What is statistical mechanics?

1

Many systems in nature are far too complex to analyze directly. Solving for the behavior of all the atoms in a block of ice, or the boulders in an earthquake fault, or the nodes on the Internet, is simply infeasible. Despite this, such systems often show simple, striking behavior. Statistical mechanics explains the simple behavior of complex systems.

The concepts and methods of statistical mechanics have infiltrated into many fields of science, engineering, and mathematics: ensembles, entropy, Monte Carlo, phases, fluctuations and correlations, nucleation, and critical phenomena are central to physics and chemistry, but also play key roles in the study of dynamical systems, communications, bioinformatics, and complexity. Quantum statistical mechanics, although not a source of applications elsewhere, is the foundation of much of physics. Let us briefly introduce these pervasive concepts and methods.

Ensembles. The trick of statistical mechanics is not to study a single system, but a large collection or *ensemble* of systems. Where understanding a single system is often impossible, one can often calculate the behavior of a large collection of similarly prepared systems.

For example, consider a random walk (Fig. 1.1). (Imagine it as the trajectory of a particle in a gas, or the configuration of a polymer in solution.) While the motion of any given walk is irregular and typically impossible to predict, Chapter 2 derives the elegant laws which describe the set of all possible random walks.

Chapter 3 uses an ensemble of all system states of constant energy to derive equilibrium statistical mechanics; the collective properties of temperature, entropy, and pressure emerge from this ensemble. In Chapter 4 we provide the best existing mathematical justification for using this constant-energy ensemble. In Chapter 6 we develop *free energies* which describe parts of systems; by focusing on the important bits, we find new laws that emerge from the microscopic complexity.

Entropy. Entropy is the most influential concept arising from statistical mechanics (Chapter 5). It was originally understood as a thermodynamic property of heat engines that inexorably increases with time. Entropy has become science's fundamental measure of disorder and information—quantifying everything from compressing pictures on the Internet to the heat death of the Universe.

Quantum statistical mechanics, confined to Chapter 7, provides the microscopic underpinning to much of astrophysics and condensed

Statistical Mechanics: Entropy, Order Parameters, and Complexity. James P. Sethna, Oxford University Press (2021). © Oxford University Press. DOI:10.1093/oso/9780198865247.003.0001

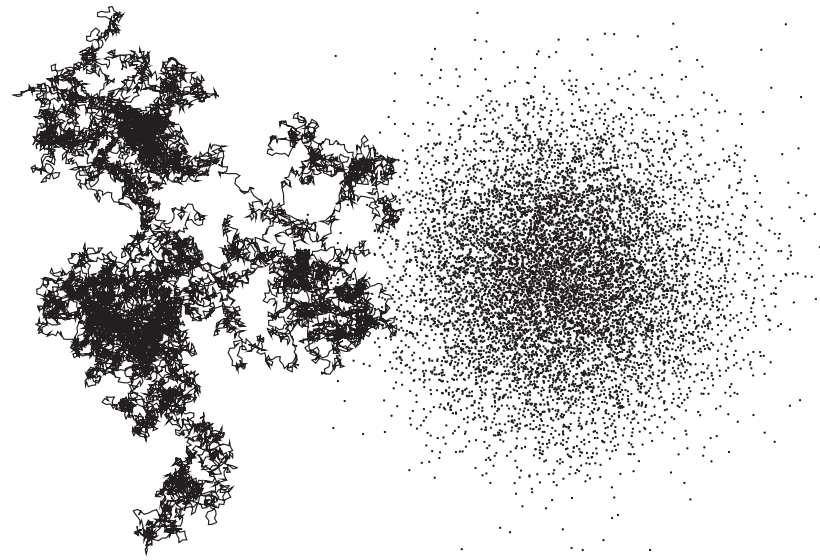

Fig. 1.1 Random walks. The motion of molecules in a gas, and bacteria in a liquid, and photons in the Sun, are described by *random walks*. Describing the specific trajectory of any given random walk (left) is not feasible. Describing the statistical properties of a large number of random walks is straightforward (right, showing endpoints of many walks starting at the origin). The deep principle underlying statistical mechanics is that it is often easier to understand the behavior of these *ensembles* of systems.

matter physics. There we use it to explain metals, insulators, lasers, stellar collapse, and the microwave background radiation patterns from the early Universe.

Monte Carlo methods allow the computer to find ensemble averages in systems far too complicated to allow analytical evaluation. These tools, invented and sharpened in statistical mechanics, are used everywhere in science and technology—from simulating the innards of particle accelerators, to studies of traffic flow, to designing computer circuits. In Chapter 8, we introduce Monte Carlo methods, the Ising model, and the mathematics of Markov chains.

Phases. Statistical mechanics explains the existence and properties of phases. The three common phases of matter (solids, liquids, and gases) have multiplied into hundreds: from superfluids and liquid crystals, to vacuum states of the Universe just after the Big Bang, to the pinned and sliding "phases" of earthquake faults. We explain the deep connection between phases and perturbation theory in Section 8.3. In Chapter 9 we introduce the *order parameter field*, which describes the properties, excitations, and topological defects that emerge in a given phase.

Fluctuations and correlations. Statistical mechanics not only describes the average behavior of an ensemble of systems, it describes the entire distribution of behaviors. We describe how systems fluctuate and evolve in space and time using *correlation functions* in Chapter 10. There we also derive powerful and subtle relations between correlations, response, and dissipation in equilibrium systems.

Abrupt phase transitions. Beautiful spatial patterns arise in statistical mechanics at the transitions between phases. Most such transitions are abrupt; ice is crystalline and solid until (at the edge of the ice cube) it becomes unambiguously liquid. We study the nucleation

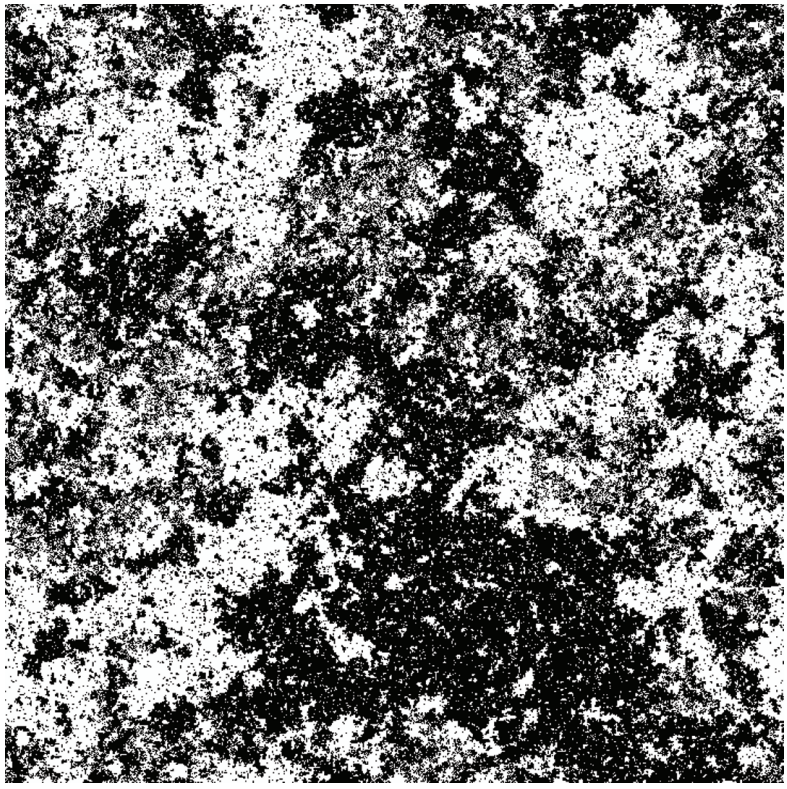

Fig. 1.2 Ising model at the critical point. The two-dimensional Ising model of magnetism at its transition temperature T_c. At higher temperatures, the system is nonmagnetic; the magnetization is on average zero. At the temperature shown, the system is just deciding whether to magnetize upward (white) or downward (black).

of new phases and the exotic structures that can form at abrupt phase transitions in Chapter 11.

Criticality. Other phase transitions are continuous. Figure 1.2 shows a snapshot of a particular model at its phase transition temperature T_c. Notice the self-similar, fractal structures; the system cannot decide whether to stay gray or to separate into black and white, so it fluctuates on all scales, exhibiting *critical phenomena*. A random walk also forms a self-similar, fractal object; a blow-up of a small segment of the walk looks statistically similar to the original (Figs. 1.1 and 2.2). Chapter 12 develops the scaling and renormalization-group techniques that explain these self-similar, fractal properties. These techniques also explain *universality*; many properties at a continuous transition are surprisingly system independent.

Science grows through accretion, but becomes potent through distillation. Statistical mechanics has grown tentacles into much of science and mathematics (see, e.g., Fig. 1.3). The body of each chapter will provide the distilled version: those topics of fundamental importance to all fields. The accretion is addressed in the exercises: in-depth introductions to applications in mesoscopic physics, astrophysics, dynamical systems, information theory, low-temperature physics, statistics, biology, lasers, and complexity theory.

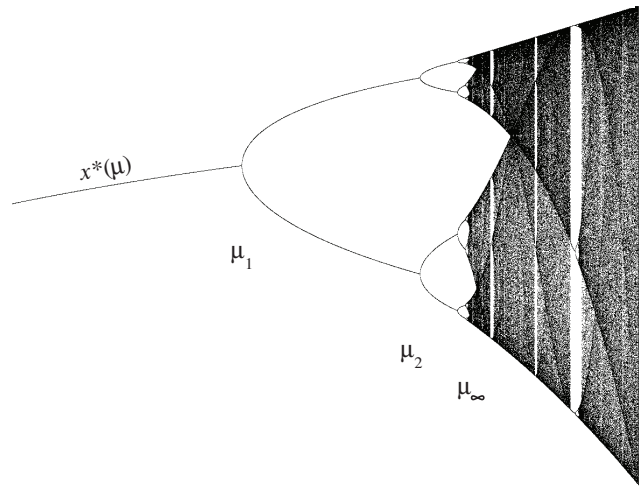

Fig. 1.3 The onset of chaos. Mechanical systems can go from simple, predictable behavior (left) to a chaotic state (right) as some external parameter μ is tuned. Many disparate systems are described by a common, *universal* scaling behavior near the onset of chaos (note the replicated structures near μ_∞). We understand this scaling and universality using tools developed to study continuous transitions in liquids and gases. Conversely, the study of chaotic motion provides statistical mechanics with our best explanation for the increase of entropy.

Exercises

Two exercises, *Emergence* and *Emergent vs. fundamental*, illustrate and provoke discussion about the role of statistical mechanics in formulating new laws of physics.

Four exercises review probability distributions. *Quantum dice and coins* explores discrete distributions and also acts as a preview to Bose and Fermi statistics. *Probability distributions* introduces the key distributions for continuous variables, convolutions, and multidimensional distributions. *Waiting time paradox* uses public transportation to concoct paradoxes by confusing different ensemble averages. And *The birthday problem* calculates the likelihood of a school classroom having two children who share the same birthday.

Stirling's approximation derives the useful approximation $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$; more advanced students can continue with *Stirling and asymptotic series* to explore the zero radius of convergence for this series, often found in statistical mechanics calculations.

Five exercises demand no background in statistical mechanics, yet illustrate both general themes of the subject and the broad range of its applications. *Random matrix theory* introduces an entire active field of research, with applications in nuclear physics, mesoscopic physics, and number theory, beginning with histograms and ensembles, and continuing with level repulsion, the Wigner surmise, universality, and emergent symmetry. *Six degrees of separation* introduces the ensemble of *small world networks*, popular in the social sciences and epidemiology for modeling interconnectedness in groups; it introduces network data structures, breadth-first search algorithms, a continuum limit, and our first glimpse of *scaling*. *Satisfactory map colorings* introduces the challenging computer science problems of *graph colorability* and *logical satisfiability*: these search through an ensemble of different choices just as statistical mechanics averages over an ensemble of states. *Self-propelled particles* discusses emergent properties of *active matter*. *First to fail: Weibull* introduces the statistical study of *extreme value statistics*, focusing not on the typical fluctuations about the average behavior, but the rare events at the extremes.

Finally, statistical mechanics is to physics as statistics is to biology and the social sciences. Four exercises here, and several in later chapters, introduce ideas and methods from statistics that have particular resonance with statistical mechanics. *Width of the height distribution* discusses maximum likelihood methods and bias in the context of Gaussian fits. *Fisher information and Cramér Rao* introduces the Fisher information metric, and its relation to the rigorous bound on parameter estimation. And *Distances in probability space* then uses the local difference between model predictions (the metric tensor) to generate total distance estimates between different models.

(1.1) **Quantum dice and coins.**[1] (Quantum) @
You are given two unusual three-sided dice which, when rolled, show either one, two, or three spots. There are three games played with these dice: *Distinguishable*, *Bosons*, and *Fermions*. In each turn in these games, the player rolls the two dice, starting over if required by the rules, until a legal combination occurs. In *Distinguishable*, all rolls are legal. In *Bosons*, a roll is legal only if the second of the two dice shows a number that is is larger or equal to that of the first of the two dice. In *Fermions*, a roll is legal only if the second number is strictly larger than the preceding number. See Fig. 1.4 for a table of possibilities after rolling two dice.
Our dice rules are the same ones that govern the quantum statistics of noninteracting identical particles.

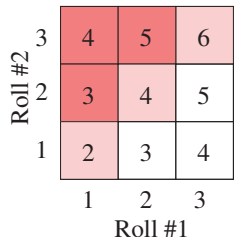

Fig. 1.4 Quantum dice. Rolling two dice. In *Bosons*, one accepts only the rolls in the shaded squares, with equal probability 1/6. In *Fermions*, one accepts only the rolls in the darkly shaded squares (not including the diagonal from lower left to upper right), with probability 1/3.

(a) *Presume the dice are fair: each of the three numbers of dots shows up* 1/3 *of the time. For a legal turn rolling a die twice in the three games (Distinguishable, Bosons, and Fermions), what is the probability* $\rho(5)$ *of rolling a 5?*
(b) *For a legal turn in the three games, what is the probability of rolling a double?* (Hint: There is a Pauli exclusion principle: when playing *Fermions*, no two dice can have the same number of dots showing.) Electrons are fermions; no two noninteracting electrons can be in the same quantum state. Bosons are gregarious (Exercise 7.9); noninteracting bosons have a larger likelihood of being in the same state.
Let us decrease the number of sides on our dice to $N = 2$, making them quantum coins, with a head H and a tail T. Let us increase the total number of coins to a large number M; we flip a line of M coins all at the same time, repeating until a legal sequence occurs. In the rules for legal flips of quantum coins, let us make $T < H$. A legal Boson sequence, for example, is then a pattern $TTTT\cdots HHHH\cdots$ of length M; all legal sequences have the same probability.
(c) *What is the probability in each of the games, of getting all the* M *flips of our quantum coin the same (all heads* $HHHH\cdots$ *or all tails* $TTTT\cdots$*)?* (Hint: How many legal sequences are there for the three games? How many of these are all the same value?)
The probability of finding a particular legal sequence in Bosons is larger by a constant factor due to discarding the illegal sequences. This factor is just one over the probability of a given toss of the coins being legal, $Z = \sum_\alpha p_\alpha$ summed over legal sequences α. For part (c), all sequences have equal probabilities $p_\alpha = 2^{-M}$, so $Z_{\text{Dist}} = (2^M)(2^{-M}) = 1$, and Z_{Boson} is 2^{-M} times the number of legal sequences. So for part (c), the probability to get all heads or all tails is $(p_{TTT\ldots}+p_{HHH\ldots})/Z$. The normalization constant Z in statistical mechanics is called the *partition function*, and will be amazingly useful (see Chapter 6).
Let us now consider a biased coin, with probability $p = 1/3$ of landing H and thus $1-p = 2/3$ of landing T. Note that if two sequences are legal in both Bosons and Distinguishable, their relative probability is the same in both games.
(d) *What is the probability* $p_{TTT\ldots}$ *that a given toss of* M *coins has all tails (before we throw out the illegal ones for our game)? What is* Z_{Dist}*? What is the probability that a toss in Distinguishable is all tails? If* Z_{Bosons} *is the probability that a toss is legal in Bosons, write the probability that a legal toss is all tails in terms of* Z_{Bosons}*. Write the probability* $p_{TTT\ldots HHH}$ *that a toss has* $M-m$ *tails followed by* m *heads (before throwing out the illegal ones). Sum these to find* Z_{Bosons}*. As* M *gets large, what is the probability in Bosons that all coins flip tails?*
We can view our quantum dice and coins as noninteracting particles, with the biased coin having a lower energy for T than for H (Section 7.4). Having a nonzero probability of having all the

[1] This exercise was developed in collaboration with Sarah Shandera.

bosons in the single-particle ground state T is Bose condensation (Section 7.6), closely related to superfluidity and lasers (Exercise 7.9).

(1.2) **Probability distributions.** ②

Most people are more familiar with probabilities for discrete events (like coin flips and card games), than with probability distributions for continuous variables (like human heights and atomic velocities). The three continuous probability distributions most commonly encountered in physics are: (i) *uniform*: $\rho_{\text{uniform}}(x) = 1$ for $0 \le x < 1$, $\rho(x) = 0$ otherwise (produced by random number generators on computers); (ii) *exponential*: $\rho_{\text{exponential}}(t) = \mathrm{e}^{-t/\tau}/\tau$ for $t \ge 0$ (familiar from radioactive decay and used in the collision theory of gases); and (iii) *Gaussian*: $\rho_{\text{gaussian}}(v) = \mathrm{e}^{-v^2/2\sigma^2}/(\sqrt{2\pi}\sigma)$, (describing the probability distribution of velocities in a gas, the distribution of positions at long times in random walks, the sums of random variables, and the solution to the diffusion equation).

(a) Likelihoods. *What is the probability that a random number uniform on $[0, 1)$ will happen to lie between $x = 0.7$ and $x = 0.75$? That the waiting time for a radioactive decay of a nucleus will be more than twice the exponential decay time τ? That your score on an exam with a Gaussian distribution of scores will be greater than 2σ above the mean?* (Note: $\int_2^\infty (1/\sqrt{2\pi}) \exp(-v^2/2)\,\mathrm{d}v = (1 - \mathrm{erf}(\sqrt{2}))/2 \sim 0.023$.)

(b) Normalization, mean, and standard deviation. *Show that these probability distributions are normalized: $\int \rho(x)\,\mathrm{d}x = 1$. What is the mean x_0 of each distribution? The standard deviation $\sqrt{\int (x - x_0)^2 \rho(x)\,\mathrm{d}x}$?* (You may use the formulæ $\int_{-\infty}^{\infty} (1/\sqrt{2\pi}) \exp(-v^2/2)\,\mathrm{d}v = 1$ and $\int_{-\infty}^{\infty} v^2 (1/\sqrt{2\pi}) \exp(-v^2/2)\,\mathrm{d}v = 1$.)

(c) Sums of variables. *Draw a graph of the probability distribution of the sum $x + y$ of two random variables drawn from a uniform distribution on $[0, 1)$. Argue in general that the sum $z = x + y$ of random variables with distributions $\rho_1(x)$ and $\rho_2(y)$ will have a distribution given by $\rho(z) = \int \rho_1(x)\rho_2(z - x)\,\mathrm{d}x$ (the* convolution *of ρ with itself).*

Multidimensional probability distributions. In statistical mechanics, we often discuss probability distributions for many variables at once (for example, all the components of all the velocities of all the atoms in a box). Let us consider just the probability distribution of one molecule's velocities. If v_x, v_y, and v_z of a molecule are independent and each distributed with a Gaussian distribution with $\sigma = \sqrt{kT/M}$ (Section 3.2.2) then we describe the combined probability distribution as a function of three variables as the product of the three Gaussians:

$$\begin{aligned}\rho(v_x, v_y, v_z) =& \frac{1}{(2\pi(kT/M))^{3/2}} \exp(-M\mathbf{v}^2/2kT) \\ =& \sqrt{\frac{M}{2\pi kT}} \exp\left(\frac{-Mv_x^2}{2kT}\right) \\ & \times \sqrt{\frac{M}{2\pi kT}} \exp\left(\frac{-Mv_y^2}{2kT}\right) \\ & \times \sqrt{\frac{M}{2\pi kT}} \exp\left(\frac{-Mv_z^2}{2kT}\right). \end{aligned} \tag{1.1}$$

(d) *Show, using your answer for the standard deviation of the Gaussian in part (b), that the mean kinetic energy is $kT/2$ per dimension. Show that the probability that the speed is $v = |\mathbf{v}|$ is given by a Maxwellian distribution*

$$\rho_{\text{Maxwell}}(v) = \sqrt{2/\pi}(v^2/\sigma^3) \exp(-v^2/2\sigma^2). \tag{1.2}$$

(Hint: What is the shape of the region in 3D velocity space where $|\mathbf{v}|$ is between v and $v + \delta v$? The surface area of a sphere of radius R is $4\pi R^2$.)

(1.3) **Waiting time paradox.**[2] ⓐ

Here we examine the *waiting time paradox*: for events happening at random times, the average time until the next event equals the average time between events. If the average waiting time until the next event is τ, then the average time since the last event is also τ. Is the mean total gap between two events then 2τ? Or is it τ, the average time to wait starting from the previous event? Working this exercise introduces the importance of different *ensembles*.

On a highway, the average numbers of cars and buses going east are equal: each hour, on average, there are 12 buses and 12 cars passing by. The buses are scheduled: each bus appears exactly 5 minutes after the previous one. On the other hand, the cars appear at random. In a short interval $\mathrm{d}t$, the probability that a car comes by is $\mathrm{d}t/\tau$, with $\tau = 5$ minutes. This

[2] The original form of this exercise was developed in collaboration with Piet Brouwer.

leads to a distribution $P(\delta^{\mathrm{Car}})$ for the arrival of the *first* car that decays exponentially, $P(\delta^{\mathrm{Car}}) = 1/\tau \exp(-\delta^{\mathrm{Car}}/\tau)$.

A pedestrian repeatedly approaches a bus stop at random times t, and notes how long it takes before the first bus passes, and before the first car passes.

(a) *Draw the probability density for the ensemble of waiting times δ^{Bus} to the next bus observed by the pedestrian. Draw the density for the corresponding ensemble of times δ^{Car}. What is the mean waiting time for a bus $\langle \delta^{\mathrm{Bus}} \rangle_t$? The mean time $\langle \delta^{\mathrm{Car}} \rangle_t$ for a car?*

In statistical mechanics, we shall describe specific physical systems (a bottle of N atoms with energy E) by considering *ensembles* of systems. Sometimes we shall use two different ensembles to describe the same system (all bottles of N atoms with energy E, or all bottles of N atoms at that temperature T where the mean energy is E). We have been looking at the time-averaged ensemble (the ensemble $\langle \cdots \rangle_t$ over random times t). There is also in this problem an ensemble average over the gaps between vehicles ($\langle \cdots \rangle_{\mathrm{gap}}$ over random time intervals); these two give different averages for the same quantity.

A traffic engineer sits at the bus stop, and measures an ensemble of time gaps Δ^{Bus} between neighboring buses, and an ensemble of gaps Δ^{Car} between neighboring cars.

(b) *Draw the probability density of gaps she observes between buses. Draw the probability density of gaps between cars.* (Hint: Is it different from the ensemble of car waiting times you found in part (a)? Why not?) *What is the mean gap time $\langle \Delta^{\mathrm{Bus}} \rangle_{\mathrm{gap}}$ for the buses? What is the mean gap time $\langle \Delta^{\mathrm{Car}} \rangle_{\mathrm{gap}}$ for the cars?* (One of these probability distributions involves the Dirac δ-function[3] if one ignores measurement error and imperfectly punctual public transportation.)

You should find that the mean waiting time for a bus in part (a) is half the mean bus gap time in (b), which seems sensible—the gap seen by the pedestrian is the sum of the $\delta_+^{\mathrm{Bus}} + \delta_-^{\mathrm{Bus}}$ of the waiting time and the time since the last bus. However, you should also find the mean waiting time for a car *equals* the mean car gap time. The equation $\Delta^{\mathrm{Car}} = \delta_+^{\mathrm{Car}} + \delta_-^{\mathrm{Car}}$ would seem to imply that the average gap seen by the pedestrian is twice the mean waiting time.

(c) *How can the average gap between cars measured by the pedestrian be different from that measured by the traffic engineer? Discuss.*

(d) *Consider a short experiment, with three cars passing at times $t = 0$, 2, and 8 (so there are two gaps, of length 2 and 6). What is $\langle \Delta^{\mathrm{Car}} \rangle_{\mathrm{gap}}$? What is $\langle \Delta^{\mathrm{Car}} \rangle_t$? Explain why they are different.*

One of the key results in statistical mechanics is that predictions are *independent of the ensemble* for large numbers of particles. For example, the velocity distribution found in a simulation run at constant energy (using Newton's laws) or at constant temperature will have corrections that scale as one over the number of particles.

(1.4) **Stirling's formula.** (Mathematics) @

Stirling's approximation, $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$, is remarkably useful in statistical mechanics; it gives an excellent approximation for large n. In statistical mechanics the number of particles is so large that we usually care not about $n!$, but about its logarithm, so $\log(n!) \sim n \log n - n + \frac{1}{2} \log(2\pi n)$. Finally, n is often so large that the final term is a tiny fractional correction to the others, giving the simple formula $\log(n!) \sim n \log n - n$.

(a) *Calculate* $\log(n!)$ *and these two approximations to it for $n = 2$, 4, and 50. Estimate the error of the simpler formula for $n = 6.03 \times 10^{23}$. Discuss the fractional accuracy of these two approximations for small and large n.*

Note that $\log(n!) = \log(1 \times 2 \times 3 \times \cdots \times n) = \log(1) + \log(2) + \cdots + \log(n) = \sum_{m=1}^{n} \log(m)$.

(b) *Convert the sum to an integral, $\sum_{m=1}^{n} \approx \int_0^n \mathrm{d}m$. Derive the simple form of Stirling's formula.*

(c) *Draw a plot of* $\log(m)$ *and a bar chart showing* $\log(\mathrm{ceiling}(n))$. (Here $\mathrm{ceiling}(x)$ represents the smallest integer larger than x.) *Argue that the integral under the bar chart is* $\log(n!)$. (Hint: Check your plot: between $x = 4$ and 5, $\mathrm{ceiling}(x) = 5$.)

The difference between the sum and the integral in part (c) should look approximately like a collection of triangles, except for the region between zero and one. The sum of the areas equals the error in the simple form for Stirling's formula.

[3] The δ-function $\delta(x-x_0)$ is a probability density which has 100% probability of being in any interval containing x_0; thus $\delta(x-x_0)$ is zero unless $x = x_0$, and $\int f(x)\delta(x - x_0)\,\mathrm{d}x = f(x_0)$ so long as the domain of integration includes x_0. Mathematically, this is not a function, but rather a distribution or a measure.

(d) *Imagine doubling these triangles into rectangles on your drawing from part (c), and sliding them sideways (ignoring the error for m between zero and one). Explain how this relates to the term* $\frac{1}{2}\log n$ *in Stirling's formula* $\log(n!) - (n\log n - n) \approx \frac{1}{2}\log(2\pi n) = \frac{1}{2}\log(2) + \frac{1}{2}\log(\pi) + \frac{1}{2}\log(n)$.

(1.5) **Stirling and asymptotic series.**[4] (Mathematics, Computation) ③

Stirling's formula (which is actually originally due to de Moivre) can be improved upon by extending it into an entire series. It is not a traditional Taylor expansion; rather, it is an *asymptotic series*. Asymptotic series are important in many fields of applied mathematics, statistical mechanics [171], and field theory [172].

We want to expand $n!$ for large n; to do this, we need to turn it into a continuous function, interpolating between the integers. This continuous function, with its argument perversely shifted by one, is $\Gamma(z) = (z-1)!$. There are many equivalent formulæ for $\Gamma(z)$; indeed, any formula giving an analytic function satisfying the recursion relation $\Gamma(z+1) = z\Gamma(z)$ and the normalization $\Gamma(1) = 1$ is equivalent (by theorems of complex analysis). We will not use it here, but a typical definition is $\Gamma(z) = \int_0^\infty \mathrm{e}^{-t} t^{z-1}\,\mathrm{d}t$; one can integrate by parts to show that $\Gamma(z+1) = z\Gamma(z)$.

(a) *Show, using the recursion relation* $\Gamma(z+1) = z\Gamma(z)$*, that* $\Gamma(z)$ *has a singularity (goes to* $\pm\infty$*) at zero and all the negative integers.*

Stirling's formula is extensible [18, p. 218] into a nice expansion of $\Gamma(z)$ in powers of $1/z = z^{-1}$:

$$\begin{aligned}\Gamma[z] &= (z-1)! \\ &\sim (2\pi/z)^{1/2}\mathrm{e}^{-z}z^z(1 + (1/12)z^{-1} \\ &\quad + (1/288)z^{-2} - (139/51840)z^{-3} \\ &\quad - (571/2488320)z^{-4} \\ &\quad + (163879/209018880)z^{-5} \\ &\quad + (5246819/75246796800)z^{-6} \\ &\quad - (534703531/902961561600)z^{-7} \\ &\quad - (4483131259/86684309913600)z^{-8} \\ &\quad + \ldots). \end{aligned} \tag{1.3}$$

This looks like a Taylor series in $1/z$, but is subtly different. For example, we might ask what the radius of convergence [174] of this series is. The radius of convergence is the distance to the nearest singularity in the complex plane (see note 26 on p. 227 and Fig. 8.7(a)).

(b) *Let* $g(\zeta) = \Gamma(1/\zeta)$*; then Stirling's formula is something times a power series in* ζ*. Plot the poles (singularities) of* $g(\zeta)$ *in the complex* ζ *plane that you found in part (a). Show that the radius of convergence of Stirling's formula applied to g must be zero, and hence no matter how large z is Stirling's formula eventually diverges.*

Indeed, the coefficient of z^{-j} eventually grows rapidly; Bender and Orszag [18, p. 218] state that the odd coefficients ($A_1 = 1/12$, $A_3 = -139/51840$, ...) asymptotically grow as

$$A_{2j+1} \sim (-1)^j 2(2j)!/(2\pi)^{2(j+1)}. \tag{1.4}$$

(c) *Show explicitly, using the ratio test applied to formula 1.4, that the radius of convergence of Stirling's formula is indeed zero.*[5]

This in no way implies that Stirling's formula is not valuable! An asymptotic series of length n approaches $f(z)$ as z gets big, but for fixed z it can diverge as n gets larger and larger. In fact, asymptotic series are very common, and often are useful for much larger regions than are Taylor series.

(d) *What is* 0!*? Compute* 0! *using successive terms in Stirling's formula (summing to* A_N *for the first few* N*).* Considering that this formula is expanding about infinity, it does pretty well! Quantum electrodynamics these days produces the most precise predictions in science. Physicists sum enormous numbers of Feynman diagrams to produce predictions of fundamental quantum phenomena. Dyson argued that quantum electrodynamics calculations give an asymptotic series [172]; the most precise calculation in science takes the form of a series which cannot converge. Many other fundamental expansions are also asymptotic series; for example, Hooke's law and elastic theory have zero radius of convergence [35, 36] (Exercise 11.15).

[4] Hints for the computations can be found at the book website [181].

[5] If you do not remember about radius of convergence, see [174]. Here you will be using every other term in the series, so the radius of convergence is $\lim_{j\to\infty}\sqrt{|A_{2j-1}/A_{2j+1}|}$.

(1.6) **Random matrix theory.**[6] (Mathematics, Quantum, Computation) ③

One of the most active and unusual applications of ensembles is *random matrix theory*, used to describe phenomena in nuclear physics, mesoscopic quantum mechanics, and wave phenomena. Random matrix theory was invented in a bold attempt to describe the statistics of energy level spectra in nuclei. In many cases, the statistical behavior of systems exhibiting complex wave phenomena—almost any correlations involving eigenvalues and eigenstates—can be quantitatively modeled using ensembles of matrices with completely random, uncorrelated entries!

The most commonly explored ensemble of matrices is the Gaussian orthogonal ensemble (GOE). Generating a member H of this ensemble of size $N \times N$ takes two steps.

- Generate an $N \times N$ matrix whose elements are independent random numbers with Gaussian distributions of mean zero and standard deviation $\sigma = 1$.
- Add each matrix to its transpose to symmetrize it.

As a reminder, the Gaussian or normal probability distribution of mean zero gives a random number x with probability

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-x^2/2\sigma^2}. \tag{1.5}$$

One of the most striking properties that large random matrices share is the distribution of level splittings.

(a) *Generate an ensemble with* $M = 1{,}000$ *or so GOE matrices of size* $N = 2$, 4, *and* 10. *(More is nice.) Find the eigenvalues* λ_n *of each matrix, sorted in increasing order. Find the difference between neighboring eigenvalues* $\lambda_{n+1} - \lambda_n$, *for* n, *say, equal to*[7] $N/2$. *Plot a histogram of these eigenvalue splittings divided by the mean splitting, with bin size small enough to see some of the fluctuations.* (Hint: Debug your work with $M = 10$, and then change to $M = 1{,}000$.)

What is this dip in the eigenvalue probability near zero? It is called *level repulsion*.

For $N = 2$ the probability distribution for the eigenvalue splitting can be calculated pretty simply. Let our matrix be $M = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$.

(b) *Show that the eigenvalue difference for* M *is* $\lambda = \sqrt{(c-a)^2 + 4b^2} = 2\sqrt{d^2 + b^2}$ *where* $d = (c-a)/2$, *and the trace* $c+a$ *is irrelevant. Ignoring the trace, the probability distribution of matrices can be written* $\rho_M(d, b)$. *What is the region in the* (b, d) *plane corresponding to the range of eigenvalue splittings* $(\lambda, \lambda + \Delta)$? *If* ρ_M *is continuous and finite at* $d = b = 0$, *argue that the probability density* $\rho(\lambda)$ *of finding an eigenvalue splitting near* $\lambda = 0$ *vanishes (level repulsion).* (Hint: Both d and b must vanish to make $\lambda = 0$. Go to polar coordinates, with λ the radius.)

(c) *Calculate analytically the standard deviation of a diagonal and an off-diagonal element of the GOE ensemble (made by symmetrizing Gaussian random matrices with* $\sigma = 1$*).* You may want to check your answer by plotting your predicted Gaussians over the histogram of H_{11} and H_{12} from your ensemble in part (a). *Calculate analytically the standard deviation of* $d = (c-a)/2$ *of the* $N = 2$ *GOE ensemble of part (b), and show that it equals the standard deviation of* b.

(d) *Calculate a formula for the probability distribution of eigenvalue spacings for the* $N = 2$ *GOE, by integrating over the probability density* $\rho_M(d, b)$. (Hint: Polar coordinates again.)

If you rescale the eigenvalue splitting distribution you found in part (d) to make the mean splitting equal to one, you should find the distribution

$$\rho_{\mathrm{Wigner}}(s) = \frac{\pi s}{2} \mathrm{e}^{-\pi s^2/4}. \tag{1.6}$$

This is called the *Wigner surmise*; it is within 2% of the correct answer for larger matrices as well.[8]

(e) *Plot eqn 1.6 along with your* $N = 2$ *results from part (a). Plot the Wigner surmise formula against the plots for* $N = 4$ *and* $N = 10$ *as well.*

Does the distribution of eigenvalues depend in detail on our GOE ensemble? Or could it be *universal*, describing other ensembles of real symmetric matrices as well? Let us define a ± 1 ensemble of real symmetric matrices, by generating an $N \times N$ matrix whose elements are independent random variables, ± 1 with equal probability.

[6] This exercise was developed with the help of Piet Brouwer. Hints for the computations can be found at the book website [181].

[7] Why not use all the eigenvalue splittings? The mean splitting can change slowly through the spectrum, smearing the distribution a bit.

[8] The distribution for large matrices is known and universal, but is much more complicated to calculate.

(f) *Generate an ensemble of* $M = 1{,}000$ *symmetric matrices filled with* ± 1 *with size* $N = 2$, 4, *and* 10. *Plot the eigenvalue distributions as in part (a). Are they universal (independent of the ensemble up to the mean spacing) for* $N = 2$ *and* 4*? Do they appear to be nearly universal*[9] *(the same as for the GOE in part (a)) for* $N = 10$*? Plot the Wigner surmise along with your histogram for* $N = 10$.

The GOE ensemble has some nice statistical properties. The ensemble is invariant under orthogonal transformations:

$$H \to R^\top H R \quad \text{with} \quad R^\top = R^{-1}. \tag{1.7}$$

(g) *Show that* $\mathrm{Tr}[H^\top H]$ *is the sum of the squares of all elements of* H. *Show that this trace is invariant under orthogonal coordinate transformations (that is,* $H \to R^\top H R$ *with* $R^\top = R^{-1}$*).* (Hint: Remember, or derive, the cyclic invariance of the trace: $\mathrm{Tr}[ABC] = \mathrm{Tr}[CAB]$.)

Note that this trace, for a symmetric matrix, is the sum of the squares of the diagonal elements plus *twice* the squares of the upper triangle of off-diagonal elements. That is convenient, because in our GOE ensemble the variance (squared standard deviation) of the off-diagonal elements is half that of the diagonal elements (part (c)).

(h) *Write the probability density* $\rho(H)$ *for finding GOE ensemble member* H *in terms of the trace formula in part (g). Argue, using your formula and the invariance from part (g), that the GOE ensemble is invariant under orthogonal transformations:* $\rho(R^\top H R) = \rho(H)$.

This is our first example of an *emergent symmetry*. Many different ensembles of symmetric matrices, as the size N goes to infinity, have eigenvalue and eigenvector distributions that are invariant under orthogonal transformations *even though the original matrix ensemble did not have this symmetry*. Similarly, rotational symmetry emerges in random walks on the square lattice as the number of steps N goes to infinity, and also emerges on long length scales for Ising models at their critical temperatures.

(1.7) **Six degrees of separation.**[10] (Complexity, Computation) ④

One of the more popular topics in random network theory is the study of how connected they are. *Six degrees of separation* is the phrase commonly used to describe the interconnected nature of human acquaintances: various somewhat uncontrolled studies have shown that any random pair of people in the world can be connected to one another by a short chain of people (typically around six), each of whom knows the next fairly well. If we represent people as nodes and acquaintanceships as neighbors, we reduce the problem to the study of the relationship network. Many interesting problems arise from studying properties of randomly generated networks. A network is a collection of *nodes* and *edges*, with each edge connected to two nodes, but with each node potentially connected to any number of edges (Fig. 1.5). A random network is constructed probabilistically according to some definite rules; studying such a random network usually is done by studying the entire ensemble of networks, each weighted by the probability that it was constructed. Thus these problems naturally fall within the broad purview of statistical mechanics.

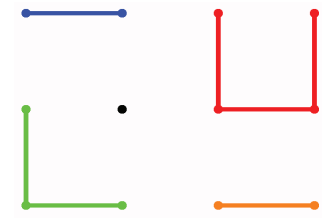

Fig. 1.5 Network. A network is a collection of nodes (circles) and edges (lines between the circles).

In this exercise, we will generate some random networks, and calculate the distribution of distances between pairs of points. We will study *small world networks* [140, 206], a theoretical model that suggests how a small number of shortcuts (unusual international and intercultural friendships) can dramatically shorten the typical chain lengths. Finally, we will study how a simple, universal scaling behavior emerges for large networks with few shortcuts.

[9] Note the spike at zero. There is a small probability that two rows or columns of our matrix of ± 1 will be the same, but this probability vanishes rapidly for large N.

[10] This exercise and the associated software were developed in collaboration with Christopher Myers. Hints for the computations can be found at the book website [181].

Constructing a small world network. The L nodes in a small world network are arranged around a circle. There are two kinds of edges. Each node has Z short edges connecting it to its nearest neighbors around the circle (up to a distance $Z/2$). In addition, there are $p \times L \times Z/2$ shortcuts added to the network, which connect nodes at random (see Fig. 1.6). (This is a more tractable version [140] of the original model [206], which rewired a fraction p of the $LZ/2$ edges.)

(a) *Define a network object on the computer. For this exercise, the nodes will be represented by integers. Implement a network class, with five functions:*

(1) `HasNode(node)`, *which checks to see if a node is already in the network;*

(2) `AddNode(node)`, *which adds a new node to the system (if it is not already there);*

(3) `AddEdge(node1, node2)`, *which adds a new edge to the system;*

(4) `GetNodes()`, *which returns a list of existing nodes; and*

(5) `GetNeighbors(node)`, *which returns the neighbors of an existing node.*

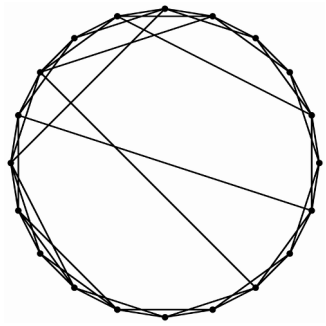

Fig. 1.6 Small world network *with* $L = 20$, $Z = 4$, *and* $p = 0.2$.[11]

Write a routine to construct a small world network, which (given L, Z, and p) adds the nodes and the short edges, and then randomly adds the shortcuts. Use the software provided to draw this small world graph, and check that you have implemented the periodic boundary conditions correctly (each node i should be connected to nodes $(i - Z/2) \bmod L, \ldots, (i + Z/2) \bmod L$).[12]

Measuring the minimum distances between nodes. The most studied property of small world graphs is the distribution of shortest paths between nodes. Without the long edges, the shortest path between i and j will be given by hopping in steps of length $Z/2$ along the shorter of the two arcs around the circle; there will be no paths of length longer than L/Z (halfway around the circle), and the distribution $\rho(\ell)$ of path lengths ℓ will be constant for $0 < \ell < L/Z$. When we add shortcuts, we expect that the distribution will be shifted to shorter path lengths.

(b) *Write the following three functions to find and analyze the path length distribution.*

(1) `FindPathLengthsFromNode(graph, node)`, *which returns for each* `node2` *in the graph the shortest distance from* `node` *to* `node2`. *An efficient algorithm is a* breadth-first traversal *of the graph, working outward from* `node` *in shells. There will be a* `currentShell` *of nodes whose distance will be set to ℓ unless they have already been visited, and a* `nextShell` *which will be considered after the current one is finished (looking sideways before forward, breadth first), as follows.*

- *Initialize $\ell = 0$, the distance from* `node` *to itself to zero, and* `currentShell = [node]`.
- *While there are nodes in the new* `currentShell`*:*
  - *start a new empty* `nextShell`*;*
  - *for each neighbor of each node in the current shell, if the distance to* `neighbor` *has not been set, add the node to* `nextShell` *and set the distance to $\ell + 1$;*
  - *add one to ℓ, and set the current shell to* `nextShell`.
- *Return the distances.*

This will sweep outward from `node`, measuring the shortest distance to every other node in the network. (Hint: Check your code with a network with small N and small p, compar-

[11] There are seven new shortcuts, where $pLZ/2 = 8$; one of the added edges overlapped an existing edge or connected a node to itself.

[12] Here $(i - Z/2) \bmod L$ is the integer $0 \le n \le L - 1$, which differs from $i - Z/2$ by a multiple of L.

ing a few paths to calculations by hand from the graph image generated as in part (a).)

(2) `FindAllPathLengths(graph)`, *which generates a list of all lengths (one per pair of nodes in the graph) by repeatedly using* `FindPathLengthsFromNode`. *Check your function by testing that the histogram of path lengths at* $p = 0$ *is constant for* $0 < \ell < L/Z$, *as advertised. Generate graphs at* $L = 1{,}000$ *and* $Z = 2$ *for* $p = 0.02$ *and* $p = 0.2$*; display the circle graphs and plot the histogram of path lengths. Zoom in on the histogram; how much does it change with* p*? What value of* p *would you need to get "six degrees of separation"?*

(3) `FindAveragePathLength(graph)`, *which computes the mean* $\langle \ell \rangle$ *over all pairs of nodes. Compute* ℓ *for* $Z = 2$, $L = 100$, *and* $p = 0.1$ *a few times; your answer should be around* $\ell = 10$. *Notice that there are substantial statistical fluctuations in the value from sample to sample. Roughly how many long bonds are there in this system? Would you expect fluctuations in the distances?*

(c) *Plot the average path length between nodes* $\ell(p)$ *divided by* $\ell(p = 0)$ *for* $Z = 2$, $L = 50$, *with* p *on a semi-log plot from* $p = 0.001$ *to* $p = 1$. (Hint: Your curve should be similar to that of with Watts and Strogatz [206, Fig. 2], with the values of p shifted by a factor of 100; see the discussion of the continuum limit below.) *Why is the graph fixed at one for small* p*?*

Large N *and the emergence of a continuum limit.* We can understand the shift in p of part (c) as a continuum limit of the problem. In the limit where the number of nodes N becomes large and the number of shortcuts $pLZ/2$ stays fixed, this network problem has a nice limit where distance is measured in radians $\Delta\theta$ around the circle. Dividing ℓ by $\ell(p = 0) \approx L/(2Z)$ essentially does this, since $\Delta\theta = \pi Z \ell / L$.

(d) *Create and display a circle graph of your geometry from part (c) (*$Z = 2$, $L = 50$*) at* $p = 0.1$*; create and display circle graphs of Watts and Strogatz's geometry (*$Z = 10$, $L = 1{,}000$*) at* $p = 0.1$ *and* $p = 0.001$. *Which of their systems looks statistically more similar to yours? Plot (perhaps using the scaling collapse routine provided) the rescaled average path length* $\pi Z \ell / L$ *versus the total number of shortcuts* $pLZ/2$*, for a range* $0.001 < p < 1$*, for* $L = 100$ *and* 200*, and for* $Z = 2$ *and* 4.

In this limit, the average bond length $\langle \Delta\theta \rangle$ should be a function only of M. Since Watts and Strogatz [206] ran at a value of ZL a factor of 100 larger than ours, our values of p are a factor of 100 larger to get the same value of $M = pLZ/2$. Newman and Watts [144] derive this continuum limit with a renormalization-group analysis (Chapter 12).

(e) *Real networks. From the book website [181], or through your own research, find a real network*[13] *and find the mean distance and histogram of distances between nodes.*

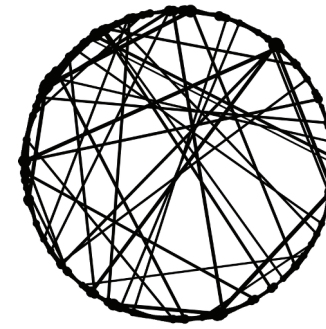

Fig. 1.7 Betweenness Small world network with $L = 500$, $K = 2$, and $p = 0.1$, with node and edge sizes scaled by the square root of their betweenness.

In the small world network, a few long edges are crucial for efficient transfer through the system (transfer of information in a computer network, transfer of disease in a population model, ...). It is often useful to measure how crucial a given node or edge is to these shortest paths. We say a node or edge is *between* two other nodes if it is along a shortest path between them. We measure the *betweenness* of a node or edge as the total number of such shortest paths passing through it, with (by convention) the initial and final nodes included in the count of between nodes; see Fig. 1.7. (If there are K multiple shortest paths of equal length between two nodes, each path adds $1/K$ to its intermediates.) The efficient algorithm to measure betweenness is a depth-first traversal quite analogous to the shortest-path-length algorithm discussed above.

[13] Examples include movie-actor costars, *Six degrees of Kevin Bacon*, or baseball players who played on the same team.

(f) *Betweenness (advanced). Read [68, 141], which discuss the algorithms for finding the betweenness. Implement them on the small world network, and perhaps the real world network you analyzed in part (e). Visualize your answers by using the graphics software provided on the book website [181].*

(1.8) **Satisfactory map colorings.**[14] (Computer science, Computation, Mathematics) ③

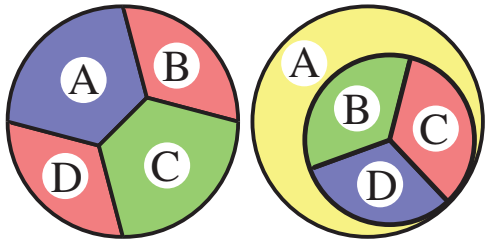

Fig. 1.8 Graph coloring. Two simple examples of graphs with $N = 4$ nodes that can and cannot be colored with three colors.

Many problems in computer science involve finding a good answer among a large number of possibilities. One example is *3-colorability* (Fig. 1.8). Can the N nodes of a graph be colored in three colors (say red, green, and blue) so that no two nodes joined by an edge have the same color?[15] For an N-node graph one can of course explore the entire ensemble of 3^N colorings, but that takes a time exponential in N. Sadly, there are no known shortcuts that fundamentally change this; there is no known algorithm for determining whether a given N-node graph is three-colorable that guarantees an answer in a time that grows only as a power of N.[16]

Another good example is *logical satisfiability* (**SAT**). Suppose one has a long logical expression involving N boolean variables. The logical expression can use the operations NOT ($\neg$), AND ($\wedge$), and OR ($\vee$). It is *satisfiable* if there is some assignment of *True* and *False* to its variables that makes the expression *True*. Can we solve a general satisfiability problem with N variables in a worst-case time that grows less quickly than exponentially in N? In this exercise, you will show that logical satisfiability is in a sense computationally at least as hard as 3-colorability. That is, you will show that a 3-colorability problem with N nodes can be mapped onto a logical satisfiability problem with $3N$ variables, so a polynomial-time (nonexponential) algorithm for the **SAT** would imply a (hitherto unknown) polynomial-time solution algorithm for 3-colorability.

If we use the notation A_R to denote a variable which is true when node A is colored red, then $\neg(A_R \wedge A_G)$ is the statement that node A is not colored both red and green, while $A_R \vee A_G \vee A_B$ is true if node A is colored one of the three colors.[17]

There are three types of expressions needed to write the colorability of a graph as a logical satisfiability problem: A has some color (above), A has only one color, and A and a neighbor B have different colors.

(a) *Write out the logical expression that states that A does not have two colors at the same time. Write out the logical expression that states that A and B are not colored with the same color.* Hint: Both should be a conjunction (AND, $\wedge$) of three clauses each involving two variables.

Any logical expression can be rewritten into a standard format, the *conjunctive normal form*. A *literal* is either one of our boolean variables or its negation; a logical expression is in conjunctive normal form if it is a conjunction of a series of clauses, each of which is a disjunction (OR, $\vee$) of literals.

(b) *Show that, for two boolean variables X and Y, that $\neg(X \wedge Y)$ is equivalent to a disjunction of literals $(\neg X) \vee (\neg Y)$.* (Hint: Test each of the four cases). *Write your answers to part (a) in conjunctive normal form. What is the maximum*

[14]This exercise and the associated software were developed in collaboration with Christopher Myers, with help from Bart Selman and Carla Gomes. Computational hints can be found at the book website [181].

[15]The famous four-color theorem, that any map of countries on the world can be colored in four colors, shows that all planar graphs are 4-colorable.

[16]Because 3-colorability is **NP**–complete (see Exercise 8.15), finding such a polynomial-time algorithm would allow one to solve traveling salesman problems and find spin-glass ground states in polynomial time too.

[17]The operations AND ($\wedge$) and NOT $\neg$ correspond to common English usage ($\wedge$ is true only if both are true, $\neg$ is true only if the expression following is false). However, OR ($\vee$) is an *inclusive or*—false only if both clauses are false. In common English usage *or* is usually *exclusive*, false also if both are true. ("Choose door number one or door number two" normally does not imply that one may select both.)

number of literals in each clause you used? Is it the maximum needed for a general 3-colorability problem?

In part (b), you showed that any 3-colorability problem can be mapped onto a logical satisfiability problem in conjunctive normal form with at most three literals in each clause, and with three times the number of boolean variables as there were nodes in the original graph. (Consider this a hint for part (b).) Logical satisfiability problems with at most k literals per clause in conjunctive normal form are called **kSAT** problems.

(c) *Argue that the time needed to translate the 3-colorability problem into a* **3SAT** *problem grows at most quadratically in the number of nodes M in the graph (less than αM^2 for some α for large M).* (Hint: the number of edges of a graph is at most M^2.) *Given an algorithm that guarantees a solution to any N-variable* **3SAT** *problem in a time $T(N)$, use it to give a bound on the time needed to solve an M-node 3-colorability problem. If $T(N)$ were a polynomial-time algorithm (running in time less than N^x for some integer x), show that 3-colorability would be solvable in a time bounded by a polynomial in M.*

We will return to logical satisfiability, **kSAT**, and **NP**–completeness in Exercise 8.15. There we will study a statistical ensemble of **kSAT** problems, and explore a phase transition in the fraction of satisfiable clauses, and the divergence of the typical computational difficulty near that transition.

(1.9) **First to fail: Weibull.**[18] (Mathematics, Statistics, Engineering) ③

Suppose you have a brand-new supercomputer with $N = 1{,}000$ processors. Your parallelized code, which uses all the processors, cannot be restarted in mid-stream. How long a time t can you expect to run your code before the first processor fails?

This is example of *extreme value statistics* (see also exercises 12.23 and 12.24), where here we are looking for the smallest value of N random variables that are all bounded below by zero. For large N the probability distribution $\rho(t)$ and survival probability $S(t) = \int_t^\infty \rho(t')\,\mathrm{d}t'$ are often given by the *Weibull distribution*

$$S(t) = \mathrm{e}^{-(t/\alpha)^\gamma},$$
$$\rho(t) = -\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1}\mathrm{e}^{-(t/\alpha)^\gamma}. \tag{1.8}$$

Let us begin by assuming that the processors have a constant rate Γ of failure, so the probability density of a single processor failing at time t is $\rho_1(t) = \Gamma\exp(-\Gamma t)$ as $t \to 0$, and the survival probability for a single processor $S_1(t) = 1 - \int_0^t \rho_1(t')dt' \approx 1 - \Gamma t$ for short times.

(a) *Using $(1-\epsilon) \approx \exp(-\epsilon)$ for small ϵ, show that the the probability $S_N(t)$ at time t that all N processors are still running is of the Weibull form (eqn 1.8). What are α and γ?*

Often the probability of failure per unit time goes to zero or infinity at short times, rather than to a constant. Suppose the probability of failure for one of our processors

$$\rho_1(t) \sim Bt^k \tag{1.9}$$

with $k > -1$. (So, $k < 0$ might reflect a breaking-in period, where survival for the first few minutes increases the probability for later survival, and $k > 0$ would presume a dominant failure mechanism that gets worse as the processors wear out.)

(b) *Show the survival probability for N identical processors each with a power-law failure rate (eqn 1.9) is of the Weibull form for large N, and give α and γ as a function of B and k.*

The parameter α in the Weibull distribution just sets the scale or units for the variable t; only the exponent γ really changes the shape of the distribution. Thus the form of the failure distribution at large N only depends upon the power law k for the failure of the individual components at short times, not on the behavior of $\rho_1(t)$ at longer times. This is a type of *universality*,[19] which here has a physical interpretation; at large N the system will break down soon, so only early times matter.

The Weibull distribution, we must mention, is often used in contexts not involving extremal statistics. Wind speeds, for example, are naturally always positive, and are conveniently fit by Weibull distributions.

[18] Developed with the assistance of Paul (Wash) Wawrzynek

[19] The Weibull distribution is part of a family of extreme value distributions, all of whom are universal. See Chapter 12 and Exercise 12.24.

(1.10) **Emergence.** ⓟ
We begin with the broad statement "Statistical mechanics explains the simple behavior of complex systems." New laws emerge from bewildering interactions of constituents.
Discuss which of these emergent behaviors is probably not *studied using statistical mechanics.*
(a) The emergence of the wave equation from the collisions of atmospheric molecules,
(b) The emergence of Newtonian gravity from Einstein's general theory,
(c) The emergence of random stock price fluctuations from the behavior of traders,
(d) The emergence of a power-law distribution of earthquake sizes from the response of rubble in earthquake faults to external stresses.

(1.11) **Emergent vs. fundamental.** ⓟ
Statistical mechanics is central to condensed matter physics. It is our window into the behavior of materials—how complicated interactions between large numbers of atoms lead to physical laws (Fig. 1.9). For example, the theory of sound emerges from the complex interaction between many air molecules governed by Schrödinger's equation. More is different [10].

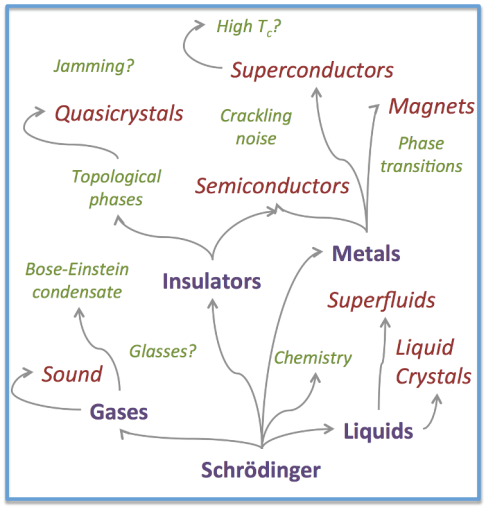

Fig. 1.9 Emergent. New laws describing macroscopic materials emerge from complicated microscopic behavior [177].

For example, if you inhale helium, your voice gets squeaky like Mickey Mouse. The dynamics of air molecules change when helium is introduced—the same law of motion, but with different constants.
(a) *Look up the wave equation for sound in gases. How many constants are needed? Do the details of the interactions between air molecules matter for sound waves in air?*
Statistical mechanics is tied also to particle physics and astrophysics. It is directly important in, e.g., the entropy of black holes (Exercise 7.16), the microwave background radiation (Exercises 7.15 and 10.1), and broken symmetry and phase transitions in the early Universe (Chapters 9, 11, and 12). Where statistical mechanics focuses on the *emergence* of comprehensible behavior at low energies, particle physics searches for the *fundamental* underpinnings at high energies (Fig. 1.10). Our different approaches reflect the complicated science at the atomic scale of chemistry and nuclear physics. At higher energies, atoms are described by elegant field theories (the *standard model* combining electroweak theory for electrons, photons, and neutrinos with QCD for quarks and gluons); at lower energies effective laws emerge for gases, solids, liquids, superconductors, ...

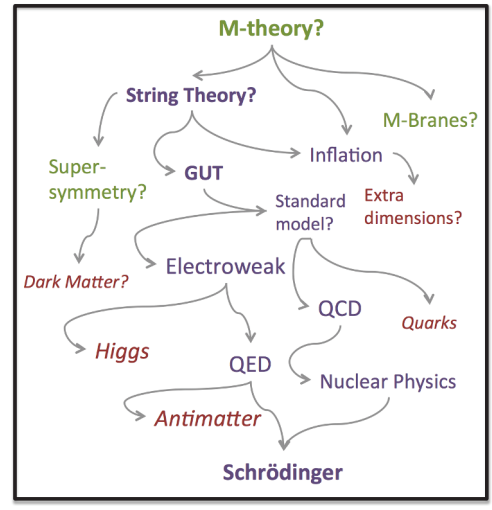

Fig. 1.10 Fundamental. Laws describing physics at lower energy emerge from more fundamental laws at higher energy [177].

The laws of physics involve parameters—real numbers that one must calculate or measure, like the speed of sound for a each gas at a given density and pressure. Together with the initial conditions (e.g., the density and its rate of change for a gas), the laws of physics allow us to predict how our system behaves.

Schrödinger's equation describes the Coulomb interactions between electrons and nuclei, and their interactions with electromagnetic field. It can in principle be solved to describe almost all of materials physics, biology, and engineering, apart from radioactive decay and gravity, using a Hamiltonian involving only the parameters $\hbar$, e, c, m_e, and the the masses of the nuclei.[20] Nuclear physics and QCD in principle determine the nuclear masses; the values of the electron mass and the fine structure constant $\alpha = e^2/\hbar c$ could eventually be explained by even more fundamental theories.

(b) *About how many parameters would one need as input to Schrödinger's equation to describe materials and biology and such?* Hint: There are 253 stable nuclear isotopes.

(c) *Look up the Standard Model—our theory of electrons and light, quarks and gluons, that also in principle can be solved to describe our Universe (apart from gravity). About how many parameters are required for the Standard Model?*

In high-energy physics, fewer constants are usually needed to describe the fundamental theory than the low-energy, effective emergent theory—the fundamental theory is more elegant and beautiful. In condensed matter theory, the fundamental theory is usually less elegant and messier; the emergent theory has a kind of parameter compression, with only a few combinations of microscopic parameters giving the governing parameters (temperature, elastic constant, diffusion constant) for the emergent theory.

Note that this is partly because in condensed matter theory we confine our attention to one particular material at a time (crystals, liquids, superfluids). To describe all materials in our world, and their interactions, would demand many parameters.

My high-energy friends sometimes view this from a different perspective. They note that the methods we use to understand a new superfluid, or a topological insulator, are quite similar to the ones they use to study the Universe. They admit a bit of envy—that we get a new universe to study every time an experimentalist discovers another material.

(1.12) **Self-propelled particles.**[21] (Active matter) ③ Exercise 2.20 investigates the statistical mechanical study of flocking—where animals, bacteria, or other active agents go into a collective state where they migrate in a common direction (like moshers in circle pits at heavy metal concerts [31, 188–190]). Here we explore the transition to a migrating state, but in an even more basic class of active matter: particles that are self-propelled but only interact via collisions. Our goal here is to both study the nature of the collective behavior, and the nature of the transition between disorganized motion and migration.

We start with an otherwise equilibrium system (damped, noisy particles with soft interatomic potentials, Exercise 6.19), and add a propulsion term

$$F_i^{\text{speed}} = \mu(v_0 - v_i)\hat{v}_i, \tag{1.10}$$

which accelerates or decelerates each particle toward a target speed v_0 without changing the direction. The damping constant μ now controls how strongly the target speed is favored; for $v_0 = 0$ we recover the damping needed to counteract the noise to produce a thermal ensemble.

This simulation can be a rough model for crowded bacteria propelling themselves around, or for artificially created *Janus* particles that have one side covered with a platinum catalyst that burns hydrogen peroxide, pushing it forward.

Launch the mosh pit simulator [32]. If necessary, reload the page to the default setting. Set all particles active (Fraction Red to 1), set the Particle count to $N = 200$, Flock strength = 0, Speed $v_0 = 0.25$, Damping = 0.5 and Noise Strength = 0, Show Graphs, and click Change. After some time, you should see most of the particles moving along a common direction. (Increase Frameskip to speed the process.) You can increase the Box size and number maintaining the density if you have a powerful computer, or

[20] The gyromagnetic ratio for each nucleus is also needed in a few situations where its coupling to magnetic fields are important.
[21] This exercise was developed in collaboration with David Hathcock. It makes use of the mosh pit simulator [32] developed by Matt Bierbaum for [190].

decrease it (but not below 30) if your computer is struggling.

(a) *Watch the speed distribution as you restart the simulation.* Turn off Frameskip to see the behavior at early times. *Does it get sharply peaked at the same time as the particles begin moving collectively? Now turn up frameskip to look at the long-term motion. Give a qualitative explanation of what happens. Is more happening than just selection of a common direction?* (Hint: Understanding why the collective behavior maintains itself is easier than explaining why it arises in the first place.)

We can study this emergent, collective flow by putting our system in a box—turning off the periodic boundary conditions along x and y. Reload parameters to default, then all active, $N = 300$, flocking $= 0$, speed $v_0 = 0.25$, raise the damping up to 2 and set noise $= 0$. Turn off the periodic boundary conditions along both x and y, set the frame skip to 20, and Change. Again, box sizes as low as 30 will likely work.

After some time, you should observe a collective flow of a different sort. You can monitor the average flow using the angular momentum (middle graph below the simulation).

(b) *Increase the noise strength. Can you disrupt this collective behavior? Very roughly, a what noise strength does the transition occur? (You can use the angular momentum as a diagnostic.)*

A key question in equilibrium statistical mechanics is whether a qualitative transition like this is continuous (Chapter 12) or discontinuous (Chapter 11). Discontinuous transitions usually exhibit both bistability and hysteresis: the observed transition raising the temperature or other control parameter is higher than when one lowers the parameter. Here, if the transition is abrupt, we should have a region with three states—a melted state of zero angular momentum, and a collective clockwise and counterclockwise state.

Return to the settings for part (b) to explore more carefully the behavior near the transition.

(c) *Use the angular momentum to measure the strength of the collective motion (taken from the center graph, treating the upper and lower bounds as ± 1). Graph it against noise as you raise the noise slowly and carefully from zero, until it vanishes.* (You may need to wait longer when you get close to the transition.) *Graph it again as you lower the noise. Do you find the same transition point on heating and cooling (raising and lowering the noise)? Is the transition abrupt, or continuous? Did you ever observe switches between the clockwise and anti-clockwise states?*

(1.13) **The birthday problem.** ②

Remember birthday parties in your elementary school? Remember those years when two kids had the same birthday? How unlikely!

How many kids would you need in class to get, more than half of the time, at least two with the same birthday?

(a) Numerical. *Write* `BirthdayCoincidences(K, C)`*, a routine that returns the fraction among C classes for which at least two kids (among K kids per class) have the same birthday.* (Hint: By sorting a random list of integers, common birthdays will be adjacent.) *Plot this probability versus K for a reasonably large value of C. Is it a surprise that your classes had overlapping birthdays when you were young?*

One can intuitively understand this, by remembering that to avoid a coincidence there are $K(K-1)/2$ pairs of kids, all of whom must have different birthdays (probability $364/365 = 1 - 1/D$, with D days per year).

$$P(K, D) \approx (1 - 1/D)^{K(K-1)/2} \tag{1.11}$$

This is clearly a crude approximation—it doesn't vanish if $K > D$! Ignoring subtle correlations, though, it gives us a net probability

$$\begin{aligned} P(K, D) &\approx \exp(-1/D)^{K(K-1)/2} \\ &\approx \exp(-K^2/(2D)) \end{aligned} \tag{1.12}$$

Here we've used the fact that $1 - \epsilon \approx \exp(-\epsilon)$, and assumed that K/D is small.

(b) Analytical. *Write the exact formula giving the probability, for K random integers among D choices, that no two kids have the same birthday.* (Hint: What is the probability that the second kid has a different birthday from the first? The third kid has a different birthday from the first two?) *Show that your formula does give zero if $K > D$. Converting the terms in your product to exponentials as we did above, show that your answer is consistent with the simple formula above, if $K \ll D$. Inverting eqn 1.12, give a formula for the number of kids needed to have a 50% chance of a shared birthday.*

Some years ago, we were doing a large simulation, involving sorting a lattice of $1{,}000^3$ random

fields (roughly, to figure out which site on the lattice would trigger first). If we want to make sure that our code is unbiased, we want different random fields on each lattice site—a giant birthday problem.

Old-style random number generators generated a random integer (2^{32} "days in the year") and then divided by the maximum possible integer to get a random number between zero and one. Modern random number generators generate all 2^{52} possible double precision numbers between zero and one.

(c) *If there are 2^{32} distinct four-byte unsigned integers, how many random numbers would one have to generate before one would expect coincidences half the time? Generate lists of that length, and check your assertion.* (Hints: It is faster to use array operations, especially in interpreted languages. I generated a random array with N entries, sorted it, subtracted the first $N-1$ entries from the last $N-1$, and then called *min* on the array.) *Will we have to worry about coincidences with an old-style random number generator? How large a lattice $L \times L \times L$ of random double precision numbers can one generate with modern generators before having a 50% chance of a coincidence?*

(1.14) **Width of the height distribution.**[22] (Statistics) ③

In this exercise we shall explore statistical methods of fitting models to data, in the context of fitting a Gaussian to a distribution of measurements. We shall find that *maximum likelihood* methods can be *biased*. We shall find that all sensible methods converge as the number of measurements N gets large (just as thermodynamics can ignore fluctuations for large numbers of particles), but a careful treatment of fluctuations and probability distributions becomes important for small N (just as different ensembles become distinguishable for small numbers of particles).

The Gaussian distribution, known in statistics as the *normal* distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \qquad (1.13)$$

is a remarkably good approximation for many properties. The heights of men or women in a given country, or the grades on an exam in a large class, will often have a histogram that is well described by a normal distribution.[23] If we know the heights x_n of a sample with N people, we can write the likelihood that they were drawn from a normal distribution with mean μ and variance σ^2 as the product

$$P(\{x_n\}|\mu,\sigma) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2). \qquad (1.14)$$

We first introduce the concept of *sufficient statistics*. Our likelihood (eqn 1.14) does not depend independently on each of the N heights x_n. What do we need to know about the sample to predict the likelihood?

(a) *Write $P(\{x_n\}|\mu,\sigma)$ in eqn 1.14 as a formula depending on the data $\{x_n\}$ only through N, $\overline{x} = (1/N)\sum_n x_n$ and $S = \sum_n (x_n - \overline{x})^2$.*

Given the model of independent normal distributions, its likelihood is a formula depending only on[24] $\overline{x}$ and S, the sufficient statistics for our Gaussian model.

Now, suppose we have a small sample and wish to estimate the mean and the standard deviation of the normal distribution.[25] *Maximum likelihood* is a common method for estimating model parameters; the estimates $(\mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}})$ are given by the peak of the probability distribution P.

(b) *Show that $P(\{x_n\}|\mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}})$ takes its maximum value at*

$$\begin{aligned} \mu_{\mathrm{ML}} &= \frac{\sum_n x_n}{N} = \overline{x} \\ \sigma_{\mathrm{ML}} &= \sqrt{\sum_n (x_n - \overline{x})^2/N} = \sqrt{S/N}. \end{aligned} \qquad (1.15)$$

(Hint: It is easier to maximize the log likelihood; $P(\boldsymbol{\theta})$ and $\log(P(\boldsymbol{\theta}))$ are maximized at the same point $\boldsymbol{\theta}_{\mathrm{ML}}$.)

If we draw samples of size N from a distribution of known mean μ_0 and standard deviation σ_0,

[22] This exercise was developed in collaboration with Colin Clement.

[23] This is likely because one's height is determined by the additive effects of many roughly uncorrelated genes and life experiences; the central limit theorem would then imply a Gaussian distribution (Chapter 2 and Exercise 12.11).

[24] In this exercise we shall use $\overline{X}$ denote a quantity averaged over a single sample of N people, and $\langle X \rangle_{\mathrm{samp}}$ denote a quantity also averaged over many samples.

[25] In physics, we usually estimate measurement errors separately from fitting our observations to theoretical models, so each experimental data point d_i comes with its error σ_i. In statistics, the estimation of the measurement error is often part of the modeling process, as in this exercise.

how do the maximum likelihood estimates differ from the actual values? For the limiting case $N = 1$, the various maximum likelihood estimates for the heights vary from sample to sample (with probability distribution $\mathcal{N}(x|\mu,\sigma^2)$, since the best estimate of the height is the sampled one). Because the average value $\langle\mu_{\rm ML}\rangle_{\rm samp}$ over many samples gives the correct mean, we say that $\mu_{\rm ML}$ is *unbiased*. The maximum likelihood estimate for $\sigma^2_{\rm ML}$, however, is biased. Again, for the extreme example $N = 1$, $\sigma^2_{\rm ML} = 0$ for every sample!

(c) *Assume the entire population is drawn from some (perhaps non-Gaussian) distribution of variance* $\langle x^2\rangle_{\rm samp} = \sigma_0^2$. *For simplicity, let the mean of the population be zero. Show that*

$$\begin{aligned}\left\langle\sigma^2_{\rm ML}\right\rangle_{\rm samp} &= (1/N)\left\langle\sum_{n=1}^{N}(x_n-\overline{x})^2\right\rangle_{\rm samp}\\ &= \frac{N-1}{N}\sigma_0^2. \end{aligned}\tag{1.16}$$

that the variance for a group of N people is on average smaller than the variance of the population distribution by a factor $(N-1)/N$. (Hint: $\overline{x} = (1/N)\sum_n x_n$ is not necessarily zero. Expand it out and use the fact that x_m and x_n are uncorrelated.)

The maximum likelihood estimate for the variance is biased on average toward smaller values. Thus we are taught, when estimating the standard deviation of a distribution[26] from N measurements, to divide by $\sqrt{N-1}$:

$$\sigma^2_{\rm N-1} \approx \frac{\sum_n (x_n-\overline{x})^2}{N-1}. \tag{1.17}$$

This correction $N \to N-1$ is generalized to more complicated problems by considering the number of independent degrees of freedom (here $N-1$ degrees of freedom in the vector $x_n - \overline{x}$ of deviations from the mean). Alternatively, it is interesting that the bias disappears if one does not estimate both σ^2 and μ by maximizing the joint likelihood, but integrating (or *marginalizing*) over μ and then finding the maximum likelihood for σ^2.

(1.15) **Fisher information and Cramér–Rao.**[27] (Statistics, Mathematics, Information geometry) ④

Here we explore the geometry of the space of probability distributions. When one changes the external conditions of a system a small amount, how much does the ensemble of predicted states change? What is the *metric* in probability space? Can we predict how easy it is to detect a change in external parameters by doing experiments on the resulting distribution of states? The metric we find will be the *Fisher information matrix* (FIM). The *Cramér–Rao bound* will use the FIM to provide a rigorous limit on the precision of any (unbiased) measurement of parameter values.

In both statistical mechanics and statistics, our models generate probability distributions $P(\mathbf{x}|\boldsymbol{\theta})$ for behaviors $\mathbf{x}$ given parameters $\boldsymbol{\theta}$.

- A crooked gambler's loaded die, where the state space is comprised of discrete rolls $\mathbf{x} \in \{1,2,\ldots,6\}$ with probabilities $\boldsymbol{\theta} = \{p_1,\ldots,p_5\}$, with $p_6 = 1 - \sum_{j=1}^{5}\theta_j$.
- The probability density that a system with a Hamiltonian $\mathcal{H}(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (T,P,N)$ giving the temperature, pressure, and number of particles, will have a probability density $P(\mathbf{x}|\boldsymbol{\theta}) = \exp(-\mathcal{H}/k_BT)/Z$ in phase space (Chapter 3, Exercise 6.22).
- The height of women in the US, $\mathbf{x} = \{h\}$ has a probability distribution well described by a normal (or Gaussian) distribution $P(\mathbf{x}|\boldsymbol{\theta}) = 1/\sqrt{2\pi\sigma^2}\exp(-(x-\mu)^2/2\sigma^2)$ with mean and standard deviation $\boldsymbol{\theta} = (\mu,\sigma)$ (Exercise 1.14).
- A least squares model $y_i(\boldsymbol{\theta})$ for N data points $d_i \pm \sigma$ with independent, normally distributed measurement errors predicts a likelihood for finding a value $\mathbf{x} = \{x_i\}$ of the data $\{d_i\}$ given by

$$P(\mathbf{x}|\boldsymbol{\theta}) = \frac{\mathrm{e}^{-\sum_i (y_i(\theta)-x_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{N/2}}. \tag{1.18}$$

  (Think of the theory curves you fit to data in many experimental labs courses.)

How "distant" is a loaded die is from a fair one? How "far apart" are the probability distributions of particles in phase space for two small system at different temperatures and pressures? How

[26] Do not confuse this with the estimate of the error in the *mean* $\overline{x}$.

[27] This exercise was developed in collaboration with Colin Clement and Katherine Quinn.

hard would it be to distinguish a group of US women from a group of Pakistani women, if you only knew their heights?

We start with the least-squares model.

(a) *How big is the probability density that a least-squares model with true parameters $\boldsymbol{\theta}$ would give experimental results implying a different set of parameters $\boldsymbol{\phi}$? Show that it depends only on the distance between the vectors $|\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}(\boldsymbol{\phi})|$ in the space of predictions.* Thus the predictions of least-squares models form a natural manifold in a behavior space, with a coordinate system given by the parameters. The point on the manifold corresponding to parameters $\boldsymbol{\theta}$ is $\mathbf{y}(\boldsymbol{\theta})/\sigma$ given by model predictions rescaled by their error bars, $\mathbf{y}(\boldsymbol{\theta})/\sigma$.

Remember that the metric tensor $g_{\alpha\beta}$ gives the distance on the manifold between two nearby points. The squared distance between points with coordinates $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta}$ is $\epsilon^2 \sum_{\alpha\beta} g_{\alpha\beta}\Delta_\alpha\Delta_\beta$.

(b) *Show that the least-squares metric is $g_{\alpha\beta} = (J^T J)_{\alpha\beta}/\sigma^2$, where the Jacobian $J_{i\alpha} = \partial y_i/\partial\theta_\alpha$.*

For general probability distributions, the natural metric describing the distance between two nearby distributions $P(\mathbf{x}|\boldsymbol{\theta})$ and $Q = P(\mathbf{x}|\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta})$ is given by the FIM:

$$g_{\alpha\beta}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \log P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\mathbf{x}} \tag{1.19}$$

Are the distances between least-squares models we intuited in parts (a) and (b) compatible with the the FIM?

(c) *Show for a least-squares model that eqn 1.19 is the same as the metric we derived in part (b).* (Hint: For a Gaussian distribution $\exp((x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, $\langle x\rangle = \mu$.)

If we have experimental data with errors, how well can we estimate the parameters in our theoretical model, given a fit? As in part (a), now for general probabilistic models, how big is the probability density that an experiment with true parameters $\boldsymbol{\theta}$ would give results perfectly corresponding to a nearby set of parameters $\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta}$?

(d) *Take the Taylor series of $\log P(\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta})$ to second order in ϵ. Exponentiate this to estimate how much the probability of measuring values corresponding to the predictions at $\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta}$ fall off compared to $P(\boldsymbol{\theta})$.* Thus to linear order the FIM $g_{\alpha\beta}$ estimates the range of likely measured parameters around the true parameters of the model.

The *Cramér–Rao bound* shows that this estimate is related to a rigorous bound. In particular, errors in a multiparameter fit are usually described by a *covariance matrix* Σ, where the variance of the likely values of parameter θ_α is given by $\Sigma_{\alpha\alpha}$, and where $\Sigma_{\alpha\beta}$ gives the correlations between two parameters θ_α and θ_β. One can show within our quadratic approximation of part (d) that the covariance matrix is the inverse of the FIM $\Sigma_{\alpha\beta} = (g^{-1})_{\alpha\beta}$. The *Cramér–Rao bound* roughly tells us that no experiment can do better than this at estimating parameters. In particular, it tells us that the error range of the individual parameters from a sampling of a probability distribution is bounded below by the corresponding element of the inverse of the FIM

$$\Sigma_{\alpha\alpha} \geq (g^{-1})_{\alpha\alpha}. \tag{1.20}$$

(if the estimator is *unbiased*, see Exercise 1.14). This is another justification for using the FIM as our natural distance metric in probability space. In Exercise 1.16, we shall examine *global* measures of distance or distinguishability between potentially quite different probability distributions. There we shall show that these measures all reduce to the FIM to lowest order in the change in parameters. In Exercises 6.23, 6.21, and 6.22, we shall show that the FIM for a Gibbs ensemble as a function of temperature and pressure can be written in terms of thermodynamic quantities like compressibility and specific heat. There we use the FIM to estimate the *path length in probability space*, in order to estimate the entropy cost of controlling systems like the Carnot cycle.

(1.16) **Distances in probability space.**[28] (Statistics, Mathematics, Information geometry) ③

In statistical mechanics we usually study the behavior expected given the experimental parameters. Statistics is often concerned with estimating how well one can deduce the parameters (like temperature and pressure, or the increased risk of death from smoking) given a sample of the ensemble. Here we shall explore ways of measuring distance or distinguishability between distant probability distributions.

Exercise 1.15 introduces four problems (loaded dice, statistical mechanics, the height distribu-

[28] This exercise was developed in collaboration with Katherine Quinn.

tion of women, and least-squares fits to data), each of which have parameters $\boldsymbol{\theta}$ which predict an ensemble probability distribution $P(\mathbf{x}|\boldsymbol{\theta})$ for data $\mathbf{x}$ (die rolls, particle positions and momenta, heights, ...). In the case of least-squares models (eqn 1.18) where the probability is given by a vector $x_i = y_i(\boldsymbol{\theta}) \pm \sigma$, we found that the distance between the predictions of two parameter sets $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ was naturally given by $|\mathbf{y}(\boldsymbol{\theta})/\sigma - \mathbf{y}(\boldsymbol{\phi})/\sigma|$. We want to generalize this formula—to find ways of measuring distances between probability distributions given by arbitrary kinds of models.

Exercise 1.15 also introduced the Fisher information metric (FIM) in eqn 1.19:

$$g_{\mu\nu}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \log(P(\mathbf{x}))}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\mathbf{x}} \tag{1.21}$$

which gives the distance between probability distributions for nearby sets of parameters

$$d^2(P(\boldsymbol{\theta}), P(\boldsymbol{\theta}+\epsilon\boldsymbol{\Delta})) = \epsilon^2 \sum_{\mu\nu} \Delta_\mu g_{\mu\nu} \Delta_\nu. \tag{1.22}$$

Finally, it argued that the distance defined by the FIM is related to how distinguishable the two nearby ensembles are—how well we can deduce the parameters. Indeed, we found that to linear order the FIM is the inverse of the covariance matrix describing the fluctuations in estimated parameters, and that the Cramér–Rao bound shows that this relationship between the FIM and distinguishability works even beyond the linear regime.

There are several measures in common use, of which we will describe three—the Hellinger distance, the Bhattacharyya "distance", and the Kullback–Liebler divergence. Each has its uses. The Hellinger distance becomes less and less useful as the amount of information about the parameters becomes large. The Kullback–Liebler divergence is not symmetric, but one can symmetrize it by averaging. It and the Bhattacharyya distance nicely generalize the least-squares metric to arbitrary models, but they violate the triangle inequality and embed the manifold of predictions into a space with Minkowski-style time-like directions [155].

Let us review the properties that we ordinarily demand from a distance between points P and Q.

- We expect it to be positive, $d(P,Q) \geq 0$, with $d(P,Q) = 0$ only if $P = Q$.
- We expect it to be symmetric, so $d(P,Q) = d(Q,P)$.
- We expect it to satisfy the *triangle inequality*, $d(P,Q) \leq d(P,R) + d(R,Q)$—the two short sides of a triangle must extend at total distance enough to reach the third side.
- We want it to become large when the points P and Q are extremely different.

All of these properties are satisfied by the least-squares distance of Exercise 1.15, because the distances between points on the surface of model predictions is the Euclidean distance between the predictions in data space.

Our first measure, the Hellinger distance at first seems ideal. It defines a *dot product* between probability distributions P and Q. Consider the discrete gambler's distribution, giving the probabilities $\mathbf{P} = \{P_j\}$ for die roll j. The normalization $\sum P_j = 1$ makes $\{\sqrt{P_j}\}$ a unit vector in six dimensions, so we define a dot product $P \cdot Q = \sum_{j=1}^{6} \sqrt{P_j}\sqrt{Q_j} = \int \mathrm{d}\mathbf{x} \sqrt{P(\mathbf{x})}\sqrt{Q(\mathbf{x})}$. The Hellinger distance is then given by the squared distance between points on the unit sphere:[29]

$$\begin{aligned} d^2_{\mathrm{Hel}}(P,Q) &= (P-Q)^2 = 2 - 2P\cdot Q \\ &= \int \mathrm{d}\mathbf{x} \left(\sqrt{P(\mathbf{x})} - \sqrt{Q(\mathbf{x})}\right)^2. \end{aligned} \tag{1.23}$$

(a) *Argue, from the last geometrical characterization, that the Hellinger distance must be a valid distance function. Show that the Hellinger distance does reduce to the FIM for nearby distributions, up to a constant factor. Show that the Hellinger distance never gets larger than* $\sqrt{2}$. *What is the Hellinger distance between a fair die* $P_j \equiv 1/6$ *and a loaded die* $Q_j = \{1/10, 1/10, \ldots, 1/2\}$ *that favors rolling 6?*

The Hellinger distance is peculiar in that, as the statistical mechanics system gets large, or as one adds more experimental data to the statistics model, all pairs approach the maximum distance $\sqrt{2}$.

(b) *Our gambler keeps using the loaded die. Can the casino catch him? Let* $P_N(\mathbf{j})$ *be the probability that rolling the die* N *times gives the sequence* $\mathbf{j} = \{j_1, \ldots, j_N\}$. *Show that*

$$P_N \cdot Q_N = (P \cdot Q)^N. \tag{1.24}$$

[29] Sometimes it is given by *half* the distance between points on the unit sphere, presumably so that the maximum distance between two probability distributions becomes one, rather than $\sqrt{2}$.

and hence

$$d^2_{\mathrm{Hel}}(P_N, Q_N) = 1 - (P \cdot Q)^N. \tag{1.25}$$

After $N = 100$ rolls, how close is the Hellinger distance from its maximum value?

From the casino's point of view, the certainty that the gambler is cheating is becoming squeezed into a tiny range of distances. (P_N and Q_N becoming increasingly orthogonal does not lead to larger and larger Hellinger distances.) In an Ising model, or a system with N particles, or a cosmic microwave background experiment with N measured areas of the sky, even tiny changes in parameters lead to orthogonal probability distributions, and hence Hellinger distances near its maximum value of one.[30]

The Hellinger overlap $(P \cdot Q)^N = \exp(N \log(P \cdot Q))$ keeps getting smaller as we take N to infinity; it is like the exponential of an extensive quantity.

Our second measure, the Bhattacharyya distance, can be derived from a limit of the Hellinger distance as the number of data points N goes to zero:

$$\begin{aligned} d^2_{\mathrm{Bhatt}}(P, Q) &= \lim_{N\to 0} \tfrac{1}{2} d^2_{\mathrm{Hel}}(P_N, Q_N)/N \\ &= -\log(P \cdot Q) \\ &= -\log\left(\sum_{\mathbf{x}} \sqrt{P(\mathbf{x})}\sqrt{Q(\mathbf{x})}\right). \end{aligned} \tag{1.26}$$

We sometimes say that we calculate the behavior of N *replicas* of the system, and then take $N \to 0$. Replica theory is useful, for example, in disordered systems, where we can average $F = -k_B T \log(Z)$ over disorder (difficult) by finding the average of Z^N over disorder (not so hard) and then taking $N \to 0$.

(d) *Derive eqn 1.26.* (Hint: $Z^N \approx \exp(N \log Z) \approx 1 + N \log Z$ for small N.)

The third distance-like measure we introduce is the *Kullback–Leibler divergence* from Q to P.

$$d_{\mathbf{KL}}(Q|P) = -\int \mathrm{d}\mathbf{x} P(\mathbf{x}) \log(Q(\mathbf{x})/P(\mathbf{x})). \tag{1.27}$$

(c) *Show that the Kullback–Liebler divergence is positive, zero only if $P = Q$, but is not symmetric. Show that, to quadratic order in ϵ in eqn 1.22, that the Kullback–Liebler divergence does lead to the FIM.*

The Kullback–Liebler divergence is sometimes symmetrized:

$$\begin{aligned} d_{\mathbf{sKL}}(Q, P) &= \tfrac{1}{2}(d_{\mathbf{KL}}(Q|P) + d_{\mathbf{KL}}(P|Q)) \\ &= \int \mathrm{d}\mathbf{x}(P(\mathbf{x}) - Q(\mathbf{x})) \log(P(\mathbf{x})/Q(\mathbf{x})). \end{aligned} \tag{1.28}$$

The Bhattacharyya distance and the symmetrized Kullback–Liebler divergence share several features, both good and bad.

(d) *Show that they are* intensive *[155]—that the distance grows linearly with repeated measurements*[31] *(as for repeated rolls in part (b)). Show that they do not satisfy the triangle inequality. Show that they does satisfy the other conditions for a distance. Show, for the nonlinear least-squares model of eqn 1.18, that they equal the distance in data space between the two predictions.*

[30] The problem is that the manifold of predictions is being curled up onto a sphere, where the short-cut distance between two models becomes quite different from the geodesic distance within the model manifold.

[31] This also makes these measures behave nicely for large systems as in statistical mechanics, where small parameter changes lead to nearly orthogonal probability distributions.