: This is quite hard! Try to remember this rule, it crops up all the time.)

Exercise 14 Find the derivatives of: (i) $x^2 + \alpha x$; (ii) $\frac{x^2}{x^3+1}$; (iii) $(x^5 + x^3)^7$. (Hint: the last one goes a lot quicker with the chain rule.)

Exercise 15 Can you find the derivative of $\sqrt[3]{x} = x^{\frac{1}{3}}$? (Hint: exercise 9.)

1.5 Fractional powers

If you look at exercises 9 and 15, it appears that the rule “exponent in front, exponent minus one” applies also when the exponent is a fraction like $\frac{1}{2}$ or $\frac{1}{3}$. In fact, extending the idea, we can show that this rule applies when the exponent is any real number: $(x^\alpha)' = \alpha x^{\alpha-1}$, $\alpha \in \mathbb{R}$.

Exercise 16 Suppose $f'(x) = x^9$. Can you find an expression for $f(x)$?

Exercise 17 Suppose $f'(x) = x^\beta$. Can you find an expression for $f(x)$? (Hint: What is the derivative of $x^{\beta+1}$? Now write x^β as $(\beta + 1)^{-1}(\beta + 1)x^\beta$.)

1.6 Antiderivatives

In the last two exercises you were asked to find an *antiderivative*, finding the “primitive” or “progenitor” function of which a given function is the derivative. If you write “ $+K$ ” after any answer you find, you still have a correct answer, since the “ $+K$ ” is replaced by “ $+0$ ” when you differentiate. This works whatever value K might have; the only thing that matters is that K does not vary with x . So, whenever something is the sought-for antiderivative, so is something plus an arbitrary constant, which is often called the *integration constant*.

¹The abbreviation *iff* means “if and only if”.

1.7 A special area operator

You are familiar with the idea of “the area under the graph of a function between the points $x = a$ and $x = b$ ”; we will denote this area as $\mathcal{I}(f, a, b)$. (There is no ‘ x ’ in this expression since the area depends on the function and the values a and b , not on the name of the variable along the horizontal axis.) You are also familiar with the sign conventions

$$\mathcal{I}(-f, a, b) = -\mathcal{I}(f, a, b), \quad \mathcal{I}(f, a, b) = -\mathcal{I}(f, b, a)$$

and the (intuitively reasonable) formula

$$\mathcal{I}(f, a, c) = \mathcal{I}(f, a, b) + \mathcal{I}(f, b, c)$$

where $a \leq b \leq c$.

Exercise 18 Complete the following equation: $\mathcal{I}(f, a, a) = \quad$.

Exercise 19 Using the sign conventions alone, evaluate $\mathcal{I}(f, -\pi, 3\pi)$ where $f(x) = \sin x$.

Exercise 20 Let f be a constant function, $f(x) = K$. Use a sketch of the function to show that $\mathcal{I}(f, a, b) = k(b - a)$.

1.8 An area function

Using a given function f and our “area operator” \mathcal{I} , we can derive a new function \tilde{f} from f by defining $\tilde{f}(x) = \mathcal{I}(f, 1, x)$. Note that while \tilde{f} is a function derived from f , it is *not* f ’s derivative f' (quite the opposite in fact, as we shall shortly see). We won’t use the symbol \tilde{f} in what follows, and just let $\mathcal{I}(f, 1, x)$ stand as the name of this function.

Exercise 21 Verify that the following formula is consistent with the sign conventions:

$$\mathcal{I}(f, a, b) = \mathcal{I}(f, 1, b) - \mathcal{I}(f, 1, a) \tag{1.2}$$

(You may assume that $b > a$ for the sake of convenience, although that is not necessary.)

1.9 Differentiating the area function

The last exercise shows that the function $\mathcal{I}(f, 1, x)$ suffices to do all areas under the curve we may want to do. Consider the derivative of this function:

$$\mathcal{I}'(f, 1, x) = \lim_{h \rightarrow 0} \frac{\mathcal{I}(f, 1, x+h) - \mathcal{I}(f, 1, x)}{h} = \lim_{h \rightarrow 0} \frac{\mathcal{I}(f, x, x+h)}{h}. \tag{1.3}$$

Exercise 22 Make sure you understand how the second equality arises.

1.10 The relationship between derivatives and areas

If f “does not change much” between x and $x + h$, it is “almost” a constant function on the interval $[x, x + h]$. Referring back to exercise 20, you might guess that $\mathcal{I}(f, x, x + h)$ is well-approximated by $f(x)(x + h - x) = f(x)h$, certainly “in the limit” $h \rightarrow 0$.

Exercise 23 Assume that you can use $\mathcal{I}(f, x, x + h) = f(x)h$ in the limit of equation (1.3), and hence verify that $\mathcal{I}'(f, 1, x) = f(x)$.

1.11 Areas and antiderivatives

A more careful argument (which looks more closely how well-behaved f is) bears out this result. Let F be an antiderivative of f , so that $F'(x) = f(x)$.

Exercise 24 Show that $F(x) = \mathcal{I}(f, 1, x) + K$ where K is the arbitrary integration constant. (Hint: $\mathcal{I}'(f, 1, x) = f(x)$.)

Exercise 25 Use equation (1.2) to show that $\mathcal{I}(f, a, b) = F(b) - F(a)$.

Exercise 26 To evaluate $\mathcal{I}(f, a, b)$, you need to find an antiderivative of f , which is not always easy. Suppose you can find two other functions $u(x)$ and $v(x)$ such that $f(x) = u'(v(x))v'(x)$ (so f looks like the result of an application of the chain rule). Explain why the following is correct:

$$\mathcal{I}(f, a, b) = u(v(b)) - u(v(a)) .$$

1.12 The fundamental theorem of calculus

Instead of $\mathcal{I}(f, a, b)$, we usually write $\int_a^b f(x)dx$ (such an expression is called an *integral*; the symbol ‘ \int ’ is an *integral sign*). Your result from exercise 25 then looks like this:

$$\int_a^b f(x)dx = F(b) - F(a) \quad \text{where } F'(x) = f(x)$$

which is known as the *fundamental theorem of calculus*.

Exercise 27 Evaluate $\int_1^3 x^2 dx$.

Exercise 28 Translate the method of exercise 26 into integral notation, using the identity $dv = v'(x)dx$. (Hint: you should arrive at the ‘substitution method’ which should be familiar from your school days. In practice, this latter method is more convenient, whereas with the method of exercise 26 it is more clear *why* it works!)

1.13 A problem with powers

Recall that the antiderivative of x^β is $(\beta + 1)^{-1}x^{\beta+1} + K$ (K is the integration constant). Here β can be any real number, except -1 .

Exercise 29 Why is the formula no good for $\beta = -1$?

1.14 A special antiderivative

So what is the antiderivative of $x^{-1} = \frac{1}{x}$? Let us first give it a name:

$$L(x) = \mathcal{I}(f, 1, x) \quad \text{where } f(x) = \frac{1}{x}. \quad (1.4)$$

Exercise 30 Complete the following equation: $L(1) = \dots$. (Hint: the definition of L , equation (1.4), plus exercise 18.)

Exercise 31 Sketch a graph of the function $f(x) = \frac{1}{x}$. Choose some value $u > 1$. Cross-hatch $L(u) = \mathcal{I}(f, 1, u)$, the area under the curve between $x = 1$ and $x = u$. Also choose a value $v > u$ and cross-hatch $\mathcal{I}(f, v, uv)$, the area under the curve between $x = v$ and $x = uv$.

Exercise 32 Can you tell from your graph that $\mathcal{I}(f, 1, u) = \mathcal{I}(f, v, uv)$? (Hint: take your first cross-hatched area, “squeeze” it vertically by a factor $\frac{1}{v}$ and “stretch” it horizontally by a factor v .)

Exercise 33 Can you prove that $L(uv) = L(u) + L(v)$? (Hint: observe that $\mathcal{I}(f, v, uv) = L(uv) - L(v)$ the result of the previous exercise.)

1.15 L is a logarithm

The result of the last exercise means that L is a *logarithmic function*.

Exercise 34 Can you see why this follows? (Hint: define a correction function by $k(x) = L(x)/\log x$; the result follows if you can establish that $k'(x) = 0$.)

1.16 The natural logarithm

Now that we have established that L is a logarithmic function, we shall often write $\ln x$ instead of $L(x)$. This logarithm is known as the *natural logarithm*; its base is denoted e . So we have answered the question posed recently: the antiderivative of $1/x$ is $\ln x$ (“up to an integration constant”). How big is e ? If $x = \ln y$, then $y = e^x$. Now $(e^x)' = \frac{1}{L'(y)}$ because, quite generally $(f^{-1}(x))' = 1/f'(x)$.

Exercise 35 Can you prove this rule $(f^{-1}(x))' = 1/f'(x)$? (Hint: how are the graphs of f and f^{-1} related? The relation $f^{-1}(f(x)) = x$ defines the inverse function f^{-1} .)

1.17 The exponential function

Also, $L'(y) = 1/y$, so $(e^x)' = y = e^x$. So: *the derivative of e^x is e^x . The exponential function e^x is “its own derivative”:*

$$\lim_{h \rightarrow 0} \frac{e^{x+h} - e^x}{h} = e^x.$$

1.18 The number e

Let the function g be defined by $g(h) = (e^{x+h} - e^x)/(he^x)$ (pay close attention: the argument of g is called h here, not x).

Exercise 36 Use the definition of $g(h)$ to derive the formula $e = (1 + hg(h))^{1/h}$.

Exercise 37 Verify that $\lim_{h \rightarrow 0} g(h) = 1$. (Hint: raise both sides to the power of h .)

Exercise 38 Now show that $\lim_{h \rightarrow 0} (1 + h)^{1/h} = e$. Evaluate the expression $(1 + h)^{1/h}$ numerically, for $h = 0.1$, $h = 0.001$, $h = 0.0000001$ (or whatever small numbers you fancy!).

1.19 Some interesting (anti)derivatives

Now that we have the logarithmic function \ln , we can do some interesting integrals, which, as you have seen, involves the determination of antiderivatives (which explains why people say “integrate” when they mean “determine an antiderivative” which is more correct, though also more long-winded).

Exercise 39 Find the derivative of 2^x . (Hint: rewrite as $e^{\ln 2^x}$, then apply the chain rule, exercise 13, with $f(x) = \ln 2^x = x \ln 2$ and $g(x) = e^x$.)

Exercise 40 Find the derivative of $\ln(5 + x^8)$. (Hint: the chain rule, with $f(x) = 5 + x^8$ and $g(x) = \ln x$.)

Exercise 41 Find an *antiderivative* of $\frac{x^2}{1 + \frac{1}{3}x^3}$. (Hint: the previous exercise.)

Exercise 42 Find an antiderivative of $\frac{x}{\alpha + x^2}$.

1.20 An approximation

Derivatives are useful in approximating functions. The idea is

$$f(x) \approx f(a) + (x - a)f'(a)$$

where $f(a)$ and $f'(a)$ are these functions evaluated at some special value a , called the *operation point*.

Exercise 43 Can you relate this approximation to equation (1.1) (which defines the derivative)? (Hint: rearrange, take $x - a = h$.)

Exercise 44 Can you explain why the above approximation “is exact at the operation point”?

1.21 Higher derivatives

More notation: the derivative of f is $f' = f^{[1]}$; the derivative of f' is $f'' = f^{[2]}$; the derivative of f'' is $f''' = f^{[3]}$; and so on, with the $f^{[n]}$ notation being useful because we may want to talk about $f^{[1000]}$, which would require more primes than would print on a single line! We call $f^{[n]}$ the *nth derivative of f* . Finally, $f^{[0]}$ is just f itself.

Exercise 45 Let m be an integer, $m > 1$. Let $f(x) = x^m$. Can you give an expression for $f^{[m]}$? For $f^{[m+1]}$?

1.22 Approximating a high derivative

A refinement of the foregoing approximation idea is to apply the approximation not to f itself, but to its n th derivative:

$$f^{[n]}(x) \approx f^{[n]}(a) + (x - a)f^{[n+1]}(a).$$

The job now is to recover an approximation for f ; our motivation for doing this is that the approximation generally becomes better for bigger n .

Exercise 46 Verify that the antiderivative of the approximation is

$$f^{[n-1]}(x) \approx f^{[n]}(a)x + \frac{1}{2}f^{[n+1]}(a)(x - a)^2 + K. \quad (1.5)$$

1.23 Many antidifferentiation steps

The integration constant is determined by the function $f^{[n-1]}$ at $x = a$ and turns out to be

$$K = f^{[n-1]}(a) - f^{[n]}(a)a. \quad (1.6)$$

Exercise 47 Check the following rearrangement:

$$f^{[n-1]}(x) \approx f^{[n-1]}(a) + f^{[n]}(a)(x - a) + \frac{1}{2}f^{[n+1]}(a)(x - a)^2.$$

(Hint: substitute equation (1.6) into equation (1.5).)

Exercise 48 Can you see how to proceed? The next step is

$$f^{[n-2]}(x) \approx f^{[n-2]}(a) + f^{[n-1]}(a)(x - a) + \frac{1}{2}f^{[n]}(a)(x - a)^2 + \frac{1}{6}f^{[n+1]}(a)(x - a)^3.$$

1.24 A useful formula

Continuing these antidifferentiation steps, we finally arrive at the following approximation:

$$f(x) \approx \sum_{i=0}^{n+1} \frac{f^{[i]}(a)}{i!} (x-a)^i. \quad (1.7)$$

This suggests that perhaps an *exact* representation of a “suitably nice” function f is given by the infinite sum

$$f(x) = \sum_{i=0}^{\infty} \frac{f^{[i]}(a)}{i!} (x-a)^i. \quad (1.8)$$

Not all functions are “suitably nice” in this sense. However, the exponential function is nice.

Exercise 49 Using $a = 0$ derive the following infinite sum formula:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Exercise 50 Deduce the following:

$$e = 2 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

Add the first 10 terms of this sum to find another numerical approximation for e , the base of the natural logarithm. Compare your result with those of exercise 38, where you also computed numerical approximations to e .

Exercise 51 Let n be a (possibly large, but finite) positive integer and define f by $f(x) = x^n$. Prove that $\lim_{x \rightarrow \infty} f(x)/e^x = 0$. (Hint: substitute in the infinite sum expression for e^x ; divide numerator and denominator both by x^n .)

1.25 Two very special functions

Suppose that two functions u and v satisfy the following properties:

$$u(0) = 0 \quad v(0) = 1 \quad u'(x) = v(x) \quad v'(x) = -u(x). \quad (1.9)$$

Exercise 52 Find higher derivatives, e.g. $u'' = v' = -u$, $u''' = v'' = -u' = -v$, $u^{[4]} = v''' = -u'' = -v' = -(-u) = u$.

Exercise 53 Let $n = 0, 1, 2, 3, \dots$. Can you confirm the following?

$$0 = u^{[2n]}(0) = v^{[2n+1]}(0); \quad +1 = u^{[4n+1]}(0) = v^{[4n]}(0); \quad -1 = u^{[4n+3]}(0) = v^{[4n+2]}(0).$$

Exercise 54 Using $a = 0$ derive the following infinite sum formula:

$$u(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots$$

and find a similar expression for $v(x)$.

Exercise 55 Combine the results of exercises 49 and 54 to deduce

$$e^{ix} = v(x) + iu(x)$$

where i has the property $i^2 = -1$.

1.26 The link with trigonometry

The functions $\sin x$ and $\cos x$ share the properties (1.9) with $u(x)$ and $v(x)$ as can be verified using definition (1.1) plus a geometrical argument, and therefore these trigonometric functions have the same infinite sum expressions. Thus we have the remarkable formula:

$$e^{ix} = \cos x + i \sin x \quad (1.10)$$

where $i^2 = -1$.

Exercise 56 Deduce the formula $\cos x = (e^{ix} + e^{-ix})/2$. Find a similar formula for $\sin x$. (Hint: the properties $\cos(-x) = \cos x$, $\sin(-x) = -\sin x$.)

Exercise 57 Find the derivative of $\sin(x) \cos(x)$.

Exercise 58 Find the derivative of $e^{\alpha x}$. (Hint: the chain rule, with $g(x) = e^x$ and $f(x) = \alpha x$.)

Exercise 59 Find the derivative of $e^{\alpha x} \cos x$.

Exercise 60 Find an *antiderivative* of $e^{\alpha x} \cos x$.

2

Vectors and matrices

2.1 A simple start

Exercise 61 Solve the system

$$\begin{cases} ax + by = u \\ cx + dy = v \end{cases} .$$

(Hint: make x the subject of the first equation, substitute the resulting expression for x in the second equation, then solve for y ; do the same “but the other way around” to obtain x .)

2.2 Matrices

For our purposes, a *matrix*¹ will be a rectangular array of numbers (or variables), like so:

$$\begin{bmatrix} 8 & 20 \\ 9 & 5 \\ 15 & 7 \end{bmatrix}$$

or so:

$$\begin{bmatrix} a & b & c & d \\ e & f & g & h \end{bmatrix}$$

and a *vector* will be a matrix of ‘width’ 1, like so:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

which is a ‘3-vector’ which we may also write as $[u, v, w]^T$ to conserve space.

Exercise 62 Write down a matrix of width 4, height 4, with zeros everywhere except on the diagonal, which is the 4 positions running from the top left corner down to the bottom right corner.

2.3 Vector length

The length of the 2-vector $[u, v]^T$ is defined to be $\sqrt{u^2 + v^2}$. The length of the 3-vector $[u, v, w]^T$ is defined to be $\sqrt{u^2 + v^2 + w^2}$.

Exercise 63 Calculate the length of $[3, 5]^T$.

Exercise 64 Give a sensible definition of the length of a 5-vector like $[u, v, w, x, y]^T$.

2.4 Vector equality

Two vectors are *equal* if the components are all the same:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \Leftrightarrow x = u, y = v, z = w . \quad (2.1)$$

Exercise 65 Can a 3-vector ever be equal to a 4-vector? Does the question even make sense?

¹Plural: matrices. The word means ‘mother’, in the sense of surrounding substratum, as in ‘mother lode’. ‘Vector’ means ‘carrier’.

2.5 A vector times a number

We can multiply a vector by a number, like so:

$$\alpha \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha x \\ \alpha y \end{bmatrix}. \quad (2.2)$$

Exercise 66 Calculate the length of $[x, y]^T$ and of $\alpha[x, y]^T$. How do these lengths compare when $|\alpha| = 1$? When $|\alpha| < 1$? When $|\alpha| > 1$?

2.6 Adding vectors

If two vectors have the same height, we can add them together, like so:

$$\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x + u \\ y + v \end{bmatrix}. \quad (2.3)$$

The thing on the right still has width 1, so it is still a vector.

Exercise 67 Determine x and y such that

$$x \begin{bmatrix} -1 \\ 2 \end{bmatrix} + y \begin{bmatrix} -3 \\ 7 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}.$$

(Hint: apply the rule (2.2), followed by rule (2.3) and finally rule (2.1), to obtain a system like the one in exercise 61.)

Exercise 68 Try to solve the following for x and y :

$$x \begin{bmatrix} -1 \\ 2 \end{bmatrix} + y \begin{bmatrix} -3 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}.$$

If you run into trouble, try to describe the nature of the problem.

Exercise 69 Try to solve the following for x and y :

$$x \begin{bmatrix} -1 \\ 2 \end{bmatrix} + y \begin{bmatrix} -3 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Can you be sure that everybody finds the same solution?

2.7 Vector equations

The system of exercise 61 involves a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, a vector $[u, v]^T$, and, perhaps less obviously, another vector $[x, y]^T$. The equations in exercise 61 tell us about a way to combine the matrix and the vector $[x, y]^T$ to obtain the vector $[u, v]^T$; at least, this is one way of looking at these equations.

Exercise 70 Use rules (2.1)—(2.3) to rewrite the equations of exercise 61 as:

$$x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}.$$

2.8 More on vector equations

We now agree that the following is another way of writing the same system:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Exercise 71 Rewrite in “matrix-dot-vector” form:

$$\begin{cases} ax + by + cz = u \\ dx + ey + dz = v \end{cases}$$

(Hint: first figure out what the matrix and the two vectors should be.)

Exercise 72 Can you find a , b , c , and d such that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

is true, no matter what values x and y take?

2.9 A shorthand notation

The advantage of the dot-product notation is that we can use a single letter to stand for an entire matrix, like so:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{2.4}$$

so the problem of exercise 61 can be written as follows:

$$\mathbf{A} \cdot [x, y]^T = [u, v]^T.$$

Exercise 73 Let \mathbf{A} be as defined in equation (2.4) (with arbitrary values, not necessarily those you found in exercise 72) and let

$$\mathbf{M} = \begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix}. \tag{2.5}$$

Now verify that the following is true:

$$\mathbf{M} \cdot [u, v]^T = [x, y]^T.$$

where the two vectors also satisfy equation (2.4). (Hint: work out the multiplication on the left, and check that it agrees with your results in exercise 61.)

Exercise 74 With the matrices and vectors as in the previous exercise, verify the following:

$$\mathbf{M} \cdot \begin{bmatrix} a \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{M} \cdot \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

2.10 Multiplying matrices

In the last exercise, you multiplied the matrix \mathbf{M} by two vectors that constitute the “columns” of the matrix \mathbf{A} . The following notation condenses these two multiplications into a single one:

$$\mathbf{M} \cdot \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

or even more briefly:

$$\mathbf{M} \cdot \mathbf{A} = \mathbf{I}$$

where \mathbf{I} is standard notation for a “square” matrix that has zeros everywhere except on the main diagonal. Condensed notation often makes the novice nervous; to assuage such worries, it helps to remember that you can always “unwrap” the notation into full form, and you should keep doing this until you do this automatically whenever you come across condensed notation.

Exercise 75 Let

$$\mathbf{P} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

and

$$\mathbf{Q} = \begin{bmatrix} k & l \\ m & n \end{bmatrix}$$

and verify that the right member of the following equation is correct:

$$\mathbf{P} \cdot \mathbf{Q} = \begin{bmatrix} \alpha k + \beta m & \alpha l + \beta n \\ \gamma k + \delta m & \gamma l + \delta n \end{bmatrix}.$$

(Hint: first, substitute the definitions of the two matrices into the equation. Then split the matrix \mathbf{Q} into two column vectors to get two systems of equations.)

Exercise 76 With the matrices \mathbf{P} and \mathbf{Q} as in the previous exercise, calculate the matrix

$$\mathbf{Q} \cdot \mathbf{P}.$$

Exercise 77 Choose numerical values of your own liking for the matrix elements of the previous two exercises (e.g. $\alpha = 3$, $\beta = 5$, \dots , $k = 1$, \dots) and substitute these into your results. Verify that $\mathbf{P} \cdot \mathbf{Q} \neq \mathbf{Q} \cdot \mathbf{P}$. (Note: if you just happen to choose values such that $\mathbf{P} \cdot \mathbf{Q}$ comes out the same as $\mathbf{Q} \cdot \mathbf{P}$, you may have a rare mathematical intuition!)

Exercise 78 With definitions (2.4) and (2.5), calculate $\mathbf{M} \cdot \mathbf{A}$ and $\mathbf{A} \cdot \mathbf{M}$. What do you notice? (Hint: you already did $\mathbf{M} \cdot \mathbf{A}$.)

Exercise 79 Consider the following two diagonal matrices: $\mathbf{D} = \begin{bmatrix} k & 0 \\ 0 & l \end{bmatrix}$ and $\mathbf{E} = \begin{bmatrix} m & 0 \\ 0 & n \end{bmatrix}$. Show that $\mathbf{D} \cdot \mathbf{E} = \mathbf{E} \cdot \mathbf{D}$.

2.11 The order of matrix multiplication matters

You have shown that the order of the two matrices that are multiplied together makes a difference as regards the result. We say that matrix multiplication does not in general *commute*. However, in some cases, pairs of matrices *do* commute, such as the ones in the last two exercises.

Exercise 80 Look back on exercise 75 and try to explain why it is not so surprising that matrix multiplication fails to be commutative in general.

2.12 Columns and rows

The ‘width’ of a matrix is called the number of *columns*, which are the vertical vectors that make up the matrix; the ‘height’ of the matrix is called the number of *rows*. If a matrix has r rows and c columns, we refer to it as an “ $r \times c$ matrix”. A *square* matrix has equal numbers of columns and rows. Identity matrices are always square.

Exercise 81 Can you show that if two square matrices are such that their dot product is an identity matrix, they must commute? (Hint: be sure to avoid assuming the very thing you are trying to prove!)

2.13 Inverse matrices

When two commuting square matrices multiply together to form an identity matrix, they are said to be each other’s *inverse*. The matrices \mathbf{A} and \mathbf{M} defined by equations (2.4) and (2.5) are an example, and we often write $\mathbf{M} = \mathbf{A}^{-1}$ (or, just as well, $\mathbf{A} = \mathbf{M}^{-1}$).

Exercise 82 With \mathbf{D} and \mathbf{E} as in exercise 79, calculate \mathbf{D}^{-1} and compare \mathbf{D}^{-1} to \mathbf{E} . (Moral: “being inverses” implies “commute” but not vice versa.)

Exercise 83 Consider the matrix multiplication $\mathbf{A} \cdot \mathbf{B} = \mathbf{C}$ where \mathbf{A} is $r_A \times c_A$, and \mathbf{B} is $r_B \times c_B$. Verify that you can only carry out the multiplication when $c_A = r_B$. Show that, if this latter condition is fulfilled, we find that the matrix \mathbf{C} is $r_A \times c_B$.

Exercise 84 Can you show that matrix multiplication is *associative*. (Hint: this means that $(\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C})$ is always true with no other requirements on the three matrices other than the obvious one that row and column numbers “match”.)

Exercise 85 Explain why we can only multiply a matrix *with itself* when it is square.

2.14 Matrix powers

More short notation: where \mathbf{A} is a square matrix, the dot product $\mathbf{A} \cdot \mathbf{A}$ is often written as \mathbf{A}^2 ; the dot product $\mathbf{A} \cdot \mathbf{A}^2$ is often written as \mathbf{A}^3 ; and so on: the n th power of a square matrix \mathbf{A} is \mathbf{A}^n ; below you will investigate the limit $\lim_{n \rightarrow \infty} \mathbf{A}^n$.

2.15 Transposition

Another sort of exponent is the ‘T’ that turns a row vector like $[x, y]$ into a column vector: $[x, y]^T = \begin{bmatrix} x \\ y \end{bmatrix}$. This also works with matrices, and it best explained by a few examples:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

and, again,

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}^T = \begin{bmatrix} \alpha & \gamma \\ \beta & \delta \end{bmatrix}.$$

This diagonal flipping is called *transposition* (hence the ‘T’).

Exercise 86 Write down the transposes of the following matrices:

$$\begin{bmatrix} k & l & m \\ n & o & p \\ q & r & s \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha & \beta & \gamma & \delta & \epsilon \\ \zeta & \eta & \vartheta & \kappa & \lambda \end{bmatrix}.$$

Exercise 87 Can you show that $(\mathbf{A} \cdot \mathbf{B})^T = \mathbf{B}^T \cdot \mathbf{A}^T$?

Exercise 88 Let $\mathbf{H} = \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ -\sqrt{1-\alpha^2} & \alpha \end{bmatrix}$. Determine \mathbf{H}^T and \mathbf{H}^{-1} . What do you notice? Can you explain this? (Hint: the ‘magical quantity’ $ad - bc$.)

2.16 The eigenvalue problem

A very important matrix equation is the so-called *eigenvalue problem*:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} u \\ v \end{bmatrix} \quad (2.6)$$

or

$$\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (2.7)$$

Exercise 89 Can you show that equation (2.7) will only have the ‘trivial’ solution $[u, v]^T = [0, 0]^T$ whenever $(a - \lambda)(d - \lambda)$ does not have the same value as bc ? (Hint: you encountered the ‘magical quantity’ $ad - bc$ before.)

Exercise 90 Solve the equation

$$(a - \lambda)(d - \lambda) - bc = 0 \quad (2.8)$$

for λ . (Hint: multiply out to obtain a quadratic equation in λ and use the “*abc*-formula”.)

2.17 Eigenvalues

The solutions (‘roots’) of equation (2.8) are called the *eigenvalues* of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, which we will denote as λ_1 and λ_2 .

Exercise 91 Find a vector $[u, v]^T$ that satisfies equation (2.7) for $\lambda = \lambda_1$. This is the *eigenvector* associated with the eigenvalue λ_1 .

2.18 Eigenvectors

Yet more (obvious) notation: given the eigenvalue problem, equation (2.7), we denote the eigenvector associated with the eigenvalue λ_1 as $[u_1, v_1]^T$ and the eigenvector associated with the eigenvalue λ_2 as $[u_2, v_2]^T$. Given these two eigenvectors, we can express any given vector $[x, y]^T$ as a sum $c_1[u_1, v_1]^T + c_2[u_2, v_2]^T$. In the earlier exercises, you saw how to work out the coefficients c_1 and c_2 .

Exercise 92 Make sure that you can follow the steps in the following derivation:

$$\begin{aligned} \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \left(c_1 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + c_2 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} \right) \\ &= c_1 \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + c_2 \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} \\ &= c_1 \lambda_1^n \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + c_2 \lambda_2^n \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} \\ &= \lambda_1^n \left(c_1 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + c_2 (\lambda_2/\lambda_1)^n \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} \right). \end{aligned}$$

Exercise 93 Assume that $|\lambda_2/\lambda_1| < 1$ and hence show that $\lim_{n \rightarrow \infty} (\lambda_2/\lambda_1)^n = 0$. Conclude that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \begin{bmatrix} x \\ y \end{bmatrix} \approx c_1 \lambda_1^n \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}$$

for large values of n .

Exercise 94 Can you evaluate the following limit? $\lim_{n \rightarrow \infty} \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \cdot \begin{bmatrix} x \\ y \end{bmatrix}$.

3

Probability

3.1 Intuitive probability

Exercise 95 What is the probability¹ of throwing a two with a fair die? What is the probability of throwing an even number? What is the probability of throwing something less than six? Less than eleven (with a *single* die)? Less than one hundred?

Exercise 96 Where p is the probability of an event, can you show that the odds ratio r of this event is given by the formula

$$r = \frac{1-p}{p} ?$$

(Hint: the odds ratio is the way of expressing probability beloved of gamblers, as in “a million to one”.)

Exercise 97 In how many different ways can you throw “two eyes” with a die? What makes you believe that this number equals the number of ways you can throw a six? (Hint: think of the throwing of a die as a physical process.)

3.2 Events

The set of all things that could happen is called the *sample space*, denoted by Ω , and its elements are *elementary* events. The question what exactly constitutes an elementary event in Nature leads to many perplexities, physical as well as philosophical². However, it suffices to assume that we can never observe any occurrence so precisely that we would be able to tell one elementary event from another; the *events* we will deal with are much cruder, much more course-grained than elementary events. Thus, we are only concerned with subsets of Ω ; an event E corresponds to a subset of Ω and the probability of this event, $\mathbb{P}(E)$, is the aggregate of the probability carried by the elementary events within this subset.

Exercise 98 Explain why you must have $\mathbb{P}(\Omega) = 1$ for a valid \mathbb{P} function.

Exercise 99 Explain why you must have $\mathbb{P}(\emptyset) = 0$ for a valid \mathbb{P} function. (Hint: the empty set \emptyset is a subset of any set.)

Exercise 100 Draw a Venn diagram of Ω and an event $E \subset \Omega$. Cross-hatch, in different colours, the sets Ω and E , and consider $\mathbb{P}(E)$ as the *ratio* between the cross-hatched area of E and Ω .

Exercise 101 In your diagram for the previous exercise, cross-hatch (in a different colour) the event $\Omega \setminus E$ (that is, all of Ω except E). This is the *complement* of E , often denoted \bar{E} . Explain the law $\mathbb{P}(\bar{E}) = 1 - \mathbb{P}(E)$.

Exercise 102 Draw a Venn diagram of Ω and two *mutually exclusive* events $E_I \subset \Omega$ and $E_{II} \subset \Omega$. Cross-hatch the event $E_I \cup E_{II}$ and explain the rule $\mathbb{P}(E_I \cup E_{II}) = \mathbb{P}(E_I) + \mathbb{P}(E_{II})$. (Hint: ‘mutually exclusive’ means ‘no overlap’.)

Exercise 103 Repeat the previous exercise allowing some overlap between the events E_I and E_{II} , and explain the rule $\mathbb{P}(E_I \cup E_{II}) < \mathbb{P}(E_I) + \mathbb{P}(E_{II})$.

¹You will have some familiarity with the concept of probability; which is often expressed as a percentage; mathematicians express probability as a number in the interval $[0, 1]$.

²Modern probability theory sidesteps this question very neatly, but its development involves some subtle set-theoretic issues; we follow a more old-fashioned, pedestrian approach here.

Exercise 104 In the Venn diagram of the previous exercise, cross-hatch the area of the *overlap*, and consider $\mathbb{P}(E_I \& E_{II})$ as the *ratio* between the newly cross-hatched area and the area of Ω .

Exercise 105 In the Venn diagram of the previous exercise, consider the *ratio* between the newly cross-hatched area and the area of E_I .

3.3 Conditional probability

The ratio you considered in exercise 105 behaves very much like a probability; indeed it *is* a probability provided that we treat E_I as the ‘universe of discourse’ instead of Ω ; we say that we “restrict” to E_I , regarding it as an surrogate omega. This probability is said to be the probability of event II *given* E_I (also: the probability of event *given* E_I , and also: the probability *conditional* on E_I). We write this probability as $\mathbb{P}(E_{II} | E_I)$.

Exercise 106 Use the various ratios of cross-hatched areas you studied in the foregoing exercises to derive the fundamental formula of conditional probability:

$$\mathbb{P}(E_I | E_{II}) = \frac{\mathbb{P}(E_I \& E_{II})}{\mathbb{P}(E_{II})} . \quad (3.1)$$

Exercise 107 (Multiplication) Explain the following expression:

$$\mathbb{P}(E_I \& E_{II}) = \mathbb{P}(E_{II} | E_I) \mathbb{P}(E_I) = \mathbb{P}(E_I | E_{II}) \mathbb{P}(E_{II}) .$$

Exercise 108 (Total probability) Draw a new Venn diagram (Ω) with two dividing lines; thus forming three mutually exclusive (non-overlapping) and *exhaustive* (covering all of Ω) events, which you call E_I , E_{II} and E_{III} . Draw in a fourth event A which may overlap with each of the first three. Now deduce the law

$$\mathbb{P}(A) = \mathbb{P}(A | E_I) \mathbb{P}(E_I) + \mathbb{P}(A | E_{II}) \mathbb{P}(E_{II}) + \mathbb{P}(A | E_{III}) \mathbb{P}(E_{III}) .$$

Exercise 109 (Bayes’ rule) With the Venn diagram of the previous exercise, show that the following formula holds:

$$\mathbb{P}(E_{\star} | A) = \frac{\mathbb{P}(A | E_{\star}) \mathbb{P}(E_{\star})}{\mathbb{P}(A | E_I) \mathbb{P}(E_I) + \mathbb{P}(A | E_{II}) \mathbb{P}(E_{II}) + \mathbb{P}(A | E_{III}) \mathbb{P}(E_{III})} .$$

Exercise 110 Generalize the results of the previous two exercises to n instead of 3 mutually exclusive, exhaustive events.

Exercise 111 (Independence) Two events E_I and E_{II} are said to be *independent* if $\mathbb{P}(E_{II} | E_I) = \mathbb{P}(E_{II})$. Show that this implies $\mathbb{P}(E_{II} \& E_I) = \mathbb{P}(E_I) \mathbb{P}(E_{II})$ and $\mathbb{P}(E_I | E_{II}) = \mathbb{P}(E_I)$.

Exercise 112 Suppose you toss two fair coins. The relevant events are then (H,H), (H,T), (T,H) and (T,T), where the letters mean ‘heads’ and ‘tails’. Let the variable H denote the number of heads. Calculate $\mathbb{P}(H = 0)$, $\mathbb{P}(H = 1)$, $\mathbb{P}(H = 2)$, $\mathbb{P}(H = 3)$, $\mathbb{P}(H \leq 2)$. (Hint: be sure to notice that both (H,T) and (T,H) map to the number 1.)

3.4 Random variables

The idea of a *random variable*³, such as H in the last exercise, is to assign numbers to events (i.e. to subsets of Ω). A random variable is thus a *map* from Ω to \mathbb{R} .

Exercise 113 The real line \mathbb{R} can itself be considered as a sample space. The events are half-open intervals, such as $(-\infty, a]$ for $a \in \mathbb{R}$, and other subsets of \mathbb{R} formed from such intervals by complementation, intersection and unions. Verify that these other subsets then include intervals of the forms $(a, +\infty)$, $[a,]$, $[a, b]$, $(a, b]$, $(a, b) \cup (c, d)$. and so forth.

³Also known as (*random*) *variate*, *aleatory variable*, *stochastic function*.

3.5 Assigning probabilities

If we have assigned probabilities to those subsets of Ω that correspond to subsets of the real line as explored in exercise 113, for a given random variable X , then we can evaluate probabilities of the kind $\mathbb{P}(X \in \mathcal{I})$ where \mathcal{I} is such a subset of \mathbb{R} .

Exercise 114 Identify the obvious problem if this condition on probabilities for subsets of Ω is not fulfilled. On the other hand, is it problematic if probabilities can be or have been assigned to events in Ω that do not correspond to events in \mathbb{R} of the sort considered in exercise 113?

3.6 The Bernoulli variate

Perhaps the simplest random variable maps the elements of Ω to either 0 or 1. This is the *Bernoulli variate* B ; it requires Ω to be divided into just two mutually exclusive, exhaustive events: $E_0 \cup E_1 = \Omega$, with $\mathbb{P}(B = 1) = \mathbb{P}(E_1) = p$ where $p \in [0, 1]$ is the (single) parameter of the Bernoulli distribution.

Exercise 115 Deduce that $\mathbb{P}(B = 0) = \mathbb{P}(E_0) = 1 - p$.

3.7 The binomial variate

The *binomial variate* X with parameters p and N is defined by $X = B_1 + B_2 + \cdots + B_N$ where the B s are N independent Bernoulli variates, all with the same parameter p .

Exercise 116 Revisit exercise 112, and view H as a Binomial variate with parameters $p = \frac{1}{2}$ and $N = 2$. Verify by direct calculation and comparison with your earlier results that

$$\mathbb{P}(H = x) = \binom{2}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{2-x}$$

for $x = 0, 1, 2$. (Hint: recall that $\binom{2}{x} = \frac{2!}{(2-x)!x!}$.)

Exercise 117 Make sure that you understand the general formula for a Binomial variate with parameters p and N :

$$\mathbb{P}(X = x) = \binom{N}{x} p^x (1-p)^{N-x} \quad \text{for } x = 0, 1, \dots, N.$$

(Hint: recall that $\binom{N}{x} = \frac{N!}{(N-x)!x!}$; consult a textbook if necessary.)

Exercise 118 Let X be a Binomial variate with parameters $p = \mu/N$ and N where μ is a fixed positive parameter. Can you prove the following result?

$$\lim_{N \rightarrow \infty} \mathbb{P}(X = x) = \frac{\mu^x e^{-\mu}}{x!}.$$

3.8 The distribution function

A random variable X can be characterized by means of its *distribution function* defined as $F_X(x) = \mathbb{P}(X \leq x)$.

Exercise 119 Show that knowledge of the distribution function suffices to evaluate probabilities for events on the real line. In particular, verify the following key formula:

$$\mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a) \quad \text{where } a < b. \quad (3.2)$$

Exercise 120 Sketch a graph of the distribution function for the random variable H of exercise 112.

Exercise 121 With F_H the distribution function you determined in the previous exercise, use equation (3.2) with $b = 2$ to evaluate

$$\mathbb{P}(H = 2) = \lim_{n \rightarrow \infty} F_H(2) - F_H(a_n) \quad \text{where } a_n = 2 - \frac{1}{n}.$$

(Hint: it may be helpful that you already know the value of this probability from exercise 112.)

Exercise 122 Repeat the previous exercise, but now with $b = 2.4$.

Exercise 123 With X a random variable, can you show that the distribution function F_X must be non-decreasing, right-continuous, and have the following limits?

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow +\infty} F_X(x) = 1$$

(Hint: non-decreasing: if this were not the case, what happens with applications of formula (3.2)? Right-continuous: reflect that we must have $\mathbb{P}(X \leq x) = \lim_{n \rightarrow \infty} \mathbb{P}(X \leq x + \frac{1}{n})$.)

3.9 Discrete and continuous random variables

The foregoing exercises show that $\mathbb{P}(X = x)$ only evaluates to a non-zero number if F_X , the distribution function of the random variable X , has a ‘jump’ at x . Random variables whose distribution functions are step functions, consisting of at most denumerably many jumps, connected by horizontal segments, are called *discrete*. Random variables without any such jumps are called *continuous*.

Exercise 124 Does this nomenclature exhaust the possibilities?

3.10 A useful property

For continuous random variables, $\mathbb{P}(X = x) = 0$. Still, we would like to have some idea of how likely we are to find X assuming a value round about x . The probability $\mathbb{P}(X = x + \Delta x)$ for some $\Delta x > 0$ gives some idea of probability ‘near’ x .

Exercise 125 Show that $\mathbb{P}(X = x + \Delta x)$ is non-decreasing in Δx .

3.11 Probability density

To get a strictly local quantity, we would like to let $\Delta x \rightarrow 0$. But this limit evaluates to zero, so we are back where we started. However, we can obtain an indication of local *probability density* if we first divide by Δx , that is, we form $\mathbb{P}(X = x + \Delta x)/\Delta x$ and then take the limit.

Exercise 126 Show that this limit is $F'_X(x)$.

Exercise 127 Show that $F'_X(x) \geq 0$. (Hint: the properties listed in exercise 123).

3.12 Examples of probability density functions

The function $F'_X(x)$ is usually denoted $f_X(x)$, which is called the *probability density function*.

Exercise 128 Show that $\int_{-\infty}^{+\infty} f(x)dx = 1$. (Hint: the properties listed in exercise 123).

Exercise 129 In view of the properties examined in exercises 127 and 128, which of the following are valid probability density functions?

(i) $f(x) = \lambda e^{-\lambda x} 1_{[0, \infty)}(x)$;

(ii) $f(x) = \exp\{-\lambda x^2 - e^{-\epsilon x}\} 1_{[0, \infty)}(x)$;

(iii) $f(x) = \frac{\exp\{x\}}{2(1 + \exp\{x\})^2}$;

(iv) $f(x) = \frac{1}{\epsilon} 1_{\mathcal{X}}(x)$ where $\mathcal{X} = \left[-\frac{\epsilon + \epsilon^2}{2\epsilon}, -\frac{1}{2\epsilon}\right] \cup \left[\frac{1}{2\epsilon}, \frac{\epsilon + \epsilon^2}{2\epsilon}\right]$;

(v) $f(x) = \frac{1}{x} 1_{\mathcal{X}}(x)$ where $\mathcal{X} = [1, 2]$;

(vi) $f(x) = \frac{\exp\{x\}}{(1 + \exp\{x\})^2}$.

(Hint: notation: $\exp\{x\} = e^x$; $1_{\mathcal{X}}(x)$ is an *indicator function* mapping to 0 for $x \notin \mathcal{X}$ and 1 for $x \in \mathcal{X}$; parameters denoted by Greek letters are strictly positive.)

3.13 Important probability density functions

Important probability density functions are: (i) the *standard uniform distribution*: $f(x) = 1_{[0, 1]}(x)$; (ii) the *exponential distribution*: $f(x) = \lambda e^{-\lambda x} 1_{[0, \infty)}(x)$; (iii) the *standard normal distribution*: $f(x) = \exp\{-x^2/2\}/\sqrt{2\pi}$.

Exercise 130 Verify that these probability density functions satisfy the properties stated in exercises 127 and 128. (Hint: if $I = \int_{-\infty}^{+\infty} e^{-x^2} dx$, then

$$I^2 = \int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dx dy$$

which can be transformed to polar coordinates: $I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\phi$.)

Exercise 131 State the probability density function of the following continuous distributions: (i) logistic; (ii) lognormal; (iii) χ^2 (chi-square); (iv) Pareto; (v) Student's t . (Hint: you can find them in any textbook on the subject.)

Exercise 132 State the formula for $\mathbb{P}(X = x)$ where the random variable X follows each of the following discrete distributions in turn: (i) geometric; (ii) hypergeometric; (iii) Poisson. (Hint: this is also called a probability *density* function by some textbooks; you may also encounter the less confusing term *mass function*.)

3.14 Transformations of random variables

Let X be a random variable, and let $g(\cdot)$ be a strictly monotone function. Then $Y = g(X)$ is also a random variable.

Exercise 133 Show that $F_X(x) = F_Y(g(x))$. Hence deduce that

$$f_X(x) = f_Y(g(x))g'(x).$$

(Hint: use the definition: $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(g^{-1}(Y) \leq x)$; remember the chain rule.)

Exercise 134 Suppose that X is exponentially distributed, $F_X(x) = 1 - \exp\{-\lambda x\}$ for $x \geq 0$, $\lambda > 0$. Let $Y = g(X) = \exp\{-\alpha X\}$, $\alpha > 0$. Find $F_Y(y)$, the distribution function of Y . (Hint: recall the properties of the distribution function from exercise 123.)

Exercise 135 In your result of the previous exercise, let $\alpha = -\lambda$. Give a formula for $f_Y(y)$, the probability density function of Y . Have you encountered this probability density function before?

3.15 Probability transforms

In the last exercise, Y followed a standard uniform distribution, and $X = g^{-1}(Y)$ some other distribution. This suggests that we can obtain a *sample* from any desired distribution, if we are given a sample from the standard uniform distribution.

Exercise 136 Let N be some (very large) number, and consider the list of numbers $1, 2, \dots, N$. Let P_i be the fraction of numbers on the list that are smaller than, or equal to, i . Show that $\lim_{N \rightarrow \infty} P_i = F_U(i)$, where F_U is the distribution function of the standard uniform distribution.

Exercise 137 In the list of the previous exercise, consider a practical way in which you could give the numbers of the list a shuffle. (Hint: e.g. felt-tip pen, flashcards.)

Exercise 138 A *pseudorandom number generator* is a computational technique to generate a series of numbers that “emulate” (in some appropriate sense) a random shuffle of the numbers from 1 to N , where N is very big, without relying on a physical (and apparently random) process. Find out more about such techniques from the web or textbooks. Indicate how you can turn this list into an emulated sample taken from the standard uniform distribution.

Exercise 139 Given a distribution function F , let

$$G(y) = \min_{u \in \mathbb{R}} \{y \leq F(u)\}. \quad (3.3)$$

Can you show that where Y is standard uniformly distributed, the random variable $X = G(Y)$ has distribution function F ? (Hint: consider first the case where F is strictly monotone and show that $G = F^{-1}$; then proceed as in exercise 133; finally consider how you must adjust the recipe in the general case where F is *not* strictly monotone.)

Exercise 140 (The probability transform technique) Indicate how you would use the transformation described by equation (3.3) to generate a sample from a given distribution F with the aid of a pseudorandom number generator.

3.16 Expectation

The *expectation* of a random variable is defined as follows:

$$\mathbb{E}_X(X) = \int x dF(x) \quad (3.4)$$

which for discrete variates works out as

$$\mathbb{E}_X(X) = \sum_{i=0}^{\infty} x_i \mathbb{P}(X = x_i) \quad (3.5)$$

where i ranges over the values which the variate can assume; whereas for continuous variates the expectation is given by

$$\mathbb{E}_X(X) = \int_{-\infty}^{+\infty} x f_X(x) dx. \quad (3.6)$$

The expectation is also called the *mean*; as with the distribution function and the probability density function, we often omit the subscript (here: X) that identifies the random variable over which the expectation is taken, when there can be no confusion.

Exercise 141 Show that the expectation of a Bernoulli variate is its parameter p .

Exercise 142 Show that the expectation of an exponentially distributed variate is $1/\lambda$.

3.17 Expectation of a transformed random variable

The expectation of a function of a random variable is defined as follows:

$$\mathbb{E}(g(X)) = \int g(x)dF(x). \quad (3.7)$$

Exercise 143 Write down this formula for (i) discrete variates and (ii) continuous variates.

3.18 Variance

An important function to take for g in formula (3.7) is the squared deviation of the mean. The resulting expectation is known as the *variance*:

$$\mathbb{V}(X) = \int (x - \mathbb{E}(X))^2 dF(x). \quad (3.8)$$

Exercise 144 Show that the variance of a Bernoulli variate with parameter p is $(1-p)p$. For which value of p is this variance maximal?

Exercise 145 Consider the *degenerate distribution* with parameter μ : $\mathbb{P}(X = \mu) = 1$. Show that its expectation is μ and its variance is zero.

Exercise 146 Show that $\mathbb{V}(X) \geq 0$ for any random variable X .

Exercise 147 Let $\mu_X = \mathbb{E}(X)$. Can you show that the following is true?

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2. \quad (3.9)$$

Exercise 148 Let $M_X(t) = \mathbb{E}(e^{tX})$. Determine $M'_X(t)$ and $M''_X(t)$, and evaluate these derivatives at $t = 0$. (Hint: determine the derivatives of $\int e^{t\xi} dF(\xi)$ with respect to t .)

Exercise 149 Combine the results of the previous two exercises to express the mean and the variance in terms of $M'_X(0)$ and $M''_X(0)$.

Exercise 150 Demonstrate the validity of the following two rules where X is any random variable (and α and β are real constants):

$$\mathbb{E}(\alpha + \beta X) = \alpha + \beta \mathbb{E}(X) \quad (3.10)$$

and

$$\mathbb{V}(\alpha + \beta X) = \beta^2 \mathbb{V}(X). \quad (3.11)$$

Exercise 151 Consider two discrete random variables X_1 and X_2 , and assume that both take values from a common set $a_1, a_2, a_3, \dots, a_k, \dots$. Show that

$$\mathbb{P}(X_1 = a_k) = \mathbb{P}(X_1 = a_k \& X_2 = a_1) + \mathbb{P}(X_1 = a_k \& X_2 = a_2) + \mathbb{P}(X_1 = a_k \& X_2 = a_3) + \dots$$

(Hint: the law of total probability, exercise 108.)

Exercise 152 For the random variables of the previous exercise, consider the expectation of their sum:

$$\mathbb{E}(X_1 + X_2) = \sum_k \sum_\ell (a_k + a_\ell) \mathbb{P}(X_1 = a_k \& X_2 = a_\ell).$$

Can you establish the following result?

$$\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2).$$

(Hint: write the sum as $\sum_k a_k \sum_\ell \mathbb{P}(X_1 = a_k \& X_2 = a_\ell) + \sum_\ell a_\ell \sum_k \mathbb{P}(X_1 = a_k \& X_2 = a_\ell)$ and then apply the result of the previous exercise.)

3.19 Expectation of a sum

The result of the last exercise, “expectation of sum is sum of expectations”, applies also for sums of three or more variates. Moreover, it also applies for continuous random variables, where a *joint probability density function* f is defined for k joint random variables, such that

$$\int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f(t_1, \dots, t_k) dt_1 \cdots dt_k = F(x_1, \dots, x_k) \quad (3.12)$$

where

$$F(x_1, \dots, x_k) = \mathbb{P}(X_1 \leq x_1 \& \cdots \& X_k \leq x_k). \quad (3.13)$$

Exercise 153 Show that the expectation of a Binomial variate with parameters p and N is Np . (Hint: look at your result for a Bernoulli variate, exercise 141, and apply the rule for the expectation of a sum.)

3.20 Independence

When two continuous random variables are independent, their joint probability distribution $f(x_1, x_2)$ is of the form $f_1(x_1)f_2(x_2)$.

Exercise 154 Let X_1 and X_2 be two independent continuous random variables, and let g and h be two functions. Deduce the following result:

$$\mathbb{E}(g(X_1) \cdot h(X_2)) = \mathbb{E}(g(X_1)) \cdot \mathbb{E}(h(X_2)). \quad (3.14)$$

Exercise 155 For the random variables of the previous exercise, let $g(x) = h(x) = e^{xt}$ where t is a non-negative parameter. Show that $\mathbb{E}(e^{t(X_1+X_2)}) = \mathbb{E}(e^{t(X_1)})\mathbb{E}(e^{t(X_2)})$.

Exercise 156 Can you generalize the result of the previous exercise to show the following result? Where X_1, X_2, \dots, X_n are independent random variables, the following is true for their sum $\sum_{i=1}^n X_i$:

$$\mathbb{E}(e^{t \sum_{i=1}^n X_i}) = \prod_{i=1}^n \mathbb{E}(e^{t(X_i)}).$$

Exercise 157 Consider a random variable X characterized by

$$\mathbb{E}(e^{tX}) = \exp\left\{\sum_{i=1}^{\infty} \kappa_i t^i\right\}$$

where $\kappa_i = \nu 2^{i-1}/i$, where ν is a positive parameter. Show that $\mathbb{E}(X) = \nu$ and $\mathbb{V}(X) = 2\nu$. (Hint: use the results of exercise 149.)

Exercise 158 Let X be the random variable of the previous exercise, with parameter ν . Can you show that

$$\mathbb{E}(e^{tX}) = (1 - 2t)^{-\nu/2} ?$$

(Hint: consult the Taylor expansion of $\ln(1+x)$.)

3.21 The chi-square variate

The random variable of the last two exercises is said to be chi-square (χ^2) distributed with ν *degrees of freedom*.

Exercise 159 Let $X_1, X_2, \dots, X_i, \dots, X_n$ be independent random variables, where the i th random variable X_i is χ^2 distributed with ν_i degrees of freedom. Show that their sum $\sum_{i=1}^n X_i$ is also χ^2 distributed, with $\sum_{i=1}^n \nu_i$ degrees of freedom. (Hint: consider $\mathbb{E}(\exp\{t \sum_{i=1}^n X_i\})$; use the results of exercises 156 and 158.)

Exercise 160 Suppose that $\{X_1, X_2, \dots, X_n\}$ is a set (sample) of independent random variables, identically distributed with distribution function F_X . Let the distribution be characterized by means of the following identity:

$$\mathbb{E}(e^{tX}) = \exp\{\kappa_1 t + \kappa_2 t^2 + \kappa_3 t^3 + \dots\}$$

where $\kappa_1, \kappa_2, \kappa_3, \dots$ are non-negative coefficients. Let Z be another random variable (the *normalized sum*) defined by:

$$Z = \frac{\sum_{i=1}^n X_i - n\kappa_1}{\sqrt{2n\kappa_2}}.$$

Can you establish the following result?

$$\lim_{n \rightarrow \infty} \mathbb{E}(e^{tZ}) = e^{t^2/2}$$

(Hint: apply the result of exercise 156.)

Exercise 161 Suppose that X is a continuous random variable whose distribution function F_X is characterized in terms of $\mathbb{E}(e^{tX})$ as in the previous exercise. Can you derive the following bound?

$$F_X(\kappa_1 + 2\kappa_2 t + 3\kappa_3 t^2 + \dots) \geq 1 - \exp\{-\kappa_2 t^2 - 2\kappa_3 t^3 - 3\kappa_4 t^4 - \dots\}$$

3.22 The Central Limit Theorem

The result of exercise 160 is a basic version of the *Central Limit Theorem*. What is remarkable about this is that you only needed the first two coefficients, κ_1 and κ_2 (which correspond to the mean and variance of F_X), to normalize the sum; then, for a large number of terms (n) you find for the normalized sum Z that $\mathbb{E}(e^{tZ}) \approx e^{t^2/2}$ which means that Z follows the standard normal distribution *regardless* of the underlying distribution F_X .

A sum of many independent identically distributed random variables is approximately normally distributed, regardless of the distribution of these variables.

Exercise 162 Try to think of situations where this central result may prove useful.

Exercise 163 Can you generalize the calculation of exercise 160 where the terms are still independent, but each with its own distribution (i.e. $\mathbb{E}(e^{tX_i}) = \exp\{\kappa_{1,i}t + \kappa_{2,i}t^2 + \kappa_{3,i}t^3 + \dots\}$)?

Exercise 164 Let Z follow a standard normal distribution, and consider the random variable Z^2 . If $f_{Z^2}(u)$ is the probability density function of Z^2 , verify that $f_{Z^2}(u) = 0$ for $u < 0$.

Exercise 165 Continuing the previous exercise, can you show that Z^2 follows a χ^2 distribution with $\nu = 1$? (Hint: work out $\mathbb{E}(\exp\{tZ^2\})$.)

Exercise 166 Let Z and V be independent random variables, where Z follows the standard normal distribution and V is χ^2 distributed with ν degrees of freedom. Also let $T = Z/\sqrt{V/\nu}$. Verify the formula

$$\mathbb{P}(T \leq t) = \mathbb{E}_V(\mathbb{P}(Z \leq t\sqrt{V/\nu})). \quad (3.15)$$

3.23 Student's t

The random variable T in the last exercise follows *Student's t* distribution with ν degrees of freedom.

Exercise 167 Can you derive the probability density distribution of Student's t distribution from equation (3.15)? (Hint: refer to exercise 131.)

3.24 Covariance

Consider two random variables X_1 and X_2 and let $\mu_1 = \mathbb{E}(X_1)$, $\mu_2 = \mathbb{E}(X_2)$. The *covariance* of the two random variables is defined as follows:

$$\text{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2)) . \quad (3.16)$$

Exercise 168 Complete the following: $\text{Cov}(X, X) = \quad$. (Hint: this is the case where $X_1 = X_2 = X$, $\mu_1 = \mu_2 = \mu$.)

Exercise 169 Show that for a pair of independent random variables X_1 and X_2 , we have $\text{Cov}(X_1, X_2) = 0$. (Hint: derive this first for continuous variables, using equation (3.14).)

Exercise 170 Suppose that for a pair of random variables X_1 and X_2 , you have $\text{Cov}(X_1, X_2) = 0$. Does it follow that X_1 and X_2 are independent?

3.25 Variance of a sum

The last two exercises stress the fact that independence implies zero covariance, but not *vice versa*.

Exercise 171 Deduce the following law:

$$\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\text{Cov}(X_1, X_2) . \quad (3.17)$$

(Hint: $\mathbb{E}((X_1 + X_2 - [\mu_1 + \mu_2])^2)$.)

Exercise 172 Complete the following: $\mathbb{V}(X_1 - X_2) = \quad$. (Hint: pay close attention to the minus sign, and remember exercise 150.)

3.26 Some commonly used statistics

The following quantities, derived from the principal ones already introduced, are commonly used:

$$\text{standard deviation } \sigma_X = \sqrt{\mathbb{V}(X)} \quad (3.18)$$

$$\text{correlation coefficient } \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} . \quad (3.19)$$

Exercise 173 Verify that the correlation coefficient of two independent variates is zero.

Exercise 174 Derive the following formula:

$$\rho_{XY} = \mathbb{E} \left(\frac{(X - \mu_X)}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y} \right)$$

Exercise 175 Can you show that $\rho_{XY}^2 \leq 1$? (Hint: show that $|\rho_{XY}| \leq 1$; begin by deducing $\mathbb{V}((X - \mu_X)/\sigma_X + (Y - \mu_Y)/\sigma_Y) = 2(1 + \rho)$; derive a similar formula for $\mathbb{V}((X - \mu_X)/\sigma_X - (Y - \mu_Y)/\sigma_Y)$; recall from exercise 146 that variances are always non-negative.)

3.27 Conditioning on a rare event

Exercises 136—140 concerned a technique, called the *probability transform technique*, to realize a sample from any given distribution, given a device that generates a sample⁴ from a standard uniform distribution, such as a pseudorandom number generator. A special difficulty arises when you want to compute the expectation of your random variable of interest, conditional on a rare event.

⁴Such a sample is then said to consist of *realizations*, synonyms *drawings* and *instantiations*.

Exercise 176 Show that, given a sample X_1, \dots, X_N from a known distribution F , you use the following approximation formula:

$$\mathbb{E}(g(X) \mid X > \xi) \approx \frac{\sum_{k=1}^N g(X_k) 1_{X_k > \xi}}{\sum_{k=1}^N 1_{X_k > \xi}} \quad (3.20)$$

(where ξ is a given constant).

Exercise 177 Show that, given a sample X_1, \dots, X_N from F , you have

$$\mathbb{E}(\sum_{k=1}^N 1_{X_k > \xi}) = \mathbb{P}(X > \xi)N.$$

Which proportion of the sampled values $\{X_k\}$ is actually used in calculating the right-hand side of equation (3.20)?

Exercise 178 Can you explain why, to achieve reasonable accuracy with formula (3.20), the sample size N must be larger (typically: very much larger) than $1/\mathbb{P}(X > \xi)$?

3.28 A tricky question

The last two exercises bring home the problem when you are conditioning on a rare event (i.e. when $\mathbb{P}(X > \xi)$ is very small): you need huge samples, and most of the calculating effort is wasted on sample points that are “thrown away”.

Exercise 179 Let h be any non-zero function and let the distribution \tilde{F} be defined by⁵

$$\tilde{F}(x) = \int_{-\infty}^x h(u) dF(u).$$

Can you show that

$$\mathbb{E}(g(X) \mid X > \xi) \approx \frac{\sum_{k=1}^N g(X_k) 1_{X_k > \xi} / h(X_k)}{\sum_{k=1}^N 1_{X_k > \xi} / h(X_k)} ? \quad (3.21)$$

(Hint: you may want to prove this first for the case where F is a discrete distribution.)

3.29 An improved approximation for rare events

The key idea is that formula (3.21) provides a much improved approximation, as compared to formula (3.20), when you have $\tilde{F}(\xi) \ll F(\xi)$.

Exercise 180 Explain the improvement of the approximation. Motivate the choice $h(x) = e^{\lambda x}$ for some suitable value of λ .

Exercise 181 What would you have to change in the above procedure if you replaced the event $X > \xi$ by any other rare event $X \in \mathcal{X}$ with $\mathbb{P}(X \in \mathcal{X}) \ll 1$?

3.30 Sampling complicated probabilistic structures

The probability transform technique is not suitable in some situations. One common situation is that the distribution of interest is not known in closed form or computationally cumbersome. This problem occurs when your random variable of interest (say, X) is itself a function of a probabilistic “state” or “configuration” characterized by some (possibly quite high-dimensional) random vector \mathbf{Y} . Typically, you will be able to compute the probability of any one given configuration,

⁵If you are unfamiliar with the notation ‘ $dF(u)$ ’ in the integrand, just read ‘ $F'(u)du$ ’ for it.

even if the distribution of X , which is a function of the configuration, is intractable. The following exercises develop a method to generate a sample⁶

$$X_1, X_2, \dots, X_N$$

from the distribution of X .

Exercise 182 Review the following key results: (i) the discrete probability distribution $x = x_1, \dots, x_n$ of a Markov chain with transition probabilities p_{ij} is stable if the *detailed balance equation*

$$x_i p_{ij} = x_j p_{ji} \tag{3.22}$$

is satisfied for all pairs (i, j) ; (ii) if the Markov chain is connected, the frequency with which an instantiation visits state i equals x_i .

3.31 Markov Chain Monte Carlo

The idea of *Markov Chain Monte Carlo* is to choose the transition probabilities in such a way that the detailed balance equations (3.22) are satisfied with x_i equalling the probability of configuration i for all i .

Exercise 183 Show that the path of the instantiation then constitutes the desired sample

$$X_1, X_2, \dots, X_N$$

provided that the transition probabilities are such that the chain is connected.

Exercise 184 Given that the k th point (X_k) of your sample is the i th configuration, discuss how to select point $(k+1)$ by enumerating p_{i1}, \dots, p_{in} and selecting the next step with a pseudorandom number generator.

Exercise 185 Verify that n is the number of configurations, and suppose that this number is astronomically large. Can you think of a way to carry out the “next step” procedure discussed in the previous exercise without enumerating the transition probabilities?

3.32 Choosing the next step

One solution to the problem posed in the last exercise is to employ a two-step procedure. *First*, select any configuration (i.e. chain state) at random; *Second*, either carry out a transition to the selected configuration (with acceptance probability a_{ij}) or reject the selected configuration (with probability $1 - a_{ij}$) and return to the first step.

Exercise 186 Show that the two steps can be carried out with a pseudorandom number generator.

Exercise 187 Verify that the two-step procedure only requires the calculation of acceptance probabilities for the configuration proposed in the first step (thus avoiding the need to enumerate p_{ij} for all j).

Exercise 188 Show that, for the two-step procedure, the probability

$$\mathbb{P}(X_{k+1} \text{ is state } j \mid X_k \text{ is state } i)$$

will be extremely well approximated by a_{ij}/n if n is very large. Hence infer that the detailed balance requirement becomes that

$$P_i a_{ij} = P_j a_{ji} \tag{3.23}$$

for all pairs (i, j) , where P_i is the probability of configuration i .

⁶Strictly speaking, a *pseudo-sample* inasmuch as you will typically use a pseudorandom number generator to furnish the necessary samples from the standard uniform distribution.

Exercise 189 Verify that the following formula for the acceptance probability:

$$a_{ij} = \min \left(1, \frac{P_j}{P_i} \right)$$

satisfies the detailed balance equation (3.23). (Hint: assume, without loss of generality, that $P_i > P_j$. Do not forget to verify that the chain is connected.)

Exercise 190 Verify that the following formula for the acceptance probability:

$$a_{ij} = P_j (1 - P_i P_j)^{n^2/2}$$

satisfies the detailed balance equation (3.23).

Exercise 191 Show that the formula of exercise 189 is optimal in the sense that a proposed transition to a less likely configuration will *always* be accepted. Can you explain why ‘optimal’ is an appropriate term in this context?

II

Core skills

4

Dynamics

4.1 State-transition functions

The *state-transition function* S of a deterministic dynamical system gives the *state* x at time t_1 given the value of the state at time t_0 :

$$x(t_1) = S(t_1, t_0, \xi) \quad \text{where } x(t_0) = \xi. \quad (4.1)$$

Thus, S “moves the state x forwards in time” by an amount $t_1 - t_0$.

Exercise 192 Let $S(t_1, t_0, \xi) = (t_0/t_1)^2 \xi$, and let $t_0 = 1$, $x(t_0) = \xi = 3$. Calculate $x(5)$. Calculate $x(8)$. Sketch a graph of $x(t)$ for $t \in [-3, 10]$. Also sketch a graph of $x(t)$ with the same state-transition function, but with $t_0 = 2$ and $x(t_0) = \xi = 3$.

Exercise 193 Motivate the *first consistency condition*:

$$S(t_0, t_0, \xi) = \xi. \quad (4.2)$$

(Hint: how far is the state “moved forwards in time”?)

Exercise 194 Motivate the *second consistency condition*:

$$S(t_2, t_0, \xi) = S(t_2, t_1, S(t_1, t_0, \xi)). \quad (4.3)$$

(Hint: use equation (4.1) to deduce that both sides of equation (4.3) must be expressions for $x(t_2)$; in a *deterministic* system $x(t_2)$ is fixed once $x(t_0)$ has been specified.)

4.2 Continuous time

In a *continuous-time* dynamical system, the values for t_0, t_1, t_2, \dots can be arbitrarily close together on the real line. Thus we can consider the partial derivative of the state-transition function with respect to its first argument:

$$S_{\{t_1\}}(t, \tau, \xi) = \lim_{\Delta t \rightarrow 0} \frac{S(t + \Delta t, \tau, \xi) - S(t, \tau, \xi)}{\Delta t}. \quad (4.4)$$

Exercise 195 Let $S(t_1, t_0, \xi) = (t_0/t_1)^2 \xi$. Verify that $S_{\{t_1\}}(t_1, t_0, \xi) = -2t_0^2 t_1^{-3} \xi$.

Exercise 196 Suppose that, in a dynamical system with a 1-dimensional state $x \in \mathbb{R}$, the state-transition function is defined in terms of two other functions f and g , as follows:

$$S(t_1, t_0, \xi) = f(t_1, t_0, \xi) \exp\{g(t_1, t_0, \xi)\} \quad (4.5)$$

and suppose, furthermore, that $g(t_0, t_0, \xi) = 0$ for all values of ξ . Then show that $f(t_0, t_0, \xi) = \xi$ for all values of ξ . (Hint: notation: $\exp\{x\} \equiv e^x$; S must satisfy the consistency conditions.)

Exercise 197 With the state-transition function given by equation (4.5), calculate $S_{\{t_1\}}$.

Exercise 198 Can you explain why we have the following for continuous-time dynamics?

$$\frac{dx(t)}{dt} = S_{\{t_1\}}(t, t, x(t)). \quad (4.6)$$

(Hint: define the derivative as $\Delta x/\Delta t$ as $\Delta t \rightarrow 0$; first show that $\Delta x = S(t + \Delta t, t, x(t)) - S(t, t, x(t))$.)

4.3 Bernoulli's method

Another common appearance of the differential equation (4.6) is

$$\dot{x}(t) = F(t, x(t)) \quad (4.7)$$

where $F(\tau, \xi) \equiv S_{\{1\}}(\tau, \tau, \xi)$.

Exercise 199 (Bernoulli's method) Can you verify that the state-transition function given by equation (4.5) with

$$g(t, t_0, \xi) = \int_{t_0}^t \phi(\tau, t_0, \xi) d\tau \quad (4.8)$$

and

$$f(t, t_0, \xi) = x_0 + \int_{t_0}^t \psi(\tau, t_0, \xi) \exp\{-g(\tau, t_0, \xi)\} d\tau \quad (4.9)$$

is associated with the following differential equation?

$$\dot{x}(t) = \psi(t, t_0, x_0) + \phi(t, t_0, x_0)x(t) \quad (4.10)$$

where $x(t_0) = \xi$. (Hint: use your results in exercises 196 and 197; notice first that $g_{\{1\}}(t, t_0, x_0)$ corresponds to $\phi(t, t_0, x_0)$ and use the convention that $g(t_0, t_0, \xi) = 0 \forall \xi$.)

Exercise 200 Find the state-transition function corresponding to the following differential equation:

$$\dot{x}(t) = \alpha t - \lambda x(t)$$

with $x(t_0) = \xi$. (Hint: refer to equation (4.10), setting $\phi \equiv -\lambda$ and $\psi \equiv \alpha t$, and evaluate f and g according to equations (4.8) and (4.9); $\frac{d}{dt}(te^{\lambda t}/\lambda) = te^{\lambda t} + e^{\lambda t}/\lambda$.)

Exercise 201 Find the state-transition function corresponding to the following differential equation:

$$\dot{x}(t) = \alpha t - \beta t x(t)$$

with $x(t_0) = \xi$. (Hint: follow the same procedure as in exercise 200 setting $\phi \equiv -\beta t$ and $\psi \equiv \alpha t$; $\frac{d}{dt} \exp\{\beta t^2/2\} = \beta t \exp\{\beta t^2/2\}$.)

Exercise 202 Find the state-transition function corresponding to the following differential equation:

$$\dot{x}(t) = \rho e^{-\lambda t} x(t)$$

with $x(t_0) = \xi$. (Hint: follow the same procedure as in exercise 200 setting $\phi \equiv \rho e^{-\lambda t}$ and $\psi \equiv 0$.)

4.4 The ubiquitous use of differential equations

For a wide range of scientific problems involving the quantitative description of "processes", it turns out that differential equations provide a natural medium for the description of these processes.

Exercise 203 Can you explain why this is so?

4.5 Finding the state-transition function

Given a differential equation, you will typically want to find (or at least characterize) the accompanying state-transition function. Finding the state-transition function corresponding to a differential equation is called *solving* the differential equation¹.

When a state-transition function $S(t_1, t_0, \xi)$ is presented as a putative solution of a differential equation $\dot{x}(t) = F(t, x(t))$, with boundary condition $x(t_0) = x_0$, two things need to be verified: the first consistency condition $S(t_0, t_0, x_0) = x_0$, and $S_{\{t_1\}}(t, t_0, \xi) = F(t, S(t, t_0, x_0))$.

Exercise 204 Perform these two checks on the solutions you found in exercises 200–202.

Exercise 205 Given a dynamical system with state $x \in \mathbb{R}$, $x(t) = S(t, t_0, x(t_0))$, and a strictly increasing function R , define the variable u by

$$x = R(u) + x(t_0) \quad (4.11)$$

and show that u satisfies the following differential equation:

$$\dot{u}(t) = \frac{S_{\{t_1\}}(t, t, R(u) + x(t_0))}{R'(u(t))} = \frac{\dot{x}(t)}{R'(u(t))}. \quad (4.12)$$

Exercise 206 (Separation of variables) Can you show that a differential equation whose right member is a product of two functions, one of which is a function of only the state variable, and the other is a function of only time, like so:

$$\dot{u}(t) = \phi(u)\psi(t) \quad (4.13)$$

has the following solution?

$$\int_{t_0}^t \psi(\tau) d\tau = \int_{u(t_0)}^{u(t)} \frac{ds}{\phi(s)}. \quad (4.14)$$

(Hint: compare equations (4.12) and (4.13); set $\phi(u) = 1/R'(u)$ and $\psi(t) = \dot{x}(t)$. Integrate these and observe that the left and right members of equation (4.14) are both expressions for $x(t) - x(t_0)$.)

Exercise 207 Use formula (4.14) to solve

$$\dot{u}(t) = u^\alpha t^\beta$$

where $t \geq t_0 = 1$, $u(t_0) = 1$, and $\alpha \neq 1$, $\beta \neq -1$.

Exercise 208 Do the previous exercise for the case where $\beta = -1$.

Exercise 209 Use the formula (4.14) to solve

$$\dot{x}(t) = \rho x(1-x)t^\alpha$$

where $\alpha \neq -1$ and $x(t_0) = \xi < 1$.

¹The ‘unknown’ for which the differential equation is solved is the *function* which maps time t to x at time t . A differential equation specifies a function by giving a relation between that function and its derivative(s).

4.6 Autonomy

A differential equation is *autonomous* if its right member depends on t through the state x , but *only* through the state:

$$\dot{x}(t) = F(x(t)) \quad (4.15)$$

(cf. equation (4.7)).

Exercise 210 Which of the following differential equations is autonomous?

- (i) $\dot{x}(t) = V \frac{x(t)}{K+x(t)}$
- (ii) $\dot{x}(t) = \alpha x(t) + \beta t^2$
- (iii) $\dot{x}(t) = \cos(t) - \lambda x(t)$
- (iv) $\dot{x}(t) = (e^{\mu t} + x(t))^{-1}$

(Hint: look out for appearances of t in the right member other than in $x(t)$; these indicate that the equation is non-autonomous.)

Exercise 211 (Barrow's formula) Use formula (4.14) to show that the autonomous differential equation is solved by:

$$t - t_0 = \int_{x(t_0)}^{x(t)} \frac{d\xi}{F(\xi)}$$

whenever $F(\xi) \neq 0$ between the integration limits.

4.7 Finding solutions

Solving differential equations is a difficult art. Fortunately, much information about the state-transition function can already be gleaned from the differential equation itself, without the need for an explicit solution.

Exercise 212 Consider the autonomous differential equation:

$$\dot{x}(t) = F(x) = x - x^2 = x(1 - x) .$$

Sketch $F(x)$ as a function of x . Mark on the x -axis the *critical points* where $F(x) = 0$; color *green* the part(s) of the x -axis where $F(x) > 0$, and *red* the parts(s) where $F(x) < 0$. If the initial condition $x(t_0)$ is in a green region, what will the solution look like (i.e. will x go up or down, will it continue to do so or approach an eventual value, *et cetera*)? Ditto for an initial condition in a red region.

Exercise 213 Draw a polynomial of your own choosing (just a smooth line that wiggles around the x -axis a couple of times is fine) which you regard as $F(x)$ as in the previous exercise and repeat the red/green segments procedure.

Exercise 214 Repeat the previous exercise, but now make sure that a critical point coincides with an extremum of your polynomial.

Exercise 215 Consider an autonomous differential equation $\dot{x}(t) = F(x(t))$ where $F(x) > 0$ for $x_\alpha < x < x_\omega$, $F(x_\alpha) = F(x_\omega) = 0$, and $|F(x)| \leq L_\omega |x_\omega - x|$ where L_ω is a finite positive constant. Can you establish the following?

$$\lim_{t \rightarrow \infty} x(t) = x_\omega \quad \text{whenever } x(t_0) \in (x_\alpha, x_\omega) .$$

If, in addition, we have $|F(x)| \leq L_\alpha |x_\alpha - x|$ can you similarly prove the following?

$$\lim_{t \rightarrow -\infty} x(t) = x_\alpha \quad \text{whenever } x(t_0) \in (x_\alpha, x_\omega) .$$

Exercise 216 Consider two distinct state-transition functions ($\xi < 0$ and $t > t_0$ in both cases):

$$S(t, t_0, \xi) = (\xi^{1/3} + \frac{1}{3}(t - t_0))^3$$

and

$$S(t, t_0, \xi) = \begin{cases} (\xi^{1/3} + \frac{1}{3}(t - t_0))^3 & \text{for } t \leq t_0 - 3\xi^{1/3} \\ 0 & \text{for } t > t_0 - 3\xi^{1/3} \end{cases} .$$

Show that *both* state-transition functions satisfy the autonomous differential equation

$$\dot{x}(t) = x^{2/3} . \quad (4.16)$$

4.8 Violation of the 'L' condition

The non-uniqueness in the last exercise arises because the critical point (here: $x = 0$) is reached in finite time. This cannot happen when the condition with L_ω given in exercise 215 is satisfied, as was shown in that exercise.

Exercise 217 Show that the right member of equation (4.16) fails to satisfy the 'L' condition. (Hint: show that there is no finite L such that $|x^{2/3}| \leq L|x|$ for all $x < 0$; it may be helpful to sketch $x^{2/3}$ as a function of x , paying particular attention to the slope at $x = 0$.)

Exercise 218 Can you sketch the solutions of

$$\dot{x}(t) = \frac{1}{1 - x(t)}$$

where $x(t_0) < 1$ and where $x(t_0) > 1$?

4.9 Arrowheads for qualitative analysis

Instead of colouring segments of the abscissa red and green, as you did in exercise 212, you can alternatively use little arrowheads pointing to the left and to the right, respectively.

Exercise 219 Explain how this arrowheads recipe allows you to identify (putative) alpha and omega points.

Exercise 220 Consider

$$\dot{x}(t) = F(x(t)) = \lambda - x(t)^2 .$$

Draw sketches of $F(x)$ for (i) $\lambda < 0$; (ii) $\lambda = 0$; (iii) $\lambda > 0$. Carry out the red/green segments (or arrowheads) procedure to identify alpha and omega points (if any).

Exercise 221 Repeat the previous exercise for the following differential equation:

$$\dot{x}(t) = F(x) = x(t)(\lambda - x(t)) .$$

Exercise 222 Consider

$$\dot{x}(t) = F(x(t)) = \lambda - \alpha x(t) + \beta \frac{x(t)^2}{1 + x(t)^2}$$

where $x \geq 0$, $\lambda \geq 0$, and $0 < \alpha < \beta < 2\alpha$. Draw sketches of $F(x)$ for various values of λ . Can you also sketch a *bifurcation* diagram, which has λ as abscissa², and the x -values for which $F(x) = 0$ (at the given λ) as ordinate?

²The *abscissa* is the horizontal axis, the *ordinate* is the vertical axis. We often also say "x-axis" and "y-axis", knowing full well that the quantities along these axes may bear different names.

4.10 The phase plane

So far the state x was 1-dimensional. Next, consider 2-dimensional systems: $x \in \mathbb{R}^2$. We write $x = (x_1, x_2)$, $\xi = (\xi_1, \xi_2)$. Notice that $S(t, t_0, \xi)$ can be represented by a point in the plane (x_1, x_2) (called the *phase plane*); for instance, $S(t_0, t_0, \xi)$ corresponds to the point (ξ_1, ξ_2) . As t ranges over \mathbb{R} , the state-transition function S ‘sweeps out’ a set of points in the phase plane, usually a smooth curve, which is the (*phase path*) (*passing through* ξ).

Exercise 223 Let $t_0 = 0$ and

$$\begin{cases} x_1(t) &= ae^{-\lambda t} \\ x_2(t) &= be^{-\lambda t} + ce^{-\gamma t} \end{cases}$$

Sketch the phase path for $a = 1$, $b = 3$, $c = 3$, $\gamma = 2\lambda$. (Hint: find x_2 as function of x_1 ; note that $\exp\{-\gamma t\} = (\exp\{-\lambda t\})^{\gamma/\lambda}$; make sure that your curve *only* contains points ‘visited’ by the system at some time t .)

Exercise 224 For the system of the previous exercise, verify that $\xi_1 = a$ and $\xi_2 = b + c$ and show that

$$\lim_{t \rightarrow \infty} S(t, t_0, \xi) = (0, 0)$$

for all $\xi = (\xi_1, \xi_2)$. Sketch a few more paths through different choices of ξ . Put arrowheads on your paths to indicate the motion toward the origin.

Exercise 225 Show that the system

$$\begin{cases} x_1(t) &= ae^{-\lambda\sqrt{t}} \\ x_2(t) &= be^{-\lambda\sqrt{t}} + ce^{-\gamma\sqrt{t}} \end{cases}$$

has the *same* collection of phase paths as the system of the previous two exercises.

4.11 The phase flow

The entirety of phase paths (which you cannot draw, since you’d fill up your phase plane with ink!) makes up the *phase flow* of the system. The last exercise showed that the phase flow only determines a dynamical system up to a monotone increasing transformation of time. In exercise 224, you saw that distinct paths can converge at the *omega point* (here: $(0, 0)$) which is attained in the infinite future. This is generally true of deterministic autonomous systems: paths can only coincide (or intersect) at omega points or *alpha points*, which are attained in the infinite past (that is, limiting points as $t \rightarrow -\infty$).

Exercise 226 Can you explain why phase paths can only coincide (intersect) in alpha or omega points?

Exercise 227 In exercise 224, you drew in arrowheads. In the system

$$\begin{cases} \dot{x}_1(t) &= F_1(x_1(t), x_2(t)) \\ \dot{x}_2(t) &= F_2(x_1(t), x_2(t)) \end{cases} \quad (4.17)$$

relate the direction of these arrows to the signs of F_1 and F_2 . (Hint: consider the ‘direction vector’ $[F_1, F_2]^T$.)

4.12 The phase portrait

A *phase portrait* is a picture of the phase plane together with selected phase paths, so as to convey an impression of the phase flow of the autonomous system (4.17). Of special importance in constructing a phase portrait are the critical points, where $F_1(x_1, x_2) = F_2(x_1, x_2) = 0$, and the *nullclines*. The x_1 -*nullcline* is the set of points where $F_1(x_1, x_2) = 0$; the x_2 -*nullcline* is the set of points where $F_2(x_1, x_2) = 0$.

Exercise 228 Explain why the critical points are at the intersection of the x_1 -nullcline and the x_2 -nullcline.

Exercise 229 Suppose that a path crosses the x_1 -nullcline at a point which is *not* a critical point. What can you say about the arrowhead at that point? (Hint: consider the above definition of the x_1 -nullcline and look at the ‘direction vector’ $[F_1, F_2]^T$.)

Exercise 230 What can you say about the arrowheads at the critical points?

Exercise 231 Consider the following autonomous system for $x_1 \geq 0, x_2 \geq 0$:

$$\begin{cases} \dot{x}_1 &= \rho x_1 - ax_1^2 - bx_1x_2 \\ \dot{x}_2 &= \mu x_2 - cx_2^2 - dx_1x_2 \end{cases} \quad (4.18)$$

(x_1 and x_2 are functions of time t , as before; this is not explicitly indicated here for the sake of brevity). Show that the x_1 -nullcline consists of (i) the x_2 -axis and (ii) a straight line connecting the point $\frac{\rho}{b}$ on the x_2 -axis to the point $\frac{\rho}{a}$ on the x_1 -axis.

Exercise 232 For system (4.18), find the x_2 -nullcline. (Hint: it consists of two straight lines, much like the x_1 -nullcline.)

Exercise 233 Sketch the phase plane (for $x_1 \geq 0, x_2 \geq 0$) and sketch the two nullclines, using different colours; assume $\frac{\rho}{a} > \frac{\mu}{d}, \frac{\mu}{c} > \frac{\rho}{b}$. Indicate the critical points.

Exercise 234 In your sketch of the previous exercise, draw short vertical (horizontal) lines on the x_1 -nullcline (x_2 -nullcline), and add arrowheads in the appropriate direction. (Hint: consider the ‘direction vector’ $[F_1, F_2]^T$.)

Exercise 235 In your sketch of the previous exercises, you should have four regions bounded by axes and nullclines. In each of those, consider the *signs* of F_1 and F_2 , and draw a little arrow pointing NE, SE, SW, or NW to indicate the approximate direction of the phase flow.

Exercise 236 In your sketch of the previous exercises, can you tell which critical points are alpha points, and which are omega points? (Hint: three are straightforward. One is a mixed alpha/omega point.)

4.13 Finding whence & whither

The last exercise shows that the classification of alpha and omega points is not always straightforward. Things are somewhat easier when, in system (4.17), F_1 and F_2 depend linearly on their arguments.

Exercise 237 Consider the system

$$\begin{cases} \dot{x}_1 &= \xi_1 e^{\lambda_1 t} \\ \dot{x}_2 &= \xi_2 e^{\lambda_2 t} \end{cases} \quad (4.19)$$

Draw a phase portrait for the case where $\lambda_1 = \lambda_2 < 0$.

Exercise 238 Draw a phase portrait for system (4.19) in the case where $\lambda_1 = \lambda_2 > 0$.

Exercise 239 Draw a phase portrait for system (4.19) in the case where $\lambda_1 \neq \lambda_2$, $\lambda_1 < 0$, $\lambda_2 < 0$.

Exercise 240 Draw a phase portrait for system (4.19) in the case where $\lambda_1 \neq \lambda_2$, $\lambda_1 < 0$, $\lambda_2 > 0$.

Exercise 241 Verify that system (4.19) satisfies the differential equation

$$\dot{x} = \mathbf{D} \cdot x$$

where $\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$.

Exercise 242 With x given by system (4.19), let

$$\begin{cases} u_1 &= \alpha x_1 + \beta x_2 \\ u_2 &= \gamma x_1 + \delta x_2 \end{cases} \quad (4.20)$$

or, briefly $u = \mathbf{R} \cdot x$, where \mathbf{R} is a 2×2 matrix collecting the coefficients of system (4.20). Verify that

$$\dot{u} = \mathbf{R}^{-1} \cdot \mathbf{D} \cdot \mathbf{R} \cdot u. \quad (4.21)$$

Exercise 243 Verify that system (4.20) can be rewritten in manifest vector form:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = c_1 \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} e^{\lambda_1 t} + c_2 \begin{bmatrix} \beta \\ \delta \end{bmatrix} e^{\lambda_2 t}$$

where the coefficients c_1 and c_2 are determined by the boundary conditions (can you give an explicit formula for this dependence?).

Exercise 244 Can you describe in qualitative terms how the transformation \mathbf{R} in the previous exercise alters the x -phase flow to give the u -phase flow?

4.14 A linear system

In exercise 242 you saw that solving a linear two-dimensional autonomous system of the form

$$\dot{x} = \mathbf{A} \cdot x \quad (4.22)$$

(where \mathbf{A} is a 2×2 matrix called the *system matrix*) is easy if we can rewrite \mathbf{A} as follows:

$$\mathbf{A} = \mathbf{S}^{-1} \cdot \mathbf{D} \cdot \mathbf{S}.$$

The critical point is the origin, and its alpha/omega status depends on the diagonal elements of \mathbf{D} , as illustrated by exercises 237—240.

Exercise 245 Verify that the solution, in manifest vector form, of system (4.22) is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c_1 \begin{bmatrix} s_{11} \\ s_{21} \end{bmatrix} e^{\lambda_1 t} + c_2 \begin{bmatrix} s_{12} \\ s_{22} \end{bmatrix} e^{\lambda_2 t} \quad (4.23)$$

where the ss are elements of \mathbf{S} , and the λs are the diagonal elements of \mathbf{D} . (Hint: cf. exercise 243.)

Exercise 246 Consider system (4.22) with initial conditions such that, in terms of system (4.23), you have $c_1 = 1$ and $c_2 = 0$. Verify that

$$\dot{x} = \lambda_1 \begin{bmatrix} s_{11} \\ s_{21} \end{bmatrix} e^{\lambda_1 t} = \lambda_1 x$$

and hence infer that $\mathbf{A} \cdot x = \lambda_1 x$.

4.15 The characteristic equation

The result of the last exercise can be written

$$(\mathbf{A} - \lambda_1 \mathbf{I}) \cdot x = 0$$

(\mathbf{I} is the 2×2 identity matrix). In order that x is not the null vector everywhere (which would be a rather boring solution), the matrix on the left has to be non-invertible, which yields the equation

$$(a_{11} - \lambda_1)(a_{22} - \lambda_2) - a_{21}a_{12} = 0. \quad (4.24)$$

The procedure for λ_2 is entirely analogous and yields the same equation.

Exercise 247 Solve equation (4.24) (Hint: it is a quadratic in λ ; solve it for λ using the *abc*-formula.)

Exercise 248 Let

$$x(t) = \begin{bmatrix} x_1(0) & \alpha t \\ x_2(0) & \beta t \end{bmatrix} e^{\mu t}$$

and verify that this solution corresponds to a differential equation of the form (4.22), by deriving an expression for the matrix \mathbf{A} and show that the procedure of the previous exercise yields a quadratic with a *repeated root* (i.e., the roots are the same and both equal λ).

4.16 Local linear approximation

The idea of *linearization* is to approximate the behaviour of the general non-linear system (4.17) in the neighbourhood of a critical point by a linear system.

Exercise 249 If \bar{x} is a critical point of system (4.17) (that is, $F_1(\bar{x}_1, \bar{x}_2) = F_2(\bar{x}_1, \bar{x}_2) = 0$) and the deviation from the critical point is z (that is, $z_1 = x_1 - \bar{x}_1$, $z_2 = x_2 - \bar{x}_2$) can you show that a first-order approximation to the dynamics of the deviation is

$$\dot{z} = \begin{bmatrix} F_{1\{1\}}(\bar{x}_1, \bar{x}_2) & F_{1\{2\}}(\bar{x}_1, \bar{x}_2) \\ F_{2\{1\}}(\bar{x}_1, \bar{x}_2) & F_{2\{2\}}(\bar{x}_1, \bar{x}_2) \end{bmatrix} \cdot z$$

where the matrix elements are the partial derivatives of the functions F_1 and F_2 ?

Exercise 250 Consider once more the system of exercise 231

$$\begin{cases} \dot{x}_1 &= F_1(x_1, x_2) = \rho x_1 - ax_1^2 - bx_1 x_2 \\ \dot{x}_2 &= F_2(x_1, x_2) = \mu x_2 - cx_2^2 - dx_1 x_2 \end{cases} \quad (4.25)$$

Calculate the derivatives of F_1 with respect to x_1 and x_2 ; do the same with F_2 , and collect them in a 2×2 matrix as in the previous exercise.

Exercise 251 System (4.18) has the critical point $(\bar{x}_1, \bar{x}_2) = (0, 0)$. Substitute these values in the matrix you obtained in the previous exercise, and solve equation (4.24) for this matrix. Show that both λ s are positive. Can you explain why this means that the critical point $(0, 0)$ is an alpha point?

Exercise 252 Repeat the previous exercise for the critical point $(\bar{x}_1, \bar{x}_2) = (0, \mu/c)$. Show that both λ s are negative. Can you explain why this means that this critical point is an omega point?

Exercise 253 Can you show, for system (4.18), that the critical point $(\bar{x}_1, \bar{x}_2) = (\rho/a, 0)$ is an omega point?

Exercise 254 Can you show, for system (4.18), that the critical point

$$\bar{x}_1 = \frac{b\mu - c\rho}{bd - ac} \quad \bar{x}_2 = \frac{d\rho - a\mu}{bd - ac}$$

is a mixed alpha/omega point?

Exercise 255 Draw the phase portrait of the following system

$$\begin{cases} \dot{x}_1 &= \mu x_1 + x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 &= -x_1 + \mu x_2 - x_2(x_1^2 + x_2^2) \end{cases}$$

for (i) the case $\mu \leq 0$ and (ii) the case $\mu > 0$. (Hint: in polar coordinates the equations become $\dot{r} = r(\mu - r^2)$, $\dot{\theta} = -1$.)

Exercise 256 Can you generalize the results of the previous exercise to systems of the following form?

$$\begin{cases} \dot{x}_1 &= \mu x_1 + x_2 - x_1 f(r) \\ \dot{x}_2 &= -x_1 + \mu x_2 - x_2 f(r) \end{cases}$$

where $r = \sqrt{x_1^2 + x_2^2}$, the functions $f(r)$ and $f'(r)$ are continuous for $r \geq 0$, $f(0) = 0$ and $f'(r) > 0$ for $r > 0$.

4.17 Discrete time

In a *discrete-time* dynamical system, only denumerably many values for t_0, t_1, t_2, \dots are considered, together with the associated *state sequence* $x(t_0), x(t_1), x(t_2), \dots$

Exercise 257 Contrast continuous-time versus discrete-time dynamical systems. How fundamental is this distinction? Can the same state-transition function arise in both a continuous-time and a discrete-time system?

4.18 Iterate maps

The state-transition function is usually written as the *k*th iterate map, defined as follows:

$$F^{[k]}(i, \xi) = S(t_{i+k}, t_i, \xi) \tag{4.26}$$

Exercise 258 Complete: $F^{[0]}(i, \xi) = \dots$

Exercise 259 Verify that the second consistency condition now reads:

$$F^{[m]}(i, \xi) = F^{[m-k]}(i+k, F^{[k]}(i, \xi))$$

for $k = 0, 1, 2, \dots, m$.

4.19 Fixpoints

Again, the autonomous case is of special interest, where $F^{[k]}(i, \xi) \equiv F^{[k]}(\xi) \forall i$ (“the same map for every step”). A *critical point (of the kth iterate)* \bar{x} satisfies $F^{[k]}(\bar{x}) = \bar{x}$. Critical points of first iterate maps are called *fixpoints*.

Exercise 260 Draw a graph of $F^{[1]}(x)$ with $x \in \mathbb{R}$ (any one you like). Also draw the diagonal line $y = x$. Indicate the fixpoints (in any) in your graph.

Exercise 261 Suppose that, given a fixpoint $\bar{x} \in \mathbb{R}$ and some $\delta > 0$, the first iterate map satisfies the following:

$$\left| \frac{F^{[1]}(\bar{x} + \epsilon) - \bar{x}}{\epsilon} \right| < 1$$

for all $|\epsilon| < \delta$. Deduce that \bar{x} is an omega point for all x_0 within a distance δ of \bar{x} . Can you find a similar condition for alpha points?

Exercise 262 Show that a fixpoint $\bar{x} \in \mathbb{R}$, satisfying $F^{[1]}(\bar{x}) = \bar{x}$, is also a critical point of $F^{[2]}$.

Exercise 263 Can you show that a fixpoint $\bar{x} \in \mathbb{R}$ is also a critical point of $F^{[k]}$ for $k \geq 2$?

Exercise 264 Suppose that $\bar{x} \in \mathbb{R}$ is a critical point of the second iterate. Then $F^{[1]}(\bar{x})$ is also a critical point of the second iterate; show this. (Hint: you have $F^{[2]}(\bar{x}) = \bar{x}$; $F^{[2]}(\bar{x})$ may, but need not, be equal to \bar{x} .)

Exercise 265 Show that a critical point $\bar{x} \in \mathbb{R}$ of the second iterate which is *not* also a fixpoint corresponds to a periodic state sequence

$$\dots, \bar{x}, F^{[1]}(\bar{x}), \bar{x}, F^{[1]}(\bar{x}), \bar{x}, F^{[1]}(\bar{x}), \dots$$

Exercise 266 Can you show that a critical point $\bar{x} \in \mathbb{R}$ of the p th iterate, where p is a prime number, which is *not* also a fixpoint corresponds to a p -periodic state sequence? (Hint: why the emphasis on primes here?)

Exercise 267 Consider a q -dimensional discrete-time dynamical system ($x \in \mathbb{R}^q$) defined by the following first-iterate map:

$$F^{[1]}(x) = \begin{bmatrix} \gamma_1 x_1 + \gamma_2 x_2 + \dots + \gamma_q x_q \\ x_1 \\ x_2 \\ \vdots \\ x_{q-1} \end{bmatrix} \quad (4.27)$$

and let μ be a root of the polynomial equation

$$z^q = \gamma_1 z^{q-1} + \gamma_2 z^{q-2} + \dots + \gamma_{q-1} z + \gamma_q .$$

Then verify that the following equation defines a possible state sequence for this system (indexed by n):

$$k \begin{bmatrix} \mu^n \\ \mu^{n-1} \\ \mu^{n-2} \\ \vdots \\ \mu^{n-q+1} \end{bmatrix}$$

where k is an arbitrary constant.

Exercise 268 Suppose that the system of the previous exercise is modified as follows, to become non-autonomous:

$$F^{n,[1]}(x) = \begin{bmatrix} \gamma_1 x_1 + \gamma_2 x_2 + \dots + \gamma_q x_q + \kappa \lambda^n \\ x_1 \\ x_2 \\ \vdots \\ x_{q-1} \end{bmatrix} \quad (4.28)$$

with an additional *forcing term* $\kappa\lambda^n$. Verify that the following equation defines a possible state sequence for this system (indexed by n):

$$k \begin{bmatrix} \kappa\lambda^{n+q}/(\lambda^q - \gamma_1\lambda^{q-1} - \gamma_2\lambda^{q-2} - \dots - \gamma_p) \\ \kappa\lambda^{n-1+q}/(\lambda^q - \gamma_1\lambda^{q-1} - \gamma_2\lambda^{q-2} - \dots - \gamma_p) \\ \kappa\lambda^{n-2+q}/(\lambda^q - \gamma_1\lambda^{q-1} - \gamma_2\lambda^{q-2} - \dots - \gamma_p) \\ \vdots \\ \kappa\lambda^{n+1}/(\lambda^q - \gamma_1\lambda^{q-1} - \gamma_2\lambda^{q-2} - \dots - \gamma_p) \end{bmatrix}.$$

Exercise 269 If $x_H(n)$ denotes the state sequence given in exercise 267 and $x_{PS}(n)$ denotes the state sequence given in exercise 268, show that $x_H(n) + cx_{PS}(n)$, where c is an arbitrary constant, is also a possible state sequence for system (4.28).

Exercise 270 In continuous-time dynamics, you encounter terms like $e^{\lambda t}$ where $\lambda > 0$ means ‘expanding’ as t increases (while $\lambda < 0$ means ‘shrinking’); whereas in discrete dynamics, you encounter terms like λ^t , where $\lambda > 1$ means ‘expanding’ as t increases (while $\lambda < 1$ means ‘shrinking’). How come the cross-over occurs at zero in one case, and at unity in the other? (Hint: the rule $e^{ab} = (e^a)^b$ might help.)

Exercise 271 Consider a 4-dimensional discrete dynamical system ($x \in \mathbb{R}^4$) defined by the following first-iterate map:

$$F^{[1]}(x) = \begin{bmatrix} \alpha x_1 + \gamma x_2 + \gamma x_3 + \gamma x_4 \\ \sigma x_2 \\ \sigma x_3 \\ \sigma x_4 \end{bmatrix}. \quad (4.29)$$

Verify the following matrix equation:

$$\begin{bmatrix} \alpha & \gamma & \gamma & \gamma \\ 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & \sigma \end{bmatrix}^n = \begin{bmatrix} \alpha^n & \gamma \frac{\alpha^n - \sigma^n}{\alpha - \sigma} & \gamma \frac{\alpha^n - \sigma^n}{\alpha - \sigma} & \gamma \frac{\alpha^n - \sigma^n}{\alpha - \sigma} \\ 0 & \sigma^n & 0 & 0 \\ 0 & 0 & \sigma^n & 0 \\ 0 & 0 & 0 & \sigma^n \end{bmatrix}$$

and hence give an expression for the state-transition function.

4.20 Stochastic dynamics

So far, we have thought of the state as a quantity which can, in principle at least, be observed at all desired points in time. Thus, if the measurement of the state is ξ at time $t = t_0$, then the state transition function predicts that the state measurement (or *observation*) at time $t = t_1$ will be $S(t, t_0, \xi)$. We say that the dynamics is *deterministic*.

By contrast, the state in a *stochastic* dynamical system is not an observable value, but a probability distribution. Thus, the state at time t does not give the value of some given quantity of interest at t , but instead specifies probabilistic quantities, such as the probability that the quantity of interest is smaller than some predefined value.

Exercise 272 Can you justify the claim that deterministic systems are just special cases of stochastic systems? (Hint: think of degenerate distributions.)

Exercise 273 Consider the stochastic system specified by the following state transition:

$$x_i(t_1) = e^{-\lambda(t_1 - t_0)} \sum_{j=0}^i x_j(t_0) \frac{[\lambda(t_1 - t_0)]^{i-j}}{(i-j)!} \quad (4.30)$$

for $i = 0, 1, 2, 3, \dots$. You may interpret $x_i(t)$ as the probability that a quantity of interest assumes the value i at time t . Accordingly $x(t_0)$ must satisfy the conditions $x_i(t_0) \geq 0 \forall i, \sum_{i=0}^{\infty} x_i = 1$. Show that, under state transition (4.30), the state at time t_1 also satisfies these conditions (hence conclude that the state remains a valid probability distribution as it is carried forwards in time by the transition function).

Exercise 274 Verify that the state transition (4.30) satisfies the first consistency condition. (Hint: see exercise 193 to remind yourself of the definition.)

Exercise 275 Can you verify that the state transition (4.30) satisfies the second consistency condition? (Hint: see exercise 194 to remind yourself of the definition; you may find the binomial formula, $(a + b)^n = \sum_{\ell=0}^n \binom{n}{\ell} a^{n-\ell} b^{\ell}$, helpful.)

Exercise 276 Consider the stochastic system specified by the following state transition:

$$x_i(t_1) = \sum_{j=i}^{\infty} x_j(t_0) \binom{j}{i} \exp\{-\lambda i(t_1 - t_0)\} (1 - \exp\{-\lambda(t_1 - t_0)\})^{j-i} \quad (4.31)$$

for $i = 0, 1, 2, 3, \dots$. Verify that the first consistency condition is verified.

Exercise 277 Show that state transition (4.31) has the following omega point:

$$x_{\omega} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

4.21 The persistence function

While the state of a stochastic dynamic system is a probability distribution, the quantity of interest is usually taken to occupy a unique, definite state value at any given moment in time; we say that the measurement (observation) is done on an *instantiation* of the stochastic system.

Such an instantiation can be described by the *persistence function*³ which describes the ‘persistence time’ spent by the instantiation occupying the i th state value; more precisely, $P_i(t, t_0)$ denotes the probability that an instantiation which assumed the i th value at time t_0 will retain that value at least until time t (where $t \geq t_0$).

Exercise 278 Show that the persistence function must satisfy the following properties:

$$P_i(t, t_0) \in [0, 1] \quad P_i(t, t) = 1 \quad P_{i, \{1\}}(t, t_0) \leq 0.$$

(Hint: the persistence has been defined to be a probability.)

4.22 The hazard rate

The *hazard rate* is defined as follows:

$$h(t, t_0) = -\frac{P_{i, \{1\}}(t, t_0)}{P_i(t, t_0)} \quad (4.32)$$

Exercise 279 Show that $h(t, t_0) \geq 0$ for all $t \geq t_0$.

Exercise 280 Consider an ‘ensemble’ of N instantiations all assuming the i th state value at time $t = t_0$. Can you argue why you can write

$$\dot{N}(t) = -h(t, t_0)N(t)$$

provided that N is very large? (Hint: the law of large numbers.)

³Also known as *reliability function* in quality control and the *survivor* in biomedical applications.

4.23 The no-memory property

In many applications, the hazard function is autonomous in the particular sense that it depends only on the time that has expired since t_0 . Thus:

$$h(t, t_0) \equiv h(t - t_0) \quad (4.33)$$

where the h on the right is a function with a single argument.

Exercise 281 ('No memory') Consider two independent instantiations entering the i state value at distinct times, the first at t_0 and the second at a later time \tilde{t}_0 . Suppose that the hazard rates for leaving the i th state value both have property (4.33) and are identical for all $t \geq \tilde{t}_0$; and, finally, suppose that this identity of hazard rates is true regardless of the choice of the second instance \tilde{t}_0 . Infer that the hazard rate is constant, and hence show that the persistence functions for the two instantiations are $\exp\{-\lambda_i(t - t_0)\}$ and $\exp\{-\lambda_i(t - \tilde{t}_0)\}$. (Hint: rewrite equation (4.32) as $\dot{P} = -\lambda_i P$ and solve this equation.)

Exercise 282 Verify that the average persistence time for an exponential persistence function $P(t, t_0) = \exp\{-\lambda(t - t_0)\}$ is $1/\lambda$. (Hint: you have to evaluate $\int_{t_0}^t (-\lambda) \exp\{-\lambda(\tau - t_0)\} d\tau$.)

Exercise 283 Suppose that an instantiation will cease to retain the i th state value at the moment when either one of two 'terminating' events happens. Call these events E_I and E_{II} . Furthermore, suppose that the time until E_I happens is exponentially distributed with parameter μ_I , and that the time until E_{II} is also exponentially distributed with parameter μ_{II} . Can you show that the time until either E_I or E_{II} , *whichever happens first*, is exponentially distributed with parameter $\mu_I + \mu_{II}$? Hence deduce that $\lambda_i = \mu_I + \mu_{II}$. (Hint: you are asked to calculate the distribution of the *minimum* of two exponential variates.)

Exercise 284 Under the assumptions of the previous exercise, show that the formula $\mu_I/(\mu_I + \mu_{II})$ gives the probability of E_I terminating the residence of the instantiation in the i th state value.

Exercise 285 Can you generalize the results of the previous two exercises to three or more terminating events?

Exercise 286 Consider an ensemble of N_i instantiations, with an exponentially distributed persistence time, that have the i th state value at time t ; the instantiations need not be synchronized, that is, they may have entered the i th state at different times in the past. Nonetheless, you can still write

$$\dot{N}_i(t) = -\lambda_i N_i(t)$$

(for large N_i) where $1/\lambda_i$ is the average persistence time. Can you explain why? (Hint: combine exercises 280 and 281.)

4.24 Accounting for an extra influx

In the last exercise, we ignored the possibility of instantiations assuming the i th state value after time t ; this "influx" of instantiations is represented by a second term:

$$\dot{N}_i(t) = -\lambda_i N_i(t) + \Phi_i(t)$$

where $\Phi_i(t)$ is the rate at which instantiations assume the i th state value at time t . In general $\Phi_i(t)$ depends on the numbers of instantiations (and their persistence functions) in all other state values $j \neq i$.

Exercise 287 Suppose that persistence time is exponentially distributed for all state variables. Show that the general equation then becomes

$$\dot{N}_i(t) = -\lambda_i N_i(t) + \sum_{j=0}^{\infty} p_{ji} \lambda_j N_j \quad (4.34)$$

where p_{ji} is the probability that an instantiation assumes the i th state value when it ceases to retain the j th state value ($p_{ii} = 0$ for all i).

Exercise 288 Under the assumptions of the previous exercise, define $N_T = \sum_{i=0}^{\infty} N_i(t)$ and show that $\dot{N}_T = 0$. (Hint: first verify that $\sum_{j=0}^{\infty} p_{ij} = 1$ for all i , and use this fact.)

4.25 Taking the ensemble limit

In an ideal ensemble, the numbers of instantiations are infinitely large. To deal with this, we normalize by defining $x_i(t) = N_i(t)/N_T$.

Exercise 289 Verify the following properties:

$$x_i(t) \in [0, 1] \quad \text{and} \quad \sum_{i=0}^{\infty} x_i(t)$$

which establish that x is a discrete probability distribution.

Exercise 290 Use equation (4.34) to obtain the following equation:

$$\dot{x}_i = -\lambda_i x_i + \sum_{j=0}^{\infty} p_{ji} \lambda_j x_j . \quad (4.35)$$

Exercise 291 Equation (4.35) provides a general recipe to formulate stochastic dynamics. Review the assumptions leading up to this equation, and discuss their general applicability.

4.26 The jump chain

We can associate a discrete-time stochastic dynamic system by forgetting the duration of residence and focussing only on the ‘jumps’, where a jump occurs whenever an instantiation changes its state value: $x_i(t_k)$ denotes the probability of finding the instantiation having the i th state value before the k th jump. The state sequence is then called the *jump chain*.

Exercise 292 Verify that the jump chain associated with equation (4.35) has the following dynamics:

$$x_i(t_{k+1}) = \sum_{j=0}^{\infty} p_{ji} x_j(t_k) . \quad (4.36)$$

4.27 The Markov chain

The jump chain obtains under the condition that $p_{ii} = 0$ for all i (retaining the state value does not constitute a “jump”). However, it is often expedient to relax this restriction, and allow “stationary jumps”. The state sequence for this (slightly) more general variant is called a *Markov chain*.

Exercise 293 Suppose that x is the state (i.e. the distribution)] of a Markov chain with discrete jump dynamics (4.36). Let \bar{x} be a state which satisfies

$$\bar{x}_i p_{ij} = \bar{x}_j p_{ji} \quad \text{for all pairs } (i, j). \quad (4.37)$$

Suppose that $x(t_k) = \bar{x}$ implies $x(t_{k+1}) = \bar{x}$. (Hint: recall the hint of exercise 288; show that $x_i(t_{k+1}) = \bar{x}_i$ for all i .)

4.28 Detailed balance

The last exercise shows that a state which satisfies the *detailed balance equation* (4.37) is a stationary distribution. A Markov chain is *connected* if a series of jumps, each with positive probability, exists between any pair (i, j) ; so either $p_{ij} > 0$, or there is at least one i' such that $p_{ii'}$ and $p_{i'j}$ are both positive, or there is at least one pair (i', i'') such that $p_{ii'}$ and $p_{i'i''}$ and $p_{i''j}$ are all positive, and so forth.

Exercise 294 Suppose we have an ensemble of N independent instantiations whose probability distribution satisfies the detailed balance equation (4.37) where the Markov chain is connected. Consider the following quantity, which defines how frequently one of these instantiations “visits” the i th state value:

$$q_i = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^K \mathbf{1}_i(t_k) \quad (4.38)$$

where $\mathbf{1}_i(t_k) = 1$ if the selected instantiation is found to have the i th state value prior to its k th jump (and $\mathbf{1}_i(t_k) = 0$ otherwise). Can you show that (i) the quantity q_i is the same for every instantiation; and (ii) you have $q_i = \bar{x}_i$ for all i ?

4.29 Ergodicity

The property $q_i = \bar{x}_i$ (“average over time for a single particle equals average over particles for a single time”) is an example of *ergodicity*.

Exercise 295 In the foregoing exercises, it was assumed that only denumerable many state values need be distinguished (which can be labelled $0, 1, 2, 3, \dots$) so that a discrete probability distribution suffices to characterise the state. Can you sketch how you would tackle the case where the state x represents a *continuous* distribution?

4.30 Controlled dynamics

Consider a discrete-time dynamical system with a finite state sequence $x(t_1), x(t_2), \dots, x(t_N)$ and the following peculiarity: at each time t_i , a *choice* is made between two alternative 1st iterate (state transition) maps F and G : if F is chosen at time t_i , $x(t_{i+1}) = F(x(t_i))$, whereas if G is chosen at time t_i , $x(t_{i+1}) = G(x(t_i))$. The initial value $x(t_1)$ is given.

Exercise 296 Suppose that the choice is made at random. Can you describe the associated stochastic dynamical system? (Hint: stochastic dynamic systems carry a distribution function forward in time; this will have to be the distribution of the state of the “alternative dynamics” system.)

4.31 Control

To render the alternative dynamics deterministic, a sequence of choices, often called the *control*, will have to be specified as a forcing function. Thus we have $\{u_i\}_{i=1}^{N-1}$ where $u_i = 0$ indicates that map F is selected at time t_i , and $u_i = 1$ indicates that map G is selected at time t_i .

Exercise 297 Why does the u -sequence only need to run to time t_{N-1} ?

Exercise 298 Suppose that you have a map J and you let $u_i = J(x(t_i))$. Show on this specification that the dynamics of x becomes fully deterministic.

4.32 Control as a state

One approach is to regard the forcing sequence as the state sequence of some other dynamic system; thus u effectively becomes a state variable in its own right.

Exercise 299 Show that the “Cartesian product” of the two dynamical systems at hand then forms a usual deterministic discrete-time system.

4.33 Optimized control

A more interesting way to fix the forcing sequence is by supplying a *optimization criterion*: $\{u_i\}_{i=1}^N$ is to be such that a given *objective function* $H(\{x(t_i), u_i\}_{i=1}^{N-1})$ is minimized. For many interesting problems, the objective function takes the form of a sum, like so:

$$H(\{x(t_i), u_i\}_{i=1}^{N-1}) = \sum_{i=1}^{N-1} h_i(x(t_i), u_i) \quad (4.39)$$

where we have split H into N ‘local’ contributions h_i .

Exercise 300 Suppose that the forcing sequence has been specified up to u_{N-2} , with only u_{N-1} remaining to be determined. Show that this nearly complete forcing sequence, together with the initial condition $x(t_1)$, determines the state sequence $x(t_1), \dots, x(t_{N-1})$, with $x(t_N)$ still open.

Exercise 301 For the problem of the previous exercise, verify that

$$H = \sum_{i=1}^{N-2} h_i(x(t_i), u_i) + h_{N-1}(x(t_{N-1}), 0) \quad \text{for } u_{N-1} = 0$$

$$H = \sum_{i=1}^{N-2} h_i(x(t_i), u_i) + h_{N-1}(x(t_{N-1}), 1) \quad \text{for } u_{N-1} = 1$$

and hence show that the difference made by the choice at time t_{N-1} is

$$\Delta h = h_{N-1}(x(t_{N-1}), 0) - h_{N-1}(x(t_{N-1}), 1).$$

Exercise 302 For the problem of the previous two exercises, show that the optimal choice is $u_{N-1} = 0$ if $\Delta h \leq 0$ and $u_{N-1} = 1$ if $\Delta h > 0$.

4.34 Optimal choice

In the last few exercises, you constructed an *optimal choice function* U_{N-1} :

$$U_{N-1}(\xi) = \begin{cases} 0 & \text{if } h_{N-1}(\xi, 0) \leq h_{N-1}(\xi, 1) \\ 1 & \text{if } h_{N-1}(\xi, 0) > h_{N-1}(\xi, 1) \end{cases}.$$

The key idea is now to apply the same procedure to determine U_{N-2} , and work your way back all the way to U_1 . Then $u_1 = U_1(x(t_1))$, $u_2 = U_2(x(t_2))$, \dots and the dynamics has become fully deterministic.

Exercise 303 Let η_{N-2} denote the last two terms of H : $h_{N-2}(x(t_{N-2}), u_{N-2}) + h_{N-1}(x(t_{N-1}), u_{N-1})$. Show that η_{N-2} *only* depends on $x(t_{N-2})$ and u_{N-2} . (Hint: u_{N-2} determines $x(t_{N-1})$, through map choice and state transition; $x(t_{N-1})$ in turn determines u_{N-1} through the map U_{N-1} .)

Exercise 304 Let $\eta_{N-2}(x(t_{N-2}), u_{N-2})$ be the map defined in the previous exercise. Show that the following defines the optimal choice function at this stage:

$$U_{N-2}(\xi) = \begin{cases} 0 & \text{if } \eta_{N-2}(\xi, 0) \leq \eta_{N-2}(\xi, 1) \\ 1 & \text{otherwise} \end{cases}.$$

Exercise 305 Let η_{N-3} denote the last three terms of H . Show that η_{N-3} depends *only* on $x(t_{N-3})$ and u_{N-3} , and establish the following result

$$U_{N-3}(\xi) = \begin{cases} 0 & \text{if } \eta_{N-3}(\xi, 0) \leq \eta_{N-3}(\xi, 1) \\ 1 & \text{otherwise} \end{cases}.$$

Exercise 306 Can you generalize the results of the last few exercise, that is, give U_{N-T} from η_{N-T} where T is the number of ‘tail’ terms considered?

4.35 Dynamic programming

In the foregoing exercises you established a basic version of the so-called *dynamic programming* method.

Exercise 307 Suppose that the state x only takes values in a finite set \mathcal{X} where \mathcal{X} has n_X elements (so the state space is discrete as well as time). Verify that the sequence of optimal choice functions $\{U_i\}_{i=1}^{N-1}$ assumes the form of a *table* with n_X rows and $N - 1$ columns.

4.36 The dynamic programming table

For a discrete state space of cardinality n_X , the sequence $\{\eta_i(\cdot), U_i(\cdot)\}_{i=1}^{N-1}$ can also be represented as a table with n_X rows and $N - 1$ columns (η_{N-1} is just h_{N-1}). You can work the problem systematically by building up this table “from the right to the left” in conjunction with the optimal control table of exercise 307.

Exercise 308 The solution of the problem is the state sequence, with the initial $x(t_1)$ given as boundary condition; compare this output of $N - 1$ data points to the amount of data that need to be calculated to compile the tables. (Hint: loosely speaking, the (i, j) th entry of the table tells you the best forcing option u if the system should ever find itself occupying the i th state value at the j th point in time; but will this circumstance typically present itself?)

Exercise 309 So far, u was either 0 or 1 and specified a binary choice between two alternative state transition maps. How do you adapt the procedure if you allow more choices (i.e. u takes values in a bigger set)? What if you even allow u to be continuous?

Modelling Methodology

5.1 What is a model?

Loosely speaking, a *model*¹ is a piece of mathematics, interpreted in terms of observable real-world phenomena. The mathematical component of the model is called the *model proper*.

Exercise 310 Could any “model proper” (uninterpreted mathematics) be made part of a model by means of a suitable interpretation?

Exercise 311 Can you define *observable phenomenon*? (Hint: how do the terms of art of (empirical) scientific discourse acquire their meaning? How are these meanings affected by advances in experimental techniques? How theory-laden are data and experimental observations?)

Exercise 312 Is it necessary that *everything* (say, every variable and parameter) in the model proper corresponds to an observable phenomenon?

Exercise 313 Can you motivate the following statement? The interpretation of a model can be split into two components, (i) a list of identifications between parts of the model proper and parts of the world; (ii) the observational interface, the experimental approach to the world.

Exercise 314 Can you explain why the identifications (in the sense of the previous exercise) are said to constitute a *theoretical hypothesis*?

5.2 Assumptions

In practice, you start with (imperfect) knowledge about the real-world, and you attempt to formulate a suitable model proper. This knowledge can take the form of a list of statements, called the *assumptions underlying the model*.

Exercise 315 The assumptions are said to be *contingent*, which means that they may be wrong, depending on what is actually the case². Is *wrong* an all-or-nothing property, or are there shades of gray? Is the model proper contingent?

Exercise 316 Comment on the claim that every model has exactly *one* assumption, namely that: Model proper \mathcal{M} represents real world phenomenon \mathcal{P} according to a list of identifications \mathcal{I} (with specifications of \mathcal{M} , \mathcal{P} and \mathcal{I} supplied).

Exercise 317 What are the advantages of breaking down the “monolithic” assumption of the previous exercise into a long list of detailed & specific assumptions? What are the disadvantages?

5.3 Hidden assumptions

An assumption that is not spelled out by the modeller is said to be *implicit* (antonym: *explicit*). When it becomes important, an implicit assumption is usually called a *hidden assumption*.

Exercise 318 Why would modellers leave assumptions implicit?

¹‘Model’ will mean ‘mathematical model’ in this worksheet, although some comments also apply to other types of model such as scale models.

²Die Welt ist alles, was der Fall ist.

5.4 Strong and weak assumptions

It could happen that the conclusions of the model analysis may change dramatically if an assumption is altered only slightly; such an assumption is said to be *strong* (antonym: *weak*).

Exercise 319 Motivate a preference for weak assumptions.

Exercise 320 An assumption may strongly confine or reduce the range of real-world phenomena described by the model; such an assumption is said to be *stringent* (antonym: *relaxed*).

Exercise 321 A common strategy is to start with stringent assumptions, and as you go round the model cycle, to relax these assumptions (i.e. replace them by more relaxed versions). Why is this a sound strategy? (Hint: stringent assumptions tend to be easier to state and to analyse.)

5.5 Consistency

Assumptions (within a given model!) should not contradict one another; this is the requirement of *consistency*.

Exercise 322 Should the consistency requirement be extended to cover the demand that the assumptions do not contradict the available data?

Exercise 323 Explain why it might well happen that the assumptions of a model are consistent with each other as well as the available data, but the conclusions of the model analysis fail to do so. Do you consider this to be a failure or a success of the modelling exercise as a whole?

Exercise 324 Discuss (in a group) the thorny issue of *coherence* between the assumptions. (Hint: should the assumptions be equally stringent, equally strong? Why (not)?)

5.6 Model complexity

As a modeller, you have to decide how complex you want your model to be.

Exercise 325 Can you define or characterize the concept of complexity for a model?

Exercise 326 Consider the following three possible aims of a model:

(i) to predict the (future) behaviour of a real-world phenomenon;

(ii) to control the behaviour of a real-world phenomenon;

(iii) to understand the behaviour of a real-world phenomenon;

and arrange these three aims in the order of required model complexity.

Exercise 327 What positive purpose may be served by deliberately *leaving out* certain bits of knowledge (i.e. deliberately not representing certain known mechanisms in the model proper)? Discuss advantages and disadvantages.

Exercise 328 Discuss ways of tackling real-world phenomena that would appear, in one sense or another, too complex to capture in a model. (Hint: think of scales in time and space, emergent properties, statistical mechanics. As regards examples of domains where this problem arises in the life sciences; think of neurones ÷ brains, individuals ÷ ecosystems, molecules ÷ cells.)

Exercise 329 Case study: the mechanics of the left ventricle heart is modelled by a finely meshed finite-element representation of the myocardium, incorporating the various layers of muscle fibers in alternating directions, the electrical impulse conduction pathway, and the coronary perfusion system. The aorta, against which the heart is working during systole, is represented as a ‘wind-kessel’ which is effectively a hydraulic resistor and a capacitor. Discuss the disparity in detail between the myocardial and the aortic representation.

Exercise 330 Discuss (in a group) the following common pitfalls:

- (i) wishful thinking—tailoring model assumptions to attractive equations;
- (ii) misleading names for model quantities—inviting overinterpretation;
- (iii) imputing goals, objectives³ and functions⁴ to living things—adopting an engineer’s point of view.

(Hint: reflect that, if used judiciously, a pitfall may turn into a useful trick.)

5.7 Simulation

To derive results & conclusions from the model, its behaviour *vis a vis* the real-world phenomenon must be probed. To this end, the model proper is subjected to mathematical analysis. An (almost obligatory) part of this is *simulation*⁵, which is a numerical (and usually graphical) demonstration of the behaviour of the model.

Exercise 331 Discuss the distinction between analytical and non-analytical, tractability, and review standard numerical techniques. (Hint: why is it permitted⁶ to define $y = J_0(x)$ as the solution to $x^2\ddot{y} + x\dot{y} = -x^2y$ whereas you cannot very well declare your problem to be solved by $y = \mathcal{S}_{\text{my problem}}(x)$ where $\mathcal{S}_{\text{my problem}}$ is defined to be just that?)

Exercise 332 Suppose that ξ , v , and ζ are related by

$$\zeta = \frac{\xi}{v} \quad (5.1)$$

but the modeller is unaware of this simple fact and generates a series graphs in which ζ is plotted against v , perhaps at a variety of values of ξ . Will the modeller be able to infer from his graphics that the three quantities are interrelated according to equation (5.1)?

5.8 Dimensional analysis

The last exercise illustrates a problem that can become very serious when you do not know what relationship you are looking for, and when there are many parameters and parameter values to study. The moral is that it is a good idea to try to penetrate the model by analytical means as far as your mathematical prowess (or that of friendly mathematicians) will take you, because it is not always easy to uncover important interdependencies. *Dimensional analysis* is a useful technique which helps to alleviate this problem.

5.9 Empirical dimension

Two quantities (P and Q) are said to have the *same (empirical) dimension*, written briefly as $P \stackrel{d}{\sim} Q$, iff:

1. Quantities P and Q can be meaningfully equated or compared ($=$, $<$, $>$).
2. Quantities P and Q have a meaningful sum or difference.

The **empirical dimension** (briefly: **dimension**⁷) as an equivalence class of quantities under $\stackrel{d}{\sim}$. Finally, we say that the dimension of given quantity Q is the particular equivalence class (under $\stackrel{d}{\sim}$) to which Q belongs.

³*Teleology* is the doctrine that everything in nature was or is designed with a view to achieving a (God-given?) set of objectives (from the Greek *telos* meaning ‘purpose’ or ‘end’).

⁴The *biological function* of a certain structure or process in an organism is constituted by whatever effects the presence of that structure or process happens to have within the context of the organism. If these effects tend to aid the lifetime reproductive success of the individuals in which it happens to find itself, it will persist on an evolutionary timescale.

⁵The term ‘simulation’ is usually taken to imply the use of numerical techniques, in particular for the solution of boundary value problems.

⁶Quod licet iovi non licet bovi.

⁷Be careful that you do not confuse this use of the word ‘dimension’ with other meanings in mathematics, such as that of linear algebra; in the proof of Buckingham’s theorem we will be using the term in both senses.

Exercise 333 Verify that ‘same dimension’ is an equivalence relation.

5.10 Equality of dimensions

To establish whether two quantities have the same dimension, we need to ascertain whether or not $P \stackrel{d}{\sim} Q$.

Exercise 334 Why is it not really a matter of mathematics whether $P \stackrel{d}{\sim} Q$ is valid? (Hint: the crux is the word ‘meaningful’, which hinges on interpretation, not model proper.)

Exercise 335 Motivate the convention

$$\frac{P}{P} \stackrel{d}{\sim} \frac{Q}{Q}$$

for any two quantities P and Q , regardless of whether $P \stackrel{d}{\sim} Q$.

5.11 Dimensionless quantities

The convention of the last exercise induces an equivalence class denoted by 1: we write $(P/P) \stackrel{d}{\sim} 1$ and call the equivalence class thus formed *pure number*.

Exercise 336 Explain why quantities in this special class are referred to as *dimensionless*.

Exercise 337 Why is the (commonly heard) statement that “quantities, naturally interpreted as numbers, are always dimensionless” in fact fallacious? (Hint: refer to absolute scales of measurement, explained below.)

Exercise 338 Suppose your model has M scalar quantities $\{Q_1, \dots, Q_M\}$; these quantities are the variables and parameters in the model proper. Using the $Q_i \stackrel{d}{\sim} Q_j$ test, you can apportion these quantities among a number of empirical dimensions. Show that the number of distinct empirical dimensions will be anywhere between 1 (inclusive) and M (inclusive).

5.12 A basis of empirical dimensions

Consider m distinct dimensions, each represented by some quantity D_i , which has the following properties:

$$\text{for every } Q_j, j = 1, \dots, M, \quad Q_j \stackrel{d}{\sim} \prod_{i=1}^m D_i^{r_i}$$

for some vector $[r_1, \dots, r_m]$, while for *no choice* of $\{r_1, \dots, r_{i-1}, r_{i+1}, \dots, r_m\}$ it is true that

$$\text{for any } D_i, i = 1, \dots, m, \quad D_i \stackrel{d}{\sim} \prod_{k \neq i} D_k^{r_k}$$

(the r_i are real numbers). Such a set $\{D_1, \dots, D_m\}$ constitutes a **basis** of empirical dimensions for your model.

Exercise 339 Verify that the first condition ensures sufficiency of the basis for the model. whereas the second is an analog of linear independence.

5.13 Dimension formulæ

We may refer to these basis dimensions (which you will recall are equivalence classes) with an appropriate sans serif name that recalls the interpretation of the quantities in the class, for example time. To indicate that the dimension of the quantity t is time one usually writes $\dim\{t\} = \text{time}$ instead of $t \in \text{time}$.

Exercise 340 Can you make sense of a formula such as $\text{velocity} = \text{length}/\text{time}$? (Hint: with $v \in \text{velocity}$, we have $v \propto s/t$ where $s \in \text{length}$ and $t \in \text{time}$.)

Exercise 341 Show that the following formulæ hold:

$$\dim\{x + y\} = \dim\{x\} = \dim\{y\} \quad (5.2)$$

$$\dim\{xy\} = \dim\{x\} + \dim\{y\} \quad (5.3)$$

$$\dim\{x^n\} = n \dim\{x\} \quad \text{where} \quad \dim\{1\} = 0. \quad (5.4)$$

Exercise 342 Show that $\ln\{x\}$ is well-defined only if $\dim\{x\} = 1$.

Exercise 343 For chemists: why is the Henderson-Hasselbalch relation dimensionally correct, despite appearances and the previous exercise? Is the claim that “the pK has no units” defensible?

Exercise 344 For physicists: verify that $\{\text{length, time, mass}\}$ is a basis of classical mechanics. Can you show that $\{\text{velocity, force, energy}\}$ is a basis, too? What about $\{\text{velocity, momentum, action}\}$?

5.14 Alternate bases

The last exercise shows that a dimension basis need not be unique. A systematic procedure is available to discover a dimension basis given a problem with M quantities apportioned over a set of dimensions, but inspection and intuition suffice in many problems, as the simple example in the following exercises makes clear.

Exercise 345 Let the following model properly represent the dynamics of the amount of mRNA molecules transcribed from a given gene in a cell:

$$\frac{d}{dt}N = M - \lambda N \quad (5.5)$$

with the following interpretations: N is the number of mRNA molecules; M is the rate at which mRNA molecules are produced by transcription; and λ is the rate at which the mRNA molecules are being degraded. At time $t = 0$, there are N_0 molecules present in the cell. Name the *three* parameters of this model.

Exercise 346 Show equation (5.5) has the following solution:

$$N(t) = \frac{M}{\lambda} + \left(N_0 - \frac{M}{\lambda}\right) \exp\{-\lambda t\}. \quad (5.6)$$

5.15 Natural units

Equation (5.6) is simple enough for its characteristic properties to be gleaned by inspection. However, the solution of a differential equation is not always that simple, or may not be analytically available and must be numerically evaluated. In both cases, you would study the behaviour of the model by plotting curves for specific numerical settings of the parameter. If you plot equation (5.6) for various settings of M , λ , and N_0 , you get a tangled mess of different curves. In models with more than three parameters, numerically exploring the behaviour at all parameter settings readily becomes a hopeless task. However, the parameter count can be reduced by as many dimensions as there are. The idea is to use the parameters as ‘natural units.’

Exercise 347 Choose time and number of molecules as the basis dimensions. Establish the following:

$$\begin{aligned}\dim\{N\} &= \text{number of molecules} \\ \dim\{t\} &= \text{time} \\ \dim\{M\} &= \text{number of molecules} \cdot \text{time}^{-1} \\ \dim\{\lambda\} &= \text{time}^{-1} \\ \dim\{N_0\} &= \text{number of molecules}\end{aligned}$$

(Hint: use the dimension test.)

5.16 Units

A *unit* for an empirical dimension d is an element of d , say u_d , which serves to report the magnitude of all other $x \in d$ in terms of the dimensionless ratio x/u_d .

Exercise 348 Verify that x/u_d is dimensionless. (Hint: the thing being measured and the unit must share empirical dimension.)

5.17 Choosing model-derived units

It is perfectly legitimate to choose quantities in the model as units.

Exercise 349 Show that $\lambda^{-1} \in \text{time}$ and that λ^{-1} can therefore serve as a unit of time.

Exercise 350 Show that $M\lambda^{-1}$ can serve as a unit of number of molecules.

Exercise 351 Show that $t^* = t\lambda$ and $N^* = NM^{-1}\lambda$ are dimensionless.

Exercise 352 Derive the following scaled version of equation (5.5):

$$\frac{d}{dt^*}N^* = 1 - N^* \quad (5.7)$$

Exercise 353 Show that one free dimensionless parameter remains: $N_0^* = N_0M^{-1}\lambda$.

Exercise 354 Sketch the solution of equation (5.7) for various values of N_0^* .

Exercise 355 Show that alternative choice of units are N_0 for number of molecules and $M^{-1}N_0^{-1}$ for time; then show that the remaining free dimensionless parameter is λ^* . (Hint: $M^{-1}N_0^{-1}t$ is dimensionless.)

Exercise 356 Can you think of reasons why you might prefer to (not) use the alternative units of the previous exercise.

5.18 How to choose units

It is noteworthy that (scaled versions of) quantities remain in the dimensionless model whenever these quantities have *not* been used to define the units. This gives you a general guideline for choosing units: do *not* involve the parameters which are of interest.

Exercise 357 Suppose that the model now also deals with the mRNA transcript of a second gene. Now you are dealing with two genes, say I and II. Would number of mRNA molecules of species I and number of mRNA molecules of species II count as two distinct empirical dimensions, or would both variables belong to the same class number of mRNA molecules?

5.19 Buckingham's theorem

The foregoing example suggests that by scaling, you can reduce the number of independent parameters by as many dimensions as there are independent dimensions in your model. *Buckingham's theorem* asserts that this is indeed always so. *Every relationship*

$$f(Q_1, \dots, Q_M) = 0 \quad (5.8)$$

among the M quantities in a model proper can be rewritten in dimensionless form

$$\bar{f}(Z_1, \dots, Z_n) = 0 \quad Z_i \stackrel{d}{\sim} 1, \quad i = 1, \dots, n \quad (5.9)$$

as a relationship \bar{f} among $n = M - m$ dimensionless quantities.

The following series of exercises establish a proof of this theorem.

Exercise 358 Show that, given a vector $\mathbf{r} \in \mathbf{R}^M$, you can form a quantity $W(\mathbf{r})$ which is expressible in the basis dimensions:

$$W(\mathbf{r}) = \prod_{j=1}^M Q_j^{r_j} \stackrel{d}{\sim} \prod_{i=1}^m D_i^{s_i} . \quad (5.10)$$

(Hint: recall the defining properties of a dimension basis, § 5.12.)

Exercise 359 Verify that equation (5.10) induces a linear transformation

$$T : \mathbf{R}^M \mapsto \mathbf{R}^m .$$

(Hint: the vector $\mathbf{r} \in \mathbf{R}^M$ is mapped to a vector $\mathbf{s} \in \mathbf{R}^m$.)

Exercise 360 Verify that the above transformation T has a kernel $\text{Nul } T$ of dimension $n = M - m$. (Warning: the term 'dimension' is used in the sense of linear algebra here!)

Exercise 361 Verify that $\mathbf{s} = \mathbf{0}$ means that the quantity $W(\mathbf{r})$ is dimensionless.

Exercise 362 Let $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ be a basis for $\text{Nul } T$ and verify that this induces n dimensionless quantities

$$Z_k = \prod_{j=1}^M Q_j^{b_j^k} \stackrel{d}{\sim} 1 \quad k = 1, \dots, n .$$

5.20 Forming a basis

Adding vectors $\{\mathbf{b}_{n+1}, \dots, \mathbf{b}_M\}$ to the basis of $\text{Nul } T$, you can form a complete basis for \mathbf{R}^M . With each of these additional basis vectors, associate a quantity

$$Z_k = \prod_{j=1}^M Q_j^{b_j^k} \not\stackrel{d}{\sim} 1 \quad k = n + 1, \dots, M .$$

Exercise 363 Show that none of these quantities Z_k is dimensionless. (Hint: the kernel of T is already spanned by the first n basis vectors.)

Exercise 364 Show that each Q_j ($j = 1, \dots, M$) can be expressed uniquely as a product of powers of the $\{Z_k\}_{k=1}^M$. (Hint: the matrix collecting the basis vectors is invertible.)

Exercise 365 Show that each relation (5.8) can be rewritten in the form $\bar{f}(Z_1, \dots, Z_M) = 0$. (Hint: the previous exercise.)

5.21 Dimensionless quantities

If we vary units arbitrarily, $f(Q_1, \dots, Q_M) = 0$ remains valid in virtue of compensatory changes among all the Q_j ; on the other hand, the first n Z_j are immune from such compensation since they are dimensionless.

Exercise 366 Argue that such immunity implies that you can derive $\bar{f} \neq 0$ by varying units, unless f only depends on Z_1, \dots, Z_n .

Exercise 367 Show that \bar{f} only depends on Z_1, \dots, Z_n . (Hint: $\bar{f} = 0$ is valid, no matter what the units are.)

Exercise 368 Show that $\bar{f}(Z_1, \dots, Z_M) = 0$ reduces to $\bar{f}(Z_1, \dots, Z_n) = 0$.

5.22 The proof concluded

In the last exercise you proved Buckingham's theorem, which tells us that we need to work with only n quantities rather than M . This greatly simplifies the amount of work to do, and, more importantly, essential relationships are preserved when non-dimensionalizing. In most cases there are at least as many parameters as there are independent dimensions in the unscaled model, and one is left with some remaining scaled parameters.

Exercise 369 Suppose a trainspotter spots a number of engines during the course of a windy autumn day. In the evening, the diligent anorak adds up all these numbers and calculates the average. Comment on the usefulness of this calculation.

5.23 Measurement strength

The solecism of the last exercise illustrates the issue of the *strength* of measurements: not all scales are measurements are strong enough to support all possible arithmetical operations.

5.24 The nominal scale of measurement

On a *nominal scale of measurement*, number is merely a name. No arithmetical operations make empirical sense on a nominal scale. The only comparison possible within a nominal scale is for equality.

Exercise 370 Give examples of quantities measured on a nominal scale. (Hint: think of situations where categories of data are represented by integers.)

5.25 The ordinal scale of measurement

On an *ordinal scale of measurement*, the order of the numbers now has a meaning. For instance, every pair of biological species may be assigned the numbers 1, 2, 3, 4, 5, or 6 according to whether they belong to the same genus, family, order, class, phylum, or kingdom.

Exercise 371 Verify ordinality: a pair of species characterized by 3 is more closely related than a pair characterized by 5 on this scale.

Exercise 372 Explain why the difference between pairs that score 4 and 5 cannot be meaningfully compared to another pair of pairs that score 2 and 3. (Hint: the numbers 1 to 6 might just as well have been 3, 12, 45, 700, 701, and 764980.)

Exercise 373 Is a species pair that scores 6 three times less closely related than a pair that scores 2?

Exercise 374 Can you meaningfully work with averages on an ordinal scale? Why (not)?

Exercise 375 (Transitivity) Let Or represent an ordinal scale, so that $Or(X)$ is the measurement ('score') of object X on this scale. Then verify that $Or(A) > Or(B)$ and $Or(B) > Or(C)$ together imply $Or(A) > Or(C)$.

Exercise 376 Can the concept of empirical dimension be applied to quantities that are measured on an ordinal scale?

5.26 The interval scale of measurement

On an *interval scale of measurement*, differences between the numbers have an empirical meaning.

Exercise 377 Show that you can meaningfully do additions, subtractions and averaging on the interval scale.

Exercise 378 Argue why it makes sense to state that $15\text{ }^{\circ}\text{C}$ is the mean of $10\text{ }^{\circ}\text{C}$ and $20\text{ }^{\circ}\text{C}$. (Hint: think about mixing equal quantities of water of different temperatures.)

Exercise 379 Is it true that a temperature of $20\text{ }^{\circ}\text{C}$ is twice as high as $10\text{ }^{\circ}\text{C}$? Why (not)?

Exercise 380 Explain why proportions (ratios, quotients) between measurements are meaningless on an interval scale. (Hint: the zero of an interval scale is arbitrary.)

5.27 The ratio scale of measurement

On a *ratio scale of measurement*, differences and proportions (ratios, quotients) between the numbers have an empirical meaning. The zero is no longer arbitrary but expresses a natural null point. All arithmetic makes sense on the ratio scale. Only the unit is still arbitrary on the ratio scale. The ratio scale is the scale of most commonly used quantities in physics, chemistry and biology.

Exercise 381 Explain why most mistakes with scale strength are due to an erroneous assumption that the measurement at hand has the strength of a ratio scale.

5.28 The absolute scales of measurement

Only *absolute scales* are stronger than ratio scales. These scales are like the ratio scale, but the unit is no longer arbitrary.

Exercise 382 Can you see why a common mistake with quantities on an absolute scale is to consider them to be dimensionless?

Exercise 383 Indicate whether the following scales are nominal, ordinal, interval, ratio, or absolute: (i) mass in kilograms; (ii) Moh's scale for the hardness of minerals; (iii) temperature on the Kelvin scale; (iv) 'semi-quantitative' scales that express the cognitive state of an animal, such as degree of aggressiveness or IQ; (v) the temperature scales of Celsius, Réaumur and Fahrenheit; (vi) population size; (vii) time in seconds; (viii) the Beaufort scale of wind; (ix) the decibel scale of sound intensity; (x) the Richter scale.

Exercise 384 Verify that only quantities on an interval scale or stronger have an empirical dimension.

Exercise 385 Explain why a quantity that admits measurement on a scale of a given strength also admits measurement on a weaker scale. (Hint: everything can be expressed on a nominal scale, if only as 'measurement number x '.)

Exercise 386 Why would you prefer a weaker scale? (Hint: reliability.)

Exercise 387 Can you prove the following claims?

- (i) A nominal scale is invariant under any bijection (1-1 transformation).
- (ii) An ordinal scale is invariant under any monotone transformation.
- (iii) An interval scale is invariant under any transformation of the form $y = ax + b$.
- (iv) A ratio scale is invariant under any transformation of the form $y = ax$.
- (iv) An absolute scale is invariant only under the identity $y = x$.

5.29 Data

The combined result of measurements performed to chart or characterize a real-world phenomenon is called *the experimental data*⁸. A major use of mathematical models is to order data and extract information from these data (cf. 326).

Exercise 388 Contrast and compare the following two ways in which a model interacts with data: data *source material* versus data as *test material*.

5.30 Testing predictions

When the model is developed independently of some particular data set, confrontation of the model with that data set amounts to the testing of a prediction.

Exercise 389 Is it essential that the test data set is collected only after the model has been developed?

5.31 Parameter estimation

In *parameter fitting*⁹, data serve both as source and test material, as parameter values are estimated while the model is being confronted with the data. Suppose that a model f describes a process variable y as a function of time t and three parameters α, β, γ :

$$y = f(t; \alpha, \beta, \gamma)$$

and suppose that a number n of data pairs (t_i, y_i) is available (the entire data set may be represented as (\mathbf{t}, \mathbf{y})).

Exercise 390 Suppose that there are i, j such that $t_i = t_j$. Do you expect $y_i = y_j$? If not, how do you account for the discrepancy?

Exercise 391 Do you expect the data point y_i to equal $f(t_i; \alpha, \beta, \gamma)$? If not, how do you account for the discrepancy?

5.32 Goodness of fit

To quantify how strongly the model prediction confirms to (or deviates from) the data, in other words, to express the **goodness-of-fit**, the following **sum of squared errors** function S is widely used:

$$S(\alpha, \beta, \gamma; \mathbf{t}, \mathbf{y}) = \sum_{i=1}^n (f(t_i; \alpha, \beta, \gamma) - y_i)^2. \quad (5.11)$$

Exercise 392 Is the goodness-of-fit high or low when S is high?

⁸The singular is *datum* which means 'that which is given'.

⁹Sometimes also called *calibration*.

5.33 Best-fit values

The best-fit values $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}\}$ of the three parameters are those for which S is minimized. These are obtained by putting $\partial S/\partial\alpha = \partial S/\partial\beta = \partial S/\partial\gamma = 0$ and solving for the parameters.

Exercise 393 Discuss (numerical) procedures to determine the best-fit values.

Exercise 394 In general, you are looking for a minimum of S in a parameter space that has as many dimensions as you have parameters. Discuss the difficulties you might expect determining the *global* minimum of S .

5.34 Simultaneous fitting

It often happens that the data comprise simultaneous measurements: thus you have data points of the form (t_i, x_i, y_i, z_i) is available. If the model describes all these state variables, you can extend the procedure to find the best fitting set of parameter values using all available data sets at once.

Exercise 395 (Simultaneous fitting can be very parsimonious) Suppose that you have, as in the above example, three parameters and three state variables. The data can be represented as three graphics, plotting, respectively, data points of the form (t_i, x_i) , (t_i, y_i) , and (t_i, z_i) . How many parameters are you estimating “per curve”? How does this compare to a straight line (often touted as “the simplest possible model”)?

5.35 A justification for least-squares

The method of minimizing the sum of squares with respect to the parameters can be justified as follows. Assume that the measurement error in y_i is normally distributed about the predicted value, while the t_i are known with perfect accuracy. Furthermore, the errors are taken to be: (i) unbiased relative to a ‘true’ model, that is, the expected value of every measurement is what is predicted by the model for the ‘true’ parameter values; (ii) independent; and of (iii) equal variance σ^2 . Then the joint probability density p of finding the measurements \mathbf{t}, \mathbf{y} given the model and its parameters is

$$p(\mathbf{y}; \mathbf{t}, \alpha, \beta, \gamma) = \prod_{i=1}^n (\sqrt{2\pi}\sigma)^{-1} \exp\{-([f(t_i; \alpha, \beta, \gamma) - y_i]/\sigma)^2/2\}.$$

Now you apply a conceptual twist: you view this p as the **likelihood** of the *parameters*, given the data. You are looking for the parameter values that maximize this likelihood.

Exercise 396 Show that you have to maximize

$$L(\alpha, \beta, \gamma; \mathbf{t}, \mathbf{y}) = -\sigma^{-2} \sum_{i=1}^n (f(t_i; \alpha, \beta, \gamma) - y_i)^2$$

with respect to the parameters. (Hint: take logarithms, discard terms that do not depend on the parameters.)

Exercise 397 Show that maximizing L amounts to minimizing the sum of squares as defined in equation (5.11).

Exercise 398 Verify that you do not need to know (or estimate) the value of σ .

5.36 Multiple data sets

The log-likelihood argument can be extended to simultaneous data sets, as follows. Let x_i, y_i, z_i be three measurements at time t_i . Suppose that f, g, h are the solutions of the model that was proposed to describe the data. The log-likelihood argument then leads us to minimize

$$L(\alpha, \beta, \gamma; \mathbf{t}, \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^n \left(\frac{(f(t_i; \alpha, \beta, \gamma) - x_i)^2}{\sigma_x^2} + \frac{(g(t_i; \alpha, \beta, \gamma) - y_i)^2}{\sigma_y^2} + \frac{(h(t_i; \alpha, \beta, \gamma) - z_i)^2}{\sigma_z^2} \right). \quad (5.12)$$

The only additional difficulty is that now the variances $\sigma_x^2, \sigma_y^2, \sigma_z^2$ of the measurement error within each of the three simultaneous data sets $\mathbf{x}, \mathbf{y}, \mathbf{z}$ must be known.

Exercise 399 How would you tackle this difficulty?

5.37 Scaling by variance

One solution is as follows: when iteratively solving the set of equations $\partial S / \partial \alpha = 0 \dots$ you use the variance found in the previous iteration. To initialize this procedure (in iteration 1 there is no previous iteration!) you use the within-dataset variances.

Exercise 400 Suppose that you obtain a reasonably good fit, or perhaps a poor fit, or an excellent fit. How do you think goodness-of-fit should influence your confidence in the model?

Exercise 401 Suppose that in your exploration of your model you have found that, as parameters are varied, the curves depicting the process variables' variation with time assume a wide range of shapes. Explain why in such circumstances even a good agreement between best fit and data does little to confirm the structure of the model. (Hint: how 'easy' is it for the parameters to find fitting values?)

Exercise 402 Suppose that, by contrast, your model can only assume a restricted set of shapes. Argue why in this case a good agreement between best fit and data does enhance confidence in the model¹⁰.

Exercise 403 Suppose once more that your model is 'versatile' in the sense of exercise 401, and suppose that you have obtained a good fit. How confident are you that the best-fit values of the parameters are close to the "true" values? Answer the same question when your model is 'restricted' in the sense of exercise 402, and you have obtained a good fit. (Hint: how large is the region of the parameter space where the fit is any good?)

5.38 A fundamental trade-off

The foregoing exercises suggest a trade-off between the confidence a good fit gives you *in the model* and the confidence a good fit gives you *in the fitted values*. You can chart how the model's behaviour changes over the parameter space to determine on which side of this trade-off your model lies.

Exercise 404 Discuss the role of dimensional analysis in this exploration of parameter space.

¹⁰Keep in mind, though, that a radically different model, based on quite distinct assumptions, may provide an equally good fit to the data.

5.39 Estimating confidence

A more direct way to assess the confidence you might have in your best-fit parameter values is to determine (by numerical means, if necessary) the second partial derivatives with respect to the parameters:

$$\frac{\partial^2 S}{\partial \alpha^2}, \quad \frac{\partial^2 S}{\partial \alpha \partial \beta} \quad \text{and so on,}$$

evaluated at $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$, $\gamma = \hat{\gamma}$.

Exercise 405 Explain why these quantities express how precisely the parameter values are fixed by the data.

5.40 Fisher information

In the more general setting of likelihood theory, the statistical expectations of the second partial derivatives of the log likelihood with respect to the parameters make up a matrix called the *Fisher information*, an apt name as it tells us how well the parameters are fixed by the data.

Exercise 406 Explain the seemingly perverse attitude of a modeller who claims that the large values in the Fisher information matrix tend to *support* the theoretical hypothesis of his model.

5.41 Out-of-sample prediction

After the best fitting parameter values have been found, you may yet decide to confront the model with an additional data set that was not used in the fitting procedure (the model then gives a so-called *out-of-sample* prediction).

6

Inferential statistics

6.1 Populations

A *population* of size N consists of N items to each of which is assigned a value: $\{a_1, a_2, \dots, a_N\}$. A *random observation* picks out one of these values; if X is the observation, then $\mathbb{P}(X = a_i) = \frac{1}{N}$ for all i .

Exercise 407 make sure you understand the following notation:

$$\mathbb{P}(X \leq x) = \frac{1}{N} \sum_{i=1}^N 1_{a_i \leq x}.$$

(Hint on notation: $1_C = 1$ if the condition C is true, otherwise $1_C = 0$.)

6.2 Observations

If *multiple observations* are carried out, the second observation picks a value out of the population, again at random, from among the items left after the first observation.

Exercise 408 If X_1 is the first observation and X_2 the second, show that

$$\mathbb{P}(X_2 \leq x) = \frac{1}{N-1} \sum_{i=1}^N 1_{a_i \leq x} - \frac{1_{X_1 \leq x}}{N-1}.$$

Exercise 409 Can you generalize the formula of the previous exercise to the case of n observations? For the k th observation ($1 \leq k \leq N$) you have

$$\mathbb{P}(X_k \leq x) = \frac{1}{N+1-k} \sum_{i=1}^N 1_{a_i \leq x} - \frac{\sum_{i=1}^{k-1} 1_{X_i \leq x}}{N+1-k}.$$

6.3 Means

The average of the n observations is the *sample mean* $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Exercise 410 What do you know about the difference between the population mean $\frac{1}{N} \sum_{i=1}^N X_i$ and the sample mean when $n = N$?

6.4 The population distribution function

Let

$$F^{[N]}(x) = \mathbb{P}(X \leq x) = \frac{1}{N} \sum_{i=1}^N 1_{a_i \leq x}.$$

This is the *population distribution function*.

Exercise 411 Consider an infinitely large population and let $F(x) = \lim_{N \rightarrow \infty} F^{[N]}(x)$. Show that you have

$$\mathbb{P}(X_k \leq x) = F(x)$$

for each of the n observations in the sample. (Hint: take the limit for the formula you found in exercise 409.)

Exercise 412 Argue, using exercise 411, that the observations can be viewed as independent realizations of a random variable with distribution function F provided that the sample is very small compared to the population ($n \ll N$). We will always work under this assumption.

6.5 The statistical population

An infinitely large population of the sort considered in the exercise 411 is sometimes called a *statistical population*. The cumbersome correction terms with which you worked in exercise 409 vanish, which accounts for the central importance of such infinite (hence idealized) populations in statistical theory.

Exercise 413 Contrast and compare this notion of a statistical population with that of a population as encountered in demographics.

Exercise 414 The sample mean is also a random variable. Derive the following results:

$$\mathbb{E}(\bar{X}) = \mu \quad \text{and} \quad \mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}$$

where μ and σ^2 denote the mean and the variance of the distribution F . (Hint: refer to the formula that defines the sample mean, and apply the rules for expectation and variance of weighted sums. Refer to the worksheet on probability if you do not remember these rules.)

Exercise 415 What can you say about the difference between the mean μ and the sample mean when $n \rightarrow \infty$?

6.6 Repeated observations

The foregoing exercises indicate why *samples of repeated observations* are a mainstay of experimental science: first, the average of the observations is a random variable whose mean coincides with that of the population (we say that it is an *unbiased estimate* of μ); and second, the variance of this estimate decreases as the sample size increases. This decreasing variance means that the sample mean gives a more *reliable* indication of the true population mean μ as more observations are carried out.

Exercise 416 If X_k is the k th observation, the quantity $|X_k - \mu|$ is customarily called the *error* associated with the k th observation. Contemplate the (in)appropriateness of this terminology for various experimental contexts you can think of.

6.7 Sample variance

An unbiased estimate of the population variance σ^2 is given by a quantity called the *sample variance*:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X} - X_i)^2. \quad (6.1)$$

Exercise 417 Can you show that the sample variance indeed provides an unbiased estimate? (Hint: you are asked to show that $\mathbb{E}(S^2) = \sigma^2$.)

Exercise 418 Show that the sample squared deviation of the mean $\frac{\sum_{i=1}^n (\bar{X} - X_i)^2}{n-1}$ is a *biased* estimator of the population variance σ^2 . (Hint: use the result of exercise 417.)

Exercise 419 Finally, show that the quantity $\frac{\sum_{i=1}^n (\mu - X_i)^2}{n}$ is an *unbiased* estimator of the population variance σ^2 .

6.8 An unbiased estimate of sample variance

Together, the last few exercises show that the sample variance contains a factor $(n - 1)$ instead of n because the sample mean is used in the formula, instead of the true population mean.

Exercise 420 Why do we need to use the sample mean? (Hint: is it always possible to use the formula of exercise 419?)

Exercise 421 Deduce that an unbiased estimator of the variance of the sample mean is:

$$\frac{1}{n(n-1)} \sum_{i=1}^n (\bar{X} - X_i)^2 .$$

(Hint: combine your results of exercises 414 and 417.)

Exercise 422 A set of replicated observations (a sample) is usually reported as the sample mean plus or minus an indication of the spread in these data. When would you use S the square root of the sample variance, and when would you use S/\sqrt{n} ? (Hint: refer to the previous exercise.)

Exercise 423 Suppose you do an observation where you know that the variable observed follows the Poisson distribution (so $\mathbb{P}(X = x) = e^{-\mu} \mu^x / x!$). However, you do not know the value of the parameter μ and you want to distinguish¹ between the case where $\mu \leq \tilde{\mu}$ and the case where $\mu > \tilde{\mu}$, where $\tilde{\mu}$ is some predefined value. Suppose the observed value X is much larger than $\tilde{\mu}$. Would this incline you to reject the hypothesis $\mu \leq \tilde{\mu}$ or rather the hypothesis $\mu < \tilde{\mu}$?

Exercise 424 Continuing the previous exercise, consider a set of integer values $\mathcal{R} = \{r, r + 1, r + 2, \dots\}$ where $r > \tilde{\mu}$. Show that for all $x \in \mathcal{R}$ you have

$$\mathbb{P}(X = x) \leq e^{-\tilde{\mu}} \tilde{\mu}^x / x!$$

under the hypothesis $\mu \leq \tilde{\mu}$, whereas

$$\mathbb{P}(X = x) > e^{-\tilde{\mu}} \tilde{\mu}^x / x!$$

under the hypothesis $\mu > \tilde{\mu}$. Hence conclude that the latter hypothesis is “more likely” to be true for observations in \mathcal{R} (where this idea of likelihood still wants some threshing out).

Exercise 425 With \mathcal{R} and X as defined in the previous two exercises, show that $\mathbb{P}(X \in \mathcal{R}) \leq \sum_{i=r}^{\infty} e^{-\tilde{\mu}} \tilde{\mu}^i / i!$ under the hypothesis $\mu \leq \tilde{\mu}$. Write down a similar expression for the “alternative” hypothesis $\mu > \tilde{\mu}$.

6.9 Hypotheses

It is useful to have a handy abbreviation for the hypotheses at hand, so that we do not have to repeat their specification in full every time we need to refer to them. Thus, in the decision problem considered in the foregoing exercises, let H_0 denote the hypothesis that $\mu \leq \tilde{\mu}$; and let H_1 denote the hypothesis that $\mu > \tilde{\mu}$.

Exercise 426 Suppose you decide to reject H_0 when the observed value is found to be in the set \mathcal{R} . Show that

$$\mathbb{P}(H_0 \text{ is rejected while } H_0 \text{ is true}) \leq \sum_{i=r}^{\infty} \frac{e^{-\tilde{\mu}} \tilde{\mu}^i}{i!} .$$

¹A typical example of such a Poisson-distributed observable in cell biology is the number of ion channels of some particular type that open in a fixed time window, under the prevailing membrane potential, neurotransmitter concentration, and suchlike. The test decision—whether or not the parameter exceeds $\tilde{\mu}$ —can help us decide whether these conditions affect the opening kinetics of the ion channel under study.

6.10 The critical region

When a set of values like \mathcal{R} is used to decide to reject H_0 on the basis of the observation, it is called the *critical region*. We say that H_0 is rejected when the observed value falls in the critical region, and that H_0 is accepted otherwise.

Exercise 427 A priori, a critical region may be any set of values. Would it be sensible to use the following set as critical region? $\{r, r+2, r+4, \dots\}$ where $r > \tilde{\mu}$. Why (not)?

6.11 Choosing alpha

It is clearly erroneous to reject H_0 when it is in fact true. We try to control this error by deciding on the maximum probability of error we are willing to live with (traditionally denoted by α). Popular choices for the value of α are 0.001, 0.01 and 0.05.

Exercise 428 For a critical region of the form $\mathcal{R} = \{r, r+1, r+2, \dots\}$, find the smallest r (i.e. the largest critical region) such that

$$\mathbb{P}(H_0 \text{ is rejected while } H_0 \text{ is true}) \leq \alpha$$

where α has a fixed value.

6.12 Alpha or P-value?

After the observation has been made, the *P-value* associated with the observed value is the smallest choice of α such that H_0 would be rejected on the basis of the observed value.

Exercise 429 Consider the above definition carefully. Is it better to decide on α (also known as the *size of the test*) beforehand, and reject or accept H_0 accordingly, *or* report the P-value (a.k.a. the *observed size*)?

Exercise 430 Let P_ξ denote the P-value associated with the observed value ξ :

$$P_\xi = \mathbb{P}(X \geq \xi \mid \mu = \tilde{\mu})$$

(where X is the observed value) and let $\xi(p)$ be the observed value for which the P-value equals p :

$$\mathbb{P}(X \geq \xi(p) \mid \mu = \tilde{\mu}) = p.$$

Finally, let F_{P_ξ} denote the distribution function of the P-value. Can you show that the P-value P_ξ follows the standard uniform distribution when $\mu = \tilde{\mu}$? (Hint: you are asked to show that $F_{P_\xi} = p$; recall that $F_{P_\xi} = \mathbb{P}(P_\xi \leq p)$.)

6.13 A useful result

The conclusion of the previous exercise holds in general: the P-value follows the standard uniform distribution under H_0 .

6.14 Another type of error

While H_0 can be falsely rejected, it could also be falsely accepted. This other type of error (*type II*) is given by the following probability

$$\mathbb{P}(H_0 \text{ is accepted while } H_1 \text{ is true}).$$

Exercise 431 Show that

$$\mathbb{P}(H_0 \text{ is accepted while } H_1 \text{ is true}) \leq \sum_{i=0}^{r-1} \frac{e^{-\tilde{\mu}} \tilde{\mu}^i}{i!}$$

where r is the boundary of the critical region, as before.

Exercise 432 Suppose that you attempt to control the first type of error by choosing α very small. Can you deduce how this affects the type II error? (Hint: in what direction does r move when α becomes smaller? How does that affect the type II error? See the previous exercise.)

6.15 Power

The type II error is traditionally denoted by β . The quantity $1 - \beta$ is known as the *power* of the test.

Exercise 433 Verify the following interpretation of statistical power: the probability of rejecting the null hypothesis, when it ought to be rejected.

6.16 Estimation

In testing, the problem is to decide whether a population's distribution parameter is above or below some cut-off value. Another approach to population parameters is *estimation*, where the unknown parameter is to be assigned some value.

Exercise 434 Suppose that the observation X is Poisson-distributed, $\mathbb{P}(X = x) = e^{-\mu} \mu^x / x!$, where the parameter μ is unknown. Suppose, moreover, that we find $X = 10$. Which of the following values would you be most inclined to assign to μ : 1, 10, 100? Why?

Exercise 435 Suppose that you have n independent, repeated observations of the random variable of the previous exercise: X_1, X_2, \dots, X_n . Show that

$$\mathbb{P}(X_1 = x_1 \ \& \ X_2 = x_2 \ \& \ \dots \ X_n = x_n) = e^{n\mu} \prod_{i=1}^n \frac{\mu^{x_i}}{x_i!}. \quad (6.2)$$

Exercise 436 How would you find the value of μ for which the right hand side of equation (6.2) is maximal?

6.17 Likelihood

The *likelihood* of the value μ for the population parameter is defined as follows:

$$\mathbb{L}(\mu \mid X_1 = x_1 \ \& \ X_2 = x_2 \ \& \ \dots \ X_n = x_n) = e^{n\mu} \prod_{i=1}^n \frac{\mu^{x_i}}{x_i!}. \quad (6.3)$$

The expression on the right is exactly the same as that of equation (6.2). However, the twist is that here this term is computed as a function of μ *given* the observed values, whereas the classical context of equation (6.2) is that the probability of observing these values is calculated, given the parameter value μ .

Exercise 437 Explain the need for new terminology, that is, why is the quantity calculated called a likelihood not probability, which may strike you as very odd since the formula is exactly the same as that of a probability. (Hint: is μ a random variable at all?)

Exercise 438 Explain why the μ -value that maximizes the likelihood \mathbb{L} is the one that is “the most reasonable” in view of the observations.

Exercise 439 Show that the μ that maximizes $\mathbb{L}(\mu)$ is identical to the μ that maximizes the *log-likelihood* $\ln \mathbb{L}(\mu)$.

Exercise 440 Denote by $\hat{\mu}$ the μ that maximizes $\ln \mathbb{L}(\mu)$. Can you show that $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$? (Hint: start with the expression $\frac{d \ln \mathbb{L}}{d\mu} = 0$; work out a formula for $\ln \mathbb{L}$ first, using right member of equation 6.3.)

6.18 Multiple observations and likelihood

When the distribution of the observations is continuous, the likelihood is formally identical to the joint probability density function.

Exercise 441 Suppose that the n observations in the sample follow the normal distribution, independently with parameters μ and σ^2 . Write down the joint probability density function. Can you show that maximizing $\ln \mathbb{L}$ is equivalent to minimizing $\sum_{i=1}^n (\mu - X_i)^2$?

Exercise 442 Consider N observations on a Bernoulli variate; the outcome 1 is observed $n \leq N$ times. Show that $\hat{p} = n/N$ is the maximum likelihood estimate. (Hint: the likelihood is $\binom{N}{n} p^n (1-p)^{N-n}$; take the logarithm, differentiate with respect to p , put equal to zero.)

6.19 Logistic regression

A typical application of maximum likelihood estimation is the estimation of a dosage threshold for all-or-none responses². The starting point is a *model function* G , which provides the probabilities of response and non-response:

$$\mathbb{P}(\text{response} \mid \text{dose } x) = G(x; \vartheta) \quad (6.4)$$

$$\mathbb{P}(\text{no response} \mid \text{dose } x) = 1 - G(x; \vartheta) \quad (6.5)$$

where ϑ is the parameter vector of the model, which we aim to estimate on the basis of observations. These observations take the following form: there are n observations of the form (x_i, Y_i) , where x_i is the dose (set by the experimenter) and Y_i is the observation, which is scored as 0 when the i th replicate failed to respond, and as 1 when the i th replicate did respond.

Exercise 443 Verify the following formula:

$$\ln \mathbb{L} = \sum_{i=1}^n 1_{Y_i=1} \ln\{G(x_i; \vartheta)\} + 1_{Y_i=0} \ln\{1 - G(x_i; \vartheta)\}.$$

Exercise 444 Can you write down the equations that determine the maximum likelihood estimate of ϑ ? (Hint: you are asked to work out $0 = \frac{\partial \ln \mathbb{L}}{\partial \vartheta_j}$ where ϑ_j is the j th component of ϑ .)

Exercise 445 Can you work out the maximum likelihood estimation equations for the following model function?

$$G(x) = \frac{1}{1 + \exp\{-\vartheta_1(x - \vartheta_2)\}}.$$

Exercise 446 The experimental data may be presented in a slightly different form: m different doses are included in the experiment (doses x_k , $k = 1, \dots, m$) and at each dose n_k measurements are taken; the *proportion* of responders is recorded as p_k , where $0 \leq p_k \leq 1$. Can you adapt the formula of exercise 443 to this data format?

Exercise 447 How would you adapt the procedure in the case where the datum x is specified by more than one number (i.e. is higher-dimensional)?

²As a typical example, consider an experiment in which cells are incubated with various concentrations of a compound that induces the expression of a cluster of genes (associated with a specific cellular function) in these cells.

6.20 A standard test problem

A common test problem is to decide whether a pair of samples contain observations on a single statistical population, or, alternatively, on two distinct populations³.

Exercise 448 If the two samples have different sample means, this is *prima facie* evidence that the two samples do correspond to two distinct statistical populations. However, there is a serious problem with this idea. Explain the problem, and explain why it would be alleviated if the two samples were both infinitely large. (Hint: how likely is it that the two sample means are *exactly* the same, even if they arise from a single, common population?)

6.21 Characterizing the samples

Denote the observations from the first sample by $X_1, \dots, X_i, \dots, X_{n_X}$, those of the second sample by $Y_1, \dots, Y_i, \dots, Y_{n_Y}$. Assume that all observations are independent and normally distributed, with mean μ_X and variance σ^2 in the first sample and with mean μ_Y and variance σ^2 in the second sample. The sample means are $\bar{X} = \frac{1}{n_X} \sum_{i=1}^{n_X} X_i$ and $\bar{Y} = \frac{1}{n_Y} \sum_{i=1}^{n_Y} Y_i$.

Exercise 449 Verify that $\mathbb{E}(\bar{X}) = \mu_X$, $\mathbb{E}(\bar{Y}) = \mu_Y$, $\mathbb{V}(\bar{X}) = \sigma^2/n_X$ and $\mathbb{V}(\bar{Y}) = \sigma^2/n_Y$. (Hint: exercise 414.)

Exercise 450 Can you show that the difference $\bar{Y} - \bar{X}$ is normally distributed with mean $\Delta\mu = \mu_Y - \mu_X$ and variance $\sigma^2(1/n_X + 1/n_Y)$?

Exercise 451 Verify that the hypothesis that the samples were taken from a single statistical population is expressed by the formula $\Delta\mu = 0$ whereas the hypothesis that the samples were taken from distinct populations is expressed by $\Delta\mu \neq 0$.

Exercise 452 If the difference $\bar{Y} - \bar{X}$ is much greater than the standard deviation evident from the data, would you be inclined to believe $\Delta\mu = 0$ or $\Delta\mu \neq 0$? What if the observed difference in means is much smaller than the standard deviation?

6.22 A test statistic

We shall use the observed difference in means $\bar{Y} - \bar{X}$ to take our test decision. Such a quantity used to base statistical inference on is called a *test statistic*.

Exercise 453 Motivate the methodological requirement that a test statistic must be capable of being calculated entirely from the observational data.

6.23 The null hypothesis

We shall take the *null hypothesis* H_0 to be $\Delta\mu = 0$. The null hypothesis fully determines the distribution of the test statistic.

Exercise 454 Why is the distribution function of the test statistic $\bar{Y} - \bar{X}$ not fixed by the *alternative hypothesis* $H_1: \Delta\mu \neq 0$?

Exercise 455 Show that the quantity

$$\frac{\bar{Y} - \bar{X} - \Delta\mu}{\sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

³For example, we want to know if a certain gene G is expressed at different levels in cancerous (transformed) cells. We measure mRNA levels of gene G in (i) a sample of transformed cells and (ii) a sample of normal cells. These measurements clearly constitute observations on two distinct populations when G changes its activity during cancer. However, if G behaves essentially the same in both kinds of cells, the observations all arise from a *single* statistical population.

follows the standard normal distribution under the null hypothesis H_0 .

Exercise 456 Motivate the following choice for the critical region

$$(-\infty, C_L] \cup [C_R, +\infty)$$

where $C_L < 0$ and $C_R > 0$. (Hint: your more intuitive answer to exercise 452 may come in helpful.)

6.24 Determining the critical region

The boundaries of the critical region are determined by

$$\mathbb{P}\left(\frac{\bar{Y} - \bar{X}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} < C_L\right) = \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}\left(\frac{\bar{Y} - \bar{X}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} > C_R\right) = \frac{\alpha}{2}$$

where α is the desired size of the test.

Exercise 457 Using a set of statistical tables, verify that the boundaries of the critical region are given by

$$C_{L,R} = \pm\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}z_{1-\alpha/2}$$

when σ^2 is known.

6.25 Using an estimate for the variance

Of course, in practice σ^2 will not be known, and the “z-table” is of no use. We are able to estimate σ^2 , though. Also, since we are concerned with the distribution of the test statistic under H_0 , we can pool the samples as if they did arise from a single statistical population.

Exercise 458 Show that the pooled estimate of the variance:

$$S^2 = \frac{\sum_{i=1}^{n_X} (X_i - \bar{X})^2 + \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2}{n_X + n_Y - 2}$$

is an unbiased estimate of σ^2 . (Hint: imitate your calculation for exercise 417.)

Exercise 459 Show that $(X_i - \bar{X})^2/\sigma^2$ follows a chi-square distribution with one degree of freedom.

Exercise 460 Can you show that $(n_X + n_Y - 2)S^2/\sigma^2$ is χ^2 distributed with $\nu = n_X + n_Y - 2$?

Exercise 461 Can you show that the following scaled version of the test statistic:

$$\frac{\bar{Y} - \bar{X}}{S\sqrt{1/n_X + 1/n_Y}}$$

follows Student’s t -distribution?

Exercise 462 Confirm that the determination of the boundaries of the critical region now proceeds as in exercise 457, but with $z_{1-\alpha/2}$ replaced by $t_{1-\alpha/2}$ (again to be found in a statistical table).

Exercise 463 Can you see how to adapt the test if the hypotheses are: $H_0: \Delta\mu = 0$ versus $H_1: \Delta\mu > 0$?

Exercise 464 Some experiments yield comparative data that are expressed as a single sample⁴. Discuss how the above procedure is to be adapted. (Hint: you have been treating the two samples as a single pooled sample (under H_0) in the previous proceedings anyway, so up to notational changes the procedure is essentially the same.)

6.26 The choice of a critical region, in general

The examples of statistical testing have so far been simple enough that the “shape” of the critical region could be determined on the basis of common sense. However, it is useful to have a general criterion for selecting a critical region. The general situation is that of two competing hypotheses, H_0 versus H_1 , which make a statement on the value of some parameter ϑ of interest:

$$H_0 : \vartheta = \vartheta_0 \quad \text{vs} \quad H_1 : \vartheta = \vartheta_1 . \quad (6.6)$$

Exercise 465 Let X_1, \dots, X_n represent the data on which the test decision is to be based. Make sure you can interpret the following:

$$\mathbb{P}((X_1, \dots, X_n) \in \mathcal{R} \mid \vartheta_0) = \alpha$$

(Hint: you should have encountered all this notation before.)

Exercise 466 Comment on the rationale of the following criterion: *The optimal critical region $\widehat{\mathcal{R}}$ of size α for the test problem (6.6) is such that for any other region \mathcal{R}_α of size α you have:*

$$\mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \mid \vartheta_1) \geq \mathbb{P}((X_1, \dots, X_n) \in \mathcal{R}_\alpha \mid \vartheta_1) . \quad (6.7)$$

(Hint: recall that H_0 is rejected when the test statistic takes a value in the critical region; recall the definition of (statistical) power and observe that the probabilities in (6.7) refer to the event of taking a *correct* test decision.)

6.27 The Neyman-Pearson lemma

The Neyman-Pearson lemma gives an explicit recipe for the construction of the optimal critical region specified by condition (6.7). If $f(x_1, \dots, x_n; \vartheta)$ is the probability density function of the test statistic,

$$\widehat{\mathcal{R}} = \left\{ (x_1, \dots, x_n) \mid \frac{f(x_1, \dots, x_n; \vartheta_0)}{f(x_1, \dots, x_n; \vartheta_1)} \leq k \right\} \quad (6.8)$$

where k is a positive constant chosen such that

$$\mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \mid \vartheta_0) = \alpha . \quad (6.9)$$

Exercise 467 Argue why a point (x_1, \dots, x_n) for which the ratio

$$f(x_1, \dots, x_n; \vartheta_0) / f(x_1, \dots, x_n; \vartheta_1)$$

is very small is a point you would be strongly inclined to include in the critical region. (Hint: what does this ratio tell you about the likelihoods of the competing hypotheses at that point?)

Exercise 468 Motivate the philosophy behind the Neyman-Pearson lemma: to include points in the critical region, starting with the ones for which the ratio

$$f(x_1, \dots, x_n; \vartheta_0) / f(x_1, \dots, x_n; \vartheta_1)$$

is smallest, until the desired size α is reached.

⁴For instance, consider an experiment in which the effect of cytokine C on the expression of gene G is determined by measuring gene G mRNA levels before and after incubation with C in n cells. For each cell, the “before” measurement is subtracted from the “after” measurement, yielding a list of n differences.

6.28 Proving the Neyman-Pearson lemma

The following exercises work up to a proof of the Neyman-Pearson lemma.

Exercise 469 Consider $\widehat{\mathcal{R}}$ as specified by equation (6.8) and any other \mathcal{R}_α of size α , and let $\widehat{\mathcal{R}} \setminus \mathcal{R}_\alpha$ denote the set of all points in $\widehat{\mathcal{R}}$ except those it has in common with \mathcal{R}_α (if any). Verify that

$$\begin{aligned} \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \mid \vartheta_1) &= \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \setminus \mathcal{R}_\alpha \mid \vartheta_1) \\ &\quad + \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \cap \mathcal{R}_\alpha \mid \vartheta_1) \end{aligned}$$

(Hint: a Venn diagram may prove helpful.)

Exercise 470 With set notations as in the previous exercise, Show that

$$\mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \setminus \mathcal{R}_\alpha \mid \vartheta_1) \geq \frac{1}{k} \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \setminus \mathcal{R}_\alpha \mid \vartheta_0),$$

Exercise 471 With set notations as in the previous exercise, show that

$$\mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \setminus \mathcal{R}_\alpha \mid \vartheta_0) = \alpha - \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \cap \mathcal{R}_\alpha \mid \vartheta_0).$$

Exercise 472 With set notations as in the previous exercise, show that

$$\alpha - \mathbb{P}((X_1, \dots, X_n) \in \widehat{\mathcal{R}} \cap \mathcal{R}_\alpha \mid \vartheta_0) = \mathbb{P}((X_1, \dots, X_n) \in \mathcal{R}_\alpha \setminus \widehat{\mathcal{R}} \mid \vartheta_0)$$

where $\mathcal{R}_\alpha \setminus \widehat{\mathcal{R}}$ denotes the set of all points in \mathcal{R}_α except those it has in common with $\widehat{\mathcal{R}}$ (if any).

Exercise 473 Can you now prove the Neyman-Pearson lemma, that is, show that condition (6.7) is satisfied by the critical region specified by (6.8)? (Hint: string the results of the previous four exercises together.)

6.29 Generalized likelihood ratio principle

The version of the Neyman-Pearson you have just proved deals with *simple* hypotheses, that is, ones that only feature equals signs in their statements on parameters. To accommodate *complex* hypotheses (ones with \neq , $>$, $<$, \geq , or \leq , such as you have already encountered) we generalize the likelihood ratio principle as follows: the test statistic has a probability density function $f(x_1, \dots, x_n; \vartheta)$ with a parameter vector $\vartheta \in \Omega$ and the competing hypotheses are:

$$H_0 : \vartheta \in \Omega_0 \quad \text{vs} \quad H_1 : \vartheta \in \Omega \setminus \Omega_0. \quad (6.10)$$

where $\Omega_0 \subset \Omega$. The *unrestricted* maximum likelihood estimate of the parameter vector is $\widehat{\vartheta}$ where

$$f(x_1, \dots, x_n; \widehat{\vartheta}) \geq f(x_1, \dots, x_n; \vartheta) \forall \vartheta \in \Omega.$$

Exercise 474 Verify that the maximum likelihood estimate of the parameter vector under H_0 is $\widehat{\vartheta}_0$ where

$$f(x_1, \dots, x_n; \widehat{\vartheta}_0) \geq f(x_1, \dots, x_n; \vartheta) \forall \vartheta \in \Omega_0.$$

6.30 The likelihood ratio

The *likelihood ratio* is now defined as follows:

$$\lambda(x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n; \hat{\vartheta}_0)}{f(x_1, \dots, x_n; \hat{\vartheta})} \quad (6.11)$$

and the critical region is defined by

$$\hat{\mathcal{R}} = \{(x_1, \dots, x_n) \mid \lambda(x_1, \dots, x_n) \leq k\} \quad (6.12)$$

where k is a positive constant chosen such that

$$\mathbb{P}((X_1, \dots, X_n) \in \hat{\mathcal{R}} \mid \hat{\vartheta}_0) = \alpha. \quad (6.13)$$

Exercise 475 Consider a sample of n independent, normally distributed observations (with parameters μ and σ^2 , the latter assumed known), where $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$. Verify that for this problem, you have $\Omega = (-\infty, +\infty)$, $\Omega_0 = \{\mu_0\}$, $\hat{\vartheta} = \bar{X}$ (the sample mean), $\hat{\vartheta}_0 = \mu$.

Exercise 476 For the test problem of the previous exercise, show that the likelihood ratio is given by the following formula:

$$\lambda = \exp\{-n(\bar{x} - \mu_0)^2 / (2\sigma^2)\}.$$

Exercise 477 The critical region for the test problem of exercise 475 is defined by $\lambda \leq k_\alpha$ (where k_α solves the equation $\mathbb{P}(\lambda \leq k) = \alpha$ for k , under H_0). Show that the following formula is equivalent to this condition on λ :

$$\frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} \geq k_\alpha. \quad (6.14)$$

Exercise 478 Show that condition (6.14) leads to the critical region

$$(-\infty, -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}, +\infty)$$

for the test statistic $(\bar{x} - \mu_0) / (\sigma\sqrt{n})$, or, equivalently, to the critical region

$$[\chi_{1-\alpha}(1), +\infty)$$

for the test statistic $(\bar{x} - \mu_0)^2 / (\sigma^2 n)$.

Exercise 479 Can you work out the procedure for the test problem of exercise 475ff. when the hypotheses are $H_0: \mu = \mu_0$, $H_1: \mu > \mu_0$? (Hint: $\Omega = [\mu_0, \infty)$.)

Exercise 480 Can you work out the likelihood ratio procedure for the test problem of exercise 475ff. in the case where σ^2 is unknown? (Hint: Ω is 2-dimensional, with parameter axes μ and σ^2 , while Ω_0 is the line corresponding to $\mu = \mu_0$ in the Ω -plane; you should arrive at the two-sided t -test considered in exercise 462.)

6.31 Analysis of variance

The generalized likelihood ratio test can be extended to the test problem where there are more than two samples of normal variates, and it is to be decided if all these samples arise from a common statistical population.

Exercise 481 Review *analysis of variance (ANOVA)* in a textbook; e.g. Bain & Engelhardt pp.423–425.

6.32 Chi-square tests

Both estimation and statistical tests concern the uncovering of information regarding parameters of interest on the basis of experimental data (observations). Test problems can often be cast as a question whether the population at hand follows a given distribution.

Exercise 482 Consider n statistically independent observations on a Bernoulli variate⁵ so that $\mathbb{P}(X_i = 1) = p$ where p is the Bernoulli parameter. Verify that $\sum_{i=1}^n X_i$ follows the Binomial distribution with parameters p and n . Consider $H_0: p = p_0$ versus $H_1: p \neq p_0$. Discuss how the likelihood ratio is used to construct the critical region.

Exercise 483 For the test problem of the previous exercise, show that, when n is large, you could equally well do a z -test, where the test statistic

$$\frac{\sum_{i=1}^n X_i - np_0}{\sqrt{np_0(1-p_0)}}$$

follows a standard normal distribution. (Hint: recall the formulæ for mean and variance of a binomial variate; remember the Central Limit Theorem.)

Exercise 484 For the test problem of the previous exercise, show that you could just as well do a χ^2 -test, where the test statistic

$$\frac{(\sum_{i=1}^n X_i - np_0)^2}{np_0(1-p_0)}$$

follows a chi-square distribution with one degree of freedom.

Exercise 485 Show that the test statistic of the previous exercise can be rewritten as follows:

$$\sum_{i=1}^2 \frac{(o_i - e_i)^2}{e_i}$$

where o_1 is the number of observed 0s, o_2 is the number of observed 1s, e_1 is the expected number of zeros under H_0 (which is $(1-p_0)n$, as you should verify) and e_2 is the expected number of 1s under H_0 (which is p_0n).

Exercise 486 Suppose you have r samples, with n_i observations on a Bernoulli variate in the i th sample⁶ and the testing problem is $H_0: p_i = p_{i0}$ versus $H_1: p_i \neq p_{i0}$ ($i = 1, \dots, r$). Show that the test statistic

$$\sum_{i=1}^r \sum_{j=1}^2 \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where $e_{i1} = (1-p_{i0})n_i$ and $e_{i2} = p_{i0}n_i$, approximately follows a chi-square distribution with r degrees of freedom when the n_i are all large. (Hint: review your results on the sum of a number of χ^2 variates.)

⁵For instance, the experiment may involve n cells which have been incubated, at a given time and a given concentration, with a mutagenic compound. Each cell is subsequently cultured and the ensuing colony is checked for evidence of transformation after some set time. Inasmuch as the outcome is binary (yes/no transformation) the fate of each cell is a Bernoulli variate, whose parameter p is the probability of mutation.

⁶Continuing the previous example on mutagenesis, the experiment is carried out at r different incubation concentrations, with a mathematical model giving the value of p_i at the i th incubation concentration.

6.33 Further chi-square tests

Besides binomials, common discrete distribution-testing problems involve multinomial distributions⁷. As with the binomial test, the multinomial test can be extended to r samples.

Exercise 487 Review the probability density function of the multinomial distribution.

Exercise 488 Let c be an integer, $c \geq 2$. Show that c independent Poisson variates with parameters μ_i ($i = 1, \dots, c$), conditioned on their sum being equal to a given constant s , follow a multinomial distribution with parameters $p_i = \mu_i / (\sum_{j=1}^c \mu_j)$. (Hint: review the closure property of Poisson variates: a sum of independent Poisson variates is again a Poisson variate.)

Exercise 489 Consider a testing problem with a single multinomially distributed sample of size n (with c distinct outcomes) where $H_0: p_i = p_{i0}$ versus $H_1: p_i \neq p_{i0}$ ($i = 1, \dots, c$). Calculate e_i , the expected number of times the i th outcome is observed, under the null hypothesis.

Exercise 490 For the testing problem of the previous exercise, can you show that the test statistic

$$\sum_{i=1}^c \frac{(o_i - e_i)^2}{e_i}$$

approximately follows a chi-square distribution with $c - 1$ degrees of freedom when the number of observations is large? (Hint: first, show that the conditioning of exercise 488 is equivalent, in the large n limit, to having $c - 1$ independent Poisson variates plus one that is correlated with the others; then derive a sum in which the first $c - 1$ terms are χ^2 , with expectation and variance both equal to 1, while the c th term has expectation and variance zero in the limit; you may want to consider trinomials first, before tackling the general case.)

Exercise 491 Consider an r -sample test of a c -multinomial⁸ where H_0 specifies the c probabilities for each of the r “ c -nomials”. Show that the test statistic

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

approximately follows a chi-square distribution with $r(c - 1)$ degrees of freedom when the n_i are all large. (Hint: the result quoted in exercise 490 plus the rule for sums of independent χ^2 variates.)

6.34 Testing for independence

Perhaps a more common variant of the “ $r \times c$ ” test problem occurs when H_0 merely states that the probabilities are the same for each of the r samples (i.e. $p_{ij} \equiv p_j \forall i \in \{1, \dots, r\}$) without specifying them⁹. This is known as *testing for independence*: the null hypothesis is that the categorization implied by the r samples does not affect the probability distribution. To calculate expected occurrence numbers, you have to estimate the (common) probability distribution for your c -multinomial. To do this, you pool the r samples and take the frequencies of the c outcomes over the r samples taken together as estimates \hat{p}_j , $j = 1, \dots, c$ of the probabilities.

⁷Multinomials arise quite naturally in basic sequence analysis, in which the occurrence frequencies of alphabet symbols in the sequence are scored; an *alphabet* is a finite set of distinct possibilities, called symbols; for instance, the DNA alphabet consists of 4 nucleotides, whereas the proteome consists of 20 amino acids.

⁸For example, the investigator may consider r regions of the genome, suspecting that these regions differ in nucleotide usage patterns; each region i has a certain length, say n_i base pairs. The investigator records the frequencies at which each of the four nucleotides occur in every region. These frequencies are then compared to the expected numbers calculated on the basis of the null hypothesis, which specifies the probabilities of finding adenosine, guanosine, cytosine, and thymidine for each region (for each region, these four probabilities should add up to 1).

⁹Continuing the previous example, the null hypothesis would merely state that the r regions of the genome under consideration contain the four nucleotides at the same frequencies. Incidentally, note that the quaternary alphabet is an obvious, but not the only choice to study DNA sequences; a binary alphabet (purine, pyrimidine) can also be useful. Similar lumped alphabets can be defined for proteins.

Exercise 492 Can you confirm that the pooling strategy is consistent with the generalized likelihood ratio approach?

Exercise 493 Can you show that the test statistic for the chi-square test for independence,

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where $e_{ij} = \hat{p}_j n_i / N$ with $N = \sum_{i=1}^r n_i$ and $\hat{p}_j = \sum_{i=1}^r o_{ij} / N$, approximately follows a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom when the n_i are all large? (Hint: how many independent Poisson variates are there, in terms of the approximation used in exercise 490? Note that now you have not only constraints of the form $\sum_{j=1}^c o_{ij} = n_i$, as before, but the totals $\sum_{i=1}^r o_{ij} = N\hat{p}_j$ are also fixed, since H_0 treats these as givens.)

6.35 Goodness-of-fit tests

The theory behind χ^2 tests is somewhat tricky, and it can be hard to arrive at the correct degrees of freedom count. Nonetheless, tests based on the χ^2 -distribution are extremely useful and straightforward to carry out.

Exercise 494 Let X denote a random variable which takes a finite number of values x_{a_1}, a_2, \dots, a_c with probabilities p_1, p_2, \dots, p_c . Suppose that a sample of n independent observations is made on X , and you want to test a null hypothesis which specifies the probability distribution (i.e. H_0 assigns specific values to p_1, p_2, \dots, p_c). How would you proceed? (Hint: exercises 489 and 490.)

Exercise 495 Continuing the previous exercise, how would you adapt your procedure if the distribution function of X (i) is discrete but non-zero for denumerably many values; or (ii) is continuous?

Exercise 496 A common *goodness-of-fit* problem is to test whether some distribution, with parameters as yet unknown, is appropriate. Suppose you estimate k parameters from the data; then how many degrees of freedom does your test statistic have?

6.36 Screening

So far, we have been working on the tacit assumption that the evaluation of a given scientific experiment involves one corresponding statistical decision problem (test). However, it is often the case that the evaluation of the data obtained in an experimental study requires a very large (say, K) number of statistical comparisons to be carried out¹⁰ A study which involves a large number of comparisons with concomitant statistical decisions is called a *screening*.

Exercise 497 Suppose that the size of each comparison is α . Show that you expect αK erroneous conclusions under the *experiment-wise null hypothesis* that H_0 is true for each comparison.

Exercise 498 Suppose that H_0 is true in $K^\circ \leq K$ cases, with each comparison of size α . Explain why you expect αK° false rejections of null hypotheses.

Exercise 499 Continuing the previous exercise, suppose that each comparison involves a test with power $1 - \beta$. Explain why you expect $\beta(K - K^\circ)$ false non-rejections of null hypotheses.

Exercise 500 Continuing the previous two exercises, let $\varkappa = (K - K^\circ)/K$ (i.e., the fraction of comparisons where the null hypothesis is false) and show that the probability of a rejection (i.e., declaration of a statistically significant result) being correct is $(1 - \beta)\varkappa / (\alpha(1 - \varkappa) + (1 - \beta)\varkappa)$.

¹⁰Recall the example of the expression of a single gene G in transformed versus non-transformed cells, where mRNA levels for gene G were determined in samples of transformed and non-transformed cells. The mRNA expression measurement may be done on a very large number of genes.

Exercise 501 Show that for small fractions of true positives ($\varkappa \ll 1$) the probability of a rejection being correct is approximately equal to $\frac{1-\beta}{\alpha}\varkappa$. Comment on this formula for the case $\alpha = 0.05$, $\beta = 0.4$, $\varkappa \leq 0.04$.

6.37 Sensitivity and specificity

In the context of multiple comparisons, the power $1-\beta$ is often referred to as the *sensitivity* of the comparison, and $1-\alpha$ is called the *specificity*. A typical application is the detection of medical conditions in a population, and then \varkappa is called the *prevalence*.

Exercise 502 Show that an erroneous rejection of the experiment-wise null hypothesis occurs with the following probability:

$$1 - \mathbb{P}(\text{for } i = 1, \dots, K, P_i > \alpha)$$

where P_i denotes the P-value in test i and α denotes the size of the individual comparison.

Exercise 503 Can you show that

$$\mathbb{P}(P_i > \alpha) = 1 - \alpha$$

where P_i is any of the K observed P-values α is the size of the individual comparison? (Hint: exercise 430.)

Exercise 504 If P_{Exp} is the probability of an experiment-wise type I error, show that you have

$$P_{\text{Exp}} = 1 - (1 - \alpha)^K$$

where α is the size of the individual comparison. Hence, if α_T is the ‘total’ experiment-wise size, derive

$$\alpha = 1 - (1 - \alpha_T)^{1/K}$$

where α is the size of the individual comparison. (Hint: the previous two exercises.)

Exercise 505 Show that you have the approximation

$$1 - (1 - \alpha_T)^{1/K} \approx \frac{\alpha_T}{K}$$

when α_T is small.

6.38 Sharing alpha

The formula $\alpha = \alpha_T/K$ suggests that α_T behaves like a valuable, scarce resource which is divided equally among all K comparisons.

Exercise 506 Suppose $\alpha_T = 0.05$ and $K = 20,000$. Calculate α . What does this mean for the statistical power of your comparisons, if all are done at size $\sim \alpha_T/K$?

6.39 Conservation of statistical power

The moral of the last exercise is that the conservation of as much statistical power as possible is clearly imperative.

Exercise 507 Among the K observed P-values, let $P_{[1]}$ denote the smallest (so that $P_{[1]} \leq P_i \forall i$) and denote by $H_0^{[K]}$ the experiment-wise null hypothesis that all K null hypotheses are true. Show that the following is a test for $H_0^{[K]}$: reject $H_0^{[K]}$ if $P_{[1]} \leq 1 - (1 - \alpha_T)^{1/K}$, accept otherwise. (Hint: if even P_1 exceeds $1 - (1 - \alpha_T)^{1/K}$, what do you know about the rest of the P-values? Observe that exercise 502 gives a formula for P_{Exp} .)

6.40 A key observation

If the experiment-wise null hypothesis $H_0^{[K]}$ has been accepted, the final conclusion has thereby already been reached: for none of the K comparisons is the individual null hypothesis rejected.

Exercise 508 Suppose that $H_0^{[K]}$ is rejected. Show that the individual null hypothesis for the comparison associated with the lowest P-value $P_{[1]}$ is rejected.

6.41 The next step

When $H_0^{[K]}$ is rejected, a conclusion has yet to be reached on the $K - 1$ comparisons other than the one associated with the lowest P-value. Let $H_0^{[K-1]}$ be the hypothesis that the null hypotheses associated with these remaining comparisons are all true. It is important to note that $H_0^{[K-1]}$ only enters considerations when $H_0^{[K]}$ is rejected.

Exercise 509 Let $P_{[2]}$ denote the next-but-one-smallest P-value ($P_{[1]} \leq P_{[2]} \leq P_i \forall i$). Show that the following is a test for $H_0^{[K-1]}$: reject $H_0^{[K-1]}$ if $P_2 \leq 1 - (1 - \alpha_T)^{1/(K-1)}$, accept otherwise. (Hint: this involves essentially the same idea as exercise 507.)

Exercise 510 Which conclusions are reached on the remaining $K - 1$ comparisons according as $H_0^{[K-1]}$ is rejected or accepted?

6.42 The step-down procedure

The general pattern should now be clear: for $0 \leq \ell \leq K - 1$, hypothesis $H_0^{[K-\ell]}$ is tested by comparing the all-but- ℓ smallest P-value $P_{[\ell+1]}$ to $1 - (1 - \alpha_T)^{1/(K-\ell)}$. If $H_0^{[K-\ell]}$ is accepted, the comparisons associated with $P_{[i]}$ for $i \leq \ell$, $i \neq 0$, have their individual null hypotheses rejected while the remaining ones have their null hypothesis accepted, and the procedure terminates. On the other hand, if $H_0^{[K-\ell]}$ is rejected and $\ell < K - 1$, ℓ is set to $\ell + 1$ and the $H_0^{[K-\ell-1]}$ is tested.

Exercise 511 Study the above algorithm for the *step-down procedure* carefully.

Exercise 512 Suppose that after the step-down procedure has terminated, K^* comparisons have had their null hypotheses rejected and the remaining $K - K^*$ have had their null hypotheses accepted. Verify that, while the experiment-wise size is α_T , the individual comparisons have been carried out at size $1 - (1 - \alpha_T)^{1/(K-\ell)}$ for $\ell = 0, \dots, K^* - 1$ and size $1 - (1 - \alpha_T)^{1/(K-K^*)}$ for $\ell = K^*, \dots, K - 1$.

Exercise 513 Can you explain why the step-down procedure will typically conserve statistical power? (Hint: the previous exercise.)

Exercise 514 How do you decide in general that a collection of statistical tests needs to be subsumed under the heading of “a single” experiment? In other words: under which circumstances should any two statistical comparisons be considered to be part of the same experiment (and hence need to be corrected for the experiment-wise type I error)?

Exercise 515 As you have seen, a test statistic must be capable of being calculated from the observational data alone; and moreover, its distribution function should be known. Can you think of other desirable characteristics in a test statistic? (Hint: revise this topic in a statistics textbook, e.g. Bain & Engelhardt, chapter 10.)

Exercise 516 Is it necessary that the distribution function of a test statistic can be evaluated analytically?

III

Advanced topics

Optimization

7.1 Simple extrema

Exercise 517 Let y depend on \mathbf{x} through a function F . Show that maximization of $y = F(\mathbf{x})$ is equivalent to minimization of $-F(\mathbf{x})$.

Exercise 518 Suppose \mathbf{x} takes a finite number of different values. Outline a procedure to find the value of \mathbf{x} for which $y = F(\mathbf{x})$ attains a minimum. (Hint: go through a list, and keep track of the lowest image value so far.)

Exercise 519 Show that the maximum of $y = -x^2$ is found at $x = 0$.

Exercise 520 Show that the minimum of $y = (x - \alpha)^{2/3}$ occurs at $x = \alpha$.

Exercise 521 Show that the maximum of $\exp\{-(x_1 - \alpha)^2(x_2 - \beta)^2\}$ is located at $(x_1, x_2) = (\alpha, \beta)$.

Exercise 522 Consider $F(x) = x/(1+x)$. Show that the maximum of $F(x)$ equals $\alpha/(1+\alpha)$ subject to the constraint $-1 < x \leq \alpha$.

Exercise 523 With $F(x) = x/(1+x)$, show that $1 > F(x)$ for $x \geq 0$. Also show that $2 > F(x)$ for $x \geq 0$.

Exercise 524 With $F(x) = x/(1+x)$, show that there is a positive value for ε such that $2 - \varepsilon > F(x)$ for $x \geq 0$. By contrast, show that there is *no* positive value for ε such that $1 - \varepsilon > F(x)$ for $x \geq 0$.

7.2 Bounds

In exercise 523, you established that both 1 and 2 are upper bounds to $F(x)$ in the range of x given. In exercise 524 you showed that, moreover, 1 is a *least upper bound*. A least upper bound is also known as a *supremum*.

Exercise 525 Can you define the term *infimum* (also known as *greatest lower bound*)?

Exercise 526 Consider, once more, $F(x) = x/(1+x)$. Show that $F(x)$ attains *no* maximum on the x -interval $(-1, \alpha)$ (where $\alpha > -1$), but that it does have a supremum (namely, $\alpha/(1+\alpha)$) on this interval.

Exercise 527 Let $F(x) = \alpha + \beta x$ where α and β are two non-negative constants. Suppose that x is restricted to some set \mathcal{X} . Show that $F(x)$ has a supremum provided that \mathcal{X} has both an infimum and a finite supremum.

Exercise 528 Find the minimum of $y = x_1^2 + x_2^2$ subject to the constraint $x_2 - \alpha + \beta x_1 = 0$. (Hint: use the constraint to eliminate x_2 from the equation for y .)

7.3 Dealing with constraints

The technique of the last exercise, of reducing the dimension by substitution, can be difficult to apply if the constraint cannot be put in explicit form (that is, in the form of an equation of which either x_1 or x_2 is the subject). An alternative trick is based on increasing the dimension of the problem with the introduction of an auxiliary parameter λ . The aim is to maximize a function $y = F(x_1, x_2)$ subject to a constraint of the form $g(x_1, x_2) = 0$.

Exercise 529 Consider the family of functions $F(x_1, x_2) + \lambda g(x_1, x_2)$ parametrized by λ . Show that the functions in this family agree (i.e. map to the same y -values) exactly at those points (x_1, x_2) where the condition is satisfied.

7.4 Maximizing the H-function

The function $F(x_1, x_2) + \lambda g(x_1, x_2)$, considered at a *fixed* value of λ , is just another function of two variables, which we may denote by $H(x_1, x_2, \lambda)$, and can be optimized in the usual manner ('unconstrained'). Let $(\hat{x}_{1\lambda}, \hat{x}_{2\lambda})$ denote the location of this maximum.

Exercise 530 Show that, for every value of λ you may write

$$H(\hat{x}_{1\lambda}, \hat{x}_{2\lambda}) \geq F(x_1, x_2) \quad \text{for every } x_1, x_2 \text{ such that } g(x_1, x_2) = 0. \quad (7.1)$$

(Hint: use the fact that $H(\hat{x}_{1\lambda}, \hat{x}_{2\lambda}) \geq H(x_1, x_2, \lambda)$ everywhere, by definition of maximum, and your finding at exercise 529.)

Exercise 531 Suppose $\hat{\lambda}$ is such that $g(\hat{x}_{1\hat{\lambda}}, \hat{x}_{2\hat{\lambda}}) = 0$. Show that $(\hat{x}_{1\hat{\lambda}}, \hat{x}_{2\hat{\lambda}})$ is in fact the constrained maximum you seek.

Exercise 532 Show that the maximum of the function $F(x_1, x_2)$, subject to the constraint $g(x_1, x_2) = 0$, if it exist, may be among the solutions of the system

$$\begin{aligned} \frac{\partial H(x_1, x_2, \lambda)}{\partial x_1} &= 0 \\ \frac{\partial H(x_1, x_2, \lambda)}{\partial x_2} &= 0 \\ \frac{\partial H(x_1, x_2, \lambda)}{\partial \lambda} &= 0 \end{aligned}$$

where $H(x_1, x_2, \lambda) = F(x_1, x_2) + \lambda g(x_1, x_2)$. (Hint: the first two equations locate an unconstrained maximum for an arbitrary value of λ ; the last one fixes λ by ensuring that the solution is on the locus of the constraint g .)

7.5 Lagrange multipliers

The technique you have established in the last few exercises is called the technique of *Lagrange multipliers*; λ is an example of a Lagrange multiplier.

Exercise 533 Rework exercise 528 using a Lagrange multiplier.

Exercise 534 Generalize the technique to multiple Lagrange multipliers: outline the solution method for the optimization of $F(\mathbf{x})$ subject to $g_1(\mathbf{x}) = 0$ and $g_2(\mathbf{x}) = 0$, based on two Lagrange multipliers λ_1, λ_2 .

7.6 Optimal process control

An important application of Lagrange multipliers is in optimal process control. The variables here are x_1, \dots, x_N and u_1, \dots, u_N , for some integer N , where the objective is to maximize the following quantity:

$$\sum_{i=1}^N h_i(x_i, u_i)$$

with x_1 fixed at some given value. The variables x_i are interpreted as the values assumed by the state variable of a one-dimensional discrete-time dynamic system at times $t_1, \dots, t_i, \dots, t_N$.

Exercise 535 Verify that x_1 is the initial condition,

7.7 Respecting the state transitions

Similarly, the u_i are the values of a forcing function at the successive instants in time t_1, t_2, \dots . The constraint is that the variables must satisfy the system's state transition function:

$$x_{i+1} = F(x_i, u_i) \quad \text{for } i = 1, \dots, N-1. \quad (7.2)$$

Exercise 536 Show that the latter constraint can be accommodated by introducing $N-1$ Lagrange multipliers, giving the objective function:

$$H = \sum_{i=1}^N h_i(x_i, u_i) + \sum_{i=1}^{N-1} \lambda_i (F(x_i, u_i) - x_{i+1}).$$

7.8 Finding an optimal solution

To determine an optimal solution, you differentiate H with respect to $x_2, \dots, x_N, u_1, \dots, u_N$ and $\lambda_1, \dots, \lambda_{N-1}$ and you set all these derivatives equal to 0.

Exercise 537 How many equations do you obtain in this way?

Exercise 538 Verify that roughly a third of these equations just give you back the system's dynamics (i.e. the state transitions according to equation—(7.2)).

Exercise 539 Show that you also obtain $N-1$ equations of the form

$$\lambda_i = h_{i+1}^{!x}(x_{i+1}, u_{i+1}) + \lambda_{i+1} F^{!x}(x_{i+1}, u_{i+1}) \quad (7.3)$$

for $i = N-1, N-2, \dots, 3, 2, 1$ where $!x$ indicates partial derivatives with respect to x (the first argument of these functions) and $\lambda_N = 0$ is a boundary condition.

Exercise 540 Interpret the Lagrange multiplier λ_i as the value assumed by a new state (the so-called *co-state*) at time t_i . What is its state transition function?

Exercise 541 Let

$$\pi_i(x_i, u_i) = h_i^{!u}(x_i, u_i) + \lambda_i F^{!u}(x_i, u_i)$$

where $!u$ indicates partial derivatives with respect to u (the second argument of these functions). Show that the final 'third' of equations is of the form

$$\pi_i(x_i, u_i) = 0. \quad (7.4)$$

Exercise 542 Suppose that for all i , u_i is constrained to an interval $[u_{\min}, u_{\max}]$, and that there is no u -value in this interval for which equation (7.4). Can you show that H is maximized by putting $u_i = u_{\max}$ whenever $\pi_i(x_i, u_i) > 0$ in the allowed interval, and $u_i = u_{\min}$ whenever $\pi_i(x_i, u_i) < 0$?

7.9 The switching function

The last exercise makes clear why π is called the *switching function*. If the switching function assumes the value zero at some point in time t_i , the associated control $u = u_i$ at that instant is said to be *singular*.

Exercise 543 Can you extend the procedure to higher-dimensional discrete-time dynamical systems?

7.10 Towards continuous time

The present approach to discrete-time optimal control problems can be applied to continuous time as well. In preparation for this, we define $h_i(x_i, u_i) = h(t, x_i, u_i)\Delta t$ and $f(x_i, u_i)\Delta t = F(x_i, u_i) - x_i$.

Exercise 544 Verify that $F^{fx}(x_i, u_i) = 1 + f^{fx}(x_i, u_i)\Delta t$ and $F^{fu}(x_i, u_i) = f^{fu}(x_i, u_i)\Delta t$.

Exercise 545 Derive the following:

$$\frac{x_i - x_{i-1}}{\Delta t} = f(x_{i-1}, u_{i-1})$$

and

$$\frac{\lambda_i - \lambda_{i+1}}{\Delta t} = -(h^{fx}(t, x_i, u_i) + \lambda_i f^{fx}(x_i, u_i)) .$$

Exercise 546 Can you now establish the following? For a process $x(t)$ described by the differential equation $\dot{x} = f(x(t), u(t))$ with $x(t_1)$ given as initial condition, where $u(t)$ is the controlling (or ‘input’) function, an optimal control regime $u(\cdot)$ with regard to the objective functional

$$J = \int_{t_1}^{t_2} h(s, x(s), u(s)) ds$$

may be sought by optimizing

$$H = h(t, x, u) + \lambda f(x, u) \tag{7.5}$$

with respect to u at each t , with $\lambda(t_2) = 0$, where

$$\dot{\lambda} = -\frac{\partial H}{\partial x} .$$

Exercise 547 With H given by equation (7.5), show that

$$\dot{x} = \frac{\partial H}{\partial \lambda} .$$

(Hint: you already know that $\dot{x} = f(x, u)$.)

Exercise 548 Suppose that $u(t)$ must satisfy $0 \leq u(t) \leq 1$ for all t , and the aim is to maximize J . Establish the following (candidate¹) optimal control: $u(t) = 0$ where $\pi(t) < 0$, $u(t) = 1$ where $\pi(t) > 0$, with $\pi(t) = \frac{\partial H}{\partial u}$.

7.11 Singular control

If $\pi(t) = 0$ only at isolated points in time (the *switching moments*), control as stipulated by the last exercise is said to be *bang-bang*. On the other hand, if there is a time interval of positive duration, say $[t_a, t_b]$ where $t_b - t_a > 0$ and $t_1 \leq t_a < t_b \leq t_2$, such that $\pi(t) = 0$ for all $t \in [t_a, t_b]$, then control is *singular* between t_a and t_b (there may be none, one or more such intervals).

Exercise 549 Suppose optimal control is singular for all $t \in [t_1, t_2]$. Show that

$$\lambda(t) = -\frac{h_{fu}(t, x(t), u(t))}{f_{fu}(x(t), u(t))} .$$

(Hint: consider $\pi(t)$.)

Exercise 550 Suppose that h does not depend explicitly on t ($h_{ft} = 0$), and that optimal control is singular everywhere, as in the previous exercise. Can you show that $\dot{H} = 0$?

¹We are not pursuing the subject with the rigour required to give assurances or detailed conditions; ‘insight’ not ‘proof’ is our watchword in these exercises.

Exercise 551 With the assumptions of the previous exercise, show that (prospective) optimal control $u(t)$ is determined by the equation

$$h(x(t), u(t)) - \frac{h_{ru}(x(t), u(t))f(x(t), u(t))}{f_{ru}(x(t), u(t))} = K \quad (7.6)$$

where K is a constant to be determined from the boundary condition $\lambda(t_2) = 0$.

7.12 Control in feedback form

An equation such as (7.6), through which $u(t)$ can be found from $x(t)$ at every t (i.e. ‘instantaneously’) is said to give control in *feedback form*.

Exercise 552 Consider the problem where an extremum of the integral $\int_{t_1}^{t_2} h(t, x, \dot{x})dt$ is sought. Show that the ‘optimal path’ $x(t)$ must satisfy

$$\frac{d}{dt} \frac{\partial h}{\partial \dot{x}} = \frac{\partial h}{\partial x}. \quad (7.7)$$

(Hint: consider $\dot{x} = u$ and assume that optimal control u is non-singular everywhere.)

7.13 The calculus of variations

The problem of the last exercise is a simple example of the sort of problems addressed by the *calculus of variations*².

Exercise 553 Can you rewrite equation (7.7) as follows?

$$\frac{d}{dt} \left(h - \dot{x} \frac{\partial h}{\partial \dot{x}} \right) = \frac{\partial h}{\partial t}.$$

Exercise 554 Suppose that $h_{tt} = 0$ everywhere. Then show that $h - \dot{x} \frac{\partial h}{\partial \dot{x}}$ is equal to some (as yet unknown) constant. (Hint: the previous exercise.)

7.14 A planar objective function

The following exercises consider a planar objective function of the form

$$F(x_1, x_2) = \alpha + \beta x_1 + \gamma x_2$$

where α , β and γ are constants, the latter two non-zero; the objective function F is to be maximized subject to $(x_1, x_2) \in \mathcal{X}$ where \mathcal{X} is a closed, simply connected subset of the (x_1, x_2) -plane.

Exercise 555 Consider a step $(\Delta x_1, \Delta x_2)$ in the (x_1, x_2) -plane. Verify that the associated change in objective value is given by $\Delta F = \beta \Delta x_1 + \gamma \Delta x_2$.

Exercise 556 Consider a *unit step* satisfying $(\Delta x_1)^2 + (\Delta x_2)^2 = 1$. If the step is at angle φ relative to the x_1 -direction, show that $\Delta x_1 = \cos \varphi$ and $\Delta x_2 = \sin \varphi$.

Exercise 557 Show that an extremum for ΔF under a unit step of angle φ is obtained for the φ -values that solve

$$\tan \varphi = \frac{\gamma}{\beta}. \quad (7.8)$$

(Hint: differentiate with respect to φ .)

²In the calculus of variations, equation (7.7) is known as *Euler’s equation* or the *Euler-Lagrange equation*.

Exercise 558 Verify that equation (7.8) has *two* solutions, one of which indicates the direction of *steepest ascent*, the other of which the direction of *steepest descent*. (Hint: draw a graph of $\tan \varphi$ over a full period.)

Exercise 559 Consider the *gradient*

$$\nabla F = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

which you should realize is a *vector*. Show that ∇F points in the direction of steepest ascent, and that $-\nabla F$ points in the direction of steepest descent³.

Exercise 560 Verify that the gradient is constant and non-zero throughout the domain \mathcal{X} . Hence conclude that the maximum exists and is unique.

7.15 Internal point

An *internal point* of the set \mathcal{X} has the property that there is a positive ε such that all points within a distance less than ε are contained in \mathcal{X} .

Exercise 561 Confirm this definition with your intuitive notion of ‘internal’.

Exercise 562 Consider a line segment PQ connecting at an arbitrary internal point P of the set \mathcal{X} with a point Q on the boundary of \mathcal{X} , such that PQ points in the direction of ∇F . Show that $F(Q) > F(P)$. (Hint: the notation means what you would expect; if x_1^Q, x_2^Q are the coordinates of the point Q , then $F(Q) = F(x_1^Q, x_2^Q)$.)

Exercise 563 Show that a maximum of F is to be found on the *boundary* of \mathcal{X} . (Hint: the previous exercise.)

Exercise 564 Let the relation $B(x_1, x_2) = 0$ define the boundary of \mathcal{X} . Show that an extremum of F on this boundary must satisfy the equation

$$\frac{B_{|x_1}(x_1, x_2)}{B_{|x_2}(x_1, x_2)} = \frac{\beta}{\gamma} \quad (7.9)$$

in addition to the condition $B = 0$. (Hint: use a Lagrange multiplier.)

Exercise 565 Suppose that the boundary of \mathcal{X} is a polygon. Argue why, generically, you will find the maximum only at one of the corner points⁴ of this polygon. (Hint: what does equation (7.9) look like for any given ‘edge’ of the polygon?)

Exercise 566 Suppose that the boundary of \mathcal{X} is a polygon. Consider a vertex V and the edge VW connecting V to a neighbouring vertex W . Show that the component of the gradient *along* VW can only point toward V if the angle between the edge VW and the gradient is greater than $\frac{\pi}{2}$.

Exercise 567 Suppose that the boundary of \mathcal{X} is a polygon. Consider a vertex V , its two neighbouring vertices W and U as well as the edges UV and VW . Suppose that the component of the gradient along UV points toward V , so that $F(V) > F(U)$, and that the component of the gradient along VW also points toward V , so that $F(V) > F(W)$. Consider the sum of the angles (i) between UV and the gradient and (ii) between Vw and the gradient; show that this sum exceeds π .

³In fact, this result holds generally, not just for planar objective functions.

⁴A corner point is usually called a *vertex*, plural *vertices*.

Exercise 568 Continuing the previous exercise, suppose that the maximum of F is achieved on some vertex M other than V ($M \neq V$). Since $F(V) > F(U)$ and $F(V) > F(W)$, you also have $M \neq U$ and $M \neq W$ (verify this). Consider the sum of the angles (i) between UV and MV and (ii) between VW and MV ; can you show that this sum, too, exceeds π ?

7.16 Convex domains

A simply connected domain is called *convex* if every line segment connecting two arbitrary points of the domain lies entirely within the domain.

Exercise 569 Show that a disc is convex.

Exercise 570 Draw the outline of a sea-star. Find two points such that the straight line segment connecting them lies only partly in your sea-star domain, thus establishing that the domain is not convex.

Exercise 571 In exercise 568 you considered the case where \mathcal{M} has a polygonal boundary; the maximum is achieved on vertex M and there are three other vertices U , V and W such that $F(V) > F(U)$ and $F(V) > F(W)$. Show that \mathcal{M} cannot be convex.

Exercise 572 Suppose that \mathcal{M} has a polygonal boundary and is convex. Let $F(M) > F(x) \forall x \in \mathcal{X}$. Show that M is a vertex and is, moreover, the only vertex V such that $F(V)$ is greater than both values of F at the two neighbouring vertices of V .

7.17 A systematic procedure

The last exercise validates the following procedure to find the maximum of a planar objective function on a convex domain with polygonal boundary: start at any vertex; if it has a neighbour where F is greater, move toward that neighbour⁵; continue moving toward ‘superior’ neighbouring vertices until you are at a vertex both of whose neighbours have a smaller F -value.

Exercise 573 Consider a problem of the following appearance:

$$\begin{aligned} \text{to maximize } F(x_1, x_2) &= \alpha_0 + \beta_0 x_1 + \gamma_0 x_2 \\ \text{subject to } 0 &\leq \alpha_i + \beta_i x_1 + \gamma_i x_2 \end{aligned}$$

for $i = 1, \dots, N$ where $N \geq 3$. All parameters are real and β_0, γ_0 are non-zero. Verify the following facts: (i) the constraints define a convex domain with a polygonal boundary; (ii) each vertex of the boundary defines a point where *two* of the N constraints are satisfied *with equality*; (iii) the problem is solved by applying the ‘vertex move’ method outlined above, with vertices translated into pairs of equations as per (i) and (ii). (Hint: for (i), first establish that a convex domain is bisected by a straight line into two convex domains, then use induction; for (ii) observe that the constraints define the edges.)

Exercise 574 Is the converse of statement (ii) in the previous exercise true? (Hint: that is, does every pair of constraints, taken with equality, define a vertex on the boundary of the domain of x -values allowed by the set of constraints?)

⁵Scanning the neighbours of your starting vertex, you may either move to the first ‘superior’ one that crops up, or you may scan both neighbours first and then, if both are superior, move to the ‘best’ of them.

7.18 Extending the technique to higher dimensions

Essentially the same method can be applied to the n -dimensional version of the problem⁶:

$$\begin{aligned} \text{to maximize } F(x_1, \dots, x_n) &= \kappa_{00} + \sum_{j=1}^n \kappa_{0j} x_j \\ \text{subject to } 0 &\leq \kappa_{i0} + \sum_{j=1}^n \kappa_{ij} x_j \end{aligned}$$

with $i = 1, \dots, N$, $N \geq n + 1$, where the parameters κ_{ij} are real numbers and the κ_{0j} , $j = 1, \dots, n$ are nonzero.

Exercise 575 Can you show that the domain is convex? (Hint: show that a straight line in \mathbb{R}^n will intersect an $n - 1$ -hyperplane at most once, and use that fact.)

Exercise 576 Let η_1, \dots, η_n be n real parameters, not all zero, and use these to define a step Δx such that $\Delta x_j = \eta_j \kappa_{0j} / \sqrt{\sum_{j=1}^n \eta_j^2 \kappa_{0j}^2}$. Verify that Δx is a unit step in n -dimensional space.

Exercise 577 Show that the change in objective value associated with the unit step defined in the previous exercise is

$$\Delta F = \frac{\sum_{j=1}^n \eta_j \kappa_{0j}^2}{\sqrt{\sum_{j=1}^n \eta_j^2 \kappa_{0j}^2}}.$$

Exercise 578 Show that an extremum of the change ΔF calculated in the previous exercise is found when η_1 through η_n are all equal, say $\eta_j \equiv \eta$. Show that for $\eta > 0$, the unit step Δx defined in exercise 576 is in the direction of the gradient ∇F , and that for $\eta < 0$, Δx points in the direction of $-\nabla F$. (Hint: set $\frac{\partial F}{\partial \eta_j}$ equal to 0 for all j to find the extremum.)

Exercise 579 Can you extend the idea of exercise 565 to show that, generically, the maximum of F is found at a vertex of the domain? (Hint: each vertex is defined by n equalities, chosen out of the set of N constraints.)

7.19 Vertex moves

The natural extension of the search method is to consider vertices defined by n -tuples of equalities chosen out of the N given constraints. As before, a move from one vertex to its neighbour corresponds to the replacing one of the equations of the n -tuples with an equality taken from the remaining $N - n$ constraints.

Exercise 580 The crux to the validity of the ‘‘vertex move’’ method is again to prove that a vertex that is superior to all of its neighbours must be the unique maximum. Can you furnish this proof? (Hint: suppose, to the contrary, that there is a vertex V other than the maximum which ‘dominates’ all its neighbours. Show that there is an $n - 1$ hyperplane \mathcal{P} such that (i) $F(x) = F(V)$ for all x on that hyperplane; (ii)—the edges connecting V to its neighbours are all on one side of \mathcal{P} ; and (iii) the maximum must be on the *other* side of \mathcal{P} . Then show that convexity enforces a contradiction.)

⁶Known to operations researchers as a *linear program*; the somewhat startling term ‘program’ hails from the fact that problems of this sort often arise as the mathematical representation of a resource allocation task faced by an organization such as a company.

Fourier series, Fourier transform, & Sampling

8.1 Functions that repeat

Exercise 581 Explain why the function

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\omega_0 t + b_k \sin k\omega_0 t \quad (8.1)$$

repeats itself every $2\pi/\omega_0$ time units, that is, $x(t + 2\pi/\omega_0) = x(t)$ for all values of t .

8.2 Periodic functions

Such functions are called T_0 -periodic where $T_0 = 2\pi/\omega_0$.

Exercise 582 Show that a T_0 -periodic function is also $2T_0$ -periodic, $3T_0$ -periodic, and so on.

8.3 The fundamental period

A periodic function thus has infinitely many periods, which are integral multiples of the *fundamental period* T_0 .

Exercise 583 Can you explain why the fundamental period of a given periodic function $x(t)$ is defined to be the *smallest* positive value of T such that $x(t + T) = x(t)$?

Exercise 584 Let \tilde{k} denote some selected value of k . Evaluate $\int_{-T_0/2}^{+T_0/2} x(t) \cos k\omega_0 t dt$ with $x(t)$ given by equation (8.1). You should obtain $a_{\tilde{k}} T_0/2$. (Hint: you can use integration by parts (twice!), or use *Euler's formulas* $\cos \theta = (e^\theta + e^{-\theta})/2$ and $\sin \theta = (e^\theta - e^{-\theta})/(2i)$ where $i^2 = -1$.)

8.4 Finding the coefficients

Your result

$$a_k = \frac{2}{T_0} \int_{-T_0/2}^{+T_0/2} x(t) \cos k\omega_0 t dt \quad (8.2)$$

indicates that the coefficients (as and bs) in equation (8.1) can be found by “integrating together” the function $x(t)$ with the cosines and sines of the corresponding frequencies (so a_k goes with a cosine of frequency $k\omega_0$, and so on).

Exercise 585 Also establish

$$a_k = \frac{2}{T_0} \int_{-T_0/2}^{+T_0/2} x(t) \cos k\omega_0 t dt \quad (8.3)$$

and

$$a_0 = \frac{2}{T_0} \int_{-T_0/2}^{+T_0/2} x(t) dt . \quad (8.4)$$

8.5 The Fourier series representation

The idea of a *Fourier series representation* is to represent *any* periodic function with fundamental period T_0 in the form of a weighed sum of cosines and sines, as in equation (8.1).

Exercise 586 Let $x(t)$ be a periodic function with fundamental period T_0 , such that $x(t) = 0$ for $-T_0/2 \leq t \leq 0$ and $x(t) = 1$ for $0 \leq t \leq T_0/2$. Sketch a graph of $x(t)$ for $x = -2T_0$ to $x = +2T_0$.

Exercise 587 With $x(t)$ as in exercise 586, show that $a_0 = 1$. (Hint: equation (8.4).)

Exercise 588 With $x(t)$ as in exercise 586, show that $a_1 = 0$. (Hint: equation (8.3); remember that $\omega_0 = 2\pi/T_0$, and refer to sketches of $\cos \omega_0 t$ and $\sin \omega_0 t$ on the interval $[-T_0/2, +T_0/2]$.)

Exercise 589 With $x(t)$ as in exercise 586, show that $a_k = 0$ for $k = 1, 2, 3, \dots$

Exercise 590 With $x(t)$ as in exercise 586, show that $b_1 = \frac{2}{\pi}$. (Hint: equation (8.2); remember that $\cos \pi = -1$.)

Exercise 591 With $x(t)$ as in exercise 586, show that $b_2 = 0$. (Hint: equation (8.2); remember that $\cos 2\pi = +1$.)

Exercise 592 Can you show, with $x(t)$ as in exercise 586, that $b_k = \frac{2}{k\pi}$ when k is odd, and $b_k = 0$ when k is even?

8.6 The amplitude spectrum

The coefficients (*as* and *bs*) which you can calculate for a given periodic function, using equations (8.2)—(8.4), contain *all* information about that function. For instance, the quantity $A_k = \sqrt{a_k^2 + b_k^2}$ tells you how much of frequency $k\omega_0$ “is present” in $x(t)$; A_k is the *amplitude* at frequency $k\omega_0$. The amplitudes together (for all values of k) make up the *amplitude spectrum* of $x(t)$.

Exercise 593 When $x(t) = \cos \omega_0 t$, show that $A_1 = 1$ and $A_k = 0$ for $k \neq 1$. (Hint: you already did most of the work in exercises 584 and 585.)

Exercise 594 Show that the *phase-shifted* function $x(t) = \cos(\omega_0 t - \varphi)$ has the same amplitude spectrum as $x(t) = \cos \omega_0 t$. (Hint: use the formulas $\cos(a - b) = \sin a \sin b + \cos a \cos b$ and $\sin^2 \varphi + \cos^2 \varphi = 1$.)

8.7 The phase spectrum

Different signals may thus have the same amplitude spectrum. To determine a function fully, another number must be provided at each frequency. This is the *phase* $\varphi_k = -\tan^{-1}(b_k/a_k)$. The phases together (for all values of k) make up the *phase spectrum* of $x(t)$.

Exercise 595 Can you tell what the phase spectrum of a function looks like if its Fourier series representation contains only sine terms (as with for instance the function of exercise 586)?

8.8 Accommodating aperiodic functions

The idea of ‘picking out the frequencies’ of a function by integrating that function together with sines and cosines, as in equations (8.2)—(8.4), is so nice that we would like to be able to apply this idea even when $x(t)$ is not periodic (or when perhaps it is, but we do not know the fundamental period). Since the function is aperiodic, we resolve to integrate from $-\infty$ to $+\infty$, and to do this for all possible frequencies (that is, for all real values of ω). This gives a function of ω :

$$X_{\cos}(\omega) = \int_{-\infty}^{+\infty} x(t) \cos \omega t dt \quad (8.5)$$

and, similarly,

$$X_{\sin}(\omega) = \int_{-\infty}^{+\infty} x(t) \sin \omega t dt . \quad (8.6)$$

Exercise 596 Assume that the aperiodic function $x(t)$ is non-zero only in a finite neighbourhood of $t = 0$:

$$x(t) = 0 \quad \text{for } |t| > \frac{T_1}{2} . \quad (8.7)$$

Show that

$$X_{\cos}(\omega) = \int_{-T_1/2}^{+T_1/2} x(t) \cos \omega t dt$$

on this assumption, and write down a similar expression for $X_{\sin}(\omega)$.

Exercise 597 Suppose that you try to assign, somewhat arbitrarily, a fundamental period T_0 to a function satisfying condition (8.7). Let $\omega_0 = 2\pi/T_0$ and explain why $X_{\cos}(k\omega_0) = a_k T_0/2$ is sure to be correct only if you choose the fundamental period such that $T_0 \geq T_1$. (Hint: combine equation (8.2) with the previous exercise.)

8.9 Toward the Fourier transform

The foregoing exercises suggest that an aperiodic function $x(t)$ satisfying condition (8.7) can be represented by a Fourier series, as follows:

$$x_{T_0}(t) = \frac{2}{T_0} \sum_{k=1}^{\infty} (X_{\cos}(k\omega_0) \cos k\omega_0 t + X_{\sin}(k\omega_0) \sin k\omega_0 t) \quad (8.8)$$

where $a_0 = (2/T_0) \int_{-T_1/2}^{+T_1/2} x(t) dt$ and care has been taken to choose $T_0 \geq T_1$.

Exercise 598 Explain why $x_{T_0}(t)$, defined by equation (8.8), is not quite the same as $x(t)$. (Hint: first explain why $x_{T_0}(t)$ is T_0 -periodic, and why $x(t)$ is not.)

Exercise 599 Can you see why we have $\lim_{T_0 \rightarrow \infty} x_{T_0}(t) = x(t)$?

Exercise 600 Show that

$$\lim_{T_0 \rightarrow \infty} x_{T_0}(t) = \lim_{T_0 \rightarrow \infty} \frac{1}{\pi} \sum_{k=1}^{\infty} (X_{\cos}(k\omega_0) \cos k\omega_0 t + X_{\sin}(k\omega_0) \sin k\omega_0 t) \omega_0 .$$

(Hint: show that the term with a_0 vanishes and work the ω_0 into the integrand.)

Exercise 601 Can you establish the following result?

$$x(t) = \frac{1}{\pi} \int_0^{\infty} (X_{\cos}(\omega) \cos \omega t + X_{\sin}(\omega) \sin \omega t) d\omega .$$

(Hint: combine the results of the previous two exercises. Let $\omega_0 = \Delta\omega$ and $\omega = k\Delta\omega$ and consider the sum as a Riemann sum.)

Exercise 602 Can you now establish the following key result?

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{i\omega t} d\omega \quad \text{where} \quad X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-i\omega t} dt . \quad (8.9)$$

(Hint: combine the results of the previous exercise with Euler's formulas $\cos \theta = (e^{\theta} + e^{-\theta})/2$ and $\sin \theta = (e^{\theta} - e^{-\theta})/(2i)$ where $i^2 = -1$.)

Exercise 603 In equation (8.9), do you have to assume that $x(t)$ satisfies the condition (8.7)? (Hint: you took the limit $T_0 \rightarrow \infty$.)

8.10 The Fourier transform and its spectra

The function $X(\omega)$ is the *Fourier transform* of $x(t)$; they are related to each other by (8.9). The function $|X(\omega)|$ is the *magnitude spectrum* of $x(t)$; the function $\arg X(\omega)$ the *phase spectrum* (for a complex number $z = Ae^{i\vartheta}$, $|z| = A$ and $\arg z = \vartheta$).

8.11 Signals

In instrumentation, t is time and $x(t)$ is interpreted as the *signal* on the *transmission line* between the probe (sensor) and the signal processing equipment. Thus the signal $x(t)$ often represents the voltage on a wire coming from the probe (or the light intensity in a fiber, or whatever physical encoding of the signal is most convenient).

In many instances, the phenomena underlying the signal $x(t)$ are much more apparent, and hence easier to tell apart, from the signal's Fourier transform $X(\omega)$ than from the signal itself. This accounts for the widespread use of Fourier transforms in instrumentation and signal processing.

The signal cannot be recorded in its entirety (this is true even of analog recording methods); the transmission line typically feeds into an accumulator which collects isolated values of the signal at regular intervals. This *sampling* process results in a sequence of *measurements* $x(kT_s)$ where $k \in \mathbb{Z}$ and T_s is the *measurement interval*, with associated *measurement frequency* $\omega_s = 2\pi/T_s$.

The basic sampling problem is to sample the signal sufficiently frequently to be able to reconstruct $X(\omega)$ and $x(t)$ from the sequence of measurements.

8.12 Sampling

The act of taking a measurement at time t_s is described by the function $\delta(t - t_s)$, which has the property that

$$\int_{-\infty}^t x(u)\delta(u - t_s)du = \begin{cases} 0 & \text{when } t < t_s \\ x(t_s) & \text{when } t \geq t_s \end{cases} \quad (8.10)$$

where this integral represents the state of the accumulator.

Exercise 604 Complete the following equation: $\int_{-\infty}^t \delta(u)du$. (Hint: comparison with equation (8.10) shows that $t_s = 0$, $x(t) = 1$.)

Exercise 605 Can you derive the Fourier series representation of the sampling process with sampling frequency ω_s represented by $\sum_{k=-\infty}^{+\infty} \delta(u - kT_s)$? (Hint: calculate the a s and b s, equations (8.2)—(8.4).)

Exercise 606 Can you rewrite your result of the previous exercise as follows?

$$\sum_{k=-\infty}^{+\infty} \delta(u - kT_s) = \frac{\omega_s}{2\pi} \sum_{k=-\infty}^{+\infty} e^{ik\omega_s t} \quad (8.11)$$

(Hint: remember, $\cos \vartheta = (e^{\vartheta} + e^{-\vartheta})/2$.)

8.13 The Fourier transform of the sampled process

The Fourier transform of the sampled process is defined to be

$$X_s(\omega) = \sum_{k=-\infty}^{+\infty} x(kT_s)e^{-i\omega kT_s} . \quad (8.12)$$

Exercise 607 Verify that $X_s(\omega)$ can be calculated from the sequence of measurements.

Exercise 608 Can you derive the following result? X_s is the Fourier transform of

$$x_s(t) = \sum_{k=-\infty}^{+\infty} x(t)\delta(t - kT_s)$$

Exercise 609 Can you derive the following key relationship between X_s and X ?

$$X_s(\tilde{\omega}) = \frac{1}{T_s} \int_{-\infty}^{+\infty} X(\omega) \delta_{\omega_s}(\tilde{\omega} - \omega) d\omega. \quad (8.13)$$

(Hint: combine equation (8.12) with the expression for $x(t)$ in (8.9); refer to exercise 606 to derive the formula $\delta_{\omega_s}(\omega) = \omega_s^{-1} \sum_{k=-\infty}^{+\infty} e^{ik2\pi\omega/\omega_s}$.)

Exercise 610 Can you explain, referring to equation (8.13), why $X_s(\omega)$ is not the same function as $X(\omega)$? (Hint: X_s is ω_s -periodic. Note that δ_{ω_s} ‘picks up’ values not just at $\tilde{\omega}$, but also at $\tilde{\omega} + k\omega_s$, where $k \in \mathbb{Z}$.)

8.14 Band-limited signals

The Fourier transform of a *band-limited* signal satisfies the following condition:

$$X(\omega) = 0 \quad \text{for } |\omega| > \omega_M \quad (8.14)$$

that is, ω_M is the highest frequency occurring in the signal $x(t)$.

Exercise 611 Can you show that for a band-limited signal, you have $X_s(\omega) = X(\omega)$ for all values of ω , *provided that* $\omega_s \geq 2\omega_M$? (Hint: verify that the integration limits $-\infty$ to $+\infty$ can be replaced by $-\omega_s/2$ to $+\omega_s/2$ in equation (8.13), and show that the problem you identified in exercise 610 is now avoided.)

Exercise 612 Conclude that for a band-limited signal with maximum frequency ω_M , the Fourier transform can be calculated from the measurements

$$X(\omega) = \sum_{k=-\infty}^{+\infty} x(kT_s) e^{-i\omega kT_s} \quad (8.15)$$

provided $\omega_s \geq 2\omega_M$. (Hint: combine equations (8.12) and the previous exercise.)

8.15 The Nyquist frequency

The sampling rate $2\omega_M$, where ω_M is the highest frequency in a band-limited signal, is known as the *Nyquist frequency*.

Exercise 613 Explain why the Nyquist frequency is the ideal sampling frequency. (Hint: consider the drawback of sampling at (much) lower and higher rates.)

Exercise 614 Can you derive the following formula? It expresses the original signal $x(t)$ in terms of only the measured values sampled at $t = kT_s$ ($k \in \mathbb{Z}$):

$$x(t) = \sum_{k=-\infty}^{+\infty} x(kT_s) \frac{\sin((t - kT_s)\omega_s/2)}{(t - kT_s)\omega_s/2} \quad (8.16)$$

where $\omega_s \geq 2\omega_M$. (Hint: combine equations (8.9) and (8.15), and use $\sin \vartheta = (e^{\vartheta} - e^{-\vartheta})/(2i)$.)

Principal Component Analysis and Clustering

9.1 A data set with a peculiar property

You consider the analysis of a data set consisting of N observations, where each observation consists of a (possibly very large) number n of observations. You wish to discover the structure of your data set. For simplicity we begin with $n = 2$; in a typical application n is much larger.

Exercise 615 Consider the following set of $N = 5$ observations:

$$\{(3, 2), (7, 4), (1, 1), (9, 5), (5, 3)\}.$$

Represent these data in a *scatterplot*, that is, as a set of points with coordinates $\{(x_i, y_i)\}_{i=1}^5$.

Exercise 616 For the data set of the previous exercise, calculate $\bar{x} = N^{-1} \sum_{i=1}^N x_i$ and $\bar{y} = N^{-1} \sum_{i=1}^N y_i$.

Exercise 617 Let $x^* = x_i - \bar{x}$ and $y^* = y_i - \bar{y}$. Draw a scatterplot of the data set $\{(x_i^*, y_i^*)\}_{i=1}^5$.

Exercise 618 Show that $\bar{x^*} = N^{-1} \sum_{i=1}^N x_i^* = 0$ and $\bar{y^*} = N^{-1} \sum_{i=1}^N y_i^* = 0$.

9.2 The mean-deviation form

The last two exercises show that a suitable translation can bring the data in a form where the coordinate averages are zero. This is called the *mean-deviation form*. Calculations involving variances are easier when the averages are zero; we shall always assume means-deviation form, and drop the superscript ‘ \star ’ to avoid unnecessarily cluttered notation.

Exercise 619 Consider the plot you drew for exercise 617. Can you match the following numbers to points in your plot?

$$1, \frac{1}{2}, -\frac{1}{2}, 0, -1.$$

9.3 Almost-linear data sets

The previous exercise suggests that data in a two-dimensional data set can be treated as “almost one-dimensional” when the data cloud is “elongated” as most of the variation occurs in one particular direction (the example of exercise 615 an extreme case where *all* variation within the data set is along a straight line).

Exercise 620 Consider the data transformation

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \quad (9.1)$$

Verify that this is a rotation.

Exercise 621 For the transformation given by equation (9.1), verify that

$$\bar{\tilde{x}} = N^{-1} \sum_{i=1}^N \tilde{x}_i = 0 \quad \text{and} \quad \bar{\tilde{y}} = N^{-1} \sum_{i=1}^N \tilde{y}_i = 0$$

(assuming means-deviation form, i.e. $\sum_{i=1}^N x_i = 0$ and $\sum_{i=1}^N y_i = 0$).

Exercise 622 For the transformation, equation (9.1), verify that the variance of \tilde{x}_i is given by $N - 1^{-1} \sum_{i=1}^N \tilde{x}_i^2$; write down similar formulæ for the variance of \tilde{y}_i and the covariance of \tilde{x}_i with \tilde{y}_i .

Exercise 623 In your scatterplot of exercise 617, draw in axes \tilde{x} and \tilde{y} so that the variance along the \tilde{x} -axis is maximal and the variance along the \tilde{y} -axis is minimal. Show that the rotation angle then satisfies $\tan \psi = \frac{1}{2}$; also show that $\overline{\tilde{x}\tilde{y}} = N - 1^{-1} \sum_{i=1}^N \tilde{x}_i\tilde{y}_i = 0$.

9.4 Mean-deviation in general

Let us apply this idea any 2-dimensional data cloud $\{(x_i, y_i)\}_{i=1}^N$. First, you make sure that the data are in mean-deviation form. Then you rotate the axis according to equation (9.1); the data cloud is described by new coordinates $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$. You now need to find the rotation angle that maximizes the variance $\overline{\tilde{x}^2}$ and minimizes $\overline{\tilde{y}^2}$.

Exercise 624 Confirm that you must solve the equation

$$\frac{d}{d\psi} \overline{\tilde{x}^2} = 0$$

for the angle ψ . Differentiate to obtain $\sum_{i=1}^N \tilde{x}_i\tilde{y}_i = 0$. Can you find the following formula?

$$\tan 2\psi = \frac{2\overline{xy}}{\overline{x^2} - \overline{y^2}}. \quad (9.2)$$

Exercise 625 For the special case where $y_i = kx_i$, $i = 1, \dots, N$, where k is a constant, show that the angle that maximizes $\overline{\tilde{x}^2}$ is given by $\tan \psi = k$. Can you verify that this agrees with equation (9.2)?

9.5 Information-rich and information-poor coordinates

The point of rotating the axes, as per equation (9.1), is that in the new coordinates (\tilde{x}, \tilde{y}) , the first coordinate \tilde{x} contains “more information” than the second coordinate \tilde{y} (in fact, the rotation angle is chosen so as to render the difference in informativeness is as pronounced as possible). Thus a minimum of information is lost if you only report the first coordinate (cf. exercise 619). Of course, this strategy of *data reduction* only really comes into its own when each observation consists of many, instead of just two, numbers. You can focus on the first few coordinates after a transformation that ensures that the first coordinate is “most informative”, the second coordinate “one-but-most informative” and so on. The transformation effectively lowers the dimensionality of your data, minimizing the loss of data involved in dropping the trailing dimensions.

Thus, the data set consists of N observations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ where each observation (*data point*) \mathbf{x}_i is an n -dimensional vector (in the foregoing you had $n = 2$). The transformation is $\tilde{\mathbf{x}} = \mathbf{A} \cdot \mathbf{x}$ where \mathbf{A} is an $n \times n$ matrix. The objective is to find the \mathbf{A} that redistributes variances along the (transformed) coordinate axes in the most extreme way possible.

Exercise 626 Write down \mathbf{A} for $n = 2$ and verify that it satisfies $\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I}$ where \mathbf{I} is the $n \times n$ identity matrix. (Hint: see equation (9.1); the superscript ‘T’ indicates matrix transposition.)

9.6 A useful property

The condition $\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I}$ generally holds for a rotation in n dimensions.

Exercise 627 Can you explain why this is so?

9.7 Generalizing to higher dimensions

In exercise 623 you showed that the covariance is zero for the desired transformation. This is our clue for the generalization.

Exercise 628 Let the data points \mathbf{x}_i be the columns of an $n \times N$ matrix which you call the *data matrix* \mathbf{X} . Consider the *covariance matrix*

$$\mathbf{Q} = \mathbf{X} \cdot \mathbf{X}^T \quad (9.3)$$

and verify that (i) $Q_{ij} = Q_{ji}$ for all pairs (i, j) ; (ii) Q_{ij} is proportional to the covariance of the i th element of the data points with the j th element of the data points; (iii) Q_{ii} is proportional to the variance of the i th element of the data points.

Exercise 629 Verify that the transformed data matrix is $\mathbf{A} \cdot \mathbf{X}$ and show that the transformed covariance matrix is given by

$$(\mathbf{A} \cdot \mathbf{X}) \cdot (\mathbf{A} \cdot \mathbf{X})^T = \mathbf{A} \cdot \mathbf{Q} \cdot \mathbf{A}^T .$$

(Hint: compare the left term to equation (9.3), and use matrix algebra rules to establish the equality.)

Exercise 630 Verify that the condition $\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I}$ implies $\mathbf{A}^T = \mathbf{A}^{-1}$.

Exercise 631 Verify that the condition of zero covariances in the transformed coordinates leads to the equation

$$\mathbf{A} \cdot \mathbf{Q} \cdot \mathbf{A}^{-1} = \mathbf{D} \quad (9.4)$$

where \mathbf{D} is a diagonal matrix.

Exercise 632 Verify that equation (9.4) leads to the following *diagonalization* of the covariance matrix \mathbf{Q} :

$$\mathbf{Q} = \mathbf{A}^{-1} \cdot \mathbf{D} \cdot \mathbf{A} . \quad (9.5)$$

Exercise 633 Verify that, if there exists a matrix \mathbf{D} such that equation (9.4) is satisfied, it follows that $\mathbf{Q} = \mathbf{Q}^T$. (Hint: use equation (9.5).)

9.8 Diagonalization

Linear algebra tells us that there exists a matrix \mathbf{D} such that equation (9.4) is satisfied iff $\mathbf{Q} = \mathbf{Q}^T$. You already established this last fact in exercise 628, since $\mathbf{Q} = \mathbf{Q}^T$ is just another way of saying that $Q_{ij} = Q_{ji}$ for all pairs (i, j) .

Exercise 634 Can you show that the following is a solution of the diagonalization equation (9.5)? Let \mathbf{D} be a diagonal matrix with the eigenvalues $\lambda_1, \dots, \lambda_n$ of the covariance matrix \mathbf{Q} on the diagonal, so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, and let the rows of \mathbf{A} be the corresponding unit eigenvectors. (Hint: if \mathbf{v} is an eigenvector, establish that $\mathbf{Q} \cdot \mathbf{v} = \lambda \mathbf{v}$ where λ is the corresponding eigenvalue; first show that $\mathbf{A} \cdot \mathbf{v}$ is a standard basis vector; then check what happens when this is premultiplied with \mathbf{D} .)

9.9 Principal components

The unit eigenvectors of the covariance matrix \mathbf{Q} are called the *principal components* of the data. The one corresponding to the largest eigenvalue is the *first principal component*; the *second principal component* is the unit eigenvector corresponding to the second largest eigenvalue of \mathbf{Q} , and so on.

Exercise 635 Now that you have characterized the matrix \mathbf{A} , consider again the transformation $\tilde{\mathbf{x}}_i = \mathbf{A} \cdot \mathbf{x}_i$. Verify that

$$\tilde{x}_{1,i} = c_1 x_{1,i} + c_2 x_{2,i} + \cdots + c_n x_{n,i}$$

where c_1, \dots, c_n are the elements of the first principal component.

Exercise 636 Consider once more the special case where $n = 2$ and $x_{2,i} = kx_{1,i}$. Show that the covariance matrix is given by the following expression:

$$\mathbf{Q} = \begin{bmatrix} 1 & k \\ k & k^2 \end{bmatrix} \sum_{i=1}^N x_{1,i}^2.$$

Exercise 637 Continuing the previous exercise, show that the eigenvalues of the covariance matrix are $\lambda = (1+k) \sum_{i=1}^N x_{1,i}^2$ and $\lambda = 0$, with corresponding unit eigenvectors

$$\begin{bmatrix} 1/(1+k^2) \\ k/(1+k^2) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} k^2/(1+k^2) \\ -k/(1+k^2) \end{bmatrix}.$$

Hence deduce that $\tilde{\mathbf{x}}_i = [x_{1,i} \ 0]^T$, confirming your earlier finding that in this special case, all information is contained in the first transformed coordinate.

9.10 The scree plot

Generally, there will be some (but less and less) information contained in subsequent transformed coordinates. We know that the transformed covariance matrix $\mathbf{A} \cdot \mathbf{Q} \cdot \mathbf{A}^{-1}$ is diagonal (equation (9.4)). Taking the trace¹ and dividing each element of the transformed covariance matrix by this trace, we obtain a series of fractions, adding up to one, which indicate how much of the variance in the data is captured by each of the components. A bar plot of these fractions (sorted by decreasing size) is called a *scree plot*.

Exercise 638 Suggest a procedure to decide on the number of dimensions (i.e. trailing coordinates in the transformed frame) that can be dropped.

Exercise 639 Suppose that mRNA expression levels are measured for a very large N_G number of genes, in cells under a variety of conditions. Suppose, furthermore, that there are $N_M \ll N_G$ “master genes” which respond independently to the conditions to which the cells are subjected, while the remaining $N_G - N_M$ genes are slaves whose expression levels are expressible as a weighted sum of the “master gene” expression levels (each “slave” having a unique set of N_G weighting coefficients). Discuss how principal component analysis could be used to identify the number of master genes.

Exercise 640 Continuing the previous exercise, discuss the efficacy of principal components when the “slave genes” depend on the masters according to non-linear functions.

Exercise 641 Suppose that the data cloud consists of a number of clusters of data, without any correlation within these clusters. Discuss what the principal components of such a data cloud will (or will not) tell you about the structure of the data.

¹The *trace* of a matrix is the sum of the diagonal elements.

9.11 Maximum Likelihood clustering

When the n -dimensional data cloud is thought to consist of a number of data clusters, you can use maximum likelihood estimation² to determine the location and scale parameters of these clusters.

Exercise 642 Consider the case $n = 1$, with $m = 2$ clouds, and assume that these two clusters are normally distributed. Motivate the following log-likelihood function, and interpret the parameters p_1 , μ_1 , μ_2 , σ_1^2 , and σ_2^2 .

$$\ln \mathbb{L} = \sum_{i=1}^N \ln \left\{ \frac{p_1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right\} + \frac{1 - p_1}{\sigma_2 \sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right\} \right\} \quad (9.6)$$

Exercise 643 The ML estimators of the five parameters in the previous exercise are found by simultaneously solving the following system:

$$\frac{\partial \ln \mathbb{L}}{\partial p_1} = 0 \quad \frac{\partial \ln \mathbb{L}}{\partial \mu_1} = 0 \quad \frac{\partial \ln \mathbb{L}}{\partial \mu_2} = 0 \quad \frac{\partial \ln \mathbb{L}}{\partial \sigma_1} = 0 \quad \frac{\partial \ln \mathbb{L}}{\partial \sigma_2} = 0.$$

Explicit work out the first three of these.

9.12 Obtaining ML estimators

The last exercise shows that the ML estimators have to be obtained numerically.

Exercise 644 How many parameters do you have to estimate in the case $n = 1$, $m = 3$? How many for general m ?

Exercise 645 Can you show how the ML estimate of m is found?

Exercise 646 Can you generalize to $n \geq 2$? (Hint: consider whether you want to incorporate a correlation structure in each of your clusters.)

Exercise 647 The use of the normal distribution in the foregoing tacitly assumes that the data are in \mathbb{R}^n . How would you adapt the procedure if the data “live” in \mathbb{Z}^n , or \mathbb{N}^n , or $[0, 1]^n$?

9.13 Assigning data points to clusters

Having found a number of clusters, plus location and scale parameters for each of these, you will want to *assign* individual data points to the clusters.

Exercise 648 Returning to the case $n = 1$, $m = 2$, consider the likelihood that data point x_i belongs to the first cluster:

$$\frac{1}{\hat{\sigma}_1 \sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2} \right\}$$

where hats on top of symbols indicate ML estimates. Write down a similar formula for the likelihood that data point x_i belongs to the second cluster. When would you assign x_i to cluster 1, and when to cluster 2?

Exercise 649 In the previous exercise, \hat{p}_1 did not figure. Motivate this choice.

²Even though this is a straightforward & obvious application of the ML principle, the clustering literature calls this the *expectation maximization* method.

9.14 k -Means clustering

An alternative procedure is to assign the N data points at random to the two clusters, and then choose points, one at a time and at random, and move them to the other cluster if this decreases the total variability within each cluster and/or increases the variability between clusters.

Exercise 650 Motivate the following formulæ ($n = 1, m = 2$):

$$\begin{aligned} \text{variability within clusters: } & \sum_{i \in C_1} (x_i - \bar{x}_{C_1})^2 + \sum_{i \in C_2} (x_i - \bar{x}_{C_2})^2 \\ \text{variability between clusters: } & (\bar{x}_{C_1} - \bar{x})^2 + (\bar{x}_{C_2} - \bar{x})^2 \end{aligned} \quad (9.7)$$

where $\bar{x}_{C_k} = (\sum_{i \in C_k} x_i) / (|C_k|)$, $k = 1, 2$ and \bar{x} is the overall (grand) mean of the data. Discuss generalizations to higher n and m .

Exercise 651 Do you expect ML clustering and k -means clustering to give similar results?

9.15 Hierarchical clustering

It is sometimes desirable to collect clusters into bigger clusters, and to collect these in turn into bigger clusters, and so on.

Exercise 652 Can you extend ML clustering to hierarchical clustering? (Hint: say you have obtained \hat{m} clusters; replace each of those clusters with a data point that represents the average of the cluster's members: you thus form a new "higher-order" data cloud.)

9.16 Distance-based clustering

Another approach to hierarchical clustering is to group data points based on some distance measure.

Exercise 653 Review the properties of the p -norm distance for various values of p :

$$\|\mathbf{x}_i - \mathbf{x}_j\| = (\sum_{\ell=1}^n |x_{i\ell} - x_{j\ell}|^p)^{1/p} . \quad (9.8)$$

Can you explain why we require $p \geq 1$ for this norm to work?

Exercise 654 Suppose the objects are strings of symbols³. Can you devise a norm?

Exercise 655 Review the *Hamming distance*; can you show that it satisfies the *triangle inequality*?

9.17 Amalgamation

The procedure is to *amalgamate* objects, two at a time, in a set of objects \mathcal{O} . The initial objects can be the data points, or the cluster averages obtained through ML or k -means clustering. When two objects are amalgamated, they are removed from \mathcal{O} and replaced by a *single* object representing the pair.

Exercise 656 Verify that $|\mathcal{O}|$ is reduced by 1 by an amalgamation. How many amalgamation steps can be done when the initial object set has N elements?

Exercise 657 Observe that an amalgamation need not involve objects from the initial set; one or both of the amalgamated objects may be the result of an earlier amalgamation step. Hence motivate the statement that hierarchical clustering forms a *nested classification*.

³These symbols might well be integers, as long as you keep in mind that they are on a nominal scale of measurement, that is, arithmetic operations are meaningless.

9.18 Distances

Associated with \mathcal{O} is a set of distances assigned to any pair of distinct elements of \mathcal{O} .

Exercise 658 Show that there are $|\mathcal{O}|(|\mathcal{O}| - 1)/2$ distances between all pairs.

9.19 The amalgamation rule

The *amalgamation rule* is simple: amalgamate the pair of objects that have the smallest distance. The distances of all other objects to the members of the amalgamated pair are deleted, and new distances are assigned between the new object (representing the amalgamated pair) and the other objects.

Exercise 659 Discuss various options to assign these new distances: (i) the amalgamated pair is assigned the average coordinates of its members, from which distances are calculated; (ii) the distance between the amalgamated pair and an object \mathbf{x} is the smallest of the two distances which the members of the pair had to \mathbf{x} (“nearest neighbour⁴”); (iii) the distance between the amalgamated pair and an object \mathbf{x} is the largest of the two distances which the members of the pair had to \mathbf{x} (“furthest neighbour⁵”). Can you think of other possibilities?

Exercise 660 Discuss how your nested classification assigns the objects in the initial object set to the terminal vertices of a *tree*⁶. Discuss applications such as *phylogenetic trees*.

9.20 Validation

You will usually require some indication of the reliability of the results of your data cloud analysis. Ideally, you would like to obtain the statistics of an ensemble of data clouds, of which the particular data set that you have analysed is just an instantiation. Unfortunately, such an ensemble exists only as a theoretical idealization. However, you can do the “next best thing⁷” which is to obtain “virtual instantiations” of the ensemble by *resampling* the original data set.

Exercise 661 There are various methods of obtaining virtual instantiations of the cloud ensemble. In *bootstrapping* you obtain samples of the same original data set size N by sampling at random from the original (given) data set, but *with replacement*. In *cross-validation*, you assign elements of the data set at random to a number ($N_V \ll N$) of mutually exclusive sub-sets. Contrast and compare these two methods in terms of their advantages and disadvantages.

Exercise 662 How can you ensure, when you use cross-validation, that you obtain V data sets of equal size?

Exercise 663 Describe in detail how you would obtain descriptive statistics for principal component analysis. (Hint: (i) agreement of principal component axes: consider the average dot product with the sample average; (ii) agreement of transform: consider the *coefficient of variation*⁸ of each transformed data point over the simulated sample.)

Exercise 664 How would you quantify agreement in a simulated sample of clusters? (Hint: consider pairwise agreement of being assigned to the same cluster.)

Exercise 665 How would you quantify agreement in a simulated sample of trees? (Hint: consider path length between two terminal objects.)

⁴Or *single linkage*; this is one of those fields where needless jargon proliferates.

⁵Also known as *complete linkage*.

⁶A tree is a graph in which any two vertices are connected by a unique succession of distinct edges.

⁷Also known as the *poor man's option*.

⁸The coefficient of variation (c.v.) is defined as the standard deviation divided by the average.

9.21 Maximum Likelihood clustering

For Maximum Likelihood clustering, you can do the following instead of simulated sampling: if ϑ_i is one of the parameters in the parameter vector ϑ , and $\hat{\vartheta}$ is the ML estimate, then an index of sensitivity is given by the expression

$$\vartheta_i^2 \frac{\partial^2 \ln \mathbb{L}}{\partial \vartheta_i^2} \Big|_{\vartheta = \hat{\vartheta}} .$$

Exercise 666 Explain why the above formula is a suitable measure of reliability, and discuss how you would (numerically) evaluate it.