MATHEMATICS IN BIOLOGY

Markus Meister, Kyu Hyun Lee, and Ruben Portugues

The MIT Press
Cambridge, Massachusetts
London, England

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Meister, Markus, author. | Lee, Kyu Hyun, 1946- author. | Portugues, Ruben, author.
Title: Mathematics in biology / Markus Meister, Kyu Hyun Lee, and Ruben Portugues.
Description: Cambridge, Massachusetts : The MIT Press, [2025] | Includes bibliographical references and index.
Identifiers: LCCN 2024017280 (print) | LCCN 2024017281 (ebook) | ISBN 9780262049405 (hardcover) | ISBN 9780262380829 (pdf) | ISBN 9780262380836 (epub)
Subjects: LCSH: Biomathematics. | Biology—Mathematical models.
Classification: LCC QH323.5 .M45 2025 (print) | LCC QH323.5 (ebook) | DDC 570.1/51—dc23/eng/20240723
LC record available at https://lccn.loc.gov/2024017280
LC ebook record available at https://lccn.loc.gov/2024017281

10 9 8 7 6 5 4 3 2 1

Contents

8 Advanced Topics in Probability and Statistics

8.1 Random Walks and Diffusion

Many processes in biology are driven at their core by random events. On the smallest scales, thermal fluctuations play an essential role: the assembly and disassembly of a protein polymer, the random stepping of a molecular motor, the thermal opening and closing of an ion channel, the random meandering of a signaling molecule through the cytoplasm. On a large scale, the dynamics of a population are governed by births and deaths among its many individuals, events that are sufficiently unpredictable to be treated as random variables.

When a variable executes many independent random steps in sequence it leads to a **random walk**. A canonical example is the position of a small particle, like a calcium ion, buffeted by thermal collisions with molecules of the surrounding fluid, as shown in figure 8.1. If we zoom out from this molecular picture to consider many such random variables collectively, such as the concentration of all calcium ion in a cell, then we observe a process of mass transport called **diffusion**. This chapter will elaborate on the dynamics of random walks and diffusion.

8.1.1 Brownian Motion

The earliest published account of random thermal motion comes from Robert Brown, a biologist interested in the process of pollination (Brown (1828)). While inspecting pollen grains suspended in water using a simple microscope, he "observed many of them very evidently in motion." The motions "arose neither from currents in the fluid, nor from its gradual evaporation, but belonged to the particle itself." Brown at first suspected the particles to be "animated," but soon confirmed that perfectly inorganic substances, when ground into a dust, produced the same type of motion. Physicists largely ignored these phenomena of "Brownian motion" until the early twentieth century, when Einstein showed how they accord with predictions from the broader framework of kinetic theory (Einstein (1905), Brush (1968)).

Suppose that we could track a molecule of oxygen suspended in water and let us just follow its movements along the x-direction. The molecule's kinetic energy is $kT/2$, where T is the temperature in degrees Kelvin and $k = 1.38 \times 10^{-23}$ J/K is Boltzman's constant, so it flies along at about 100 m/s. However, it doesn't get very far: every 10^{-13} s or so, it bangs into a water molecule that changes its speed and direction. In fact, the mean free path during which it flies straight is only 10^{-11} m, about 1/10 the size of a hydrogen atom. All this is to say that the individual step of such a Brownian particle is so small in size and duration that for all practical purposes, we will only ever have to worry about the accumulated effect of many steps.

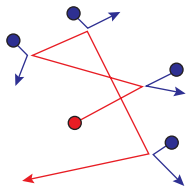

Figure 8.1
A Brownian particle (red) buffeted by collisions with molecules of the surrounding medium (blue).

8.1.2 Random Walk in One Dimension

The simplest mathematical approximation of Brownian motion is a discrete random walk (figure 8.2). Consider a particle moving in one dimension with discrete steps. The particle starts at $x = 0$. At every time step, it moves either right to $x + 1$, with probability p, or left to $x - 1$, with probability $q = 1 - p$.[1] So if

$$x_n = \text{ position of the particle after } n \text{ steps,} \tag{8.1}$$

then

$$x_{n+1} = \begin{cases} x_n + 1 \text{ with probability } p \\ x_n - 1 \text{ with probability } q = 1 - p. \end{cases} \tag{8.2}$$

What is the probability distribution $P(x_n)$ for the position of the particle after n time steps? Suppose that the n steps included m steps right and $n - m$ steps left. Then m follows the binomial distribution given in equation (6.20):

$$m \sim \text{Bin}(n, p). \tag{8.3}$$

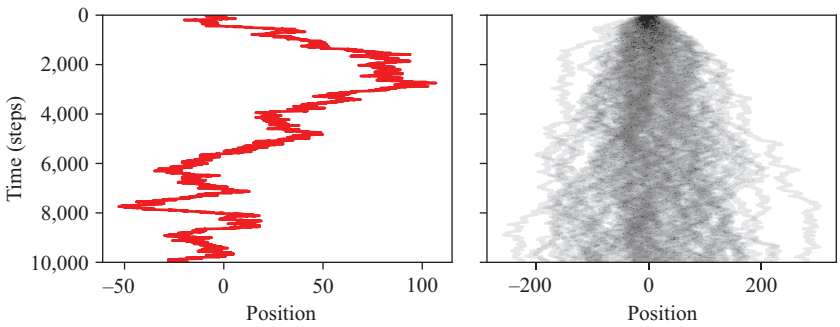

Figure 8.2
Left: Position as a function of time for a particle that performs an unbiased random walk moving right or left at each time step with equal probability. Right: 100 such random walks superposed.

1. For a Brownian particle, steps to the left and right are equally probable, but with a little extra effort, we may as well consider this more general case, where $p \neq q$.

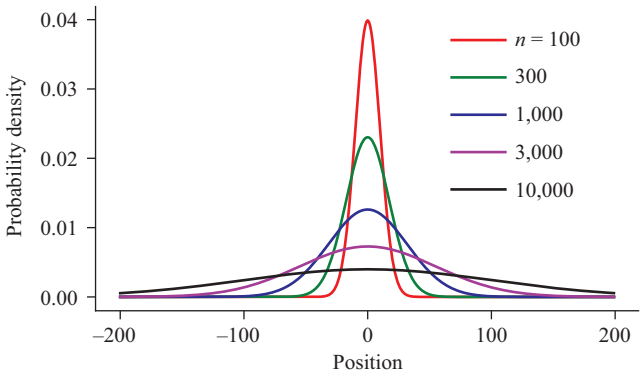

Figure 8.3
Probability density of a random walker with $p = 1/2$ after a large number of n steps.

The corresponding position of the particle is

$$x_n = m - (n - m) = 2m - n. \tag{8.4}$$

So the probability of being at position x after n steps is

$$P(x; n) = \binom{n}{\frac{n+x}{2}} p^{\frac{n+x}{2}} q^{\frac{n-x}{2}}. \tag{8.5}$$

This distribution is shown in figure 8.3. It has a mean μ and variance σ^2, given by

$$\mu = n(p - q), \qquad \sigma^2 = 4npq. \tag{8.6}$$

For large n, we can invoke the central limit theorem to state that

$$P(x; n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{8.7}$$

So after many steps, the random walker has a bell-shaped probability distribution that follows a Gaussian profile. The width of that Gaussian grows as the square root of the number of steps.

8.1.3 The Diffusion Coefficient

Returning to the real world, what can we conclude about the motion of a Brownian particle? The number of collisions that it undergoes is huge, but strictly proportional to time. After a few nanoseconds, there have been so many collisions that the central limit theorem kicks in. So we can immediately conclude that the particle's position has a Gaussian distribution whose width σ grows proportionally to the square root of time:

$$\sigma = \sqrt{2Dt}. \tag{8.8}$$

The proportionality constant D is called the **diffusion coefficient**. It completely characterizes the particle's behavior under thermal motion.

Table 8.1
Approximate distance versus time for a small molecule diffusing in water

Time	Distance
1 ms	1 μm
100 ms	10 μm
10 s	100 μm
1,000 s	1 mm
1 day	10 mm

For biological applications, it is useful to remember a couple of order-of-magnitude numbers:

- For a small molecule (molecular weight up to a few hundred) in water, $D \approx 10^{-5}\text{cm}^2/\text{s}$
- For a protein moving laterally in a cell membrane, $D \approx 10^{-9}\text{cm}^2/\text{s}$

Note the physical dimensions of the diffusion coefficient: distance^2/time. Clearly, this is not a velocity! The typical distance traveled via diffusion is proportional *not* to time, but to the square root of time. Table 8.1 lists those distances for a small molecule in water.

So a small signaling molecule can equilibrate across a typical cell body in 0.1 s, but if it needs to get 1 cm down the axon of a neuron that would take forever. Clearly, thermal transport is not sufficient there.

8.1.4 Fick's Laws of Diffusion

Let us now imagine a very large number of molecules, all executing Brownian motion independently of each other. Again, we will model this as a discrete random walk along the x-axis. Suppose that after j time steps, there are $N_{i,j}$ particles located at position i (figure 8.4):

$$N_{i,j} = \text{number of particles at position } i \text{ after step } j. \tag{8.9}$$

In the next step, half the particles at location i step to the right and the other half to the left. So the net number of particles moving across the border from i to $i+1$ is:

$$\begin{aligned} M_{i,j} &= \text{number of particles moving from position } i \text{ to } (i+1) \text{ during step } (j+1) \\ &= \frac{1}{2}\left(N_{i,j} - N_{i+1,j}\right). \end{aligned} \tag{8.10}$$

So the new number of particles becomes

$$N_{i,j+1} = N_{i,j} + M_{i-1,j} - M_{i,j}. \tag{8.11}$$

To connect to real-world units, we define

$$\begin{aligned} \Delta x &= \text{size of a step along the } x\text{-axis} \\ \Delta t &= \text{duration of a step.} \end{aligned} \tag{8.12}$$

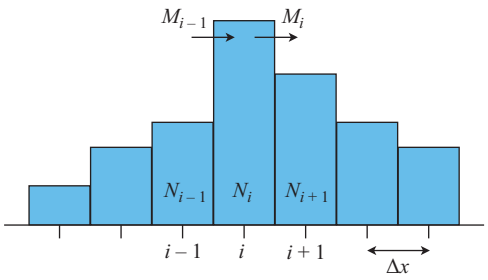

Figure 8.4
A distribution of particles undergoing independent random walks.

Then we take the continuum limit by allowing $\Delta x \to 0$ and $\Delta t \to 0$ while keeping the diffusion coefficient constant $\frac{1}{2}\frac{(\Delta x)^2}{\Delta t} = D$. In that limit,

$$\frac{N_{i,j}}{\Delta x} \to C(x,t) = \text{concentration of particles at position } x = i\Delta x \text{ and time } t = j\Delta t \quad (8.13)$$

and

$$\frac{M_{i,j}}{\Delta t} \to J(x,t) = \text{flux of particles at position } x = i\Delta x \text{ and time } t = j\Delta t. \quad (8.14)$$

Equation (8.10), after dividing by Δt, is

$$\frac{M_{i,j}}{\Delta t} = \frac{1}{2}\frac{(\Delta x)^2}{\Delta t}\frac{\left(N_{i,j} - N_{i+1,j}\right)}{(\Delta x)^2}, \quad (8.15)$$

which becomes, in the continuum limit

$$J(x,t) = -D\frac{\partial C(x,t)}{\partial x}. \quad (8.16)$$

Similarly, equation (8.11), after dividing by Δt and Δx, becomes in the continuum limit

$$\frac{\partial C(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}. \quad (8.17)$$

Equations (8.16) and (8.17) are called **Fick's laws of diffusion**. They relate the local flux of particles to the concentration. These two partial differential equations can be combined to produce the **diffusion equation**:

$$\frac{\partial}{\partial t}C(x,t) = D\frac{\partial^2}{\partial x^2}C(x,t). \quad (8.18)$$

8.1.5 Qualitative Behavior of the Diffusion Equation

Qualitatively, diffusion acts so as to "smooth" the concentration profile over time, as illustrated in figure 8.5. Around a peak in the profile of $C(x,t)$, the second spatial derivative is negative, so according to equation (8.18), the concentration here will decrease. The simple reason is that this region is flanked on both sides by an outward-sloping

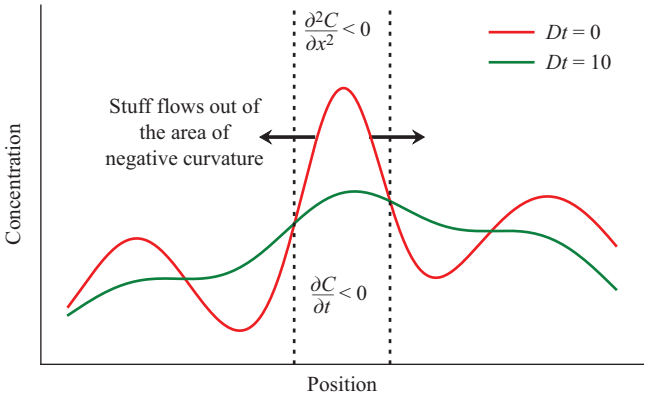

Figure 8.5
The evolution of a concentration profile under diffusion.

concentration gradient, which leads to particles flowing out of that region. On the other hand, at a local minimum, the second spatial derivative is positive, there is an inward sloping gradient on both sides, so this concentration will increase. The net effect is that **peaks of concentration get flattened and valleys get filled in**. The final state at long times tends to have no peaks or valleys, unless some special boundary conditions apply.

8.1.6 Random Walks and Diffusion in Higher Dimensions

Many biological motions take place in two or three dimensions. One can model three-dimensional (3D) Brownian motion as a random walk on a 3D coordinate grid, with the particle taking steps to a neighboring grid point simultaneously in all three directions. So during one step, x, y, and z all change by ± 1. Because the three random walks take place independently, we can consider each coordinate on its own, each of which behaves just like the one-dimensional (1D) case discussed in section 8.1.4.

Back in the real world, if a particle with diffusion coefficient D starts at the origin, then after time t, all three position variables will be distributed like a Gaussian with variance $2Dt$:

$$\begin{aligned} P_x(x) &= \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}} \\ P_y(y) &= \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{y^2}{4Dt}} \\ P_z(z) &= \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{z^2}{4Dt}}. \end{aligned} \tag{8.19}$$

The Euclidean distance from the starting point is $r = \sqrt{x^2 + y^2 + z^2}$ and its variance is

$$\left\langle r^2 \right\rangle = \left\langle x^2 + y^2 + z^2 \right\rangle = 6Dt. \tag{8.20}$$

Fick's laws relate the flux of particles to the concentration. The 3D versions are

$$\begin{aligned} \mathbf{J} &= -D\nabla C(\mathbf{r}, t) \\ \frac{\partial}{\partial t} C(\mathbf{r}, t) &= -\nabla \cdot \mathbf{J}. \end{aligned} \tag{8.21}$$

Here, $\mathbf{r} = (x, y, z)$ refers to the position in three dimensions; C is the concentration and $\mathbf{J}$ is the flux of particles. The term

$$\nabla C = \left(\frac{\partial C}{\partial x}, \frac{\partial C}{\partial y}, \frac{\partial C}{\partial z} \right) \tag{8.22}$$

is the gradient (i.e. multidimensional derivative) of the concentration and

$$\nabla \cdot \mathbf{J} = \frac{\partial J}{\partial x} + \frac{\partial J}{\partial y} + \frac{\partial J}{\partial z} \tag{8.23}$$

is the divergence of the flux field (note that this is *not* the same as the gradient; it is the dot product of the gradient operator with the flux $\mathbf{J}$).

Again, one can combine the two laws into one diffusion equation:

$$\begin{aligned} \frac{\partial}{\partial t} C(\mathbf{r}, t) &= D \nabla \cdot \nabla C(\mathbf{r}, t) \\ &= D \nabla^2 C(\mathbf{r}, t), \end{aligned} \tag{8.24}$$

where $\nabla \cdot \nabla \equiv \nabla^2 \equiv \Delta$ is the **Laplacian operator**. In 3D Cartesian coordinates, this is simply

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{8.25}$$

Depending on the spatial symmetries of the problem at hand, it can be more convenient to work in a different coordinate system, and some caution is required around differential operators. For example, in the 3D spherical coordinate system (section 1.5.1) with coordinates (r, θ, ϕ), the Laplacian is

$$\nabla^2 \equiv \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2}{\partial \phi^2}. \tag{8.26}$$

8.1.7 Solving the Diffusion Equation

In addition to the transport of particles by Brownian motion, the diffusion equation covers other phenomena of mass transport, like the conduction of heat, or the movement of electric charge in an electrolyte. In a typical problem, one is given an **initial condition** of the profile $C(x, t = 0)$ and wants to know the future concentration profile $C(x, t)$. Beside the initial condition, one has to also deal with **boundary conditions** that specify what happens at the edges of the space or other special locations. In some cases, one is mostly interested in the final **steady-state solution** at very long times $C(x, t = \infty)$. Here, we touch on some of these methods for solving the diffusion equation. These may help you devise at least an approximate solution to your problem. For tough problems, one can always resort to lookup via the Google search bar. Back when people read books, a classic collection of solutions could be found in Carslaw and Jaeger (1986).

8.1.7.1 Superposition The diffusion equation falls in the class of **linear partial differential equations**. This simply means that the function of interest $C(x, t)$ and its derivatives appear only with a power of 1. As a consequence, the solutions of the differential equation obey the **superposition principle**: If two functions $C_1(x, t)$ and

$C_2(x,t)$ are both solutions to the diffusion equation, then any linear combination $\lambda_1 C_1(x,t) + \lambda_2 C_2(x,t)$ is also a solution.

We encountered this superposition idea before in the more general treatment of linear systems in section 3.1.2. Here again, we will see that it has powerful consequences.

A simple way to understand superposition is to remember that at time $t=0$, the initial concentration profile $C(x,0)$ is made of many independent particles. Each of these executes a random walk, independent of the others. If we arbitrarily divide those particles into a red group $C_1(x,0)$ and a blue group $C_2(x,0)$, they will still produce the exact same profile $C(x,t)$ later on. Note that this argument relies on there being no interaction between the particles. If they do interfere with each other, the differential equation will not be linear, and superposition no longer applies.

8.1.7.2 Green's function This argument leads to another conclusion: We can solve for the future profile $C(x,t)$ if we simply know what happens to the probability density of each individual particle over time. After all, the full profile is simply the sum of the individual particle densities.

So let us suppose that the probability density of a particle that starts at location x' develops according to

$$G(x;x',t) = \text{probability that a particle is at location } x \text{ at time } t \text{ if it started from } x' \text{ at time 0.} \tag{8.27}$$

Technically, this is called the **Green's function** of the diffusion problem. Obviously, at time $t=0$, the particle is certain to be at $x=x'$, so

$$G\left(x;x',t=0\right) = \delta\left(x-x'\right), \tag{8.28}$$

where $\delta(x)$ is the delta function. We can write the initial profile as a sum over these delta functions:

$$C\left(x,0\right) = \int_{x'} C\left(x',0\right) G\left(x;x',0\right) \mathrm{d}x'. \tag{8.29}$$

Then we allow each of the particles to evolve according to its Green's function and sum again to get the solution:

$$C\left(x,t\right) = \int_{x'} C\left(x',0\right) G\left(x;x',t\right) \mathrm{d}x'. \tag{8.30}$$

A simple example is diffusion in free space. Suppose that there are no boundaries anywhere, so the entire x-axis is available. Then we already know the Green's function: A particle will diffuse according to the spreading Gaussian of equation (8.7):

$$G\left(x;x',t\right) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{\left(x-x'\right)^2}{4Dt}}. \tag{8.31}$$

All places along the x-axis behave the same way, so this same Green's function applies no matter where the particle starts. Therefore, the solution with initial condition $C(x,0)$

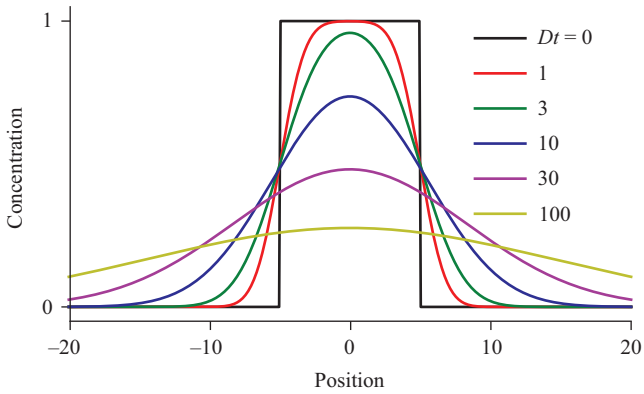

Figure 8.6
Diffusion from an initial square bolus of particles.

is simply

$$C(x,t) = \int_{x'} C(x',0)\,\frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-x')^2}{4Dt}}\,\mathrm{d}x'. \tag{8.32}$$

Figure 8.6 shows the time-dependent solution when the initial condition is a bolus of particles with a square concentration profile.

8.1.7.3 Boundary conditions At boundaries in the space, one generally considers two kinds of conditions:

- Reflecting boundary: Particles bounce off this surface. That means that there can be no flux of particles into or out of the surface. So the boundary condition is that everywhere on the surface,

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n} = 0 \tag{8.33}$$

  where $\mathbf{n}$ is the normal vector to the surface.
- Absorbing boundary: Particles get swallowed by this surface, never to appear again. That means the concentration of particles is zero everywhere on the surface:

$$C(\mathbf{r},t) = 0. \tag{8.34}$$

Some simple boundary problems can be solved with so-called **mirror sources**. For example, suppose that a particle diffuses in one dimension, starting at $x = a$, but there is a reflective boundary at $x = 0$, so its motion is constrained to the right half of the x-axis only. We can imagine instead that there is no boundary at all, but a second particle starts out at $x = -a$, in the mirror-reflected position of the true particle (figure 8.7). For every time that the true particle random-walks through the boundary to $x < 0$, the mirror particle random-walks out of the boundary in the opposite direction.

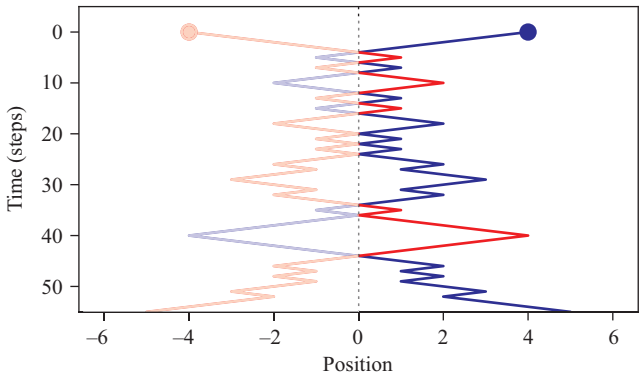

Figure 8.7
Mirror sources: To simulate the random walk of a particle with a reflecting boundary at $x = 0$, we imagine two mirror particles (blue and red) executing mirror-symmetric walks, but without a barrier. Whenever the red particle crosses into the right half, it looks like the blue particle bounced off the barrier.

So in the right half of the space, we can simply add the density of the true and the virtual particles to get the solution:

$$C_{\text{ref}}(x,t) = \frac{1}{\sqrt{4\pi Dt}}\left(e^{-\frac{(x-a)^2}{4Dt}} + e^{-\frac{(x+a)^2}{4Dt}}\right). \tag{8.35}$$

This perfectly emulates the reflection of a particle at a reflecting boundary.

Similarly, for an absorbing boundary, we add an antiparticle in the mirror position (i.e., one with "negative probability"). At the surface, the densities of the two particles precisely cancel each other out, thus enforcing the condition $C(x,t) = 0$:

$$C_{\text{abs}}(x,t) = \frac{1}{\sqrt{4\pi Dt}}\left(e^{-\frac{(x-a)^2}{4Dt}} - e^{-\frac{(x+a)^2}{4Dt}}\right). \tag{8.36}$$

8.1.8 Steady-State Solutions

After a long time t, diffusion systems typically settle into a steady state where nothing changes anymore. Based on equation (8.24), that means

$$\nabla^2 C(\mathbf{r}) = 0. \tag{8.37}$$

The solutions to this equation[2] depend entirely on the boundary conditions. Here are some examples.

Example 8.1 (Diffusion in a box) Suppose that a volume is entirely enclosed by a reflecting boundary. Then the concentration within the volume will eventually settle down to a constant value everywhere:

$$C(\mathbf{r}) = c. \tag{8.38}$$

2. This is called **Laplace's equation** and also appears in electrostatics, where it governs the electric potential in charge-free space. Sometimes you can crib a diffusion solution from an electrostatics book.

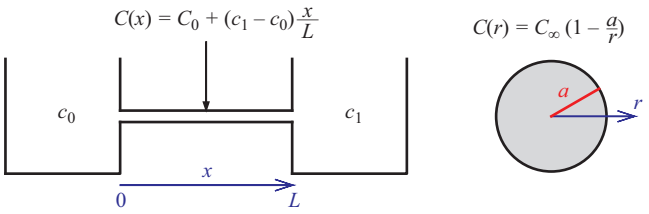

Figure 8.8
Geometry of diffusion examples. Left: Pipe between two stirred tanks. Right: An absorbing sphere in an infinite tank.

Clearly, this satisfies $\nabla^2 C(\mathbf{r}) = 0$. Also, the flux is zero everywhere: $\mathbf{J}(\mathbf{r}) = -D\nabla C(\mathbf{r}) = 0$, which satisfies the reflecting condition at the boundary of the space. □

Example 8.2 (Diffusion between two stirred compartments) Imagine a thin pipe between two water tanks (figure 8.8). Each tank is kept at a constant concentration of the solute. If $x \in [0, L]$ is the position along the pipe, then the boundary conditions for $C(x)$ are

$$C(0) = c_0, C(L) = c_1. \tag{8.39}$$

Along the pipe, the concentration is

$$C(x) = c_0 + (c_1 - c_0)\frac{x}{L}. \tag{8.40}$$

The gradient $\frac{\partial C}{\partial x} = \frac{c_1 - c_0}{L}$ is constant along the pipe, so $\frac{\partial^2 C}{\partial x^2} = 0$ satisfies equation (8.37). Also, there is a constant flux of solute along the pipe of strength $J = -D\frac{\partial C}{\partial x} = -D\frac{c_1 - c_0}{L}$ from the high-concentration to the low-concentration tank. □

Example 8.3 (Diffusion to an absorbing sphere) Imagine a sphere of radius a immersed in an infinite tank (figure 8.8). The sphere absorbs all the solute particles that hit its surface. Because of spherical symmetry, the concentration $C(r)$ will depend only on the distance r from the center of the sphere. At the surface of the sphere, $C(a) = 0$. Far from the sphere, the concentration is maintained at $C(\infty) = C_\infty$. In between, the solution to equation (8.37) is

$$C(r) = C_\infty \left(1 - \frac{a}{r}\right). \tag{8.41}$$

To verify this, recall the form of the Laplacian differential operator in spherical coordinates in equation (8.26). □

Example 8.4 (What is the "diffusion-limited reaction rate"?) For two molecules A and B to react, they must diffuse to within molecular dimensions of each other. Suppose that molecule A is held fixed at the origin and molecules of type B are present at an average concentration C_∞. We want to know at what rate per unit time molecules of type B get to within the reaction radius a of the origin. So imagine an absorbing sphere of radius a that destroys all the molecules that touch its surface. Given the result in equation (8.41), the steady-state concentration profile is

$$C(r) = C_\infty \left(1 - \frac{a}{r}\right). \tag{8.42}$$

The resulting flux of particles at the surface of the sphere is

$$J(a) = -D\frac{\partial}{\partial r}C(a) = \frac{D}{a^2}C_\infty. \tag{8.43}$$

and the rate at which particles hit the surface is

$$R = J(a)4\pi a^2 = 4\pi aDC_\infty. \tag{8.44}$$

This rate is proportional to the concentration of molecules C_∞ and the proportionality constant is called the **diffusion-limited reaction rate**, k_D. If we choose a to be a typical molecular dimension of 0.1 nm, and D a typical diffusion coefficient of 10^{-5} cm^2/s, then

$$k_\mathrm{D} = R/C_\infty = 4\pi aD \approx 10^9 \mathrm{M}^{-1}\mathrm{s}^{-1}. \tag{8.45}$$

□

8.2 Random Time Series

Experimental measurements often involve recording a quantity over time and trying to infer some structure from these measurements. Typically, the measurements are taken at discrete times t_i, yielding values y_i. Such a sequence of measurements $\{(y_1, t_1), (y_2, t_2), \ldots, (y_n, t_n)\}$ is called a "time series." A **random time series** is a function $\{(y_i, t_i)\}$ whose evolution is stochastic and not uniquely determined by the initial conditions. Examples include the position of a particle following Brownian motion, the number of mutations on a chromosome over time, the number of bacteria in a growing population, the electric current flowing across a cell membrane, and the fluorescence intensity of a chromophore, just to name a few.

A random time series can be characterized completely by specifying the joint probability distribution for its values at the various times:

$$\begin{aligned} P_n\left(y_1, t_1; \ldots; y_n, t_n\right) \mathrm{d}y_1 \ldots \mathrm{d}y_n = \\ = \mathrm{Prob}\left(y(t_1) \in [y_1, y_1 + \mathrm{d}y_1], \ldots, y(t_n) \in [y_n, y_n + \mathrm{d}y_n]\right). \end{aligned} \tag{8.46}$$

Of course, this is a huge object with almost infinitely many parameters. Fortunately, there are special conditions under which the probability distribution simplifies, to the point where one can capture its essence in a finite experiment and use it to make practical predictions.

8.2.1 Stationary Process

A **stationary process** is one whose rules don't change over time. This means that any given sequence of measurements $\{(y_1, t_1), (y_2, t_2), \ldots, (y_n, t_n)\}$ is as equally probable now as it was some time ago. The joint probability distribution depends only on time differences, not on the absolute time:

$$P_n\left(y_1, t_1; y_2, t_2; \ldots; y_n, t_n\right) = P_n\left(y_1, 0; y_2, t_2 - t_1; \ldots; y_n, t_n - t_{n-1}\right). \tag{8.47}$$

Often, one can argue from first principles that a process should be stationary, for example because none of the external constraints have changed in a long time, and the system has somehow found equilibrium.

8.2.2 Markov Process

A **Markov process** is a special type of stationary process whose future evolution is completely determined by the most recent value. How the system arrived at that value is not important: the history of the random variable plays no role in its future. This applies to many important processes, like random walks, protein state transitions, and US foreign policy (to good approximation):

$$P_n\left(y_n, t_n | y_1, t_1; y_2, t_2; \ldots; y_{n-1}, t_{n-1}\right) = P_2\left(y_n, t_n | y_{n-1}, t_{n-1}\right). \tag{8.48}$$

This is a very powerful simplification, as we only need to consider transitions from the current time point to the next. A Markov process is completely determined by the **instantaneous distribution** $P_1(y_1)$, where

$$P_1(y_1)\mathrm{d}y_1 = \mathrm{Prob}\ \left(y \in [y_1, y_1 + \mathrm{d}y_1]\right), \tag{8.49}$$

and the **transition probability**

$$P_2\left(y_2, t | y_1\right) = \frac{P_2\left(y_1, 0; y_2, t\right)}{P_1(y_1)}, \tag{8.50}$$

where

$$P_2\left(y_2, t | y_1\right)\mathrm{d}y_2 = \mathrm{Prob}\left(y \in [y_2, y_2 + \mathrm{d}y_2] \text{ at time } t, \text{ given that } y = y_1 \text{ at time } 0\right). \tag{8.51}$$

Example 8.5 (Random telegraph signal) This is a simple random process that nonetheless serves as a useful model in many situations of practical importance—namely, any time a system flips back and forth between two states in a historyless fashion (figure 8.10). Examples are chemical binding sites flipping between bound and empty, an enzyme flickering on and off, or an ion channel switching between open and closed. Here, the variable $y(t)$ is binary:

$$y(t) \in \{0, 1\}, \tag{8.52}$$

and it performs transitions from one value to the other at a constant rate: If $y = 0$, then in the next short interval $\mathrm{d}t$, it will switch to 1 with probability $\alpha_{01}\mathrm{d}t$ as shown in figure 8.9. Similarly, if $y = 1$, then it will switch to 0 with probability $\alpha_{10}\mathrm{d}t$. Note that this process is both stationary and Markov: The switching rates α_{01}, α_{10} are constant in time and transitions depend only on the current state of the variable, not on its history.

Based on this definition of the process, we compute the transition probability as follows: Call $P_2\,(1, t|0)$ the probability that $y = 1$ at time t, given that $y = 0$ at time 0. Now let us consider how that probability changes between t and $t + \mathrm{d}t$. One can get $y = 1$ at time $t + \mathrm{d}t$ in two ways: either $y = 0$ at time t, and then the value switches to 1 in the following small interval $[t, t + \mathrm{d}t]$; or $y = 1$ at time t, and there is no switch in the interval $[t, t + \mathrm{d}t]$. The two possibilities are mutually exclusive, so we can write

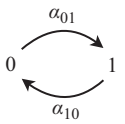

Figure 8.9
A random telegraph signal with transitions between values of 0 and 1.

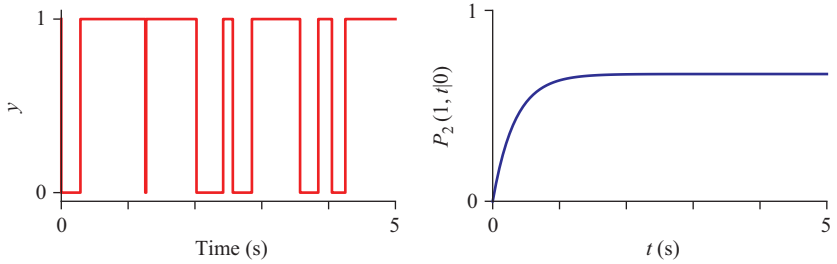

Figure 8.10
Left: Time course of a random telegraph signal with $\alpha_{01} = 2\ \mathrm{s}^{-1}$ and $\alpha_{10} = 1\ \mathrm{s}^{-1}$. Right: Probability that the signal will be 1 at time t, given that it was 0 at time $t = 0$.

$$\begin{aligned} P_2\left(1, t+\mathrm{d}t|0\right) &= P_2\left(0, t|0\right)\alpha_{01}\mathrm{d}t + P_2\left(1, t|0\right)\left(1-\alpha_{01}\mathrm{d}t\right) \\ &= \left(1-P_2\left(1, t|0\right)\right)\alpha_{01}\mathrm{d}t + P_2\left(1, t|0\right)\left(1-\alpha_{01}\mathrm{d}t\right). \end{aligned} \tag{8.53}$$

So

$$\frac{\mathrm{d}}{\mathrm{d}t}P_2\left(1, t|0\right) = \alpha_{01} - (\alpha_{01}+\alpha_{10})P_2\left(1, t|0\right), \tag{8.54}$$

with the solution

$$P_2\left(1, t|0\right) = \frac{\alpha_{01}}{\alpha_{01}+\alpha_{10}}\left(1-\mathrm{e}^{-(\alpha_{01}+\alpha_{10})t}\right). \tag{8.55}$$

From symmetry, one gets the other transition probabilities:

$$\begin{aligned} P_2\left(0, t|0\right) &= 1 - P_2\left(1, t|0\right) \\ P_2\left(0, t|1\right) &= \frac{\alpha_{10}}{\alpha_{01}+\alpha_{10}}\left(1-\mathrm{e}^{-(\alpha_{01}+\alpha_{10})t}\right) \\ P_2\left(1, t|1\right) &= 1 - P_2\left(0, t|1\right). \end{aligned} \tag{8.56}$$

Finally, the instantaneous probability that $y = 1$ is obtained from the transition probabilities after a very long time is:

$$P_1(1) = P_2\left(1, t=\infty|0\right) = \frac{\alpha_{01}}{\alpha_{01}+\alpha_{10}} \tag{8.57}$$

and obviously, $P_1(0) = 1 - P_1(1)$. Because this is a Markov process, everything about it can be computed from functions P_1 and P_2. □

8.2.3 Moments of a Random Process

The mean of a random process is defined as

$$\text{Mean} = \langle y(t) \rangle \tag{8.58}$$

and the variance as

$$\text{Variance} = \langle \big(y(t) - \langle y(t) \rangle\big)^2 \rangle = \langle y^2(t) \rangle - \langle y(t) \rangle^2, \tag{8.59}$$

where the angle brackets denote the **ensemble average** over different instantiations of the random process that start from the same initial conditions. **If the process is stationary**, then the ensemble average is equal to the time average and no longer depends on absolute time:

$$\langle y(t) \rangle = \bar{y} = \lim_{T \to \infty} \frac{1}{T} \int_0^T y(t) \mathrm{d}t. \tag{8.60}$$

8.2.4 Correlation Function and Power Spectrum

Another second moment of a random process is the **correlation function** $C(\tau)$, which relates values over time:

$$C(\tau) = \langle y(t) y(t+\tau) \rangle. \tag{8.61}$$

For a stationary process, one can again compute the averages over time, and the correlation function depends only on the time difference τ, not on absolute time t:

$$C(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T y(t) \cdot y(t+\tau)\, \mathrm{d}t. \tag{8.62}$$

An important result relates the correlation function of a random process to its power spectrum. Extending the definition in equation (3.28), the power spectrum of a random process is the expectation value of the square modulus of the Fourier transform:

$$P(\omega) = \left\langle |\hat{y}(\omega)|^2 \right\rangle, \tag{8.63}$$

where the expectation is over many instances of the same process.

For a stationary process, the **Wiener-Khintchin theorem** states that

$$P(\omega) = \int_{\tau=-\infty}^{+\infty} C(\tau) e^{-i\omega\tau} \mathrm{d}\tau. \tag{8.64}$$

Stated in words: The power spectrum is the Fourier transform of the correlation function.

Example 8.6 (Random telegraph signal) To illustrate these concepts, let us return to the random telegraph process of section 8.5. Recall that this random variable y switches between the values of 0 and 1. Transitions from 0 to 1 happen at constant probability per unit time α_{01} and from 1 to 0 at a rate α_{10}. What is the correlation function for this system?

We need to calculate

$$C(\tau) = \langle y(0)y(\tau)\rangle. \tag{8.65}$$

As the product vanishes when either $y(0)$ or $y(\tau)$ are 0, the only remaining contribution is

$$C(\tau) = \text{Prob}[y(0) = 1 \text{ and } y(\tau) = 1]. \tag{8.66}$$

Using conditional probability, we see that

$$\begin{aligned}\text{Prob}[y(0) = 1 \text{ and } y(\tau) = 1] &= \text{Prob}[y(0) = 1] \cdot \text{Prob}[y(\tau) = 1 | y(0) = 1)] \\ &= P_1(1) \cdot P_2\,(1, \tau|1)\,.\end{aligned} \tag{8.67}$$

From the results in section 8.5, one finds

$$C(\tau) = \frac{\alpha_{01}}{\alpha_{01}+\alpha_{10}} \cdot \left(\frac{\alpha_{01}}{\alpha_{01}+\alpha_{10}} + \frac{\alpha_{10}}{\alpha_{01}+\alpha_{10}} e^{-(\alpha_{01}+\alpha_{10})\tau} \right). \tag{8.68}$$

The correlation function consists of a decaying exponential. The power spectrum is the Fourier transform of that function. Note that we encountered the power spectrum of a decaying exponential previously in equation (3.30). Here,

$$\begin{aligned}P(\omega) &= \int_{-\infty}^{+\infty} C(\tau) e^{-i\omega\tau} \mathrm{d}\tau \\ &= 2\,\mathrm{Re}\left[\int_0^{+\infty} C(\tau) e^{-i\omega\tau} \mathrm{d}\tau \right], \\ &= 2b(1-b)\frac{\alpha}{\alpha^2+\omega^2}\end{aligned} \tag{8.69}$$

where

$$\begin{aligned}b &= P_1(1) = \frac{\alpha_{01}}{\alpha_{01}+\alpha_{10}} \\ \alpha &= \alpha_{01}+\alpha_{10},\end{aligned} \tag{8.70}$$

and we have ignored the divergence of the power at $\omega = 0$. Figure 8.11 illustrates the correlation function and power spectrum of this process. □

8.2.5 Discrete Markov Process

Frequently, one approximates a system as taking on a discrete set of states. For example, a protein might exist in one of several discrete conformations, or an animal may be in one of a few behavioral states. If these discrete states are long-lived compared to the transitions between them, then we can take the transitions to be instantaneous. This defines a **discrete stochastic process**.

If, in addition, the process $X(t)$ is stationary and Markovian, then it is called a **discrete Markov process**. This means that transitions from state $X = i$ to state $X = j$ happen at constant probability per unit time α_{ij}, and that transition rate is independent of the prior history of the process (figure 8.12). Note that this is a generalization of the random telegraph described in example 8.5, which exists in only two states: $X \in \{0, 1\}$.

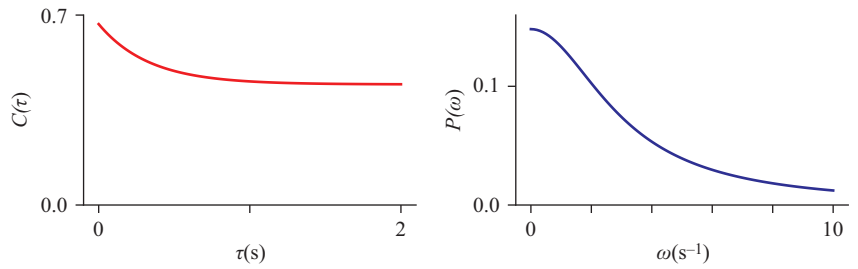

Figure 8.11
Left: Correlation function of a random telegraph signal with $\alpha_{01} = 2\ \mathrm{s}^{-1}$ and $\alpha_{10} = 1\ \mathrm{s}^{-1}$. Right: Power spectrum of that same random process.

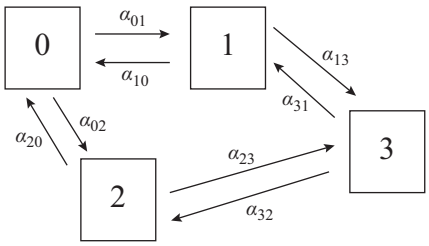

Figure 8.12
A discrete Markov process takes on one of a set of states i, with first-order transitions happening at rates α_{ij}. In this example, four states are possible.

To understand the evolution of $X(t)$ from any given starting state, one can follow the same approach as for the random telegraph process. First, define a **transition probability**:

$$P_{ij}(t) = \text{Probability that } X = j \text{ at time } t, \text{ given that } X = i \text{ at time } 0. \tag{8.71}$$

Again, by considering what happens in the last short time interval dt, one finds that

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{ij}(t) = \sum_k P_{ik}(t)\alpha_{kj} - P_{ij}(t)\sum_l \alpha_{jl}. \tag{8.72}$$

Here, the first term includes all the transitions from other states into state j and the second term are transitions away from state j. Equation (8.72) is called the **master equation** of the process.

To solve the master equation, note that it can be written in matrix form as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}(t) = \mathbf{P}(t)\cdot\mathbf{Q}, \tag{8.73}$$

where $\mathbf{P}$ is the matrix of all transition probabilities P_{ij}, and $\mathbf{Q}$ is given by

$$\mathbf{Q}_{ij} = \alpha_{ij} - \delta_{ij}\sum_l \alpha_{jl}. \tag{8.74}$$

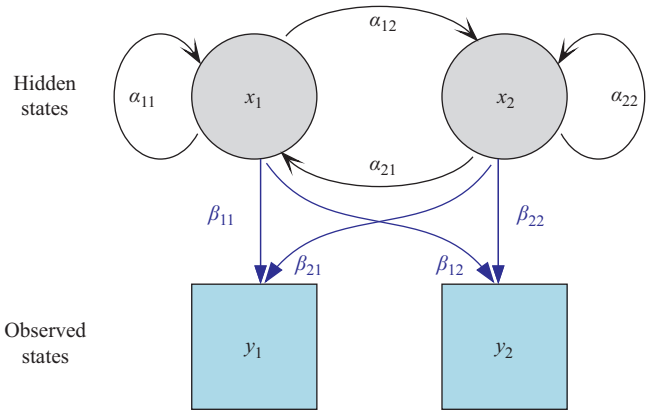

Figure 8.13
An HMM with two hidden states and two observable outcomes.

This is simply the matrix version of the rabbit equation (1.34), so the general solution is

$$\mathbf{P}(t) = \mathbf{P}(0) \cdot \mathrm{e}^{\mathbf{Q}t} = \mathrm{e}^{\mathbf{Q}t} \tag{8.75}$$

because $\mathbf{P}(0)$ is the identity. As usual, this is most easily evaluated in the eigenbasis of the matrix $\mathbf{Q}$; see section 2.11.2. Using the diagonalizing transform $\mathbf{S}$,

$$\mathbf{P}(t) = \mathbf{S} \cdot \begin{bmatrix} \mathrm{e}^{\lambda_1 t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathrm{e}^{\lambda_n t} \end{bmatrix} \cdot \mathbf{S}^{-1}, \tag{8.76}$$

where λ_k are the eigenvalues of the matrix $\mathbf{Q}$. So the transition probabilities take the general form of a sum of exponentials:

$$P_{ij}(t) = \sum_k c_k \mathrm{e}^{\lambda_k t} \tag{8.77}$$

where the c_k can be computed from equation (8.76). This fully describes the stochastic evolution of the system from any initial distribution of states.

With these transition probabilities, one can further compute the correlation function or power spectra of the process, following the approach elaborated here for the random telegraph signal (see example 8.6).

8.3 Hidden Markov Models

In section 8.2.5, we considered a system that exists in discrete states X_i, with transition rates α_{ij} among these states constant in time. Sometimes we cannot observe the system's state directly, but rather have to guess it based on some observable Y that is produced in a way that depends on the state X. A formalization of this concept is the **hidden Markov model (HMM)**.

Figure 8.13 shows an example of an HMM with two states $X \in \{x_1, x_2\}$ and an observable that takes one of two values $Y \in \{y_1, y_2\}$. We will treat this as a discrete-time

process, where time proceeds in discrete steps. If the system is in state $X = x_i$, then in the next time step, it transitions to state x_j with probability α_{ij}. Also, the system emits an observable $Y = y_k$ with probability β_{ik}. The probabilities are normalized, such that

$$\begin{aligned} \sum_j \alpha_{ij} &= 1 \qquad \text{for all } i \\ \sum_k \beta_{ik} &= 1 \qquad \text{for all } i. \end{aligned} \tag{8.78}$$

A sample sequence generated by the HMM in the figure would be

$$\begin{array}{rcccccccccc} \text{Hidden state sequence } X(t): & x_1 & x_1 & x_2 & x_2 & x_2 & x_1 & x_2 & x_1 & x_1 & x_2. \\ \text{Observable sequence } Y(t): & y_2 & y_2 & y_2 & y_1 & y_1 & y_2 & y_2 & y_2 & y_1 & y_1. \end{array} \tag{8.79}$$

An HMM is fully determined by the set of states X, the set of observables Y, and the probabilities for transitions α_{ij} and those for emissions β_{ik}:

$$[X, Y, \{\alpha_{ij}, \beta_{ik}\}]. \tag{8.80}$$

There are three types of problems that one wants to solve in the context of HMMs:

- Given a model $[X, Y, \{\alpha_{ij}, \beta_{ik}\}]$, what is the probability of a particular sequence of observations $Y(t)$? Note that several different hidden state sequences may generate the same sequence of observations, and one must take all of them into account. This is done using the **forward algorithm**.
- Given a model $[X, Y, \{\alpha_{ij}, \beta_{ik}\}]$ and a sequence of observations $Y(t)$, what is the most likely sequence of states $X(t)$ that generated it? This is done by the **Viterbi algorithm**, and this sequence of hidden states is called the **Viterbi path**.
- Given an HMM architecture $[X, Y]$ and a sequence of observations $Y(t)$, what are the most likely transition and emission probabilities $[\{\alpha_{ij}, \beta_{ik}\}]$? This is typically done using the **Baum-Welch algorithm**, which in turn uses the **forward-backward algorithm**.

In all these cases, the challenge is that the number of possibilities grows exponentially with the sequence length. These algorithms have been developed to make the calculations computationally feasible. Although a detailed explanation of these algorithms is beyond the scope of this text, we illustrate the ideas with some examples here.

8.3.1 Finding Genes in DNA Sequence

The central dogma in biology states that deoxyribonucleic acid (DNA) is transcribed into messenger ribonucleic acid (mRNA), which is translated into protein. However, during preparation of mRNA from DNA, some sections of sequence are removed by splicing. These sections are called "introns," whereas the remaining segments that ultimately encode the protein are known as "exons." Given a DNA sequence such as

$$\ldots \text{ATGCGACTGCATAGCACTT} \ldots, \tag{8.81}$$

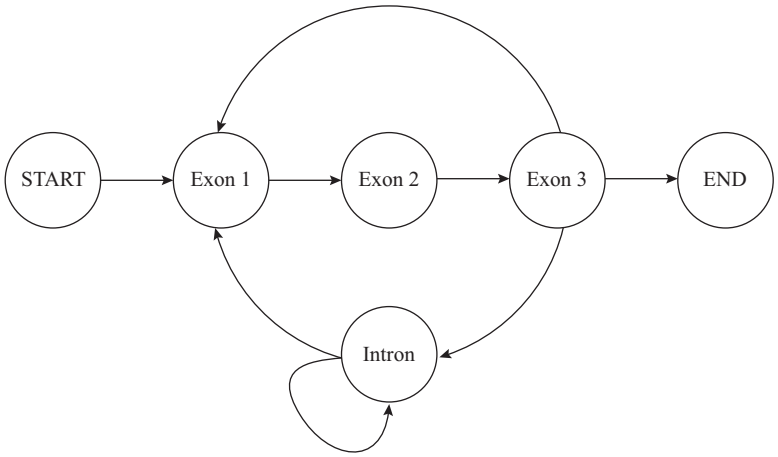

Figure 8.14
An HMM for modeling introns and exons within genes.

it is a challenge to determine which sections are exons and which are introns. Some clues are available because the sequence in exons has to meet certain constraints. Each triplet of nucleotides encodes an amino acid, and thus the frequency of triplets inside exons follows certain statistical regularities. Within introns, the frequencies are different. Effectively, the DNA sequence has a hidden state—exon or intron—at each location. We cannot observe the state directly, but we can see the nucleotides that were "emitted," and their probabilities depend on the state.

One can formalize this argument by imagining that the DNA sequence was produced by a machine that has the mechanics of an HMM (figure 8.14). The machine has three exon states and one intron state. With every step along the DNA, the machine makes a state transition. In state "Exon 1," it emits the first nucleotide of a codon. Then it transitions deterministically to state "Exon 2" and emits the second nucleotide. Next, it moves to state "Exon 3" to emit the last nucleotide of that codon. At the next step, the machine may return to "Exon 1" to deliver another codon of three nucleotides. Or it may switch to state "Intron" and deliver one or more intron nucleotides before returning to "Exon 1."

We would like to determine **what is the most likely sequence $X(t)$ of intron/exon states, given the observed sequence $Y(t)$ of nucleotides**. This requires that we know the transition probabilities α_{ij} among the states of the model and the emission probabilities β_{ik} with which each state generates nucleotides. Both have been tabulated from a large number of genes where the ground truth about introns and exons is known. Then one can use these probabilities to infer the intron/exon state on a novel sequence. This can be accomplished by the Viterbi algorithm.

8.3.2 Sequence Alignment

In the analysis of genetic sequences, either protein or DNA, one often wants to evaluate the similarity between two sequences. For example, we may be interested in understanding how the protein sequence from several species diverged from that of a common ancestor, so as to place those species on a philogenetic tree. In measuring the similarity of two sequences, the first issue is the problem of alignment: Which

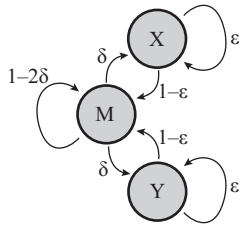

Figure 8.15
An HMM for sequence alignment.

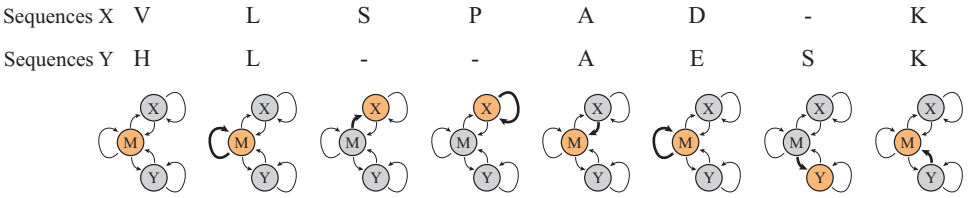

Figure 8.16
A sequence alignment as a path through the HMM.

amino acid in one sequence corresponds to which in the other? Two sequences may differ for three reasons: (1) an amino acid has been mutated to a different one; (2) an extra amino acid has been inserted; or (3) an amino acid has been deleted.

As shown in section (8.3.1), we imagine that the two amino acid sequences $X(t)$ and $Y(t)$ were produced by a machine that acts like an HMM (figure 8.15). The model has three states, M, X and Y. In state M, the model generates a pair of symbols (amino acids) and adds one to sequence $X(t)$ and the other to sequence $Y(t)$. If there is no mutation, the amino acids are the same, and this is the most likely scenario. But sometimes a mutation occurs, so the machine may add, say, alanine to one sequence and glycine to the other. The probability of emitting a particular pair of amino acids (x, y) is $P(x, y)$.

Sometimes, though, the model will switch to one of the other states, X or Y. In this state, the model generates an amino acid for one of the sequences, but not for the other. The rate at which this switch between states happens is controlled by the transition probabilities ϵ and δ, and once the system is in one of these states, it will generate an amino acid with probability $Q(x)$ or $Q(y)$. This allows the model to account for amino acid insertions and deletions.

For a given pair of sequences, the particular path $S(t) \in \{M, X, Y\}$ of the state of the model proposes an alignment between the two sequences—namely, the location of mutations, insertions, and deletions (figure 8.16). Different paths will generate the observed sequences with different likelihoods, and the maximum likelihood path represents the optimal alignment.

The model parameters are the transition probabilities ϵ and δ and the emission probabilities $P(x_i, y_j)$ and $Q(x_i)$. These parameters are determined from experience, following many successful protein alignments. For example, the emission probabilities $P(x, y)$ will depend on factors such as the evolutionary distance between the two species or the specific protein in question. Tables have been compiled empirically that capture these factors, and one such table, called *BLOSUM 50*, is shown in figure 8.17.

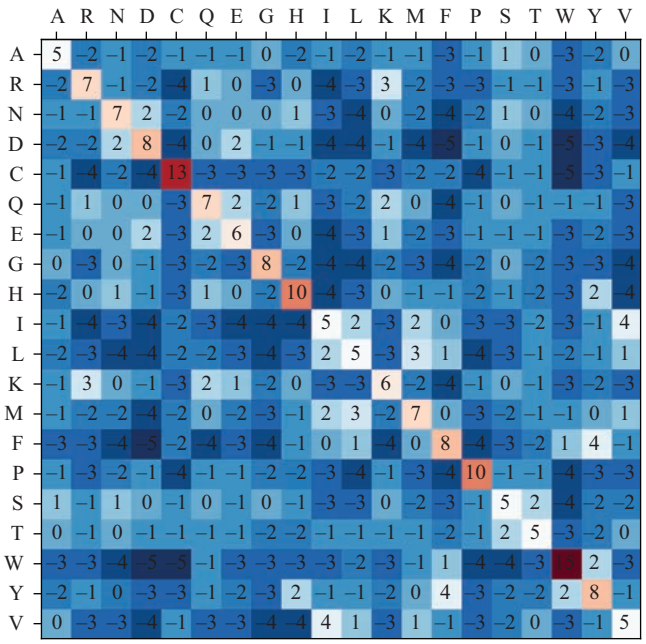

Figure 8.17
The BLOSUM 50 matrix used for protein sequence alignment. The entry (i, j) gives the log-likelihood of adding amino acid i and amino acid j to the sequences X and Y, respectively.

The simple three-state HMM presented here is only a first approximation to the myriad of complex models and refinements that have been developed in the field of sequence alignment. Proteins have different domains, and it is natural to assume that different models will apply to different domains. This can also be incorporated into the analysis.

8.3.3 Further Reading

See Henderson et al. (1997) for an early application of HMMs to gene finding. The general application to sequence analysis is reviewed in Durbin et al. (1998) and Eddy (2004). An application to problems of learning in animals is found in Smith et al. (2004). More theory and applications can be found in Dymarski (2011).

8.4 Point Processes

A **point process** is a series of identical events that are point like in time. A sample from a point process is completely specified by listing the event times $\{t_1, t_2, \ldots, t_n\}$. Some examples encountered in biological research include the following:

- The arrival times of photons at a camera or at a photoreceptor cell
- Times of collision between a ligand and a binding site
- Times of opening of an ion channel
- Times of random mutations occurring in a genome
- Times of action potentials fired by a neuron

Point processes can also be defined in spatial dimensions, such as the locations of all trees of a given species in a field, or the locations of all nerve cells of a given type on the surface of the retina. The following sections will focus on temporal processes, but the treatment translates easily to space.

8.4.1 Intensity Function

A random point process is fully specified by the **conditional intensity function**, which spells out the probability of getting an event in the next small time interval. In general, that probability depends on time, but it also depends on the entire preceding history of the process:

$$\begin{aligned} P(t|\ldots,t_{-2},t_{-1})\mathrm{d}t = {} & \text{probability of getting an event in } [t, t+\mathrm{d}t] \\ & \text{as a function of } t \text{ and the entire history} \\ & \{\ldots,t_{-2},t_{-1}\} \text{ of event times prior to } t. \end{aligned} \tag{8.82}$$

8.4.2 Stationary Point Process

As introduced in section 8.2.1, a stationary process does not depend on absolute time, but only on time differences. A point process is **stationary** if a time shift by τ leaves the conditional intensity unchanged; namely,

$$P(t+\tau|\ldots,t_{-2}+\tau,t_{-1}+\tau) = P(t|\ldots,t_{-2},t_{-1}). \tag{8.83}$$

8.4.3 Poisson Process

This is the simplest case of a point process, in which the events happen with constant probability per unit time and independent of history:

$$P(t|\ldots,t_{-2},t_{-1}) = \lambda. \tag{8.84}$$

Classic examples are the arrival of photons from a constant light source and the clicks in a Geiger counter from a radioactive sample. Some characteristics of this process have been elaborated earlier in this book, in sections 6.3.7 and 6.4.6.

The **number of events N observed in a time interval** of length T is a discrete random variable, which follows the Poisson distribution

$$P(N,T) = e^{-\mu}\frac{\mu^N}{N!}, \tag{8.85}$$

where

$$\mu = \langle N\rangle = \lambda T \tag{8.86}$$

is the expectation value of N.

The **time interval τ between successive events** is a continuous random variable that follows the exponential distribution

$$P(\tau) = \lambda e^{-\lambda\tau}. \tag{8.87}$$

Successive time intervals are statistically independent, so the time τ_n to the nth event follows a gamma distribution:

$$\tau_n \sim \text{Gamma}(n, \lambda)$$
$$P(\tau_n) = \frac{\tau_n^{n-1} e^{-\beta\tau_n} \beta^n}{\Gamma(n)}. \tag{8.88}$$

8.4.4 Inhomogeneous Poisson Process

Here, the intensity λ varies with time, but independent of the history of the process:

$$P(t|\ldots, t_{-2}, t_{-1}) = \lambda(t). \tag{8.89}$$

In a simple example, consider a lamp whose intensity $I(t)$ is getting modulated up and down with a dimmer knob. The photon stream from that light source follows an inhomogeneous Poisson process with $\lambda(t) \propto I(t)$.

One can obtain such an inhomogeneous Poisson process from a homogeneous one with $\lambda = 1$ by warping the time axis. Imagine that warp time τ flows faster and slower relative to t according to the intensity $\lambda(t)$:

$$\frac{d\tau}{dt} = \lambda(t). \tag{8.90}$$

Then events that are at a constant density of 1 per unit time on the τ-axis get compressed or expanded on the t-axis, exactly so as to achieve the density $\lambda(t)$.

From this argument, it follows that the number of events in any given time interval $[t_a, t_b]$ is again Poisson distributed. The mean number is

$$\mu = \int_{t_a}^{t_b} \lambda(t) dt, \tag{8.91}$$

and the distribution is

$$N \sim \text{Poiss}(\mu). \tag{8.92}$$

8.4.5 Spectral Analysis of a Point Process

To use the range of frequency-analysis tools developed in section 3.2, it helps to convert the point process into a continuous function of time. For that purpose, each event in the process $\{t_k\}$ contributes a dirac delta function:

$$R(t) = \sum_k \delta(t - t_k). \tag{8.93}$$

This function can be seen as the rate of occurrence of events because its integral delivers the cumulative number of events:

$$\int^t R(t')dt' = N(t) = \text{number of events before time } t. \tag{8.94}$$

The Fourier transform at frequency ω becomes

$$\hat{R}(\omega) = \int R(t) e^{-i\omega t} dt = \sum_k e^{-i\omega t_k}. \tag{8.95}$$

This expression has a useful geometric interpretation: $e^{-i\omega t_k}$ is a phasor of unit length in the complex plane, oriented at phase ωt_k. If the t_k happen with a periodicity of $2\pi/\omega$, then all the phasors point in the same direction and their sum will be very large. If, on the other hand, the t_k happen at random times, then the phasors are oriented randomly and their sum will be small. In this way, the Fourier coefficient $\hat{R}(\omega)$ reflects the degree of periodicity of the point process at frequency ω.

8.4.6 Power Spectrum of a Point Process

By applying the definition in equation (8.63) for the power spectrum to the rate function $R(t)$, one finds the power spectrum of a random point process to be

$$P(\omega) = \left\langle \left| \hat{R}(\omega) \right|^2 \right\rangle = \left\langle \left| \sum_k e^{-i\omega t_k} \right|^2 \right\rangle \tag{8.96}$$

where the expectation is over instantiations of the process.[3]

8.4.6.1 Power spectrum of a Poisson process For example, consider a Poisson point process with intensity λ, extending over the time period $[0, T]$. As discussed in section 3.2.6, we will want to evaluate the power spectrum at frequencies that are multiples of the fundamental $\omega_j = j \cdot 2\pi/T$, $j = 0, 1, \ldots$. Say that N is the number of events observed in any instantiation of the process. We know that N follows the Poisson distribution with mean $\mu = \lambda T$:

$$N \sim \text{Poiss}(\lambda T). \tag{8.97}$$

Suppose now that N is large, so we can ignore the fluctuations of order $\sqrt{N}$. Then the power at zero frequency is simply

$$P(\omega = 0) = \left\langle N^2 \right\rangle \approx (\lambda T)^2. \tag{8.98}$$

To evaluate power at the nonzero frequencies, note that each of the event times t_k is distributed uniformly throughout the interval $[0, T]$. So the phase $\omega_j t_k$ will be uniformly distributed in $[0, 2\pi]$, and thus the phasors $e^{-i\omega_j t_k}$ all point in random and independent directions. The sum of those phasors,

$$z = \sum_k e^{-i\omega t_k}, \tag{8.99}$$

is the sum of N independent random unit vectors in the complex plane. If N is large, one can invoke the central limit theorem to argue that this sum vector has a Gaussian distribution around the origin with a variance that is the sum of the individual variances—namely,

$$\text{Var}[z] = \left\langle |z|^2 \right\rangle = \langle N \rangle. \tag{8.100}$$

3. As usual, there are alternative definitions that differ by some normalization factor. If the only goal is to compare power at different frequencies, that doesn't matter.

In conclusion,

$$P(\omega_j) \approx \begin{cases} (\lambda T)^2, & \text{if } j = 0 \\ \lambda T, & \text{j} > 0. \end{cases} \tag{8.101}$$

So the homogeneous Poisson process has equal power at all frequencies (except $\omega = 0$): a **white noise** spectrum.[4]

8.4.7 Shot Noise

Often the discrete events in a random point process are not observed directly, but rather through some signal caused by each event. For example, every photon captured by a photo-detector tube produces a short unitary blip of electric current. We observe the time course of the current and want to infer the occurrence time of the blips. In such a case, the time course of the unitary signal is called the **shot**, and the superposition of all the shots is called **shot noise**.

Note that the shot noise signal $F(t)$ results from a convolution of the point process rate function $R(t)$ and the shape of the individual shot $S(t)$:

$$F(t) = \sum_k S(t - t_k) = \int R(t')S(t - t')\mathrm{d}t' = R(t) * S(t). \tag{8.102}$$

Recall that the Fourier transform of a convolution is equal to the product of the two Fourier transforms (as discussed in section 3.2.4.4). Consequently, the same is true for the power spectrum and

$$P_F(\omega) = P_R(\omega)P_S(\omega). \tag{8.103}$$

This relationship gets used in both ways: Sometimes we know the shape of the individual shot (e.g. for the photo-detector tube), and this allows us to infer something about the point process that produces the shots. Other times, we are confident about the spectrum of the point process, and thus we can learn something about the shape of the individual shot. In particular, if $R(t)$ is a Poisson point process, then the spectrum $P_R(\omega)$ is white, and therefore the spectrum of the shot $P_S(\omega)$ has the same shape as that of the measured shot noise $P_F(\omega)$. See section 9.4.2 for an example.

8.4.8 Converting a Point Process to a Time Series

For practical calculations, one often converts a point process $\{t_k\}$ into a discrete time series R_i with values of 1 or 0. A common form of conversion is **binning** of the point process: choose a bin width Δt and count the number of events in each bin

$$R_i = \frac{N((i+1)\Delta t) - N(i\Delta t)}{\Delta t}. \tag{8.104}$$

Clearly, the timing of an event within each bin is lost in the process, so Δt effectively sets the time resolution of any subsequent analysis. Now one can apply the full battery

4. If λT is not large, one can follow the same logic to find the exact power spectrum. It will still be white (namely, equal power at all frequencies except $\omega = 0$).

of time-series analysis methods, including spectral analysis and cross-correlation analysis.

However, a drawback of this approach is that it represents the point process very inefficiently: Suppose that there are 100 events in 10 s, and we want to preserve their timing to 1 ms. That produces a time series R_i with 10,000 values, even though $\{t_k\}$ has only 100 values. To compute a correlation of two such processes (by brute force) requires $\sim 10^8$ operations, compared to $\sim 10^4$ if one worked with the event times directly. In the old days when computations were done by hand, no one would have dreamed of making such an inefficient change in representation from point process to time series. These days, when computer speed is hardly a constraint for most scientific computing, the Fast Fourier transform speeds up linear operators, and optimized routines exist that work on sparse arrays, the cost can be negligible. On the other hand, if you are operating in a big data regime that involves lots of events and very high time resolution, you may reconsider your options.

8.4.9 Further Reading

Daley and Vere-Jones (2013) introduce point processes with the full mathematical armamentarium. Brown et al. (2004) discuss additional point process methods in the context of analysis of neural signals.

8.5 Dimensionality Reduction

Several subfields of biology have decidedly entered the area of big data owing to revolutionary new methods for large-scale measurements. Today, one can measure the expression levels of thousands of genes across thousands of different cells; or the activity of thousands of neurons over many thousands of timepoints. To gain any understanding from such high-dimensional data, one must somehow reduce the number of dimensions.

One goal of dimensionality reduction is to find structure in the data. For example, gene expression patterns of 10,000 genes may reduce to a few modules that group together genes with similar dynamics. Another goal is to separate signal from noise: the most dominant patterns in the data should get attention first. Another goal is visualization: we have no way of representing 10,000-dimensional space, but we can draw figures in two dimensions. Finally, one could argue that the whole process of scientific understanding itself is one of dimensionality reduction, such that eventually the meaning of a huge data set can be captured by a few equations interspersed with words of text.

For the purpose of this section, we will assume that the data consist of T data points $\mathbf{x}_j$, $j = 1, \ldots, T$. Each data point consists of N measured variables $\mathbf{x}_j = [x_{1j}, \ldots, x_{Nj}]^\top$. For example, x_{ij} might be the activity of neuron i at time j or the expression of gene i in cell j. Sometimes we will write these data in the form of a single $N \times T$ data matrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1T} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N1} & \cdots & \cdots & x_{NT} \end{pmatrix}. \tag{8.105}$$

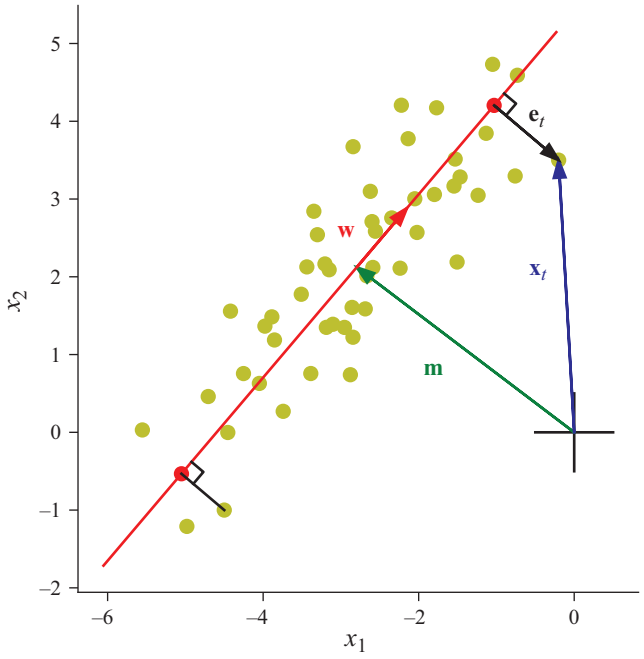

Figure 8.18
A two-dimensional (2D) data set with 50 data points (yellow) and a 1D approximation, illustrating the mean $\mathbf{m}$, the first principal component $\mathbf{w}$, the line corresponding to the 1D approximation (red), two examples of data points projected on that line (red circles), a random data point $\mathbf{x}_t$, and the residual for that data point $\mathbf{e}_t$.

8.5.1 Principal Component Analysis

Consider the $2 \times T$ data set in figure 8.18. Clearly, the two variables x_1 and x_2 vary together, a strong pattern in the data. One is tempted to just draw a line through this data cloud that gets as close as possible to all the data points. That line captures the direction along which the data vary the most, so it serves as a first approximation of the data set. With the proper definitions, discussed next, that line is called the "first principal component" of the data. The direction perpendicular to it is the "second principal component"; this is the direction along which the data vary the least. The algorithm for finding those special directions is called "principal component analysis" (PCA).

To formalize the goal of dimensionality reduction in the language of linear algebra, we want to approximate the data vectors by a linear superposition of just a few basis vectors with the smallest amount of error:

$$\mathbf{x}_j = \mathbf{m} + c_{1,j}\,\mathbf{w}_1 + \cdots + c_{D,j}\,\mathbf{w}_D + \mathbf{e}_j^{(D)}. \tag{8.106}$$

Here, $\mathbf{m}$ is a constant offset vector, $\mathbf{w}_k$ is the kth basis vector or **principal component** of the data cloud, $c_{k,j}$ is the coefficient of component k in data point j, and $\mathbf{e}_j^{(D)}$ is the residual of the approximation with D components for data point j. The goal is to choose $\mathbf{m}$, the $\mathbf{w}_k$, and the $c_{k,j}$ so as to minimize the average squared residual, which represents the **error of the D-dimensional approximation**:

$$E^{(D)} = \frac{1}{T}\sum_{j=1}^{T} \mathbf{e}_j^{(D)\top} \cdot \mathbf{e}_j^{(D)}. \tag{8.107}$$

If D is much smaller than N, and the error $E^{(D)}$ is acceptably small, then one has achieved successful dimensional reduction from N to D dimensions.

How can we find the best choices of the parameters $\mathbf{m}$, $\mathbf{w}_k$, and $c_{k,j}$? We do this by differentiating the error in equation (8.107) with respect to the parameters, as shown in exercise 10.18.

The solution tells us how to perform principal component analysis:

1. Compute the mean of the data cloud and subtract it from each data point

$$\begin{aligned} \mathbf{m} &= \frac{1}{T}\sum_{j=1}^{T} \mathbf{x}_j \\ \mathbf{y}_j &= \mathbf{x}_j - \mathbf{m}. \end{aligned} \tag{8.108}$$

2. Compute the covariance matrix as in equation (6.69) of the data

$$\mathbf{C} = \frac{1}{T}\sum_{j=1}^{T} \mathbf{y}_j \cdot \mathbf{y}_j^{\top}. \tag{8.109}$$

   This is an $N \times N$ matrix. It is symmetric (see section 2.11.1) and positive semidefinite, so it is guaranteed to have N eigenvalues, and they are all nonnegative.

3. Find the eigenvalues λ_k of the covariance matrix $\mathbf{C}$ and the associated eigenvectors $\mathbf{w}_k$. Sort them in decreasing order of the eigenvalues: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Normalize all the eigenvectors, such that $\mathbf{w}_k^{\top}\mathbf{w}_k = 1$.
4. Then the **principal component representation** of the data is

$$\mathbf{y}_j = \sum_{k=1}^{N} c_{k,j}\,\mathbf{w}_k, \tag{8.110}$$

   where

$$c_{k,j} = \mathbf{w}_k^{\top}\mathbf{y}_j. \tag{8.111}$$

Equation (8.110) is an exact representation of the data, and there is no residual error. Every data vector $\mathbf{y}_j$ gets mapped into a coefficient vector $\mathbf{c}_j = [c_{1,j}, \ldots, c_{N,j}]^{\top}$. This coefficient vector also has N dimensions.

To achieve some dimensional reduction, let us cut off the sum in equation (8.110) after the first D terms:

$$\mathbf{y}_j^{(D)} = \sum_{k=1}^{D} c_{k,j} = \mathbf{m} + \sum_{k=1}^{D} \mathbf{w}_k^{\top}\,\mathbf{y}_j\,\mathbf{w}_k. \tag{8.112}$$

Each data point is now mapped onto just D coefficients. The resulting approximation $\mathbf{x}_j^{(D)}$ corresponds to the orthogonal projection of $\mathbf{x}_j$ onto the space spanned by the principal components $\mathbf{w}_1, \ldots, \mathbf{w}_D$ (see the red dots in figure 8.18). **This D-dimensional approximation to the data in equation** (8.112) **is guaranteed to have the smallest possible residual.** That is the special property of the principal component representation.

How large is the error incurred by going from N to D dimensions? The **total variance of the data set** is

$$V = \frac{1}{T} \sum_{j=1}^{T} \mathbf{y}_j^\top \mathbf{y}_j. \tag{8.113}$$

This variance is equal to the sum of all the eigenvalues λ_k:

$$V = \sum_{k=1}^{N} \lambda_k. \tag{8.114}$$

Furthermore, the error $E^{(D)}$ of the D-dimensional approximation in equation (8.107) is the sum of the "unused" eigenvalues:

$$E^{(D)} = \sum_{k=D+1}^{N} \lambda_k. \tag{8.115}$$

This is also called the **unexplained variance** of the D-dimensional approximation. Vice versa, one says the **explained variance** is

$$V^{(D)} = \sum_{k=1}^{D} \lambda_k. \tag{8.116}$$

How should one choose D? Of course, this depends on the research goals that motivated the PCA. The trade-off is between lower dimensionality (low D) and lower error (high D). One often shows a plot of λ_k versus k, called the **eigenvalue spectrum** or **scree plot** (figure 8.19). If that plot shows a sudden break to lower eigenvalues, that can be a reason to set the cut-off D at the break.

We illustrate these procedures with two example data sets.

8.5.1.1 Example: Spearman's data In 1904, Spearman published "'General Intelligence,' Objectively Determined and Measured." This paper has historical importance as the first notable application of factor analysis, a close relative of PCA. Second, it put the psychological concept of "general intelligence" on a quantitative basis. Figure 8.20 reproduces just one data set from this study. The boys in an English village school were ranked according to their performance in various subjects. Then Spearman measured something seemingly unrelated—namely, their ability to distinguish sounds of different pitches, as well as lights of different intensities and weights of different mass.

Using the notation introduced here, each boy j is represented by a data vector $\mathbf{x}_j$ corresponding to a row in the table that contains the grades in the various subjects

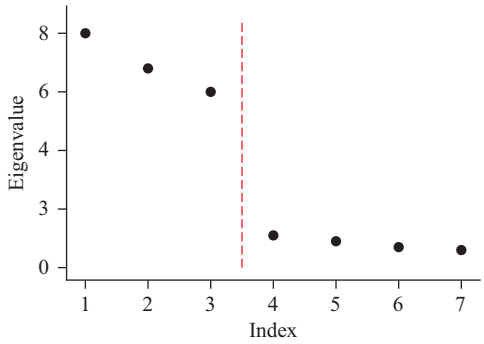

Figure 8.19
Sample scree plot, which suggests keeping only the first three principal components.

EXPERIMENTAL SERIES IV.

High Class Preparatory School for Boys.

A. Original Data.

Age		Pitch	Place in School (*before modification to eliminate Age*).												Music
			Classics			French			English			Mathem.			
Years	Months	Discrim. Thres. in ⅛ v. d. October, 1902	Xmas, 1902	Easter, 1903	July, 1903	Xmas, 1902	Easter, 1903	July, 1903	Xmas, 1902	Easter, 1903	July, 1903	Xmas, 1902	Easter, 1903	July, 1903	Ranked by Music Master
12	6	2	8	7	4	5	3	3	4	3	3	4	2	3	8
12	4	3	11	12	10	13	13	10	13	13	11	12	13	11	9
9	8	3	19	18	15	21	19	16	23	21	18	21	19	17	6
13	7	4	2	2	1	2	2	1	2	2	1	7	7	7	3
10	4	4	21		19	22		23	22		20	21		24	16
10	7	4	23	23	22	26	23	22	28	25	23	29	25	23	1
13	6	5	3			3			3			3			21
11	10	5	6	4	3	7	6	5	6	6	2	9	8	6	
10	1	5	29	26	24	23	25	21	27	26	22	25	23	19	7
11	1	6	20	20	18	20	21	18	21	20	19	17	16	15	14
13	4	7	1	1		1	1		1	1		1	1		5
10	6	7	26	24	21	27	16	13	26	19	17	22	18	16	11
12	3	7	18	17	16	17	20	19	25	23	21	19	17	14	20
13	1	8	5	5	5	4	4	2	5	8	5	5	4	1	4
11	1	10	22	19	17	19	18	17	20	17	15	23	21	21	18
9	9	10	33	29	27	33	29	27	33	27	27	32	29	27	17
10	4	11	28	25	23	30	27	24	18	18	13	30	27	22	
13	0	11	4	3	2	6	5	4	7	4	4	2	3	4	
10	2	11	7	6	6	12	7	6	8	5	8	11	9	8	
13	0	11	12	11	11	11	11	12	15	16	16	6	5	2	12
12	0	11	17	16		16	15		24	22		24	24		15
12	11	12	9	8	7	8	8	7	9	7	7	14	12	12	
13	1	14	10	9	8	10	9	8	11	10	9	10	10	9	13
10	4	14	27	21	14	24	22	15	17	11	10	26	20	18	2
10	1	15	24	22	20	18	17	14	29	24	24	18	15	13	
12	6	15	14	13	12	15	14	11	10	9	6	8	6	5	10
10	8	15	30	27		29	26		30	29		28	26		
12	8	18	16	15	13	25	24	20	14	14	12	20	21	20	19
9	5	20	32		25	31		25	32		26	33		26	
11	2	24	15	14	9	14	12	9	16	15	14	13	11	10	
10	9	50	25			28			19			15			
10	11	> 60	31	28	26	32	28	26	31	28	25	31	28	25	22
13	7	> 60	13	10		9	10		12	12		16	14		

Figure 8.20
An excerpt from Spearman's 1904 data set.

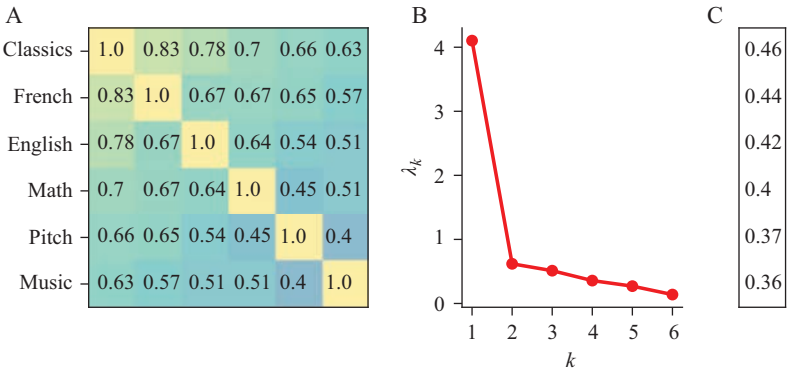

Figure 8.21
Principal component analysis of data in Spearman (1904), "Experimental Series IV." A: Correlation matrix of the scores of $T = 33$ boys in $N = 6$ school subjects including pitch discrimination. B: Scree plot of the eigenvalues λ_k. The first eigenvalue accounts for 68 percent of the variance. C: Coefficients of the first principal component arranged as in panel A.

and sensory tests. Figure 8.21 presents a principal component analysis of Spearman's "Experimental Series IV": The correlation matrix **C** (figure 8.21A) shows strong covariance of the scores across all subjects, including pitch discrimination. In other words, the typical boy tended to fare well or poorly in all subjects. This is reflected in the eigenvalue spectrum (figure 8.21B) which shows that a single principal component accounts for 68 percent of the variance. The coefficients of that component (figure 8.21C) are indeed positive along all the subjects. [5]

Spearman concluded that a single factor explains much of the boys' performance in class, but also on seemingly unrelated tests of perceptual discrimination. That factor eventually became known as "general intelligence."

8.5.1.2 Example: Dimensionality reduction of neuronal population activity Consider the example data set shown in figure 8.22A, which corresponds to the normalized activity of 50 neurons[6]. It is an $N \times T$ data matrix $\mathbf{Y}$, with N being the number of neurons and T being the number of time points. The activity is measured using a calcium indicator, a fluorescent probe expressed inside neurons that increases its fluorescence when the calcium concentration increases. Because the intracellular concentration of calcium increases when a neuron is active, the fluorescence can be used as a proxy for neuronal activity.

At first sight, one gets the impression that several neurons share the same pattern of activity. Instead of considering 50 neurons, can we collect then into a smaller number of "neuronal components" that are linear combinations of the original neurons

5. The literature on PCA suffers from a good amount of redundant and confusing nomenclature, including terms like "factors," "weights," "loadings," and "scores." In this book, we use the term "principal component" to refer to one of the eigenvectors of the covariance matrix. We use "PC coefficient" for the coefficient of a data point along that principal component.
6. Normalized activity implies that the fluorescence time series for each neuron is mean subtracted and has unit standard deviation—see section (8.5.1.3).

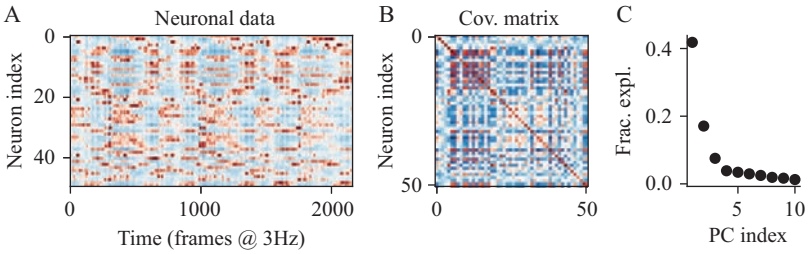

Figure 8.22
A: The activity of 50 neurons reported by a genetically encoded calcium indicator as a function of frame number, where frames were collected at 3 Hz. B: The covariance matrix computed from (A). C: Fraction of the variance explained by each of the first ten principal components.

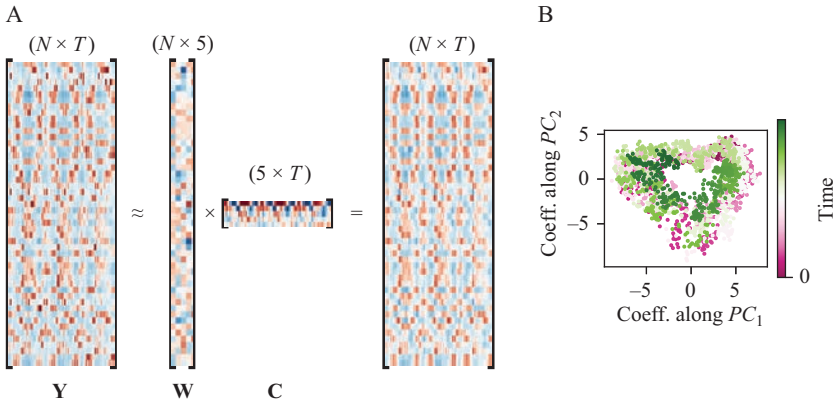

Figure 8.23
Dimensionality reduction of the activity of $N = 50$ neurons using PCA. A: The images illustrate the matrices $\mathbf{Y} \approx \mathbf{W}\,\mathbf{C}$. The dimensions of the matrices are shown on the top. The first principal component is the first column of $\mathbf{W}$. The coefficient of the data along the first principal component is given by the first row of $\mathbf{C}$. This row has larger coefficients than those in the second row, and so on. This is a manifestation that the first PC explains more of the variance than the second PC, and so on. B: We now keep just the first two neuronal components. The temporal activity of the 50 neurons is reduced to a trajectory in this 2D space. The time course is color-coded from magenta to green.

that explain most of the activity? In order to test this, one might perform principal component analysis.

Following the procedure described in the beginning of this section referring to 8.5.1, we can compute the covariance matrix, shown in figure 8.22B. The eigenvectors of this matrix are the neuronal principal components, and the eigenvalues report the variance explained by each one. We plot these in the scree plot shown in figure 8.22C. The first five principal components together contribute almost 75 percent of the variance in the data set (with the first two contributing almost 59 percent).

Using the first five principal components we can dimensionally reduce our data as shown in figure 8.23A. This shows that although although there are 50 neurons, they are strongly correlated in their time-varying activity. About 75 percent of the variance in

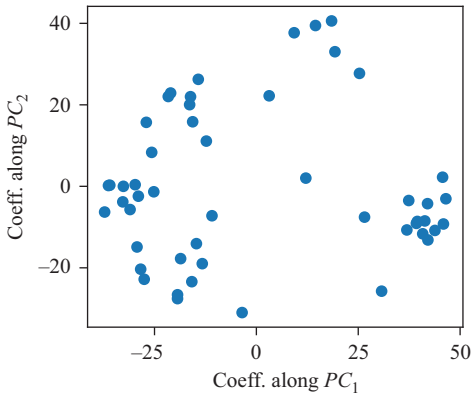

Figure 8.24
Scatterplot of the 50 neurons showing their coefficients along the first two temporal principal components.

their activity occurs in a five-dimensional subspace, and 59 percent in just two dimensions. In figure 8.23B we plot the activity as a trajectory in the space spanned by the first two principal components. Here, we can observe that it circles around the origin with changing angular velocity. [7]

What we have performed here is **neuronal PCA**. We could have also performed **temporal PCA** by computing the covariance matrix of the N-dimensional data vectors $\mathbf{y}_j$ (one for each time point j) and then performing PCA on that matrix. That covariance matrix is a $T \times T$-dimensional matrix so its eigenvectors, the temporal principal components, would be T-dimensional. Temporal PCA would give us the time courses that explain most of the variance across all the neurons. In fact, we will cluster the neurons according to the coefficients of their fluorescence along the first two temporal principal components in section 8.5.3 (see also figure 8.24).

8.5.1.3 Normalization and other preprocessing in PCA Sometimes, the components x_i that make up the data vector $\mathbf{x} = [x_1, \ldots, x_N]^\top$ represent very different variables. For example, Spearman's measurement of pitch discrimination obviously uses a different scale from the grades of the mathematics teacher. In the example involving neuronal data, the fluorescence of every neuron will depend on its size and how much fluorophore it expresses. In other cases, the measurements may be of entirely different physical quantities, like temperature and precipitation. Obviously, one needs to account for such differences in units before computing the covariance matrix.

One popular method for normalization adjusts the scale on each variable so they all have the same sample variance in the data set. This is known as **z-scoring** the data: subtract the sample mean and divide by the sample standard deviation. This leads to a preprocessed data set:

$$\mathbf{y}_j = [y_{1,j}, \ldots, y_{N,j}]^\top, \tag{8.117}$$

7. These data are from Petrucco et al. (2023) and in fact correspond to the heading direction neuronal network of larval zebrafish, which keeps track of the direction the fish is heading toward as it swims.

where

$$y_{i,j} = (x_{i,j} - m_i)/s_i, \tag{8.118}$$

and

$$m_i = \frac{1}{T}\sum_j x_{i,j}, \tag{8.119}$$

is the sample mean of the ith component and

$$s_i = \sqrt{\frac{1}{T}\sum_j (x_{i,j} - m_i)^2} \tag{8.120}$$

is the sample standard deviation of the ith component. Mathematically, this is equivalent to using the correlation matrix rather than the covariance matrix for the eigenvalue analysis. This is the method that we used for Spearman's data.

A different approach comes from considering experimental uncertainties: If component x_i of the data vector is affected by measurement error σ_i, then it may make sense to normalize each component by its uncertainty. That is, preprocess the data to

$$y_{i,j} = (x_{i,j} - m_i)/\sigma_i. \tag{8.121}$$

In that case, the total residual E in equation (8.107) takes on the character of a χ^2 statistic (see section 7.6.1), such that minimizing E is like maximizing the likelihood.

Other preprocessing steps include nonlinear transforms. For example, if x_i has a log-normal distribution in the data set, then a logarithmic transform $y_i = \log x_i$ will produce a more nearly normal variable, and thus a nicer shape to the data cloud.

Whatever preprocessing steps you apply, try to understand why you are doing it and what the consequences are for structures that might appear in the processed data.

8.5.2 Other Dimensionality Reduction Techniques: NNMF and ICA

Both linear regression and PCA can be interpreted as reducing the dimension of an $N \times T$ data matrix $\mathbf{X}$ according to

$$\mathbf{X} \approx \mathbf{WC}, \tag{8.122}$$

where $\mathbf{W}$ is $N \times D$ and $\mathbf{C}$ is $D \times T$. The rows of $\mathbf{C}$ form the basis of the new D-dimensional space and $\mathbf{W}$ are the components of the data in terms of this basis.

Linear regression and PCA impose different constraints on the basis set C. Linear regression minimizes the unexplained variance in the dependent variables, whereas PCA minimizes the total unexplained variance along all dimensions. PCA leads to an orthogonal basis set, consisting of the principal components, which can be computed as the eigenvectors of the correlation matrix.

There are versions of dimensionality reduction that call for different conditions on the basis vectors. For example, **independent component analysis (ICA)** tries to explain the data as the weighted sum of a small number of signals, just like PCA, but

in this case, with the requirement that these signals should be statistically independent of each other, as opposed to uncorrelated, which was the condition imposed by PCA. Another version, called **nonnegative matrix factorization (NNMF)**, imposes the constraint that the matrices **W** and **C** contain no negative elements. This arises when the data in question are naturally constrained nonnegative, such as intensities, concentrations, neuronal firing rates, and probabilities. These methods don't come with an analytical closed-form solution, like PCA, but efficient algorithms exist for deriving the components numerically.

8.5.3 Clustering

Dimensionality reduction simplifies the data distribution by allowing us to focus on a subspace of the original measurement space. However, within that subspace, the data are still widely distributed. An even greater simplification could be accomplished if the data form discrete clusters within the subspace. Hence there is broad interest in techniques that identify discrete clusters in a distribution.

Continuing with the 50-neuron data set here, we now replot the neuronal responses in the space of the first two temporal principal components (figure 8.24).[8] Note that these two components already explain 59 percent of the variance in the data set. One does get the vague impression that the points bunch together in certain regions of the space.

A popular method to identify clusters is called **k-means clustering**. This algorithm asks the user what number k of clusters should be found, and given k, it identifies the optimal allocation of data points to k clusters. Its criterion is to minimize the sum of the squared distances between every point and the cluster centroid to which it is assigned.

Choosing the number of clusters for k-means is not a trivial matter. One way is to apply the **elbow method**, similar to the interpretation of scree plots in PCA: repeat the analysis for a range of cluster numbers k, plot the sum of the squared distances s as a function of k, and check whether there is a critical number k beyond which s no longer decreases very much. This transition from a steep decrease in $s(k)$ to a more gradual decrease is called "an elbow." Figure 8.25 shows $s(k)$ for the neuronal data presented here. In this case, the elbow is not obviously apparent, so we show the clustering into three and six clusters, respectively, to serve as a comparison. Figure 8.26 shows that each of the clusters represents a different time course of neural activity.

8.5.3.1 Example: Otsu's image background separation method Image processing is important in many biological applications, and many experiments rely on the correct quantification of "particles" within these images. An important step in this analysis is separating the background of the image from the "particles" that need to be counted. These "particles" can be fluorescent cells, stained mitochondria, or birds flying in the sky.

Otsu's method is a simple method that uses k-means clustering to do that. It assumes that the pixel intensities cluster into two groups, one corresponding to the background and the other to the foreground. [9] Figure 8.27 shows how this is applied.

---

8. Note that in section (8.5.1.2), we performed neuronal PCA, in which each PC was a linear combination of the 50 neurons. Here, we have performed temporal PCA, in which each PC is a fluorescence time series, and the fluorescence time series of each individual neuron can be expressed as a linear combination of these PCs.

9. Note that this is a nontrivial assumption. A continuum of values can always be clustered into two clusters, although this does not necessarily mean that there are two clusters.

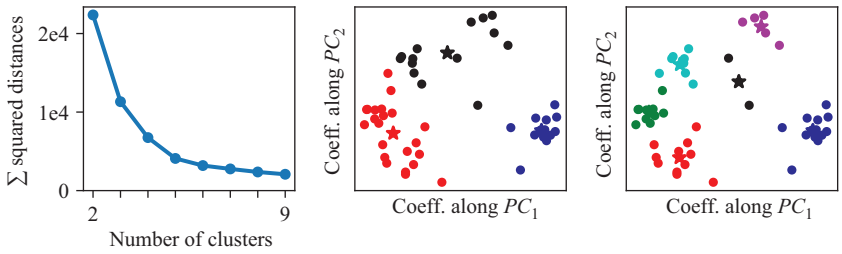

Figure 8.25
K-means clustering of the 50 neurons in the space of the first two PCs. Left: Unexplained variance as a function of the number of clusters. Middle: Three clusters and their centroids (stars). Right: Six clusters and their centroids (stars).

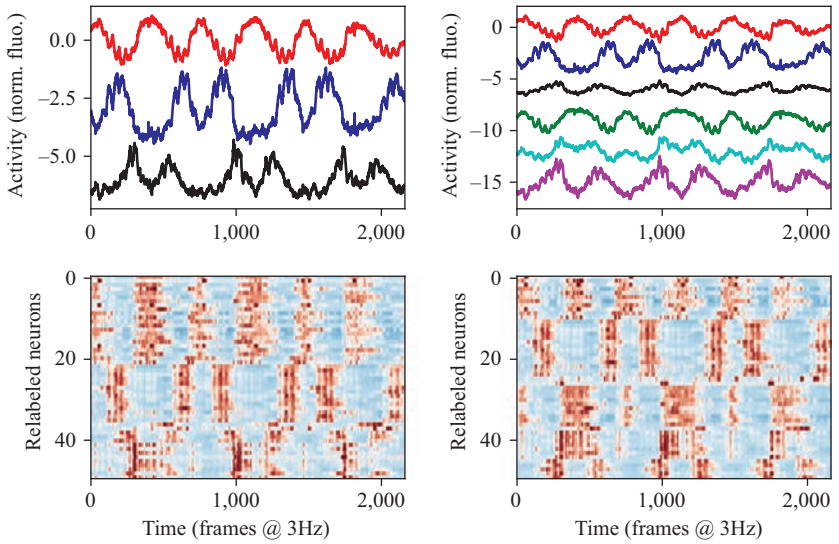

Figure 8.26
Activity traces corresponding to the cluster centroids of the figure 8.25, for both three and six clusters. Activity traces of neurons are reordered in such a way that neurons belonging to the same cluster appear together for the three and six clusters shown here.

There are many extensions of this simple algorithm; the easiest one to understand is a two-Gaussian mixture model (discussed next). Nevertheless, they all rely to some extent on clustering pixel intensities, taking into account assumptions on the distributions of these intensities or the morphology of the particles of interest.

8.5.3.2 Other clustering algorithms The k-means clustering algorithm presented here is by no means the only one. In fact, k-means implicitly assumes certain features that are not always assured, such as that all clusters have a spherical shape and the same variance, and similar numbers of members.

Other clustering methods relax some of these assumptions. Gaussian mixture models (GMMs), for example, do not assume a spherical distribution or equal variance. They return the probability that each sample belongs to each cluster. Nevertheless, GMMs

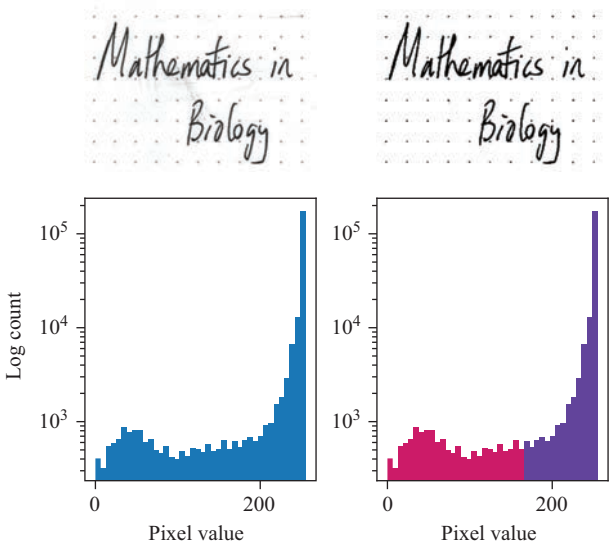

Figure 8.27
Left: Grayscale image (top) and the associated histogram (bottom). Right: Otsu's method assumes that pixel intensities cluster into two values, corresponding to background and signal, respectively. It divides the pixel intensities into these two clusters (shown in magenta and purple) and returns a binarized image (top).

do use the expected cluster number as an input into the algorithm, which needs to be decided a priori or explored empirically, just as in k-means.

Other algorithms, such as hierarchical clustering, return a branching tree where each sample is a leaf. By considering clusters to be small twigs, or small or large branches, the data can be separated into different numbers of clusters in a graded fashion. It is also possible to use more exotic distance metrics, other than the standard Euclidean metric.

On the whole, clustering is a bit of an art form. It is always possible to perform clustering, even when the data are drawn from a continuous distribution. A number of quantitative criteria have been proposed to evaluate the significance of clusters. In practice, it is important to find some visualization of the clusters, so the user can gain intuition for the results and evaluate visually whether the clusters make sense and can be interpreted usefully for the research purpose at hand.

8.6 Information Theory

Life is an interplay of energy, entropy, and information.

—Eigen (2019)

Public discourse these days is awash with loose talk of "information," often with numbers thrown in the mix, measured in gigabits or terabytes. Typically, this arises when a message needs to be transmitted from one place to another, such as to stream a television show on your monitor or to store a large document in a file. Such signal transmission also is ubiquitous in biological systems: the genome communicates through cellular machinery to specify the cell's proteome; cells signal to each other in the course

of development or an immune response; the eye signals to the brain with messages about our visual surroundings; the brain signals to muscles in order to respond. Understanding these processes (and more) benefits from a rigorous quantitative treatment of this substance called **information**.

Fundamentally, information leads to the **removal of uncertainty** (Shannon and Weaver, 1964). For illustration, consider a simple children's game: A says "I am thinking of a number between 1 and 16." B has to find the number by asking A yes/no questions. At the outset, B is uncertain about the number. With every question (if appropriately posed), B gains more information until there is no remaining uncertainty, and B knows the number exactly.

How many questions does B need to ask? An inefficient strategy would be: "Is it 1?," "Is it 2?," and so on. On average, this requires about eight questions to achieve success. A more streamlined approach is to ask each question so that it splits the remaining range of possible answers in half, such as starting with "Is it 8 or less?" This strategy requires only four questions to achieve success. In general, if A's number ranges from 1 to 2^n, then n questions are needed to nail it down precisely. In this case, we say that B had an uncertainty about A's number equal to n bits. During the question-and-answer game, that uncertainty was completely removed, so A transferred n bits of information to B.

8.6.1 Entropy

This leads us to a quantitative definition of uncertainty: The uncertainty about a random variable X is equal to the minimal number of yes/no questions required to determine X precisely. The uncertainty is also called **entropy**, denoted as $H(X)$ and is measured in units of bits.

If $X \in \{x_1, \ldots, x_n\}$ is a discrete random variable and all the outcomes are equally likely a priori, as in the guessing game, then

$$H(X) = \log_2 n. \tag{8.123}$$

Generally, X does not follow a uniform distribution. For example, X might be the outcome of a roll of two dice (as discussed in section 6.3.1). If X follows the probability mass function $P(X)$, then

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i). \tag{8.124}$$

Note that equation (8.123) is just a special case of equation (8.124).

This choice for measuring **entropy** is the only mathematical expression that satisfies two common-sense expectations of such a measure: (1) It should be positive. (2) If two variables X and Y are statistically independent, the uncertainty about both should be the sum of the uncertainties about each individually.

Example 8.7 (Bernoulli distribution) Remember that a Bernoulli random variable is one that can take two outcomes, with probability p and $1-p$, respectively, such as the outcome of a coin toss (as in section 6.3.5). If $X \sim \text{Bern}(p)$, then the entropy is (figure 8.28)

$$H(X; p) = -p \log_2 p - (1-p) \log_2 (1-p). \tag{8.125}$$

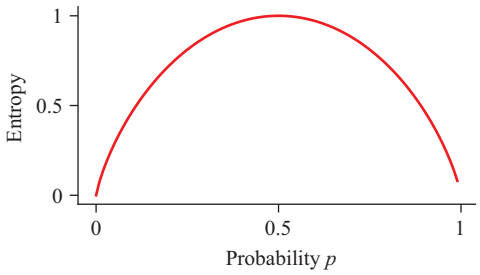

Figure 8.28
The entropy of a Bernoulli variable with bias p.

For what value of p is this entropy a maximum? Of course, one can look for the extremum of $H(X;p)$ with respect to p. For a simpler argument, note that the entropy must be symmetric under the exchange of p with $1-p$. Therefore, the maximum has to be at $p=0.5$, when the two values are equally likely. On the other hand, when $p=0$ or $p=1$, then the value of X is certain and the entropy $H(X)=0$. For these calculations, it is useful to remember that

$$x \log x \xrightarrow[x\to 0]{} 0. \tag{8.126}$$

□

Example 8.8 (What is the entropy of English?) It is well worth reading Shannon's paper on this topic (Shannon, 1951). He asks: When reading an English text, what is the uncertainty about the next character on the page?

A naive estimate goes as follows: Ignoring spaces and puctuation, there are 26 letters, so using equation (8.123), the entropy $H(C)$ of the next character C is

$$H_0(C) = \log_2 26 = 4.70 \text{ bits}. \tag{8.127}$$

However, the characters don't appear at equal frequency, so one should measure those frequencies and use the more general expression in equation (8.124) to get

$$H_1(C) = -\sum_{i=1}^{26} p_i \log_2 p_i = 4.08 \text{ bits}. \tag{8.128}$$

As it turns out, consecutive characters are not independent of each other: Certain letter pairs (like "QU") happen much more often than expected from the product of their individual frequencies. By tabulating the frequencies of letter pairs, one gets to $H_2(C)=3.56$. Shannon pursues this further, estimating the frequencies of tri-grams and words, and with each step obtains a lower entropy estimate. Eventually, he engages human subjects in a guessing game: after reading 100 characters in a book, they must guess the next one. From the number of guesses required, Shannon estimated the true entropy of English as somewhere between 0.6 and 1.3 bits/character:

$$0.6 < H_\infty(C) < 1.3. \tag{8.129}$$

This exercise foreshadows two interesting insights: First, the entropy of a symbol string depends not only on the frequency of each symbol, but on the correlations across symbols. Second, it should be possible to store English text in a very efficient way: the calculations suggest that one only needs about 1 bit per character. Instead, the popular ASCII code for English uses 8 bits per character—a great waste of bits. □

Example 8.9 (The entropy of DNA) Genetic material is stored in chromosomes, each one consisting of a long macromolecule of double-stranded DNA. Each strand of DNA is a long sequence of nucleotides that have one of four possible values: adenine (A), thymine (T), cytosine (C), or guanine (G). What is the entropy per base-pair of a long DNA sequence?

As in the case of English, we can start with the naive estimate, assuming that all 4 nucleotides appear at equal frequency, $p_i = \frac{1}{4}$, where $i \in \{A, T, G, C\}$:

$$H_0 = -\sum p_i \log_2(p_i) = 2 \text{ bits per base-pair.} \tag{8.130}$$

In actuality, the four bases do not appear at equal frequency. For example, in human chromosome 11, one finds that $p_i = [0.289, 0.289, 0.211, 0.211]$. With that knowledge, the entropy is

$$H_1 = -\sum p_i \log_2(p_i) = 1.9822 \text{ bits per base-pair.} \tag{8.131}$$

Further, it turns out that two successive nucleotides (di-grams) are not statistically independent. Again, one can estimate the frequencies of di-grams from the human chromosome 11 data to find a lower estimate:

$$H_2 = 1.9350 \text{ bits per base-pair.} \tag{8.132}$$

As in the case of English, knowledge of the statistical structure of the signal serves to reduce the uncertainty.

In coding regions of the chromosome, DNA carries the instructions for assembling amino acids into proteins. You can further explore the information-theoretic aspects of this genetic code in exercise 10.21. □

8.6.2 Communication Channel

Shannon (1948) formalized the process of communication between a source and a destination as follows (figure 8.29):

On the source side, the message gets encoded into a signal to be conveyed on a channel. During transmission on the channel, that signal may be corrupted by noise. On the destination side, the received signal gets translated back into an interpretable message. The context of Shannon's work was telecommunications, so the channel in question was typically a telephone transmission line or a wireless connection. Each of these channels suffers from a different kind of noise corruption. Ideally, the transmitters and receivers must be adapted to the kind of noise encountered, so as to allow error-free transmission regardless.

As it turns out, this framework lends itself to illuminate a vast number of phenomena, including many cases of signaling and communication in biology.

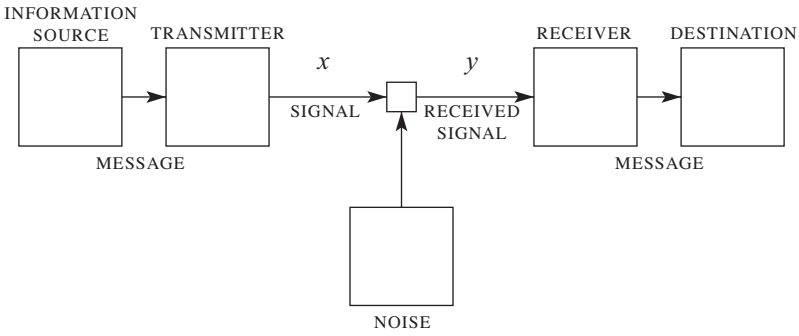

Figure 8.29
Shannon's framework for communication.

8.6.3 Mutual Information

Within this framework, suppose that the transmitter puts signal X on the channel and the receiver observes signal Y. Because of transmission noise, it is possible that $X \neq Y$. Suppose that X and Y follow a joint distribution $P_{XY}(X, Y)$. How much information about X does the receiver get from observing Y?

As discussed previously, information corresponds to a reduction of uncertainty. Prior to receiving signal Y, the receiver's uncertainty about X is the entropy $H(X)$, which depends only on the marginal probability distribution $P_X(X)$:

$$H(X) = -\sum_{x \in X} P_X(x) \log_2 P_X(x). \tag{8.133}$$

After the receiver sees the particular symbol $Y = y$, the probability distribution of X shifts from $P_X(X)$ to $P_{X|Y}(X|Y = y)$—namely, the probability conditional on observation of y. Now the remaining entropy is

$$H(X|y) = -\sum_{x \in X} P_{X|Y}(x|y) \log_2 P_{X|Y}(x|y) \tag{8.134}$$

and the information gained in the process is

$$I(X|y) = H(X) - H(X|y). \tag{8.135}$$

To assess the average gain over many transmissions, one averages this expression over all possible outcomes of Y:

$$\begin{aligned} I(X, Y) &= \sum_{y \in Y} P_Y(y) I(X|y) \\ &= -\sum_{x \in X,\, y \in Y} P_{XY}(x, y) \log_2 \left(\frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \right). \end{aligned} \tag{8.136}$$

The quantity $I(X, Y)$ is called the **mutual information** between the random variables X and Y. It tells us how much uncertainty about X is removed by measuring Y. Note that the expression for $I(X, Y)$ is symmetric in X and Y. So the information gained by the receiver about the transmitted message is equal to the information that the transmitter has about what appears at the receiver.

Note the special case in which X and Y are statistically independent. Then the joint probability factors into the product of the marginal probabilities as in equation (6.5.4), so the log term vanishes and the mutual information is zero. This is the extreme of a lousy communication channel, in which noise completely dominates the signal.

8.6.3.1 Mutual information for continuous variables Suppose that the symbol X placed on the channel and the receiver symbol Y are both continuous random variables, such as an electric voltage. Now the joint distribution of X and Y is a probability density function $P(x, y)$. The mutual information extends in a straightforward way by simply converting the sum to an integral:

$$I(X, Y) = -\int_{x,y} P_{XY}(x, y) \log_2 \left(\frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \right) \mathrm{d}x\mathrm{d}y. \tag{8.137}$$

8.6.4 Channel Capacity

The **capacity** of a communications channel is the maximum rate of information that can be transmitted down the channel, measured either in bits per symbol or bits per unit time. This maximum value of the mutual information between output and input is taken over all possible distributions of the signals, given some constraint.

8.6.4.1 Example: Binary channel with noise For example, consider the transmission of binary symbols $x \in \{0, 1\}$ across a noisy channel that occasionally changes a 0 into a 1 or vice versa. Suppose that the probability of error (of either kind) is q. So the probability of the output Y conditional on the input X is

$$P_{Y|X}(y|x) = \begin{cases} 1 - q, & \text{if } y = x \\ q, & \text{if } y \neq x. \end{cases} \tag{8.138}$$

To optimize the use of this channel, we have only one degree of freedom—namely, the fraction of time we use 0 and 1 for X. Because the channel properties are symmetric with respect to swapping 0 and 1, the only plausible optimum is when we use both symbols at equal frequency[10]. In that case, the channel capacity becomes

$$C = I(X, Y)_{\text{opt}} = 1 + q \log_2 q + (1 - q) \log(1 - q). \tag{8.139}$$

In this case, the constraint is that the signal can be either 0 or 1, and the free parameter is the fraction of 0s and 1s that are present. You can explore what happens if the channel's errors are asymmetric in exercise 10.20.

8.6.4.2 Example: Gaussian channel with noise Now consider a continuous channel with Gaussian noise. The signals X and Y are continuous variables, and the channel adds a random noise with Gaussian distribution, such that $y = x + n$ with $n \sim \mathcal{N}(0, N)$. Also, we suppose that the transmitter has limited signal power, so the variance of X is fixed: $\mathrm{Var}[X] = S^2$. Then one can show that the capacity is

$$C = \frac{1}{2} \log_2 \left(1 + \frac{S^2}{N^2} \right). \tag{8.140}$$

10. You can save a lot of effort with symmetry arguments like this.

To transmit at this limit, the optimal symbol distribution of X is Gaussian with variance S: $X \sim \mathcal{N}(0, S)$.

Note the result in equation (8.140) is logarithmic in the signal-to-noise (SNR) ratio on the channel. It has a simple interpretation: For large $\frac{S^2}{N^2}$, $C \approx \log_2 \frac{S}{N}$. But $\frac{S}{N}$ is the number of signal levels that are separated by the noise amplitude. So C is simply the $\log_2$ of the number of distinguishable signals (i.e. the number of bits needed to specify the signal to within the noise amplitude).

8.6.4.3 Redundancy The redundancy R of a communication link is the degree to which it fails to use the full channel capacity. Redundancy is generally a consequence of nonoptimal symbol use, namely, when the transmitter fails to encode the message appropriately for the channel. Redundancy is expressed as a fraction of the capacity wasted:

$$R = 1 - \frac{I(X, Y)}{C}. \tag{8.141}$$

For example, the ASCII code that uses eight binary digits to encode a character is a rather inefficient representation of English. Using Shannon's estimate that the entropy of English is about 1 bit/character, one concludes that the ASCII code has a redundancy of $R \approx 7/8$.

8.6.5 The Channel Coding Theorem

So far, we have mostly engaged in definitions of information-theoretic quantities. But what is the payback for using this way of measuring information? One powerful result is the **channel coding theorem**: Given a noisy channel with capacity C, one can use it to transmit error-free messages at an information rate up to C.

It seems counterintuitive that one can use a noisy channel for error-free communication at all. Obviously, this requires an ingenious encoding and decoding scheme that makes the message robust to the kinds of disruption that occurs on the channel. Notably, the theorem does not spell out how to achieve this, and much engineering effort goes into devising clever encoders and decoders to match messages to a channel. However, the theorem tells you when to stop trying. In many cases, it is easy to compute the capacity of the channel (see sections 8.6.4.1 and 8.6.4.2), and you can stop improving your encoders once the information rate that they support gets close to C.

In a number of biological applications, it is possible to compute the capacity C for a given signaling pathway, starting from an understanding of signal and noise in the system. Then one can ask how much information actually passes through that channel. In a few cases, the information rate seems to approach the capacity of the channel, suggesting a certain optimization of the encoding and decoding mechanisms (Tkacik and Bialek, 2016).

8.6.6 The Data-Processing Inequality

Consider a signaling chain from X to Z through an intermediate signal Y:

$$X \to Y \to Z. \tag{8.142}$$

Along such a chain, the mutual information can only decrease. In particular,

$$I(X,Z) < I(X,Y) \quad \text{and} \quad I(X,Z) < I(Y,Z). \tag{8.143}$$

In other words, along a signaling chain information can only be lost, not created. This has consequences for the capacity: if a signaling chain should have capacity C, then each individual link must have capacity $\geq C$.

In biology, we find that a signal frequently changes its physical identity along the way. For example, communication in the nervous system alternates between electrical voltage across the membrane, calcium concentration at a synapse, neurotransmitter concentration in the synaptic cleft, ionic current into the dendrite, and back to membrane voltage. Information theory is agnostic to the physical embodiment of a signal, and this is one of its chief attractions. At every stage along the way, one can measure rates and capacities in the universal unit of bits, and interpretation of those results is supported by the theorems of information theory.

8.6.7 Further Reading

The founding document of information theory is still one of the most readable introductions. Shannon and Weaver (1964) present the original papers with additional didactic material. A good technical reference is Cover and Thomas (2012), and a survey of biological applications can be found in Tkacik and Bialek (2016). Nelson (2022) explains many of the physical mechanisms by which living organisms gain information about their surroundings.

References

Abramowitz, M., and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 9th Dover printing, 10th GPO printing edition.

Alon, U. (2019). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. 2nd ed. Chapman and Hall/CRC.

American Institute of Physics. (1973). Interview of Richard Feynman by Charles Weiner on 1973 February 4, Niels Bohr Library & Archives, American Institute of Physics, College Park, MD USA. www.aip.org/history-programs/niels-bohr-library/oral-histories/5020-5

Bartol, T. M., Bromer, C., Kinney, J. P., et al. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife*, 4.

Bastian, J., Schniederjan, S., and Nguyenkim, J. (2001). Arginine vasotocin modulates a sexually dimorphic communication behavior in the weakly electric fish Apteronotus leptorhynchus. *Journal of Experimental Biology*, 204(11): 1909–1923.

Baylor, D. A., Lamb, T. D., and Yau, K. W. (1979). Responses of retinal rods to single photons. *The Journal of Physiology*, 288: 613–634.

Berry, M. J., Warland, D. K., and Meister, M. (1997). The structure and precision of retinal spike trains. *Proc Natl Acad Sci U S A*, 94: 5411–6.

Block, S. M., Segall, J. E., and Berg, H. C. (1982). Impulse responses in bacterial chemotaxis. *Cell*, 31: 215–26.

Borst, A., and Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience*, 2(11): 947–957.

Brown, E. N., Kass, R. E., and Mitra, P. P. (2004). Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nature Neuroscience*, 7(5): 456–461.

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21): 161–173.

Brush, S. G. (1968). A history of random processes: I. Brownian movement from Brown to Perrin. *Archive for History of Exact Sciences*, 5(1): 1–36.

Carslaw, H. S., and Jaeger, J. C. (1986). *Conduction of Heat in Solids*. 2nd ed. Oxford University Press.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network*, 12: 199–213.

Chichilnisky, E. J., and Kalmar, R. S. (2002). Functional asymmetries in ON and OFF ganglion cells of primate retina. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 22(7): 2737–2747.

Clark, D. A., Benichou, R., Meister, M., and Azeredo da Silveira, R. (2013). Dynamical adaptation in photoreceptors. *PLoS Computational Biology*, 9: e1003289.

Council, N. R. (2003). *BIO2010: Transforming Undergraduate Education for Future Research Biologists*. National Academies Press.

Cover, T. M., and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

Crevier, D. W., and Meister, M. (1998). Synchronous period-doubling in flicker vision of salamander and man. *Journal of Neurophysiology*, 79(4): 1869–1878. PMID: 9535954.

Czeisler, C. A., Richardson, G. S., Coleman, R. M., et al. (1981). Chronotherapy: Resetting the circadian clocks of patients with delayed sleep phase insomnia. *Sleep*, 4(1): 1–21.

Dahlquist, F. W., Lovely, P., and Koshland, D.E. (1972). Quantitative Analysis of Bacterial Migration in Chemotaxis. *Nature New Biology*, 236: 120–123.

Daley, D. J., and Vere-Jones, D. (2013). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. 2nd ed. Springer (2003).

Dayan, P., and Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

de Ruyter van Steveninck, R., Bialek, W., and Barlow, H. B. (1988). Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 234(1277): 379–414.

Digman, M. A., and Gratton, E. (2011). Lessons in fluctuation correlation spectroscopy. *Annual Review of Physical Chemistry*, 62(1): 645–668.

Drenth, J. (2007). *Principles of Protein X-ray Crystallography*. Springer.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Dymarski, P. (2011). *Hidden Markov Models, Theory and Applications*. IntechOpen.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, 22(10): 1315–1316.

Edwards, A. W. (1986). Are Mendel's results really too close? *Biological Reviews of the Cambridge Philosophical Society*, 61(4): 295–312.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Eigen, M. (2019). *From Strange Simplicity to Complex Familiarity: A Treatise on Matter, Information, Life and Thought*. Oxford University Press.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8): 549–560.

Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767): 335–338.

Entrenas Castillo, M., Entrenas Costa, L. M., Vaquero Barrios, J. M., et al. (2020). Effect of calcifediol treatment and best available therapy versus best available therapy on intensive care unit admission and mortality among patients hospitalized for COVID-19: A pilot randomized clinical study. *Journal of Steroid Biochemistry and Molecular Biology*, 203: 105751.

Fain, G. L., Matthews, H. R., Cornwall, M. C., and Koutalos, Y. (2001). Adaptation in vertebrate photoreceptors. *Physiological Reviews*, 81: 117–151.

Fisher, R. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1(2): 115–137.

Foster, R. G. (2020). Sleep, circadian rhythms and health. *Interface Focus*, 10(3): 20190098.

Franklin, A., Edwards, A. W. F., Fairbanks, D. J., and Hartl, D. L. (2008). *Ending the Mendel-Fisher Controversy*. University of Pittsburgh Press.

Frey, T. G., Chan, S. H., and Schatz, G. (1978). Structure and orientation of cytochrome c oxidase in crystalline membranes. Studies by electron microscopy and by labeling with subunit-specific antibodies. *Journal of Biological Chemistry*, 253(12): 4389–4395.

Geffen, M. N., Broome, B. M., Laurent, G., and Meister, M. (2009). Neural encoding of rapidly fluctuating odors. *Neuron*, 61: 570–586.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press.

Goentoro, L., Shoval, O., Kirschner, M. W., and Alon, U. (2009). The incoherent feedforward loop can provide fold-change detection in gene regulation. *Molecular Cell*, 36(5): 894–899.

Goodman, J. (2005). *Introduction to Fourier Optics*. 3rd ed. Roberts & Co.

Hamada, H., Watanabe, M., Lau, H. E., et al. (2014). Involvement of delta/notch signaling in zebrafish adult pigment stripe patterning. *Development*, 141(2): 318–324.

Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). Energy, quanta, and vision. *Journal of General Physiology*, 25: 819–840.

Henderson, J., Salzberg, S., and Fasman, K. H. (1997). Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, 4(2): 127–141.

Hille, B. (2001). *Ion Channels of Excitable Membranes*. 3rd ed. Sinauer Associates.

Hiscock, T. W., and Megason, S. G. (2015). Mathematically guided approaches to distinguish models of periodic patterning. *Development*, 142(3): 409–419.

Hodgkin, A. (1994). *Chance and Design: Reminiscences of Science in Peace and War*. Illustrated ed. Cambridge University Press.

Hodgkin, A., and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544.

Josef, K., Saranak, J., and Foster, K. W. (2005). Ciliary behavior of a negatively phototactic Chlamydomonas reinhardtii. *Cell Motility and the Cytoskeleton*, 61(2): 97–111.

Kanji, G. K. (2006). *100 Statistical Tests*. 3rd ed. SAGE Publications.

Kawasaki, M., Rose, G., and Heiligenberg, W. (1988). Temporal hyperacuity in single neurons of electric fish. *Nature*, 336: 173–176.

Konopka, R. J., and Benzer, S. (1971). Clock mutants of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 68(9): 2112–2116.

Lipson, E. D. (1975). White noise analysis of Phycomyces light growth response system. I. Normal intensity range. *Biophysical Journal*, 15(10): 989–1011.

Liu, R. T., Liaw, S. S., and Maini, P. K. (2006). Two-stage Turing model for generating pigment patterns on the leopard and the jaguar. *Physical Review E*, 74: 011914.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2): 130–141.

Ludwig, D., Jones, D. D., and Holling, C. S. (1978). Qualitative analysis of insect outbreak systems: The spruce budworm and forest. *Journal of Animal Ecology*, 47(1): 315–332.

Luria, S. E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6): 491–511.

Machens, C. K., Romo, R., and Brody, C. D. (2005). Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science*, 307(5712): 1121–1124.

MacKinnon, R. (2004). Potassium channels and the atomic basis of selective ion conduction. *Bioscience Reports*, 24(2): 75–100.

Mainen, Z. F., and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science (New York, N.Y.)*, 268(5216): 1503–1506.

Meister, M., and Tessier-Lavigne, M. (2021). Low-level visual processing: The retina. In Kandel, E., Koester, J. D., Mack, S. H., and Siegelbaum, S., editors, *Principles of Neural Science* 6th ed. (pp. 521–544). McGraw Hill / Medical.

Müller, P., Rogers, K. W., Jordan, B. M., et al. (2012). Differential diffusivity of Nodal and Lefty underlies a reaction-diffusion patterning system. *Science (New York, N.Y.)*, 336(6082): 721–724.

Murray, J. D. (2002). *Mathematical Biology: I. An Introduction*. 3rd ed. Interdisciplinary Applied Mathematics, Mathematical Biology. Springer-Verlag, New York.

Murray, J. (2003). *Mathematical Biology II: Spatial Models and Biomedical Applications*. Springer.

Neher, E., and Stevens, C. F. (1977). Conductance fluctuations and ionic pores in membranes. *Annual Review of Biophysics and Bioengineering*, 6(1): 345–381.

Nelson, P. (2022). *Physical Models of Living Systems: Probability, Simulation, Dynamics*. 2nd ed. Philadelphia: Chiliagon Science.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4): 243–262.

Petrucco, L., Lavian, H., Wu, Y., Svara, F., Stih, V., and Portugues, R. (2023). "Neural dynamics and architecture of the heading direction circuit in zebrafish". *Nature Neuroscience*, 26: 765–773.

Phillips, R., Kondev, J., Theriot, J., and Garcia, H. G. (2012). *Physical Biology of the Cell*. 2nd ed. CRC Press.

Pitkow, X. and Meister, M. (2014). Neural computation in sensory systems. In Gazzaniga, M. S. and Mangun, G. R., editors, *The Cognitive Neurosciences* 5th ed. (pp. 305–318). MIT Press.

Pugh, E. N. Jr. (2018). The discovery of the ability of rod photoreceptors to signal single photons. *Journal of General Physiology*, 150(3): 383–388.

Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*. 3rd ed. Cengage Learning.

Rieke, F., Bodnar, D. A., and Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings: Biological Sciences*, 262(1365): 259–265.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: Exploring the Neural Code*. MIT Press.

Roubinet, B., Bischoff, M., Nizamov, S., et al. (2018). Photoactivatable rhodamine spiroamides and diazoketones decorated with "universal hydrophilizer" or hydroxyl groups. *Journal of Organic Chemistry*, 83(12): 6466–6476.

Sakitt, B. (1972). Counting every quantum. *Journal of Physiology*, 223: 131–150.

Sakmann, B. and Neher, E., eds. (1995). *Single-Channel Recording*. 2nd ed. Springer.

Schnapf, J. L., Nunn, B. J., Meister, M., and Baylor, D. A. (1990). Visual transduction in cones of the monkey Macaca fascicularis. *Journal of Physiology*, 427: 681–713.

Schneeweis, D. M., and Schnapf, J. L. (2000). Noise and light adaptation in rods of the macaque monkey. *Visual Neuroscience*, 17(5): 659–666.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1): 50–64.

Shannon, C. E., and Weaver, W. (1964). *The Mathematical Theory of Communication*. University of Illinois Press.

Smith, A., Frank, L., Wirth, S., et al. (2004). Dynamic analysis of learning in behavioral experiments. *Journal of Neuroscience*, 24(2): 447–461.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15: 201–293.

Spellman, P., Sherlock, G., Zhang, M., et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9: 3273–3297.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, and Engineering*. 2nd ed. CRC Press.

Svoboda, K., Schmidt, C. F., Schnapp, B. J., and Block, S. M. (1993). Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 365: 721–727.

Tkacik, G., and Bialek, W. (2016). Information processing in living systems. *Annual Review of Condensed Matter Physics*, 7(1): 89–117.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, 237(641): 37–72.

Weinstein, M. S. (1977). Hares, lynx, and trappers. *The American Naturalist*, 111(980): 806–808.

Werner, G., and Mountcastle, V. B. (1965). Neural activity in mechanoreceptive cutaneous afferents: Stimulus-response relations, Weber functions, and information transmission. *Journal of Neurophysiology*, 28: 359–397.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. Wiley.

Index

Page numbers followed by *f* denote figures, *t* denote tables.