# Sloppy Models, Information Geometry, and Emergent Simplicity

James P. Sethna
Laboratory of Atomic and Solid State Physics
Cornell University, Ithaca, NY 14853


March 17, 2024

# Contents

# Chapter 1

# A new kind of emergence

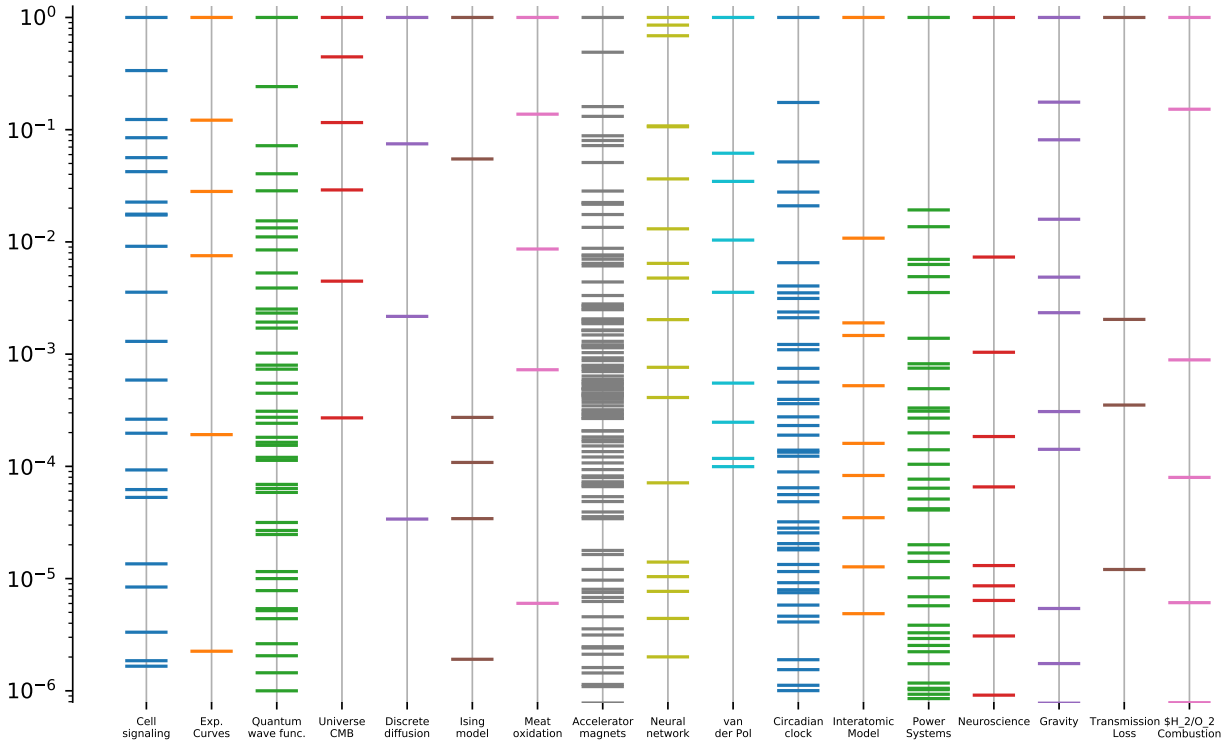



**Fig. 1.1 Sloppy spectra**. The eigenvalues of the (approximate) cost Hessian for models from a variety of fields. Each has an enormous range of eigenvalues, with roughly uniform density in log. We see also that models from physics with small parameters, like diffusion and the Ising model, also show sloppy spectra. Diffusing on a lattice has a continuum limit with an emergent law given as a partial differential equation. The Ising model has an emergent scale invariance explained by the renormalization group. Their emergent parameters (diffusion constants, or temperature and field) form the largest, stiff eigenvalues, and they both have a hierarchy of corrections with smaller eigenvalues and smaller effects on the behavior (from [11]).

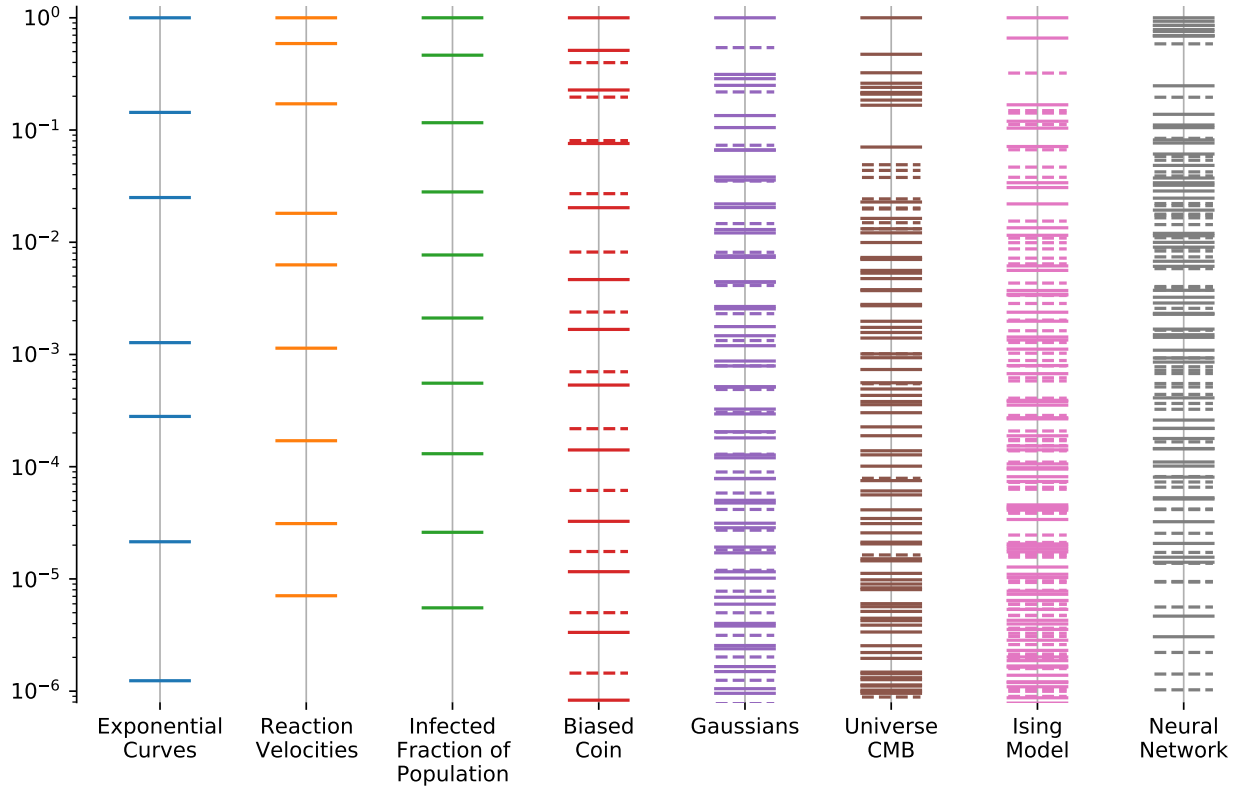



**Fig. 1.2 Hyperribbon lengths**. Here we preview our information geometry results on the hierarchy of ranges in behavior space spanned by our physical models – the widths of their model manifolds. We call these model manifolds *hyperribbons*, because each shorter direction is roughly down by the same factor. (Ribbons are longer than they are wide, and wider than they are thick . . . ) Ignoring the thin directions allows one to understand the overall behavior without encompassing all the details at once – an emergent simplicity [11].

# Exercises

S1.1 **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. S1.3). For example, the theory of sound emerges from the complex interaction between many air molecules governed by Schrödinger's equation. More is different [2].

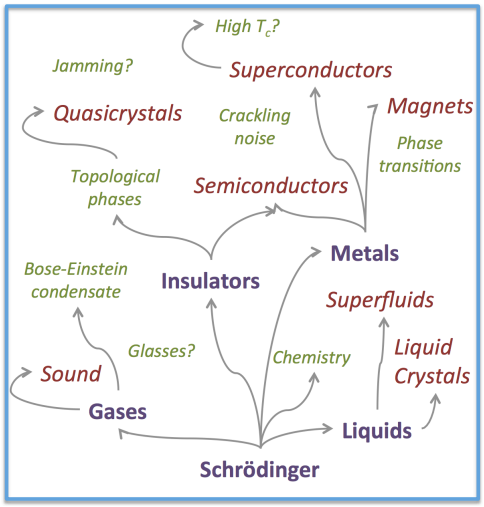



**Fig. S1.3 Emergent**. New laws describing macroscopic materials emerge from complicated microscopic behavior [16].

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. S1.4). 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, ...

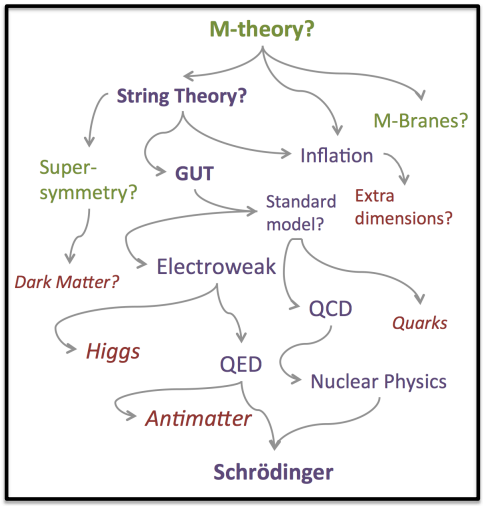



**Fig. S1.4 Fundamental**. Laws describing physics at lower energy emerge from more fundamental laws at higher energy [16].

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.[1] 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.

[1] The gyromagnetic ratio for each nucleus is also needed in a few situations where its coupling to magnetic fields are important.

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.

S1.2 **Width of the height distribution.**[2] (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

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \tag{S1.1}$$

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.[3] 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 $\mu$ and variance $\sigma^2$ as the product

$$P(\{x_n\}|\mu,\sigma) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2). \tag{S1.2}$$

We first introduce the concept of *sufficient statistics*. Our likelihood (eqn S1.2) 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 S1.2 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[4] $\overline{x}$ and $S$, the sufficient statistics for our Gaussian model.

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

[3] 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).

[4] 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.

Now, suppose we have a small sample and wish to estimate the mean and the standard deviation of the normal distribution.[5] *Maximum likelihood* is a common method for estimating model parameters; the estimates $(\mu_{\rm ML}, \sigma_{\rm ML})$ are given by the peak of the probability distribution $P$.

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

$$\begin{aligned} \mu_{\rm ML} &= \frac{\sum_n x_n}{N} = \overline{x} \\ \sigma_{\rm ML} &= \sqrt{\sum_n (x_n - \overline{x})^2/N} = \sqrt{S/N}. \end{aligned} \tag{S1.3}$$

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

If we draw samples of size $N$ from a distribution of known mean $\mu_0$ and standard deviation $\sigma_0$, 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*

$$\begin{aligned} \left\langle \sigma^2_{\rm ML} \right\rangle_{\rm samp} &= (1/N) \left\langle \sum_{n=1}^{N} (x_n - \overline{x})^2 \right\rangle_{\rm samp} \\ &= \frac{N-1}{N} \sigma_0^2. \end{aligned} \tag{S1.4}$$

*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[6] from $N$ measurements, to divide by $\sqrt{N-1}$:

$$\sigma^2_{\rm N-1} \approx \frac{\sum_n (x_n - \overline{x})^2}{N-1}. \tag{S1.5}$$

[5] 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 $\sigma_i$. In statistics, the estimation of the measurement error is often part of the modeling process, as in this exercise.

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

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 $\sigma^2$ and $\mu$ by maximizing the joint likelihood, but integrating (or *marginalizing*) over $\mu$ and then finding the maximum likelihood for $\sigma^2$.

S1.3 **Statistical mechanics and statistics.**[7] (Statistics) ③

Consider the problem of fitting a theoretical model to experimentally determined data. Let our model predict a time-dependent function $y^\theta(t)$, where $\boldsymbol{\theta}$ are the model parameters. Let there be $N$ experimentally determined data points $d_i$ at times $t_i$ with errors of standard deviation $\sigma$. We assume that the experimental errors for the data points are independent and Gaussian distributed, so that the probability that a given model produced the observed data points (the probability $P(D|\boldsymbol{\theta})$ of the data given the model) is

$$P(D|\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\left(y^\theta(t_i)-d_i\right)^2/2\sigma^2}. \tag{S1.6}$$

(a) *True or false: This probability density corresponds to a Boltzmann distribution with energy $H$ and temperature $T$, with $H = \sum_{i=1}^{N}(y^\theta(t_i) - d_i)^2/2$ and $k_B T = \sigma^2$.*

There are two approaches to statistics. Among a family of models, the *frequentists* will pick the parameters $\boldsymbol{\theta}$ with the largest value of $P(D|\boldsymbol{\theta})$ (the *maximum likelihood estimate*); the ensemble of best-fit models is then deduced from the range of likely input data (deduced from the error bars $\sigma$). The *Bayesians* take a different point of view. They argue that there is no reason to believe a priori that all models have the same probability. (In model parameter space, there is no analogue of Liouville's theorem, Section 4.1.) Suppose the probability of the model (the *prior*) is $P(\boldsymbol{\theta})$. They use the theorem

$$P(\boldsymbol{\theta}|D) = P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})/P(D). \tag{S1.7}$$

(b) *Prove Bayes' theorem (eqn S1.7) using the fact that $P(A \text{ and } B) = P(A|B)P(B)$ (see note 39 on p. 113).*

The Bayesians will often pick the maximum of $P(\boldsymbol{\theta}|D)$ as their model for the experimental data. But, given their perspective, it is even more natural to consider the entire *ensemble* of models, weighted by $P(\boldsymbol{\theta}|D)$, as the best description of the data. This ensemble average then naturally provides error bars for the parameters as well as for the predictions of various quantities.

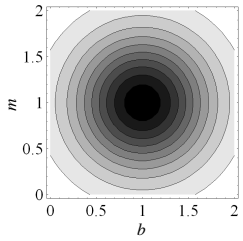
Consider the problem of fitting a line to two data points. Suppose the experimental data points are at $t_1 = 0$, $d_1 = 1$ and $t_2 = 1$, $d_2 = 2$, where both $y$-values have uncorrelated Gaussian errors with standard deviation $\sigma = 1/2$, as assumed in eqn S1.6 above. Our model $M(m, b)$, with parameters $\boldsymbol{\theta} = (m, b)$, is $y(t) = mt + b$. Our Bayesian statistician

[7] This exercise was developed with the help of Robert Weiss.

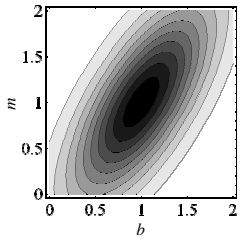
has prior knowledge that $m$ and $b$ both lie between zero and two, and assumes that the probability density is otherwise uniform; $P(m, b) = 1/4$ for $0 < m < 2$ and $0 < b < 2$.

(c) *Which of the contour plots shown accurately represent the probability distribution* $P(\boldsymbol{\theta}|D)$ *for the model, given the observed data? (The spacing between the contour lines is arbitrary.)*

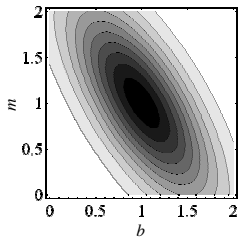
(A)



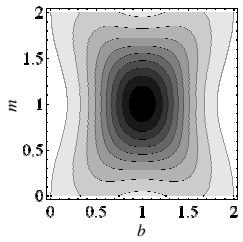

(B)

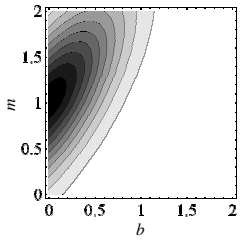



(C)




(D)



(E) 

# Chapter 2

# Sloppy models

Inverse problems in multiparameter scientific models are challenging. A systems biology model may have tens to thousands of unknown reaction constants; a climate model may have tens to hundreds; the cosmic microwave background radiation patterns (the echo of the Big Bang) is described by ΛCDM (cold-dark-matter) models with six or more parameters. The "inverse problem" is to extract the to extract these parameters from experimental or simulation data. Unless you have a lot of high-quality experimental data (as we do for the cosmic microwave background), it is typically impossible to solve the inverse problem.

One ubiquitous problem is *sloppiness*—the parameters in the model are poorly constrained by the data (the inverse problem is *ill-posed*). In this book, we shall explore sloppiness in parameter space, the model manifold, and its hyperribbon structure in two different contexts. In the first few chapters, we shall focus on *nonlinear least-squares models*. Consider a model $\mathbf{y}_{\boldsymbol{\theta}}(\mathbf{t})$ depending on $N$ parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_N\}$, making predictions about $M$ experiments under conditions given by one or more control variables $\mathbf{t}_i$, with measurements $\mathbf{d}_i(\mathbf{t}_i)$ with standard deviations $\boldsymbol{\sigma}_i$. Our least-squares models presume all the errors have Gaussian errors, so the probability we predict for the observed data given our model is

$$\rho_{\boldsymbol{\theta}}(\mathbf{d}) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}} \exp(-(y_{\boldsymbol{\theta}}(t_i) - d_i)^2/2\sigma_i^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp(-C(\boldsymbol{\theta}, \mathbf{d}, \mathbf{t})), \tag{2.1}$$

where the cost (called $\chi^2$ in statistics[1] ) is

$$C(\boldsymbol{\theta}, \mathbf{d}, \mathbf{t}) = \sum_{i=1}^{M} \frac{(y_{\boldsymbol{\theta}}(t_i) - d_i)^2}{2\sigma^2}. \tag{2.2}$$

where for simplicity we take all the data uncertainties to be equal. Here $y_{\boldsymbol{\theta}}$ is a nonlinear function, and the cost is a sum of squares, and the best fit is where the cost is smallest, hence the name nonlinear least-squares (NLLS). In later chapters, we shall consider *probabilistic*

[1] Up to a possible factor of two?

*models* that cannot be written as a sum of squares, such as the Ising model and the ΛCDM model mentioned above.

Two important quantities that will be central to our future discussion. First, the Jacobian of the model predictions

$$J_{i\alpha} = \partial y(t_i)/\partial\theta_\alpha \tag{2.3}$$

linearizes the map from parameter space $\boldsymbol{\theta}$ to prediction space $\mathbf{y}$. It is the skewness of this map that embodies the sloppiness in parameter space, and the hierarchy of widths for the model manifold in Chapter 3.

Second, expanding the cost up to quadratic order in the distance to the best fit, we get the cost Hessian

$$\mathcal{H}_{\alpha\beta} = \frac{\partial^2 C}{\partial\theta_\alpha \partial\theta_\beta}. \tag{2.4}$$

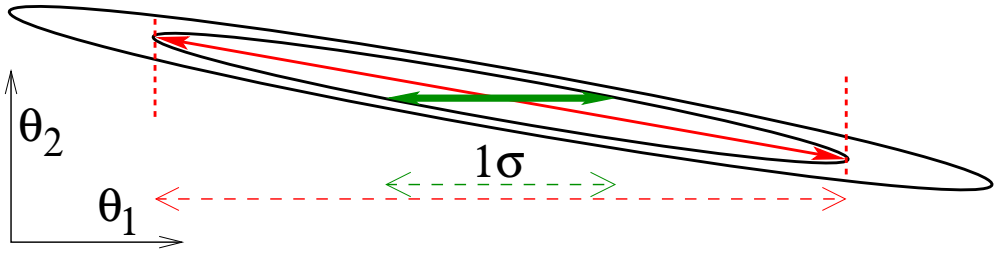



**Fig. 2.1 Errors in estimating parameters**. In a sloppy model, each parameter can be important and yet all parameters can be challenging to estimate from collective behavior. Here we see for two eigenvalues of the cost Hessian, varying one parameter keeping another fixed until a given cost allows much less variation than varying a parameter allowing the others to compensate. (See Exercise S2.3.)

Exercise S2.3 discusses the role of the Hessian in estimating parameter sensitivity and parameter uncertainty, if the uncertainties in the parameters are small enough to ignore the higher-order terms[2] as illustrated in Fig. S2.5. How important is parameter $\theta_\alpha$ to the behavior? You find that the diagonal elements of $\mathcal{H}$ gives the variance $(\sigma_\alpha^2)_{\rm fixed} = 1/\mathcal{H}_{\alpha\alpha}$ allowed in parameter $\alpha$ if all the other parameters are fixed at their best-fit values. How uncertain is the value of parameter $\theta_\alpha$, given the uncertainties in the data? You find that their variance is given by the diagonal elements of the inverse of the Hessian, $(\sigma_\alpha^2)_{\rm collective} = (\mathcal{H}^{-1})_{\alpha\alpha}$.

We usually discuss sloppiness in parameter space in terms of the eigenvalues and eigenvectors of the cost Hessian $\mathcal{H}$. Eigenvectors of $\mathcal{H}$ with large eigenvalues correspond to *stiff* directions: parameter combinations that are well constrained by the data. Eigenvectors of $\mathcal{H}$ with small eigenvalues are poorly constrained by the data. The striking result, confirmed in

[2] We shall see, however, that the sloppy directions will almost always violate this assumption, often allowing variations all the way to infinity for most parameter combinations.

models across many fields of physics, science, and engineering, is that the logarithms of the eigenvalues of $\mathcal{H}$ are roughly equally spaced. The eigenvalues are often a factor of three or more apart, and with 48 parameters (in our original model [3, 4]), $3^{50}$ is a huge range. This implies that only a few stiff parameter combinations realistically matter for model behavior: a central result that is connected to the success of simpler models. It furthermore implies that most eigenvector parameter combinations have a huge uncertainty, which typically translates into large uncertainties in all the model parameters. The inverse problem, working backward from the behavior to find the behavior, is ill-posed. But, since the sloppy parameter combinations are not crucial for the behavior, the good fits, constraining the stiff directions, can often be relied upon to make accurate predictions about the real world.

One final note. The cost Hessian for NLLS models can be written in terms of derivatives of the predictions $\mathbf{y}$:

$$\begin{aligned}\frac{\partial^2 C}{\partial\theta_\alpha\partial\theta_\beta} &= \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\left(\sum_i (y_i-d_i)^2/2\sigma^2\right)\\ &= (1/\sigma^2)\frac{\partial}{\partial\theta_\alpha}\left(\sum_i (y_i-d_i)\frac{\partial y_i}{\partial\theta_\beta}\right)\\ &= (1/\sigma^2)\left(\sum_i \frac{\partial y_i}{\partial\theta_\alpha}\frac{\partial y_i}{\partial\theta_\beta} + \cancel{(y_i-d_i)\frac{\partial^2 y_i}{\partial\theta_\alpha\partial\theta_\beta}}^{\,?}\right)\\ &\approx (1/\sigma^2)(J^TJ)_{\alpha\beta},\end{aligned} \tag{2.5}$$

where the second derivative term is often neglected, giving an approximate Hessian that is often used.

There are practical reasons for this approximation. A good fit has $y_i \approx d_i$ and makes it small. The second derivatives are expensive to calculate. And the resulting Hamiltonian is positive definite, which can be useful (especially in sloppy models, when so many of the directions have near zero eigenvalue).

Fundamentally, this approximation becomes exact when the data is perfectly described by the model. It is the quadratic approximation for the distance between points in behavior space $\mathbf{y}$ as parameters are varied. Viewing $\boldsymbol{\theta}$ as coordinates for the surface of predicted behaviors, this will be the *metric tensor* $g_{\alpha\beta} = (1/\sigma^2)(J^TJ)_{\alpha\beta}$ on the model manifold (see eqn **??**).

# Exercises

Exercise S2.1, *Sloppy exponentials*, is the first of a series of exercises exploring the ill-conditioned problem of extracting the decay parameters from sums of decaying exponentials. Exercise S2.3 discusses several different topics (nonlinear fits, model manifolds, and in particular derives the variance for one parameter when the others are fixed and when

they are allowed also to fluctuate. Exercise S2.4 discusses the relation between the cost Hessian in NLLS models, the Fisher information matrix (FIM) for probabilistic models, and the Cramér-Rao bound on the difficulty of estimating parameters from experiments. Finally, Exercises S2.5, S2.6, and S2.7 derive fascinating relations between the FIM and derivatives of the free energy in statistical mechanics, and explicitly computes the model manifold of a piston filled with ideal gas.

S2.1 **Sloppy exponentials.**[3] (Information geometry, Statistics) ②

The problem of extracting the decay rates from a sum of exponential decays is a famously difficult inverse problem, from the early days of radioactivity to modern simulations of lattice quantum chromodynamics [**?**]. In a series of exercises, we shall use our information geometry ideas to study the simplest version of this problem: the sum of $N$ exponential decays:

$$y_{\boldsymbol{\theta}}(t) = (1/N)\sum_{\alpha=1}^{N} \exp(-\theta_\alpha t). \tag{S2.6}$$

We anticipate that it will be challenging to disentangle decay rates $\theta$ which are close to one another, unless one has high-precision data over large ranges of time. All the decay curves are smoothly monotonically decreasing, and one could imagine modeling a sum of two decays with a single intermediate decay rate. You shall find in these exercises that this simple model illustrates the behavior we have found widespread in multiparameter models in physics, engineering, biology, and other fields.

In this first exercise, we presume we have perfect experimental data for the decay $d(t)$ at $M$ points $t_i$ equally spread for $t$ between 0 and 10, with separation $\Delta t = 10/M$. We shall be considering how well this data can be represented by other values of the parameters $\boldsymbol{\theta}$, so our cost (eqn 2.2) is:

$$C(\boldsymbol{\theta}, \boldsymbol{\theta}^{[0]}) = \sum_{i=1}^{M} \left(y_{\boldsymbol{\theta}}(t_i) - y_{\boldsymbol{\theta}^{[0]}}(t_i)\right)^2 / 2\sigma^2 \quad \approx \int_0^\infty \tfrac{1}{2}\left(y_{\boldsymbol{\theta}}(t) - y_{\boldsymbol{\theta}^{[0]}}(t)\right)^2 \, \mathrm{d}t. \tag{S2.7}$$

where for convenience (since our data is perfect) we set $\sigma^2 = 1/\Delta t$. We shall use the continuum approximation to evaluate the Hessian at the best fit.

To start, suppose $d(t)$ has two decay rates $\boldsymbol{\theta}^{[0]} = [1, 2]$, so the data $d(t) = \frac{1}{2}(\exp(-t) + \exp(-2t))$.

(a) *Write a function that returns $y_{\boldsymbol{\theta}}(t)$, and a function that computes the cost for $\Delta t = 0.01$. Draw a contour plot of $C$ in the square $0.5 < \theta_\alpha < 2.5$, with contours at $C = \{2^{-12}, 2^{-11}, \ldots, 2^0\}$. Set the number of grid points per side to 40 (so $\Delta\theta = 0.02$) to see the two minima.*

[3] Hints for the computations can be found at the book website [17].

The diagonal in this plot gives single exponential decays. How well does a single exponential capture the behavior at $\boldsymbol{\theta}^{[0]}$?

(b) *Constraining $\theta_1 = \theta_2$, find the point of minimum cost $\theta_{\min}$. Where is the point on the contour plot? Compare the two curves $y_{\theta^{[0]}}(t)$ and $y_{\theta_{\min}}(t)$, and also plot their difference.*

One can see from the contour plot that measuring the two rate constants separately would be a challenge. This is because the two exponentials have similar shapes, so increasing one decay rate and decreasing the other can almost perfectly compensate for one another.

This clearly is not a deep truth for two exponentials. But the effect is hugely magnified when we have many parameters. We can see this by computing the eigenvalues of the cost Hessian.

(c) *Analytically calculate the Jacobian $J_{t\alpha} = \partial y_{\boldsymbol{\theta}}(t)/\partial\theta_\alpha$ in the continuum approximation (eqn S2.7). Using the Jacobian, show that the Hessian for the cost evaluated at the best fit is*

$$\mathcal{H}_{\alpha\beta} = \left.\frac{\partial^2 C(\boldsymbol{\theta},\boldsymbol{\theta}_0)}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}^{[0]}} = \frac{2}{N^2}\frac{1}{(\theta_\alpha+\theta_\beta)^3}. \tag{S2.8}$$

(Hint: See the discussion below eqn 2.5.)

(d) *Using your answer from part (c), write a routine to calculate the entire array $H(\boldsymbol{\theta})$. Check it by examining the eigenvectors and eigenvalues for the $N = 2$ case of part (b). What do you predict the ratio $R$ = (long axis/short axis) to be, in terms of the two eigenvalues $\lambda_{\text{stiffer}}$ and $\lambda_{\text{sloppier}}$? Are the directions roughly in line with the eigenvectors?*

(e) *For a sum of seven exponentials, with $\boldsymbol{\theta}^{[0]} = [1, 2, 3, \ldots, 7]$, construct the Hessian, and find its eigenvalues. Are they sloppy (roughly equally spaced in log)? By roughly what factor does each successive eigenvalue shrink?*

This sloppiness makes it strikingly difficult to extract the parameter values from the data.

(f) *Argue that the number of measurements $n_{\text{measure}}$ needed to estimate a parameter scales inversely with its variance ($n_{\text{measure}} \sim 1/\sigma^2$). Given that the eigenvalues of the Hessian give the variance along the various eigendirections, by what factor $n_{\text{sloppy}}/n_{\text{stiff}}$ is it harder to measure the parameters along the sloppy directions, for your sum of seven exponentials?*

(g) *Given that the diagonal elements of the inverse cost Hessian, $(\mathcal{H}^{-1})_{\alpha\alpha}$ are proportional to the variance in parameter $\alpha$ for one sampling of the Gaussian given by the cost, what are the variances in the seven parameters $\theta_\alpha^{[0]}$?*

## S2.2 **Sloppy monomials.**[4] (Statistics) ③

The same function $f(x)$ can be approximated in many ways. Indeed, the same function can be fit in the same interval by the same type of function in several different ways! For

[4] Thanks to Joshua Waterfall, whose research is described here.

example, in the interval $[0,1]$, the function $\sin(2\pi x)$ can be approximated (badly) by a fifth-order Taylor expansion, a Chebyshev polynomial, or a least-squares (Legendre[5] ) fit:

$$\begin{aligned} f(x) &= \sin(2\pi x) \\ f_{\text{Taylor}} &\approx 0.000 + 6.283x + \ 0.000x^2 - 41.342x^3 \\ &\quad + \ 0.000x^4 + 81.605x^5 \\ f_{\text{Chebyshev}} &\approx 0.0066 + 5.652x + \ 9.701x^2 - 95.455x^3 \\ &\quad + 133.48x^4 - 53.39x^5 \\ f_{\text{Legendre}} &\approx 0.016 + 5.410x + 11.304x^2 - 99.637x^3 \\ &\quad + 138.15x^4 - 55.26x^5 \end{aligned}$$

It is not a surprise that the best fit polynomial differs from the Taylor expansion, since the latter is not a good approximation. But it is a surprise that the last two polynomials are so different. The maximum error for Legendre is less than 0.02, and for Chebyshev is less than 0.01, even though the two polynomials differ by

$$\begin{aligned} \text{Chebyshev} - \text{Legendre} = & \\ & - 0.0094 + 0.242x - 1.603x^2 \\ & + 4.182x^3 - 4.67x^4 + 1.87x^5 \end{aligned} \tag{S2.9}$$

a polynomial with coefficients two hundred times larger than the maximum difference!

(a) *Plot $f(x)$, $f_{\text{Legendre}}$, and $f_{\text{Chebyshev}}(x)$ between zero and one on the same graph. Plot $f(x) - f_{\text{Legendre}}$ and $f(x) - f_{\text{Chebyshev}}(x)$ on the same graph with the same range. The first minimizes the squared difference on $[0,1]$ (eqn S2.10), but it has large errors near the edges. If you were writing a routine to use for calculating $\sin(2\pi x)$ to machine precision in this range, would it be better to use the Legendre or the Chebyshev approximation? Now plot Plot $f_{\text{Chebyshev}}(x) - f_{\text{Legendre}}$ in the range $-1, 2$. Does it indeed get much flatter than you would expect given the coefficients?*

This flexibility in the coefficients of the polynomial expansion is remarkable. We can study it by considering the dependence of the quality of the fit on the parameters. Least-squares (Legendre) fits minimize a cost $C$, the integral of the squared difference between the polynomial and the function:

$$\begin{aligned} C &= (1/2) \int_0^1 (f(x) - y_{\boldsymbol{\theta}}(x))^2 \, \mathrm{d}x, \\ y_{\boldsymbol{\theta}}(x) &= \sum_{\alpha=0}^{N-1} \theta_\alpha x^\alpha \end{aligned} \tag{S2.10}$$

[5]The orthogonal polynomials used for least-squares fits on [-1,1] are the Legendre polynomials, assuming continuous data points. Were we using orthogonal polynomials for this exercise, we would need to shift them for use in [0,1].

How quickly does this cost increase as we move the $N$ parameters $\theta_\alpha$ away from their best-fit values? Varying any one monomial coefficient will of course make the fit bad. But apparently certain coordinated changes of coefficients do not cost much—for example, the difference between least-squares and Chebyshev fits given in eqn S2.9.

How should we explore the dependence in arbitrary directions in parameter space? We can use the eigenvalues of the Hessian to see how sensitive the fit is to moves along the various eigenvectors...

(b) *Note that the first derivative of the cost $C$ is zero at the best fit. Analytically (paper and pencil) show that the Hessian second derivative of the cost in eqn S2.10 is*

$$\mathcal{H}_{\alpha\beta} = \frac{\partial^2 C}{\partial\theta_\alpha \partial\theta_\beta} = \frac{1}{\alpha+\beta+1}. \tag{S2.11}$$

This Hessian is the Hilbert matrix, famous for being ill-conditioned (having a huge range of eigenvalues). Tiny eigenvalues of $\mathcal{H}$ correspond to directions in polynomial space where the fit does not change.

(c) *Numerically calculate the eigenvalues of the $6\times6$ Hessian for fifth-degree polynomial fits. Do they indeed span a large range? How big is the condition number (the ratio of the largest to the smallest eigenvalue)? Are the ratios all approximately equal (a characteristic of sloppy models)?*

Notice from Eqn S2.11 that the dependence of the polynomial fit on the monomial coefficients is *independent of the function $f(x)$* being fitted. We can thus vividly illustrate the sloppiness of polynomial fits by considering fits to the *zero* function $f(x) \equiv 0$. A polynomial given by an eigenvector of the Hilbert matrix with small eigenvalue must stay close to zero everywhere in the range $[0, 1]$. Let us check this.

(d) *Calculate the eigenvector corresponding to the smallest eigenvalue of $\mathcal{H}$, checking to make sure its norm is one (so the coefficients are of order one). Note that the elements of this vector are the coefficients of a polynomial pertubation $\delta f(x)$ that changes the cost the smallest amount for a unit vector $\boldsymbol{\theta}$. What is that polynomial? Plot the corresponding polynomial in the range $[0, 1]$: does it stay small everywhere in the interval?*

Especially for larger $M$, the monomial coefficients of the best fit to a function become sloppy—they can vary over large ranges without damaging the fit, if the other coefficients are allowed to compensate. Only a few combinations of coefficients (those of the largest Hessian eigenvalues) are well determined. This turns out to be a fundamental property that is shared with many other multiparameter fitting problems. Many different terms are used to describe this property. The fits are called *ill-conditioned*: the parameters $\theta_n$ are not well constrained by the data. The *inverse problem* is challenging: one cannot practically extract the parameters from the behavior of the model. Or, as our group describes it, the fit is *sloppy*: only a few directions in parameter space (eigenvectors corresponding to the largest eigenvalues) are constrained by the data, and there is a huge space of models (polynomials) varying along sloppy directions that all serve well in describing the data.

At root, the problem with polynomial fits is that all monomials $x^n$ have similar shapes on $[0, 1]$: they all start flat near zero and bend upward. Thus they can be traded for one another; the coefficient of $x^4$ can be lowered without changing the fit if the coefficients of $x^3$ and $x^5$ are suitably adjusted to compensate.

One should note that, were we change basis from the coefficients $\theta_n$ of the monomials $x^n$ to the coefficients $\ell_n$ of the orthogonal (shifted Legendre) polynomials, the situation completely changes. The Legendre polynomials are designed to be different in shape (orthogonal), and hence cannot be traded for one another. Their coefficients $\ell_n$ are thus well determined by the data, and indeed the Hessian for the cost $C$ in terms of this new basis is the identity matrix. This puzzled us for some time—is the sloppiness intrinsic, or just a sign of a poor choice of variables. Later work, examining the predictions of nonlinear models using *information geometry*, resolved this question: sloppiness is under rather general conditions expected for the collective predictions of multiparameter nonlinear models.

S2.3 **Nonlinear fits.** (Statistics, Information geometry) ③

In this exercise, we briefly introduce some geometrical features of nonlinear model fits to data. These fits involve unknown parameters $\theta_\alpha$, control parameters $t_i$ describing different experimental conditions, experimental data $d_i$ taken under these different conditions, and a nonlinear function $y_i(\boldsymbol{\theta})$ that makes a prediction for the data given values for the parameters. As an example, we might fit a sum of two decaying exponentials to (say) the decay of radiation from a mixture of radioactive elements with unknown decay rates (see [21, 22] and Exercise S1.3.) Our model is

$$\mathbf{y}_{\boldsymbol{\theta}}(t) = \exp(-\theta_1 t) + \exp(-\theta_2 t). \tag{S2.12}$$

Here the parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ are the decay rates, the control parameter is $t$ the time elapsed, and the data $\mathbf{d} = \{d_i\}$ are the counts from a Geiger counter. We shall assume that the experimental data points $\{d_i \pm \sigma_i\}$ have independent measurement errors with Gaussian distributions of standard deviation $\sigma_i$.

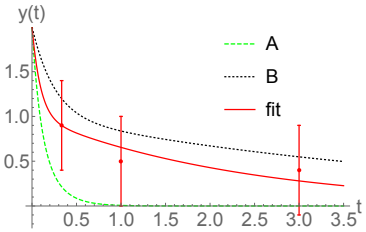



**Fig. S2.2 Fitting a nonlinear function to data**, here a sum of two exponentials to three data points $y(1/3) = 0.9 \pm 0.5$, $y(1) = 0.5 \pm 0.5$, and $y(2) = 0.4 \pm 0.5$. Fit A decays too quickly and fit B too slowly, although both are within statistical errors.

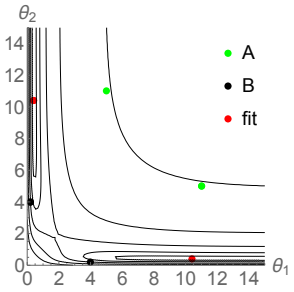



**Fig. S2.3 Contours of constant cost** $C = \chi^2/2$ in parameter space. Notice the symmetry reflecting around $\theta_1 = \theta_2$. Notice also the narrow canyons—one can fit the data well with $\theta_2 = \infty$ (a single exponential decaying from $y(0) = 1$), a point on the edge of the model manifold.

A nonlinear least-squares fit varies the parameters to minimize a cost

$$C(\boldsymbol{\theta}) = \chi^2/2 = \sum_i (y_i(\theta) - d_i)^2/2\sigma_i^2. \tag{S2.13}$$

The cost is half of what the statisticians call $\chi^2$ (pronounced "chi squared").

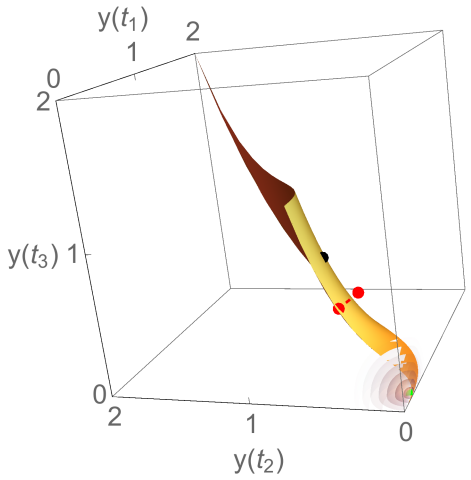



**Fig. S2.4 Nonlinear model predictions in data space**. The curved surface represents the *model manifold*—the surface of predictions in data space formed by varying the parameters of our nonlinear model. (We rescale the axes by the associated error bars.) The upper dot represents fit B. The dot at the lower right represents fit A, with the fuzzy sphere representing the range of experimental predictions around the fit. The two other dots represent the data and the best fit (the nearest point to the model manifold in data space). Note that the best fit is nearly at an edge of the model manifold.

First, let us provide a few interpretations of the cost.

(a) *[i] Interpret $C$ as half the squared distance in a* data space *(Fig. S2.4) which has one coordinate per experimental measurement. What is the metric tensor $g_{ij}$ in data space, in terms of the error bars $\sigma_i$?*[6] *[ii] Suppose the experimental data points $d_i$ have errors that are distributed as independent Gaussians of standard deviation $\sigma_i$. How is our cost related to the log-likelihood that the data would have arisen from our model? [iii] View $C$ as a Hamiltonian, and the likelihood $P(\mathbf{d}|\boldsymbol{\theta})$ giving the probability of observing data $\mathbf{d}$ in data space as being a Boltzmann distribution. What is the temperature $T$?*

Statistical mechanics focuses on predicting the behavior (probability distribution in phase space) for a Hamiltonian $\mathcal{H}(T, P, V, N)$ with known parameters. At fixed temperature, the probability density is proportional to $\exp(-\mathcal{H}/k_BT)$, where Liouville's theorem tells us how to measure phase-space volume. Statistics predicts the distribution of data points for a model $\mathbf{y}(\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. The probability distribution is proportional to $\exp(-C)$ per unit volume in data space, where the distance between two points in data space is determined by the error bars on each data point.

Finding the distribution of data points for a given set of parameters in statistics is not a challenge.

(b) *Argue, for equal Gaussian measurement errors, that the predicted distribution of data points for a given set of parameters $\boldsymbol{\theta}$ is just a blurred, Gaussian sphere in data space (as in the lower right corner of Fig. S2.4). For general $\sigma_i$, make an analogy with the momentum distribution of classical particles with different masses to describe the probability distribution.*

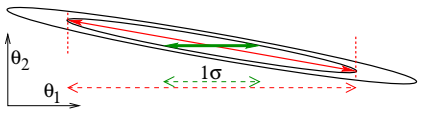



**Fig. S2.5 Error estimates for fit parameters**. Contours of constant cost $C$ in parameter space $\boldsymbol{\theta}$ near the best fit, ignoring anharmonicity. The ellipse axes are $\mathbf{e}_v = (1/6, 1)$ and $\mathbf{e}_h = (1, -1/6)$. The $1\sigma$ range of $\theta_1$ keeping $\theta_2$ fixed is the solid arrow. The total uncertainty $\sigma_1$ for $\theta_1$ includes fluctuations of $\theta_2$ (solid diagonal arrow), $\sigma_1^2 = \Sigma_{11} = (H^{-1})_{11}$ (long dashed range).

Our job in nonlinear fitting is to estimate the probabilities of different choices of parameters given the experimental data. Surely we expect the true parameters to have a large probability $P(\mathbf{d}|\boldsymbol{\theta})$ of generating the experimental data—the true $\boldsymbol{\theta}$ will be somewhere near the best fit $\boldsymbol{\theta}^{\text{best}}$ that minimizes the cost $C = \chi^2$. Let us assume that the probability $P(\boldsymbol{\theta}|\mathbf{d})$ of finding a set of parameters given the data is proportional to the probability $P(\mathbf{d}|\boldsymbol{\theta})$ that the model would have generated that data. (See Exercise S1.2 for a discussion of *priors* in Bayesian statistics.) Let us also assume that we are estimating the parameters well enough that we may approximate the cost by a Taylor

[6] The metric tensor $g_{ij}$ on a Riemannian manifold gives the distance between nearby points. If the two points have coordinates $\mathbf{x}$ and $\mathbf{x} + \boldsymbol{\Delta}$ and $\boldsymbol{\Delta}$ is small, then the squared distance is $\sum_{ij} g_{ij}\Delta_i\Delta_j$.

expansion up to second order about the maximum likelihood. If $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{best}} + \boldsymbol{\Delta}$ for small $\boldsymbol{\Delta}$,

$$C(\boldsymbol{\theta}) \approx C^{\text{best}} + \tfrac{1}{2} H_{\alpha\beta} \Delta_\alpha \Delta_\beta, \tag{S2.14}$$

where we shall call

$$H_{\alpha\beta} = \frac{\partial^2 C}{\partial \theta_\alpha \partial \theta_\beta} \tag{S2.15}$$

the cost Hessian.

(c) *Which of the eigenvectors* $\mathbf{e}_v$ *or* $\mathbf{e}_h$ *in Fig. S2.5 corresponds to a stiff direction (larger eigenvalue of* $H$*)? Which is sloppier? Verify that the probability distribution of* $\theta_1$ *holding all other variables fixed is a normal distribution with variance* $H_{11}$ *(short horizontal dashed range).*

It is more of a challenge to calculate the error in our estimate of $\theta_1$ *allowing the other variables* to vary freely (long horizontal dashed range). The variance of the estimate of a variable is given by the corresponding diagonal element of the *covariance matrix* $\Sigma = H^{-1}$, the inverse of the Hessian.

If $P(\boldsymbol{\theta})$ is approximately a multidimensional Gaussian, then the variance in $\theta_1$ is given by

$$\begin{aligned} \langle \Delta_1^2 \rangle &= \int \Delta_1^2 P(\boldsymbol{\Delta}) \mathrm{d}\boldsymbol{\Delta} \\ &= \int \frac{\Delta_1^2}{Z} \mathrm{e}^{-\frac{1}{2}\sum_{\alpha\beta} \Delta_\alpha H_{\alpha\beta} \Delta_\beta} \mathrm{d}\boldsymbol{\Delta}, \end{aligned} \tag{S2.16}$$

where

$$Z = \int \exp\left( -\tfrac{1}{2} \sum_{\alpha\beta} \Delta_\alpha H_{\alpha\beta} \Delta_\beta \right) \mathrm{d}\boldsymbol{\Delta}. \tag{S2.17}$$

is the normalization factor.

In statistical mechanics, a key method for calculating expectation values $\langle X^n \rangle$ in a Boltzmann distributions is to add a *source term* $\lambda X$ to the Hamiltonian, shifting the partition function to $Z(\lambda) = \sum \exp(-(H + \lambda X)/k_B T)$, with free energy $F(\lambda) = -k_B T \log Z$. Then

$$\begin{aligned} \left. \frac{\mathrm{d}F}{\mathrm{d}\lambda} \right|_{\lambda=0} &= \frac{-k_B T}{Z} \left. \frac{\mathrm{d}Z}{\mathrm{d}\lambda} \right|_{\lambda=0} \\ &= \frac{\cancel{-k_B T}}{Z} \int \frac{X}{\cancel{-k_B T}} \mathrm{e}^{-H/k_B T} \\ &= \langle X \rangle \end{aligned} \tag{S2.18}$$

and

$$\begin{aligned}\left.\frac{\mathrm{d}^2F}{\mathrm{d}\lambda^2}\right|_{\lambda=0} &= \left.\frac{k_BT}{Z^2}\left(\frac{\mathrm{d}Z}{\mathrm{d}\lambda}\right)^2\right|_{\lambda=0} - \left.\frac{k_BT}{Z}\frac{\mathrm{d}^2Z}{\mathrm{d}\lambda^2}\right|_{\lambda=0} \\ &= \frac{\langle X\rangle^2}{k_BT} - \frac{k_BT}{Z}\int \frac{X^2}{(k_BT)^2}\mathrm{e}^{-H/k_BT} \\ &= \frac{\langle X^2\rangle - \langle X\rangle^2}{-k_BT} \\ &= \frac{\left\langle (X-\langle X\rangle)^2\right\rangle}{-k_BT}\end{aligned} \tag{S2.19}$$

We can use this method to calculate $\langle \Delta_1^2\rangle$.

(d) *Add the source term* $\boldsymbol{\lambda}\cdot\boldsymbol{\Delta} = \lambda\Delta_1$ *to our cost, where* $\boldsymbol{\lambda}^T = (\lambda, 0, 0, \dots)$ *is* $\lambda$ *times a unit vector in the shared* $\theta_1$ *and* $\Delta_1$ *direction, so*

$$Z(\lambda) = \int \mathrm{e}^{-\frac{1}{2}\boldsymbol{\Delta}^T H\boldsymbol{\Delta} - \boldsymbol{\lambda}\cdot\boldsymbol{\Delta}}\mathrm{d}\boldsymbol{\Delta}. \tag{S2.20}$$

*Complete the square, and show that* $Z(\lambda) = \exp(\frac{1}{2}\lambda^2\Sigma_{11})Z(0)$. *Use eqn S2.19 to show that* $\langle\Delta_1^2\rangle = \Sigma_{11}$.

There is a commonly used approximation to the cost Hessian that has important geometrical significance.

(e) *[i] Write the cost Hessian in eqn S2.13 in terms of first and second derivatives of* $y_i(\boldsymbol{\theta})$. *[ii] If we take the cost Hessian at a point where* $\mathbf{d} = \mathbf{y}(\boldsymbol{\theta})$ *on the model manifold, show that* $H_{\alpha\beta} = \sum_i(\partial y_i/\partial\theta_\alpha)(\partial y_i/\partial\theta_\beta) = (J^TJ)_{\alpha\beta}$, *where* $J_{i\alpha} = (1/\sigma_i)\partial y_i/\partial\theta_\alpha$ *is the Jacobian. [iii] Show that the squared distance in data space between two model predictions* $\mathbf{y}(\boldsymbol{\theta})$ *and* $\mathbf{y}(\boldsymbol{\theta}+\boldsymbol{\Delta})$ *is given for small* $\boldsymbol{\Delta}$ *by the metric tensor* $g_{\alpha\beta} = (J^TJ)_{\alpha\beta}$.

$g_{\alpha\beta} = (J^TJ)_{\alpha\beta} = J_{i\alpha}J_{i\beta}$ is the *induced* metric on the model manifold, inherited from the embedding data space metric $g_{ij}$ of part (a). $g = J^TJ$ is called the *Fisher information matrix* in the statistics community.

S2.4 **Fisher information and Cramér–Rao.**[7] (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 $\boldsymbol{\theta}$.

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

- 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 $P(\mathbf{x}|\boldsymbol{\theta}) = \exp(-\mathcal{H}/k_B T)/Z$ in phase space (Chapter 3, Exercise S2.7).
- 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 S1.2).
- 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

$$P(\mathbf{x}|\boldsymbol{\theta}) = \frac{\mathrm{e}^{-\sum_i (y_i(\theta) - x_i)^2/2\sigma^2}}{(2\pi\sigma^2)^{N/2}}. \tag{S2.21}$$

  (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 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:

$$g_{\alpha\beta}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \log P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\mathbf{x}} \tag{S2.22}$$

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 3.2 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* $\epsilon$. *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* $\Sigma$, where the variance of the likely values of parameter $\theta_\alpha$ is given by $\Sigma_{\alpha\alpha}$, and where $\Sigma_{\alpha\beta}$ gives the correlations between two parameters $\theta_\alpha$ and $\theta_\beta$. 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

$$\Sigma_{\alpha\alpha} \geq (g^{-1})_{\alpha\alpha}. \tag{S2.23}$$

(if the estimator is *unbiased*, see Exercise S1.2). This is another justification for using the FIM as our natural distance metric in probability space.

In Exercise S6.4, 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 S6.5, S2.6, and S2.7, 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.

S2.5 **Gibbs for pistons.** (Thermodynamics) ④

The degrees of freedom in a piston are $\mathbf{X} = \{\mathbb{P}, \mathbb{Q}, V\}$, where $\mathbb{P}$ and $\mathbb{Q}$ are the $3N$ positions and momenta of the particles, and $V$ is the current volume of the piston. The Gibbs ensemble for a piston is the probability density

$$\rho = (1/\Gamma)\exp(-\beta\mathcal{H}(\mathbb{P}, \mathbb{Q}) - \beta PV). \tag{S2.24}$$

Here $\Gamma$ is the partition function for the Gibbs ensemble, normalizing the distribution to one.

*Let our piston be filled with an ideal gas of particles of mass $m$. What is the partition function $Z(V,\beta)$ for the canonical ensemble? (Be sure to include the Gibbs factor $N!$; the quantum phase-space refinements are optional.) Show that the partition function for the Gibbs ensemble is*

$$\Gamma(P,\beta) = (2\pi m/\beta)^{3N/2}(\beta P)^{-(N+1)}, \tag{S2.25}$$

*Show that the joint probability density for finding the $N$ particles with $3N$ dimensional momenta $\mathbb{P}$, the piston with volume $V$, and the $3N$ dimensional positions $\mathbb{Q}$ inside $V$ (eqn S2.24), is*

$$\rho_{\text{Gibbs}}(\mathbb{P},\mathbb{Q},V|P,\beta) = (1/\Gamma(P,\beta))\mathrm{e}^{-\beta\mathbb{P}^2/2m-\beta PV}. \tag{S2.26}$$

## S2.6 Pistons in probability space.[8] (Mathematics, Information geometry) ④

Fig. 5.3 shows the Carnot cycle as a path in the $P$–$V$ space of pressure and volume—parameters varied from the outside. One could draw a similar diagram in the space of pressure and temperature, or volume and temperature. Here we shall explore how to describe the path in the space of *probability distributions*. In the process, we shall compute the *model manifold* of the ideal gas, and show that it is a two-dimensional plane.

As discussed in Exercise S2.4, there is a natural distance, or metric, in the space of probability distributions:

$$g_{\mu\nu} = -\left\langle \frac{\partial^2 \log(\rho)}{\partial\theta_\mu \partial\theta_\nu} \right\rangle, \tag{S2.27}$$

the *Fisher information metric*. So, a system in the Gibbs ensemble is described in terms of two parameters, usually $P$ and $T$. We shall instead use the "natural" parameters $\theta_1 = p = \beta P$ and $\theta_2 = \beta$, where $\beta = 1/k_B T$ (see Exercise S2.7). The squared distance in probability space between two systems with tiny changes in pressure and temperature is then

$$d^2(\rho(\mathbf{X}|\boldsymbol{\theta}), \rho(\mathbf{X}|\boldsymbol{\theta}+\mathrm{d}\boldsymbol{\theta})) = g_{\mu\nu}\mathrm{d}\theta_\mu \mathrm{d}\theta_\nu. \tag{S2.28}$$

(a) *Compute $g^{(\mathrm{p},\beta)}_{\mu\nu} = -\langle \partial^2 \log(\rho)/\partial\theta_\mu\partial\theta_\nu\rangle$ using eqn S2.26 from Exercise S2.5.*

The metric tensor $g^{(\mathrm{p},\beta)}$ for the Gibbs ensemble of the piston tells us the distance in probability space between neighboring pressures and temperatures. What kind of surface (the *model manifold*) is formed by this two-parameter family of probability distributions? Does it have an intrinsic curvature?

(b) *Show that one can turn the metric tensor into the identity $g^{(\mathrm{x},\mathrm{y})}_{\mu\nu} = \delta_{\mu\nu}$ by a coordinate transformation $(p,\beta) \to (x = A\log(p), y = B\log(\beta))$. What are the necessary scale factors $A$ and $B$?*

[8] This exercise was developed in collaboration with Ben Machta, Archishman Raju, Colin Clement, and Katherine Quinn

Hence the model manifold of the piston in the Gibbs ensemble is a plane! We can draw our control paths in the $(x, y)$ plane. We label the four steps of the Carnot cycle as in Fig. 5.3.

(c) *Draw the Carnot cycle path in as a parameterized curve in* $(x, y)$*, with* $P_a = 1$*,* $P_b = 0.5$*,* $T_1 = 1$ *and* $T_2 = 0.8$*, for* $N = 1$*.* (Hint: eqn 5.8 will be helpful in finding the adiabatic parts of the path $p(\beta)$.) *Is the length of the expansion at fixed pressure the same as you calculated in Exercise S6.5?*

S2.7 **FIM for Gibbs.**[9] (Mathematics, Thermodynamics, Information geometry) ④

In this exercise, we study the geometry in the space of probability distributions defined by the Gibbs ensemble[10] of a general equilibrium system. We compute the Fisher Information Metric (FIM, Exercises S2.4 and S2.6)

$$g_{\mu\nu} = -\left\langle \frac{\partial^2 \log(\rho)}{\partial\theta_\mu \partial\theta_\nu} \right\rangle, \tag{S2.29}$$

of the Gibbs phase space ensemble $\rho(\mathbb{P}, \mathbb{Q})$ in terms of thermodynamic properties of the system.

In Exercise S2.6 we calculated $g_{\mu\nu}$ for the ideal gas, using the "natural" variables $\theta_1 = p = \beta P$ and $\theta_2 = \beta$, rather than $P$ and $T$. Why are these coordinates special? The log of the Gibbs probability distribution for an arbitrary interacting collection of particles with Hamiltonian $\mathcal{H}$ (eqn S2.24) is

$$\begin{aligned} \log(\rho) &= -\beta\mathcal{H}(\mathbb{P}, \mathbb{Q}) - \beta P V - \log\Gamma \\ &= -\beta\mathcal{H}(\mathbb{P}, \mathbb{Q}) - pV - \log\Gamma. \end{aligned} \tag{S2.30}$$

This is the logarithm of the partition function $\Gamma$ plus terms *linear* in $p = \beta P$ and $\beta$.[11] So the second derivatives with respect to $p$ and $\beta$ only involve $\log(\Gamma)$. We know that the Gibbs free energy $G(p, \beta) = -k_B T \log(\Gamma) = -(1/\beta)\log(\Gamma(p, \beta))$, so $\log(\Gamma) = -\beta G(p, \beta)$. The first derivatives of the Gibbs free energy $\mathrm{d}G = -S\mathrm{d}T + V\mathrm{d}P + \mu\mathrm{d}N$ are related to things like volume and entropy and chemical potential; our metric is given by the second derivatives (compressibility, specific heat, ...)

(a) *For a collection of particles interacting with Hamiltonian* $\mathcal{H}$*, relate the four terms* $g_{\mu\nu}^{(\mathrm{p},\beta)}$ *in terms of physical quantities given by the second derivatives of* $G$*. Write your answer in terms of* $N$*,* $p$*,* $\beta$*, the particle density* $\rho = N/\langle V \rangle$*, the isothermal*

[9] This exercise was developed in collaboration with Ben Machta, Archishman Raju, Colin Clement, and Katherine Quinn

[10] The Fisher information distance is badly defined except for changes in *intensive* quantities. In a microcanonical ensemble, for example, the energy $E$ is constant and so the derivative $\partial\rho/\partial E$ would be the derivative of a $\delta$ function. So we study pistons varying $P$ and $\beta = 1/k_B T$, rather than at fixed volume or energy.

[11] In statistics, log probability distributions which depend on parameters in this linear fashion are called *exponential families*. Many common distributions, including lots of statistical mechanical models like ours, are exponential families.

*compressibility* $\kappa = -(1/\langle V\rangle)(\partial\langle V\rangle/\partial P)|_T$, *the thermal expansion coefficient* $\alpha = (1/\langle V\rangle)(\partial\langle V\rangle/\partial T)|_P$, *and the specific heat per particle at constant pressure,* $c_{\mathrm{P}} = (T/N)(\partial S/\partial T)|_P$. (Hint: $G(P,T) = G(p/\beta, 1/\beta)$. Your answer will be a bit less complicated if you pull out an overall factor of $N/(\rho\beta^2)$.)

The metric tensor for a general Hamiltonian is a bit simpler in the more usual coordinates $(P,\beta)$ or $(P,T)$.

(b) *Show that*

$$g^{(\mathrm{P},\beta)} = N \begin{pmatrix} \beta\kappa/\rho & \alpha/\beta\rho \\ \alpha/\beta\rho & c_{\mathrm{P}}/\beta^2 \end{pmatrix}$$

*and*

$$g^{(\mathrm{P},\mathrm{T})} = N \begin{pmatrix} \kappa/\rho T & -\alpha/\rho T \\ -\alpha/\rho T & c_{\mathrm{P}}/T^2 \end{pmatrix}.$$

(c) *Calculate* $g^{(\mathrm{p},\beta)}$ *for the ideal gas using your answer from part (a). Compare with your results calculating* $g^{(\mathrm{p},\beta)}$ *directly from the probability distribution in Exercise S2.6. Is the difference significant for macroscopic systems?* (Hint: If you use $G = A + PV$ directly from eqn 6.24, remember that the thermal de Broglie wavelength $\lambda$ depends on temperature.)

The standard formulas for an ideal gas do not include the piston wall as a degree of freedom, so part (c) has one fewer positional degree of freedom than in Exercise S2.6. That is, the macroscopic calculation neglects the entropic contribution of the fluctuations in volume (the position of the piston inside the cylinder).

# Chapter 3

# Model manifolds and hyperribbons

Information geometry studies the shape and topology of the model manifold (a) to find better algorithms for fitting experimental data [**?**], (b) to understand why complex microscopic models often exhibit simple emergent behavior [**?**], and (c) to derive reduced, effective models that describe the experiments just as well with fewer parameters [**?**]. It has been applied to complex models in many fields [**?**]. Visualizations of the model manifold are useful because they often are well approximated with low-dimensional projections, like principal component analysis (PCA). Unlike other low-dimensional representations, PCA preserves the pairwise distances between the prediction vectors $\mathbf{y}$. Three-dimensional visualizations are useful because many models are *sloppy*; most of their behavior is governed by a few combinations of parameters. Their allowed predictions form the model manifold: for $N$ parameters and $M$ predictions they form $N$-dimensional volumes in the $M$-dimensional behavior space. Sloppy systems have model manifolds that form *hyperribbons* – volumes that are longer than they are wide, wider than they are thick, ... hierarchically thinner and thinner $N-1$ times.

## 3.1 The Jacobian and the metric tensor $g_{\alpha\beta}$.

So far, we have been characterizing sloppy models by their behavior near one point in parameter space. We have mostly discussed the eigenvalues of the Hessian $\mathcal{H}_{\alpha\beta} = \partial_\alpha \partial_\beta C(\theta)$. We have also discussed the Jacobian $J_{i\alpha} = \partial y(x_i)/\partial\theta_\alpha$, which describes the variation in behavior as parameters are changed. The eigenvalues of the Hessian for sloppy models show a hierarchy of scales (Fig. 1.1) in models drawn from a variety of scientific disciplines. This Hessian is usually well approximated using the Jacobian $\mathcal{H} \approx J^T J$. The hierarchy of eigenvalues of the Hessian is reflected in the *skewness* of the Jacobian (Fig. 3.1).

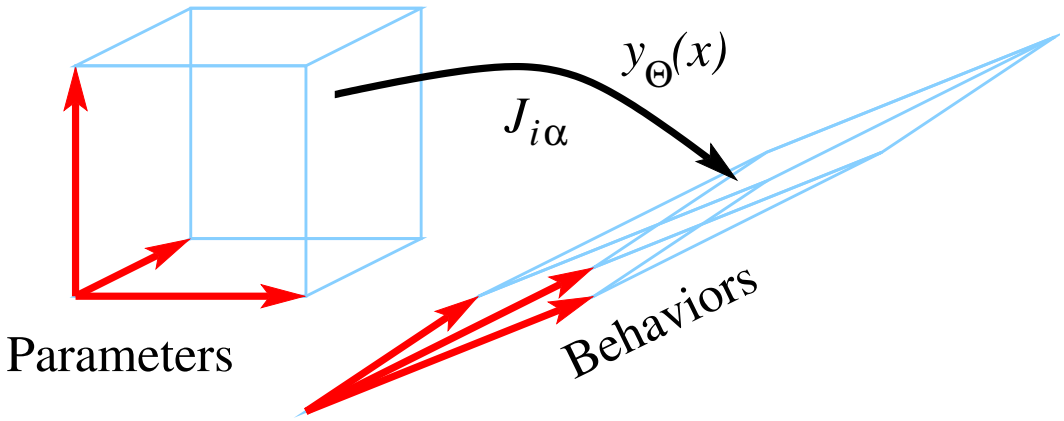



**Fig. 3.1 Skewness and sloppiness**. The mapping $y_{\boldsymbol{\theta}}$ from parameter space to behavior space in sloppy models is extremely skewed. Many parameter directions lead to similar changes in behavior. This skewness, in a small region around one point, is reflected in the Jacobian $J_{i\alpha} = \partial y(x_i)/\partial\theta_\alpha$. $J$ takes a cube formed by the parameter axes to a volume in parameter space that gets squeezed into a long, thin volume that gets thinner and thinner as more parameters are added. In nonlinear models, it is this skewness that leads to the hyperribbon structure of the model manifold.

In this section, we shall move our focus from parameter space to behavior space. Recall that, for least-squares models, the model manifold is the surface $\mathbf{y}(\boldsymbol{\theta})$ in behavior space swept out as the parameters $\boldsymbol{\theta}$ are varied through all allowed values. The parameters can be viewed as coordinates on the model manifold.

The metric tensor $g_{\alpha\beta}$ tells us how far apart two predictions are when the parameters are changes by a small distance $\boldsymbol{\theta}' - \boldsymbol{\theta} = \boldsymbol{\delta}$: $|\mathbf{y}(\boldsymbol{\theta}') - \mathbf{y}(\boldsymbol{\theta})|2 = \delta_\alpha\delta_\beta g_{\alpha\beta}$. For our NLLS model, this is written in terms of the Jacobian $J_{i\alpha} = \partial y_i/\partial\theta_\alpha$:

$$\begin{aligned}\sum_i (y_i(\boldsymbol{\theta}') - y_i(\boldsymbol{\theta}))^2 &= \sum_i \left(\sum_\alpha \delta_\alpha \frac{\partial y_i}{\partial \delta_\alpha}\right)^2 \\ &= \sum_i \left(\sum_\alpha \delta_\alpha \frac{\partial y_i}{\partial \delta_\alpha}\right)\left(\sum_\beta \delta_\beta \frac{\partial y_i}{\partial \delta_\beta}\right) \\ &= \sum_{\alpha\beta} \delta_\alpha\delta_\beta J^T_{\alpha i} J_{i\beta},\end{aligned} \tag{3.1}$$

so $g_{\alpha\beta} = (J^T J)_{\alpha\beta}$ is our metric. Remember from eqn 2.5 that this is also our approximation for the Hessian for NLLS models. The skewness of the linearized mapping $J$ from parameter space into behavior space (Fig. 3.1) gives the large range of eigenvalues for the cost Hessian, and hence the sloppiness we observe. It will also yield the flat, hyperribbon structure of the model manifold.

## 3.2 Model manifolds and behavior space.

The easy case is a model that predicts scalar experimental results with error bars. If a model with $N$ parameters $\theta_\alpha$ predicts $M$ results $y_i(\boldsymbol{\theta})$ with standard deviation $\sigma_i$, and $\sigma_i$ is independent of the parameters, the model manifold is an $N$-dimensional surface swept out in $\mathbb{R}^M$ (Fig. 3.2).

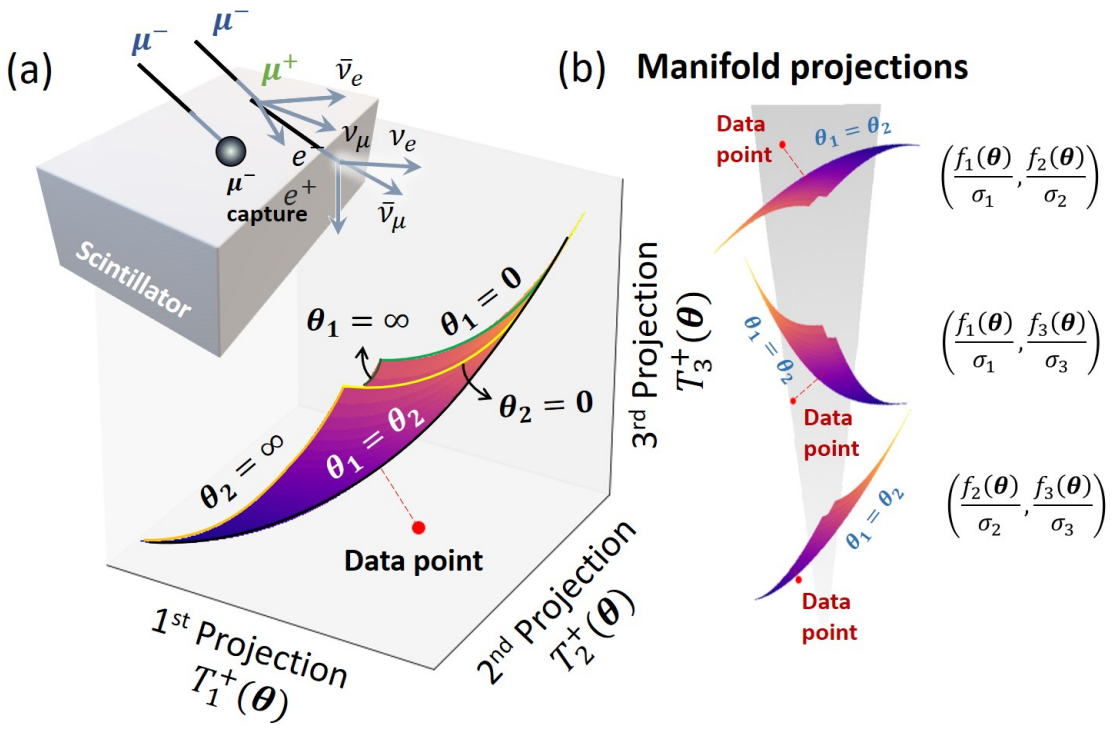



**Fig. 3.2 Model manifold for a least-squares muon decay model**. The 2D surface of predictions as two muon decay rates $(\theta_1, \theta_2)$ are varied to fit experimental measurements $f_i$ at three time-points $t_i$ (from Teoh et al. [19]). One can think of the model manifold as the surface of predictions in $\mathbb{R}^M$ at particular values of the the experimental control variables. (So here (a) shows the model manifold for the three times $t_1$, $t_2$, and $t_3$, and (b) shows the model manifolds ignoring the third time.) Or, one can think of the model manifold as a surface in the infinite-dimensional space of predictions for all possible experimental measurements. (So here (a) and (b) show projections of the model manifold onto three of the axes.)

We are interested not in just any visualization, but in a visualization that separates points generated by parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in a natural way that respects the experiment. This is easy for NLLS models, where, because the errors $\sigma_i$ on the individual data points are normally distributed and independent of parameters, the distances (measured in units of the $\sigma_i$) form

a natural metric in behavior space which immediately gives us the distances between points on the model manifold.

## 3.3 Sloppiness in physics and the Fisher Information Metric

What about more general models? In this section, we shall discuss the natural metric tensor for any model that predicts a probability distribution of possible observations: the Fisher Information Metric (FIM). This metric properly treats non-Gaussian noise and error distributions, and systems whose errors depend on parameters. It covers models like the Ising model for magnets and the ΛCDM model for the microwave background radiation, which predict the probabilities of various snapshots of complex interacting and evolving systems. The eigenvalues of this metric tensor for sloppy probabilistic models also forms a hierarchy, spanning many decades. We shall return to studying the model manifolds for probabilistic models in Chapter 6. There we will search for *isometric* embeddings of model predictions, which preserve the local distances between predictions on the model manifold given by the FIM.

Recall that the natural metric in the space of probability distributions is the Fisher Information matrix,

$$\begin{aligned} g_{\alpha\beta} &= -\left\langle \frac{\partial^2 \log \rho(\mathbf{x})}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle \\ &= -\int \mathrm{d}\mathbf{x} \rho(\mathbf{x}) \frac{\partial^2 \log \rho(\mathbf{x})}{\partial\theta_\alpha \partial\theta_\beta}. \end{aligned} \tag{3.2}$$

In Exercise S6.1, we give a rationalization for the FIM as the natural metric in parameter space, by noting that the space of probability distributions is naturally viewed as the positive "octant" of a hypersphere.

# Exercises

S3.1 **Plotting the model manifold.**[1] (Information geometry, Statistics) ③

In this exercise, we shall use our $N$ parameter model of decaying exponentials explore ways of visualizing the resulting behavior. Remember eqn S2.6 from Exercise S2.1, $y_{\boldsymbol{\theta}}(t) = (1/N)\sum_{\alpha=1}^{N} \exp(-\theta_\alpha t)$.

One way of visualizing the behavior space is to pick two or three quantities of interest, and explore how they vary with one another. This is a *projection* of the model manifold onto the three coordinate axes of interest.

[1] Hints for the computations can be found at the book website [17].

(a) *Taking $N = 2$ different exponents, draw the projection of the the model manifold onto the axes corresponding to $t = \{1/3, 1, 3\}$.* (That is, do a 3D parametric plot of $\{y_{\boldsymbol{\theta}}(1/3), y_{\boldsymbol{\theta}}(1), y_{\boldsymbol{\theta}}(3)\}$, varying $0 < \theta_1 < \infty$ and $0 < \theta_2 < \infty$.) *You'll want some of the $\theta$s to be small and some large.* (I had a range $\sim (10^{-2}$–$10^2$, with equally spaced points in log if on a grid). *Identify the two values of $\theta$ for the three pointy endpoints. What simpler model (with only one parameter) is associated with the three edges of the manifold?*

What happens when we take $N$ exponentials? The model manifold will no longer be a surface – it will fill an $N$-dimensional manifold, and its projection into 3D will also fill a volume. We cannot expect to do this with a grid of points: 10 points in each direction for $N = 7$ $\theta$s would be $10^7$ curves. Let us choose random vectors to get an idea of the shape.

(b) *Select a large number of random vectors in the space of parameters $0 < \theta_\alpha < \infty$. Starting with $N = 2$, reproduce the model manifold you found in part (a). You'll want some of the $\theta$s to be small and some large to get to the edges: choose values with probability $\rho(\theta) \propto 1/\theta$ in the same range you used in part (a). Try $N = 7$. Rotate the 3D plot to see how "thick" the model manifold is.*

You should find a thin sheet with what appears to be pointy-tipped scallops along one "edge". For $N = 1$, there were three points on the manifold, including the two at the ends of the edge.

(c) *How many points do you observe for $N = 7$? From part (a), argue that the three cuspy points for $N = 2$ correspond to values where $y(t)$ is a constant except perhaps at $t = 0$. Is the same (likely) true for $N = 7$?*

The cusps in the model manifold are simpler models with no adjustable parameters! We shall find in general that the edges, corners, and hyper-edges of the model manifold form emergent, simpler models. In Exercise **??**, we shall use noise to sample the edges of the model manifold.

Using three predictions is not an exhaustive study for a complex model. Can we create a 3D view of the entire behavior? In Fig. 6.2, we used principal component analysis to rotate our 5000 dimensional stock price information so that the most important few directions could be separated out and viewed. Let us apply principal component analysis to our data. There are packaged routines you can use for this.

(d) *Test your implementation of PCA. Generate random trajectories $y(t)$ for pairs of $\theta$s as in part (a), but now for 20 timepoints $y(t)$ evenly spaced with $t$ between zero and ten. Find the first three principal components from these trajectories. You should get a similar manifold (perhaps flipped), except rotated so that the longest axis is along the first component and the narrowest axis is along the third component.*

(e) *Now generate a random set of trajectories with $N = 7$ for $t \in (0, 10)$, and plot the first three principal components. Do they appear to be thinning by roughly a constant factor for each new component? Plot the next three components. Does the manifold continue to get thinner?*

(e) *Now generate a random set of trajectories with $N = 7$ for $t \in (0, 10)$, and plot the first three principal components. Do they appear to be thinning by roughly a constant factor for each new component? Plot the next three components. Does the manifold continue to get thinner?*

The surface swept out by $y_{\boldsymbol{\theta}}(t)$ in the space of trajectories is the *model manifold*. You have found that it forms a *hyperribbon* – a geometrical object that is longer than it is wide, wider than it is thick, and so on for as many perpendicular directions as there are parameters. In practice, most multiparameter models share this behavior [6, 20, 11], and for NLLS models with certain smoothness conditions this hyperribbon behavior can be rigorously proven [14, 21]. And, just as the edges and corners of our exponential decay model correspond to models with fewer exponentials, so too the hyper-edges of the hyperribbons for models in these other fields give rapidly converging approximate models for systems with complex microscopic laws.

S3.2 **Monomial hyperribbons.**[2] (Statistics) ③

We saw in Exercise S2.2 that the monomial coefficients for polynomial fits to data are ill-determined, and have sloppy eigenvalues for their Hessian. While linear fits with unconstrained coefficients have an unbounded model manifold (an infinite hyperplane, so not a hyperribbon), they allow arbitrarily large gradients in the resulting fit, which are not usually expected in practice and often suppressed by nonlinearities in realistic models (where parameters can often go to infinity in ways that keep the predictions bounded).

Here we consider the model manifold for polynomials $y_{\boldsymbol{\theta}}(x) = \sum_{\alpha=0}^{N-1} \theta_\alpha x^\alpha$ with bounds on the parameters $\theta_\alpha$. The Jacobian

$$J_{m\alpha} = \left.\frac{\partial y_{\boldsymbol{\theta}}}{\partial \theta_\alpha}\right|_{x_m} \tag{S3.3}$$

can be viewed as mapping small vectors $\boldsymbol{\delta}$ in parameter space onto vectors $\Delta y(\mathbf{x})$ in prediction space, where $\mathbf{x} = \{x_1, \ldots, x_M\}$ are the locations of the data points being fit. That is, $y_{\boldsymbol{\theta}+\boldsymbol{\delta}}(x_m) = y_{\boldsymbol{\theta}}(x_m) + \sum_\alpha J_{m\alpha}\delta_\alpha$ (see Fig. 3.1).

(a) *Show (or note) that $J_{m\alpha} = x_m^\alpha$.*

Thus

$$J = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{N-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{N-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_M & x_M^2 & \cdots & x_M^{N-1} \end{pmatrix} \tag{S3.4}$$

is the famous Vandermonde matrix. If $M = N$, this matrix is square, and its determinant is the ratio of the volume of the allowed parameters $\boldsymbol{\theta}$ and the volume of the resulting model manifold. The famous result is that this determinant is given by $\det(J) = \prod_{1\le i<j\le N}(x_j - x_i)$. This can be seen by observing that $\det J$ obeys the rule

[2] This exercise embodies the results of Quinn, Wilber, et al. [14].

that swapping rows of $J$ swaps the sign of the determinant, that the determinant has the correct net degree in the $x_i$ (the same degree $N(N-1)/2$ as the product of the diagonal entries), and then checking for the overall multiplicative constant.

Let us consider the constrained system where the monomial coefficients lie in a sphere

$$\sum_{\alpha} \theta_{\alpha}^2 < N, \tag{S3.5}$$

and consider fits over the unit interval $x \in (0,1)$. Each $\theta$ on average then has a variance of order one, and we are considering the behavior over a distance of order one. This bound corresponds to controlling the derivatives of the function at zero, since

$$\theta_\alpha = y_{\boldsymbol{\theta}}^{(\alpha)}(0)/\alpha! = (1/\alpha!)\, \partial^\alpha y_{\boldsymbol{\theta}}/\partial x^\alpha|_{x=0} = R^\alpha \tag{S3.6}$$

where $R$ would correspond to a radius of convergence of the function according to a ratio test. (Our assumption in eqn S3.5 that the monomial coefficients lie in a sphere can be made to work by rescaling the length until $R = 1$.) Our general theorem for nonlinear least-squares models [14] demands a stronger constraint on the functions $f(x)$ – that the sum of the squares of the $m$th derivatives be less than $N$ *for every* $x$ in the interval. They also use the fact that polynomials have the biggest range of predictions given the Taylor series bounds, so your calculation is reproducing much of the qualitative physics of the rigorous proof.

If the typical distance between the points $x_i$ is $\Delta x$, then the determinant $\det J$ is $\sim (\Delta x)^{N(N-1)/2}$, which becomes really, really tiny as $N$ gets large and the minimal spacing $L/M$ gets small. As soon as the number of data points per radius of convergence becomes larger than two, the volume of the model manifold gets progressively smaller.

(b) *Taking $M = 6$ equally spaced points on $[0,1]$ and $N = 6$ parameters, numerically check that the determinant of our Vandermonde matrix $J$ is tiny.*

Is this volume small because the the predictions are squeezed into a hyperribbon? If the widths along the $n$th direction scale as $w_n = (\Delta x)^n$, then this would work, since $\prod_{n=1}^{N} w_n = (\Delta x)^{\sum_{n=1}^{N} n} = (\Delta x)^{N(N-1)/2}$. But usually the number of predictions $M$ is larger than the number of parameters $N$, so $J$ isn't a square matrix. What mathematical operation gives us the shape of the image of the unit sphere? Singular value decompositions is a powerful generalization of eigenvector decomposition, and precisely serves this purpose.

Singular value decomposition is not well studied in physics, where we usually care about square matrices that are symmetric or Hermitian. See the excellent Wikipedia article [23] on SVD. The theory says that any matrix of real numbers can be decomposed into a product of three matrices:

$$\begin{aligned} J &= U\Sigma V^T \\ J_{i\alpha} &= U_{ij}\Sigma_{j\beta}(V^T)_{\beta\alpha}. \end{aligned} \tag{S3.7}$$

Here $U$ is $M \times M$, $V$ is $N \times N$, and $\Sigma$ is $M \times N$ and diagonal (until the diagonal hits one of the far sides of the rectangle). Here the columns of $U$ and $V$ (and hence the rows of $V^T$) are an orthonomal basis for the behavior space and the parameter space, and are called the left-singular vectors and the right-singular vectors of J, respectively. (This makes $U^T U$ and $V^T V$ the $M \times M$ and $N \times N$ identity matrices: they are unitary.) Assuming $M > N$, the first $N$ basis vectors of $U$ span the tangent to the model manifold. The $\alpha$th right singular vectors of $V$ maps onto the $\alpha$th left singular vector of $U$ after being stretched an amount given by $\sigma_{\alpha\alpha}$.

(c) *Taking $M = 11$ equally spaced points and $N = 6$ parameters, dig up the appropriate SVD routine and find $U$, $\Sigma$, and $V$ for our our Vandermonde Jacobian $J$.* The left singular vectors of $U$ are unit vectors on our sphere in parameter space. In behavior space, our sphere becomes an ellipsoid, with axes along the right singular vectors. *By how much are the unit axes of our sphere of parameters being squashed in behavior space? Is our hyperellipsoid model manifold roughly thinner by a constant factor for each new parameter?*

Finally, let us connect the skewness of $J$ and its singular values $\Sigma_{\alpha\alpha}$ with the Hessian $\mathcal{H}_{\alpha\beta} = g_{\alpha\beta} = (J^T J)_{\alpha\beta}$.

(d) *Show analytically using the singular value decomposition that the eigenvalues of $\mathcal{H}$ are the squares of the singular values of $J$. What are the eigenvectors, in terms of the left and right singular vectors?*

# Chapter 4

# Nonlinear fits: Challenges and algorithms

## Exercises

S4.1 No exercises yet

# Chapter 5

# Model boundaries and simpler emergent models

## Exercises

S5.1 **First-digit law and priors.** (Statistics) ⓟ

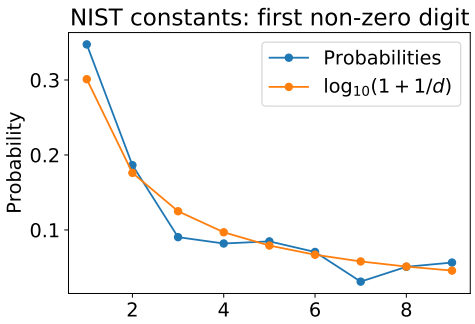



**Fig. S5.1 Fraction of first digits** for 354 fundamental physical constants. (2019 CODATA internationally recommended values [1]).

Bayesian statistics, like statistical mechanics, incorporates known experimental results into a probabilistic prediction for the behavior of the system in the future (see Exercise S1.3). In statistical mechanics, if we only know the energy of a system then Liouville's theorem tells us that all points in the energy shell are equally likely *a priori*. In Bayesian statistics, they have no theorem like Liouville's, so they need to assume a *prior*. For example, if you want to estimate a time constant $\tau$ for a chemical reaction (which can range from nanoseconds to years), you might want a prior $P_\tau(\tau)$ that gives equal weight to each decade: finding $\tau$ in the range $(10^{-9}, 10^{-8})$ seconds is equally plausible as finding $\tau$ in the range $(10^5, 10^6)$ seconds.

*Show that* $P_\tau(\tau) \propto 1/\tau$ *has this reasonable property. Show that this choice also makes the decay rates* $\Gamma = 1/\tau$ *have this same nice property:* $P_\Gamma(\Gamma) \propto 1/\Gamma$. (Hint: If $\tau$ lies in a small range $\Delta\tau$, then $\Gamma$ will lie in a corresponding small range $\Delta\Gamma$, so $P_\Gamma(\Gamma)|\Delta\Gamma| = P_\tau(\tau)|\Delta\tau|$.) *Show that this distribution predicts that the first non-zero decimal digit* $d$ *of* $\tau$ *will have probability* $\log_{10}(1+1/d)$ *(Fig. S5.1).* (Hint: Do it assuming $\tau$ lies in one decade first.) *Show your steps.* (Note: Feel free to consult the extensive discussions on the Web.)

Simon Newcomb, using a book of logarithms in 1881[1] discovered this by noticing that the pages in the beginning (1.000001, 1.000002, ...) were dirtier than the ones at the end (9.000001, 9.000002, ...). Frank Benford fleshed this out in 1938, showing that areas of rivers, molecular weights of compounds, and physical constants like the proton mass, Planck length, and Avogadro's constant (Fig. S5.1) also obey this law.

## S5.2 Bayesian priors.[2] (Statistics) ③

In this exercise, we shall explore an analogy between statistical mechanics and Bayesian statistics. As in Exercise S1.2, we consider the problem of fitting a Gaussian probability distribution to a collection of measurements. (See also Exercise S2.4 for an information-geometry analysis of this same problem.)

Consider the population the heights of women in the United States. Several websites quote a mean height of $\mu_0 = 162$ cm for US women, but neglect to mention the variance. We will assume $\mu = \mu_0$ is known, and we would like to estimate the probability distribution of the unknown standard deviation $\sigma$, given a single uncorrelated sample of $N$ women. We know

$$\begin{aligned} P(\{x_n\}|\sigma) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x_n-\mu_0)^2/(2\sigma^2)} \\ &= \frac{\mathrm{e}^{-\sum_{n=1}^{N}(x_n-\mu_0)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}^N} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(-S_N/2\sigma^2\right), \end{aligned} \tag{S5.1}$$

where the value $S_N = \sum_{n=1}^{N}(x_n - \mu_0)^2$ provides sufficient statistics for estimating $\sigma$.

In statistical mechanics, we care not only about the average behavior, but the *distribution* of behaviors. If our sample size is small, we should care not only about being correct on average, but also what the distribution will be of the true answers given the data we have. In our case, given the model $P(\{x_n\}|\sigma)$ with unknown parameter $(\boldsymbol{\theta} = (\sigma))$, and knowing only one sample of $N$ data points $\{x_n\}$, what is the probability that the standard deviation of the unknown distribution is in the range $(\sigma, \sigma + \Delta)$?

[1] Before calculators, people used printed books of logarithms, which allow one to multiply and divide quickly.

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

In Bayesian statistics, we estimate the probability of a given set of parameters $\boldsymbol{\theta}$ given data $\mathbf{d}$ by using Bayes' theorem (see Exercise S1.3):

$$P(\boldsymbol{\theta}|\mathbf{d}) = P(\mathbf{d}|\boldsymbol{\theta})P(\boldsymbol{\theta})/P(\mathbf{d}). \tag{S5.2}$$

with $P(\{x_n\}|\sigma)$ from eqn S5.1. Here the probability of the data, $P(\mathbf{d})$, is independent of the parameters and basically acts to normalize $P(\boldsymbol{\theta}|\mathbf{d})$ to one. The probability density $P(\boldsymbol{\theta})$ is called the *prior*.

There is a close relationship between Bayesian statistics and statistical mechanics. The unknown parameters are analogous to the degrees of freedom in a physical system (say, momenta and positions of the particles). The probability density $P(\boldsymbol{\theta}|\mathbf{d})$ is analogous to the Boltzmann factor $\exp(-\mathcal{H}/k_BT)/Z$ of Chapter 6. The data $\mathbf{d}$ is analogous to the known external conditions (energy, volume, pressure, ...). Statisticians do Monte Carlo in parameter space (stochastic Bayesian analysis [15]) using the same techniques we discuss in Chapter 8.

But what is the prior $P(\boldsymbol{\theta})$? It represents knowledge you had about the parameters before the data is taken, or perhaps about how parameters *should* be distributed, if no measurements have yet been taken. In the statistical mechanics of classical particles (Chapter 3), our presumption about the relative probability of different positions and momenta is given by Liouville's theorem—a uniform prior, weighting all regions of phase space equally. (It is only after we know the temperature or the energy that high momenta become less probable than low momenta.)

Uniform priors in Bayesian statistics *seem* unbiased. We shall compare several priors of the form $P_\alpha(\sigma) \sim \sigma^\alpha$.

There are three values for $\alpha$ of particular interest.

- $\alpha = 0$, the *uniform prior* for $\sigma$ where every interval $(\sigma, \sigma + \Delta\sigma)$ is equally likely.
- A value for $\alpha$, where every interval $(v, v + \Delta v)$ in the variance $\sigma^2$ is equally likely (uniform prior for $\sigma^2$).
- Jeffrey's prior $P(\sigma) = 1/\sigma$, where every *fractional* change $(\sigma, (1 + \Delta)\sigma)$ is equally likely.

Suppose three competing investigators took each took a single sample of women, with $N = 4$, $N = 40$, and $N = 400$, from a population with known mean $\mu_0$. Suppose for simplicity that in each case their sample gave the population average[3] $S_N = N\sigma^2_{\text{pop}}$.

(a) *Plot* $P_0(\sigma|S_N)$ *versus* $\sigma/\sigma_{\text{pop}}$ *for these three samples* $N = 4$, 40, *and* 400, *assuming uniform prior* $\alpha = 0$ *for* $\sigma$ *and using* $S_N = N\sigma^2_{\text{pop}}$ The normalization ($P(\mathbf{d})$ in eqn S5.2) can be computed either numerically, or analytically in terms of $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)\mathrm{d}x$. *How does the maximum likelihood* $\sigma_{\text{ML}}$, *where* $P_0(\sigma_{\text{ML}})$ *is maximum, vary with* $N$*? Is it biased, compared to the naive estimate* $\sigma_{\text{pop}}$*? Finally, explain why the curve appears so asymmetric for small* $N$*. Is the average* $\sigma$ *for this probability*

[3]The standard deviation of women's heights in the US turns out to be about $\sigma_{\text{pop}} = 6.9$ cm.

*distribution biased? In what direction?* (Hint: Is it more likely for a narrow Gaussian to give a widely distributed sample of four points, or for a wide Gaussian to happen to give a tight cluster of four points?)

So the bias in statistical estimates depends on whether you are interested in the mean (average) or the mode (maximum likelihood). From a Bayesian perspective, choosing any single number to represent the probability distribution of the quantity of interest is perhaps misguided.

Note that the bias $\langle X \rangle_{\text{samp}}$ we found in Exercise S1.2, eqn S1.4 is quite different than the bias $\langle X \rangle_{\text{BayesAv}}$ we discuss here. There we compared height variations over repeated samples of $N$ women; here we use a single sample of $N$ heights and average over the underlying true distributions that could have produced the data.

As we mentioned earlier, uniform priors *seem* unbiased. But a prior uniform in the standard deviation $\sigma$ is not uniform in the variance $\sigma^2$!

(b) *Consider a uniform prior in the variance $v = \sigma^2$, so $P(v)$ is constant. What is $P(\sigma)$?* (Hint: The probability of being in corresponding intervals must agree, so $P(\sigma)\mathrm{d}\sigma = P(v)\mathrm{d}v$.) *What is $\alpha$ for a uniform prior on the variance? Calculate the mode and the average for our three samples $S_N = N\sigma^2_{\text{pop}}$, ($N = 4$, 40, and 400, $\mu = \mu_0$) for a uniform prior on the variance. Compare with the naive estimate.*

Unmeasured rates of biochemical reactions are examples of parameters that are often uncertain over many orders of magnitude. Surely our prior expectation that the rate is in the range $\Gamma \in (10^{-3}, 10^{-4})$ should not be a million times smaller than the rate is in the range $\Gamma \in (10^3, 10^4)$ (as a uniform prior would suggest). Now consider using instead the time-scale $\tau = 1/\Gamma$ as the parameter—a uniform prior in $\tau$ would weight the two intervals differently by a factor of a million *in the opposite direction*. Jeffrey's prior, uniform in the logarithm, fixes this problem.

(c) *Consider a uniform prior in the log of the width* $\log(\sigma)$. *Show that $P(\sigma) \propto 1/\sigma$, so $\alpha = -1$. Check that the prior $P_\alpha(\sigma)$ integrated over a range $(\sigma_0, 10\sigma_0)$ is indeed constant, independent of $\sigma_0$. Calculate the mode and the average for our three samples $S_N = N\sigma^2_{\text{pop}}$ ($N = 4$, 40, and 400, $\mu = \mu_0$) for Jeffrey's prior.*

# Chapter 6

# Visualizing model behavior

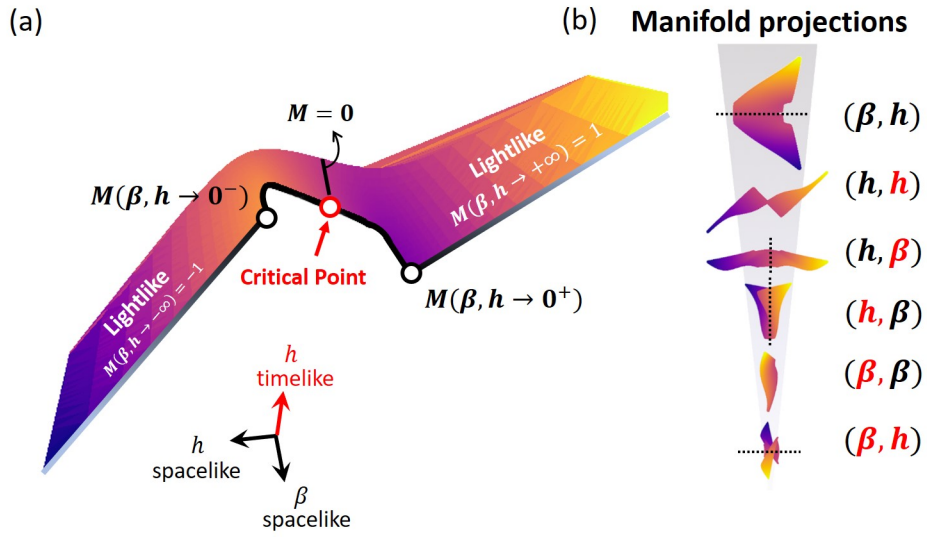



**Fig. 6.1 The model manifold for the 2D Ising model** isometrically embedded in a 4D Minkowski-like space using isKLe (intensive symmetrized Kullback–Leibler embedding), from Teoh et al. [19] (see Exercise S6.3). (a) Three-dimensional embedding, showing the critical point at $M = 0$, $T = T_c$, the jump in magnetization opening below $T_c$, and the way the magnetization grows at fixed temperature as the external field is raised. Note that the isKL embedding is a 45 degree rotation of the 4D graph of energy and magnetization versus temperature and field. (b) Six projections into the various pairs of coordinates. The red coordinates are *time-like* – they contribute negative distance to the differences between points.

The field of *information geometry* generally studies the space of behaviors for probabilistic models. For example, the Ising model makes predictions for the *probability distribution* of

spin configurations, and thus for all equilibrium behavior of a system. In this chapter, we shall discuss methods for properly visualizing model manifold for probabilistic model (e.g., the 2D surface of predictions in Fig. 6.1, swept out by varying the two parameters $T$ and $H$ in the Ising model). This is in general a surface in the infinite-dimensional space of probability distributions. (See Fig. 6.1.)

## 6.1 Visualizing least-squares models

Let us first summarize the methods we have been using to visualize the model manifolds for least-squares models. Recall from Chapter 2 that that least-squares model manifolds are the surface $\mathbf{y}(\boldsymbol{\theta})$ in prediction space swept out as the parameters $\boldsymbol{\theta}$ are varied through all allowed values. The parameters can be viewed as coordinates on the model manifold, and the metric tensor is $g_{\alpha\beta} = J^T_{\alpha i} J_{i\beta}$, where $J_{i\alpha} = (\partial_\alpha y_i)$ is the Jacobian of the map from parameters into data space.

How have we been visualizing the behavior of these least-squares models? One might think that we can visualize only three predictions at a time. Our discovery that many systems are sloppy, with predictions lying on hyperribbons, suggests that we could plot a few of the longest axes of the model manifold.

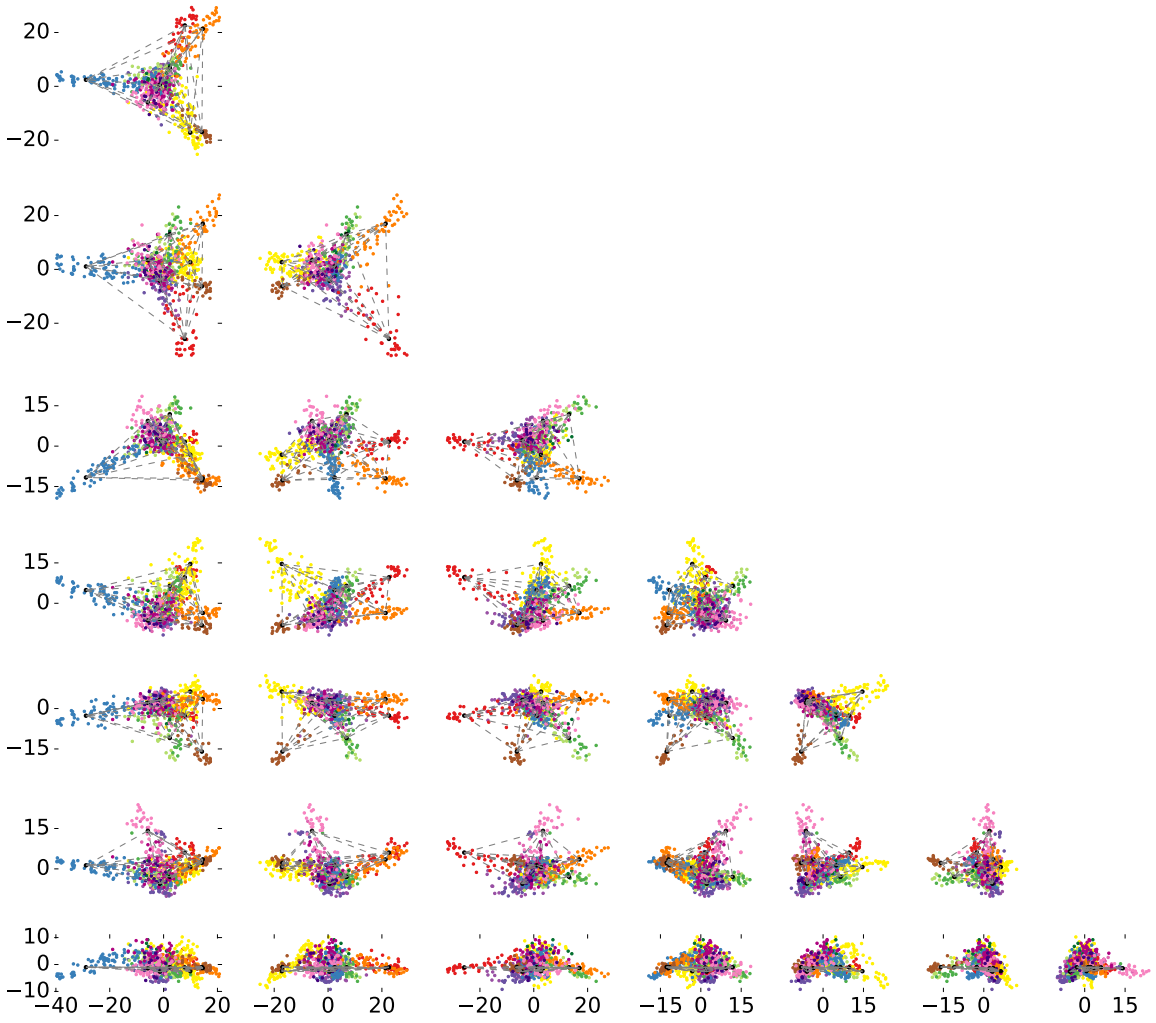



**Fig. 6.2 PCA projections of daily stock prices** for 705 large companies over ten years, from [7]. The upper left figure shows the two largest principal components; the figure in column M row N shows the N+1st principal component plotted against the Mth. Note that the ranges get smaller for the higher principal components fairly rapidly, and quite rapidly become dominated by noise. The dashed lines form a hypertetrahedron whose vertices are well correlated with sectors of the economy (depicted in different colors). The best fit hypertetrahedron is notably more hyperribbon-like. One might argue that our ability to categorize companies by whether they are in the energy sector or the tech sector is only possible because of this hyperribbon structure. Of course, one could also imagine that the traders who buy and sell stocks have agreed upon which companies are utilities, and vary the utility prices in synchrony.

Unlike other ways of visualizing high-dimensional data with low-dimensional representations (t-SNE, UMAP, manifold learning, ...) PCA preserves the pairwise distances between the

prediction vectors $\mathbf{y}$. That is, it rotates a geometrical structure that faithfully represents the relation between the points, both at short distances and at long distances. We shall emulate this feature in our visualizations of probabilistic model manifolds in later sections of this chapter.

The fact that high-dimensional data from a variety of fields can so often be effectively visualized using PCA naively would seem incomprehensible. Why would the first few dimensions capture the important behavior? Our information geometry work suggests that this hyperribbon structure is natural, at least for multiparameter models fit to data. Perhaps the world is sloppy, generically providing low-dimensional behaviors. Or, perhaps, scientists specialize in studying comprehensible systems, amenable to low-dimensional projections.

## 6.2 Visualizing probabilistic model manifolds

In recent times we have been exploring the model manifold for cases like the Ising model where the results of a model are best viewed not as a vector of predictions $\mathbf{y}(\boldsymbol{\theta})$, but as a probability distribution $\rho_{\boldsymbol{\theta}}(\mathbf{x})$ for the different states $\mathbf{x}$ of the system. Here $\mathbf{x}$ could be the possible spin configurations $\{s_{ij}\}$ of the 2D Ising model (Fig. 6.1), or the temperatures $T(\Theta_i, \Phi_j)$ of the cosmic microwave background radiation at the measured directions $\Theta, \Phi$ in the sky [13], or (more simply) the mean and standard deviation $\overline{h}, \sigma_h$ of the heights of people in Finland (Fig. 6.7, Exercise S1.2).

We confine ourselves to visualization methods which faithfully reproduce the natural geometry. As we discussed in Section 6.1, PCA faithfully preserves both the short-distance and long-distance geometry of the vectors it visualizes (although the projection onto the main principle components loses some of this information). The short-distance geometry for probabilistic models is governed by the Fisher Information Metric (FIM) metric in probability space—the Fisher Information Metric (FIM) $g_{\alpha\beta} = -\left\langle \frac{\partial^2 \log \rho(\mathbf{x})}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle$ (eqn 3.2). Exercise S2.4 discusses how the FIM distance is related to the differences in model behavior. In particular, it measures how hard it is to work backward from the behavior to distinguish one model from the other. It gives a bound on how well one can measure the parameters in a model using experimental tests.

Preserving the long-distance geometry is more of a challenge. The model manifold for nonlinear least squares inherits the $\mathbb{R}^M$ metric of the space of experimental data vectors scaled with their error bars (Section 3.2). But what is the shape of the space of probability distributions?

Exercise S6.1 explains that probability space for a distribution $\rho_{\boldsymbol{\theta}}(\mathbf{x})$ where $\mathbf{x}$ has $N$ values is geometrically an "octant" of a hypersphere of radius 2 in dimension $N-1$. The Hellinger embedding maps $\rho_{\boldsymbol{\theta}}(\mathbf{x})$ to a point $2(\sqrt{\rho_{\boldsymbol{\theta}}(x_1)}, \cdots, \sqrt{\rho_{\boldsymbol{\theta}}(x_i)}, \sqrt{\rho_{\boldsymbol{\theta}}(x_N)})$, which lies on the sphere since $\rho(\mathbf{x})$ sums to one. You will argue that this is the natural embedding, and you will use it to derive the formula for the FIM.

The Hellinger embedding is not useful for visualizing large, complicated probabilistic models: it is doomed by the curse of dimensionality, as illustrated by Figure 6.3. As soon there is enough data to make several distinguishable probability distributions, the model manifold must crumple in order to fit into flat Euclidean space (Fig. 6.4).

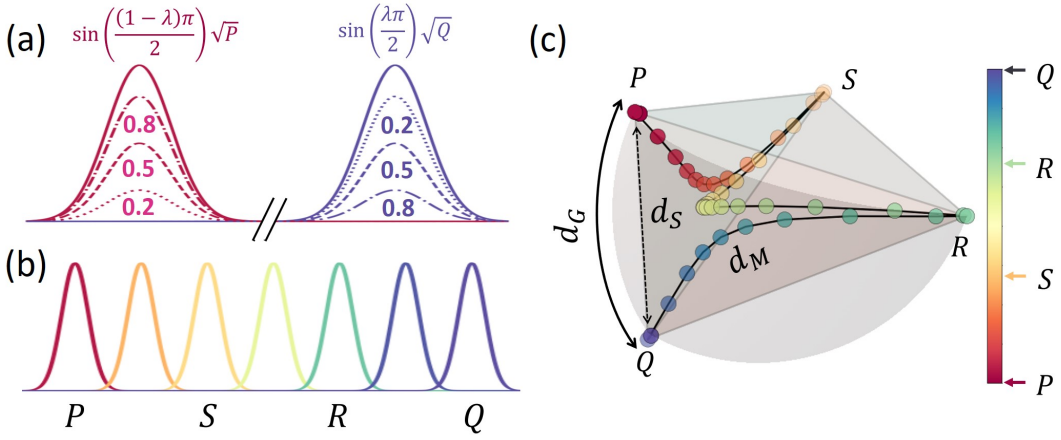



**Fig. 6.3 Euclidean embeddings are doomed**. In (b) and (a) we see the shortest path between two distant Gaussians $P(x)$ and $Q(x)$ in $x$ space and in probability space. No matter how distant two distributions are in real space, their geodesic distance (using the FIM metric) cannot be larger than $\pi$. (The longest arc in the positive octant has angle $\pi/2$, and the FIM sphere is of radius two.) In (c), we see a 3D projection of the path in probability space given by sliding the Gaussians from $P$ to $Q$. Every time a new distribution becomes nearly orthogonal to the others, it adds a dimension to the projection. For a large Ising model (or the cosmic microwave background radiation), the snapshots at even slightly different parameters quickly become easily distinguished. Any effective low-dimensional visualization will need to crumple the model manifold to fit it into Euclidean space. (see Fig. 6.4). From [19].

**Fig. 6.4 A crocheted representation of the Gaussian fit model manifold**. A Gaussian whose mean $\mu$ and standard deviation $\sigma$ are fit to data is a probabilistic model whose FIM yields a constant negative curvature: it is a hyperbolic space, $\mathbb{H}^2$. A portion of $\mathbb{H}^2$ can be isometrically embedded in 3D, but since the circumference grows faster than the radius, it must necessarily get more wiggly as the patch grows. Leafy vegetables like kale do this, and it is also possible to reproduce it more quantitatively using knitting (shown here). As described in Fig. 6.3, probabilistic models like Gaussian fits need to crumple their embeddings to project faithfully into Euclidean spaces. Fig. 6.7 shows how this may be bypassed with our intensive embeddings. (With permission from Daina Taimina [18].)

Quinn et al. [13] confronted this challenge in her quest to analyze the ΛCDM (cold-dark matter) model of the early Universe, as it applies to the cosmic microwave background radiation experiments of Niemack and colleagues. She could compute the FIM (eqn 3.2) at the best fit of the model to our Universe, whose eigenvalues are sloppy (as shown on the left of Fig. 6.5). But the Hellinger embedding was intrinsically high dimensional: PCA showed just a bit of dust with a blob in the center, where the variance was in higher dimensions. As the ΛCDM model predicts the probability of the pattern found in our Universe, it suffers from the curse of dimensionality. Quinn used the *replica trick* to take the limit of the Hellinger metric in the limit of zero data (see Exercise S6.2 and Figs 6.5 and 6.6), giving our first *intensive* embedding method, which we called inPCA (Fig. 6.5).

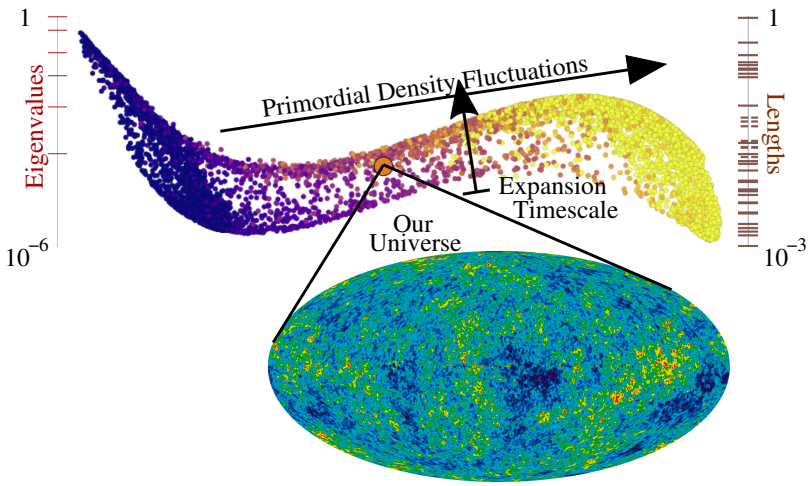



**Fig. 6.5 A portion of the model manifold for the Universe**. Taking the limit of zero data, we generalized PCA to use the Bhattacharyya divergence instead of the Hellinger distance, yielding the model manifold here shown projected onto the largest two inPCA components. On the left find the sloppy eigenvalues evaluated at the best fit to our Universe (dot near center). On the right, find the lengths of the model manifold as measured by the variance of the inPCA components (not as sloppy). (From Quinn et al. [13].)

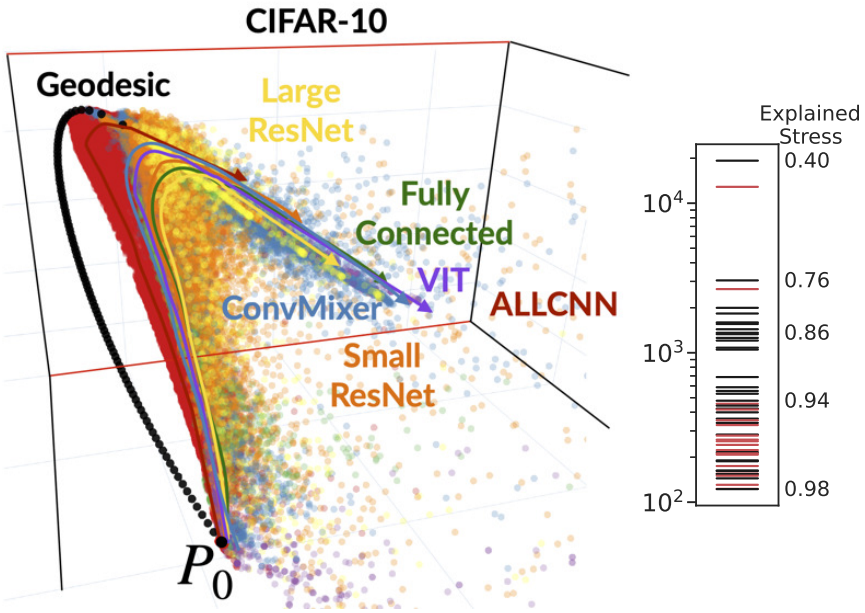



**Fig. 6.6 Deep neural network training trajectories with inPCA**. We trained many types of deep neural networks to classify images. Shown are the probability-space trajectories they took, starting from ignorance $P_0$ (all categories equally likely) towards truth (certainty for correct categories). The probability space has around half a million dimensions. Three inPCA components capture 76% of the information about their learning paths. What is amazing is not that we can tell the differences between different network architectures. It is that we can view them at all for such a complex, nonlinear, high-dimensional learning process. (From Mao et al. [10].)

InPCA rederived the classic Bhattacharyya divergence between probability distributions in the limit of zero data. (It is not a distance because it does not satisfy the triangle inequality). The Bhattacharyya divergence is one of a large family of 'intensive' divergences that we find lift the curse of dimensionality, allowing a few dimensions to capture most of the behavior of complex models while preserving the local FIM metric. Other intensive divergences (Rényi, Kullback–Leibler), violating the triangle inequality and by using a logarithm, bypass the difficulties of the Hellinger distance. They agree with least-squares fits by giving a distance between Gaussians equal to the number of standard deviations separating them. Using them to search for low-dimensional visualizations, however, demand a Minkowski-like embedding space (Fig. 6.7). See Exercise S6.3, and references [19, 11].

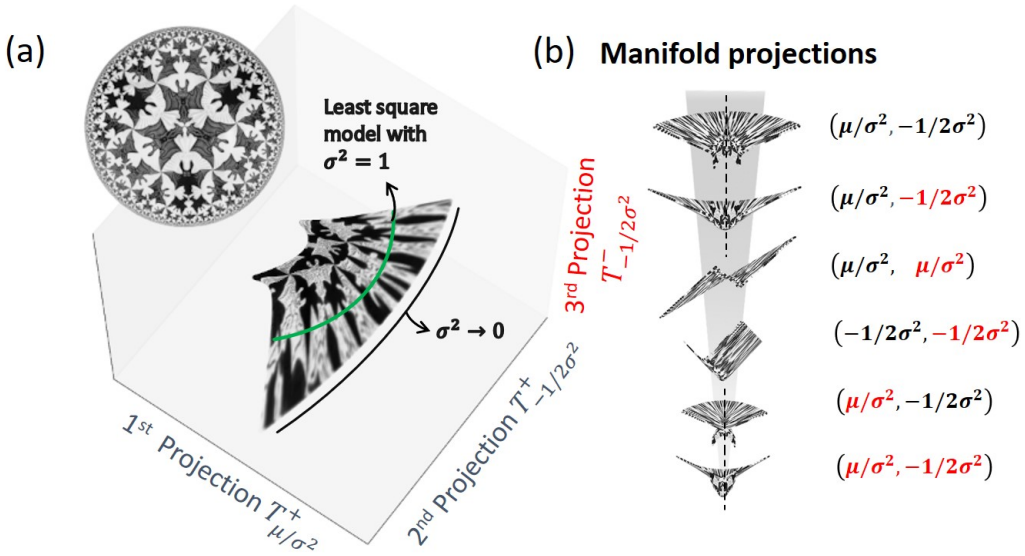



**Fig. 6.7 The model manifold for a Gaussian fit: the hyperbolic space $\mathbb{H}^2$.** Here we use the Gaussian fit model of Fig. 6.4 with mean $\mu$ and standard deviation $\sigma$ to illustrate how using Minkowski space can bypass the curse of dimensionality (Fig. 6.3) and its need to crumple the model manifold. The model manifold with a constant negative curvature (the Poincaré half-plane, $\mathbb{RP}2$) is entertainingly (but not isometrically) illustrated in the upper left of panel (a), Escher's "Heaven and Hell" (Circle Limit IV). It has a smooth, intuitive isKL embedding (Exercise S6.3) with one (vertical) time-like axis. Our isKLe embedding using a symmetrized Kullback–Leibler divergence is a four-dimensional hyperboloid in our Minkowski space. From [19]. .

Calculating the divergence between to parameter values of a large system can be daunting: the probability vector for a 10×10 Ising model with 100 spins has $N = 2^{100}$ components. Exercise S6.3 describes one particular divergence which we discover has a deep connection to statistical mechanics. You will show that the symmetrized Kullback–Leibler divergence (which we introduce briefly) in many cases yields explicit formulas for the coordinates of an isometric embedding in terms of familiar quantities easily estimated. In particular for the Ising model in Fig. 6.1, the four coordinates for the model at a temperature $T$ and external field $H$ are $H/k_BT \pm M$ and $-J/k_Bt \pm (E + MH)/J$ (see Exercise S6.3).

Finally, Fig. 6.8 shows the resulting model manifold for the ideal gas (which you can show in Exercise S2.6 has zero curvature and can also be drawn on the plane).

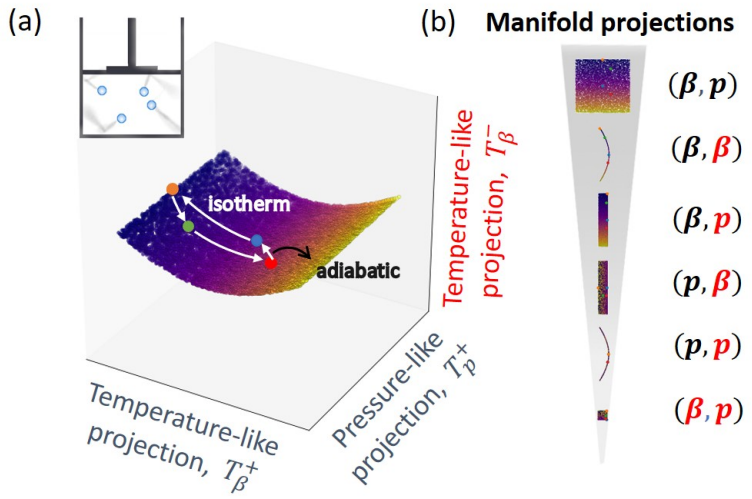



**Fig. 6.8 The space of predictions for the ideal gas**. In Exercise S2.6, we found that the model manifold could be embedded in the plane, $\mathbb{R}^2$. The ideal gas Boltzmann distribution, as a function of $P$ and $T$, is an exponential family. The isKLe embedding is a bent sheet in Minkowski space. As it must, it has no extrinsic curvature; isometric embeddings for the same model can differ in this way. The path taken in the Carnot cycle is illustrated. From [19].

# Exercises

S6.1 **Hellinger and the FIM.** (Information geometry, Statistics) @

What is the shape of the space of probability distributions?

First, the space of probability distributions is often very high dimensional. The prediction is sometimes discrete: if there are 100 spins in an $10\times10$ Ising model, there are $2^100$ probabilities $\rho(\mathbf{s})$ which sum to one. It is often continuous: if one fits a Gaussian $\rho_{\overline{x},\sigma}(x)$ to data, the model manifold is a two-dimensional surface in an infinite-dimensional space of possible probability functions.

Second, even if the prediction is discrete, it is not natural to treat the prediction as a vector, because the natural distance between two predictions is not the sum of squares of the differences between the individual probabilities. We can see this by considering how difficult it is to measure a small change in probability of one of the predictions.

(a) *Consider flipping a coin to measure the small probability $\rho$ that it lands on its edge. After $F \gg 1/\rho$ flips, what is the error in your estimate of $\rho$, to lowest order in $\rho$? How*

*many flips do you need to estimate $\rho$ to an accuracy $\epsilon$? Is it equally easy to measure rare events and common events to an absolute error of $\epsilon$? Which is harder?*

The probability $\rho(\mathbf{x})$ is normalized, $\sum_{\mathbf{x}} \rho(\mathbf{x}) = 1$. Taking the square root of each entry gives a point on the unit hypersphere (in the positive "octant", with all components greater than zero). Let us check that it is uniformly challenging to estimate the square roots of the components with differing probabilities.

(b) *In the above experiment, what is the error in your estimate of $\sqrt{\rho}$, to lowest order in $\rho$? Is it equally easy to measure the square root of the probability of rare events and common events?*

So, we conjecture that a natural measure of distance on the sphere of $\sqrt{\rho}$ is the Euclidean distance in the embedding space

$$|\rho_1 - \rho_2|_{\text{sphere}} = \sqrt{(\sqrt{\rho_1} - \sqrt{\rho_2})^2} = \sqrt{\sum_{\mathbf{x}} (\sqrt{\rho_1(\mathbf{x})} - \sqrt{\rho_2(\mathbf{x})})^2}. \tag{S6.1}$$

This is, up to a constant, the *Hellinger distance*, sometimes called the Hellinger divergence (because all the other measures for separations between probabilities are not proper distances.)

Suppose we now have a model $\rho_{\boldsymbol{\theta}}(\mathbf{x})$ depending on parameters $\boldsymbol{\theta}$. The metric tensor $g_{\alpha\beta}$ is defined to be the dependence of the distance on small changes in parameters, $\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon\boldsymbol{\delta}$. The squared distance to quadratic order should be

$$|\rho_{\boldsymbol{\theta}+\epsilon\boldsymbol{\delta}} - \rho_{\boldsymbol{\theta}}|^2 = \epsilon^2 g_{\alpha\beta}\delta_\alpha\delta_\beta. \tag{S6.2}$$

(c) *Show that the distance on the sphere implies that*

$$g^{\text{sphere}}_{\alpha\beta} = \int \mathrm{d}\mathbf{x} \tfrac{1}{4}\rho(\partial_\alpha \log\rho)(\partial_\beta \log\rho) = \tfrac{1}{4}\langle(\partial_\alpha \log\rho)(\partial_\beta \log\rho)\rangle_{\mathbf{x}}. \tag{S6.3}$$

But we have been told that the natural distance in probability space is given by the Fisher Information Matrix (FIM). Equation 3.2 tells us

$$\begin{aligned} g^{\text{FIM}}_{\alpha\beta} &= -\left\langle \frac{\partial^2 \log\rho(\mathbf{x})}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle \\ &= -\int \mathrm{d}\mathbf{x}\rho(\mathbf{x}) \frac{\partial^2 \log\rho(\mathbf{x})}{\partial\theta_\alpha \partial\theta_\beta} \end{aligned} \tag{S6.4}$$

Are these different? The version $g^{\text{FIM}}$ in eqn 3.2 is more convenient for us, but it equals $4g^{\text{sphere}}$, which is also commonly used.

(d) *Show this.*

The Hellinger distance perhaps should have been defined to be twice as big (the distance on a sphere of radius two), so the squared distance for nearby points would agree

with the FIM. Or, even better, the FIM should have been defined to be a factor of four smaller. Instead, the Hellinger distance is sometimes the sphere distance, and sometimes the squared distance divided by square-root of two (so with a metric tensor one eighth that of the FIM).

S6.2 **Bhattacharyya and the inPCA embedding.**[1] (Information geometry, Statistics) @

In Exercise S6.1, we saw that the space of probability distributions is naturally viewed as a hypersphere, where $\sqrt{\rho(\mathbf{x})}$ is viewed as a vector with components labeled by $\mathbf{x}$. We also saw that the distances between points on this sphere gave the natural Fisher information metric for local distances between probability distributions. A probabilistic model like the Ising model then has a natural, isometric embedding as a surface on this sphere. Unfortunately, we also saw that this embedding is intrinsically high dimensional for interesting models: there is no way to visualize a large Ising model by projecting it into a small dimensional space.

This is in sharp contrast with the nonlinear least squares models we have studied in the rest of this book! There the predictions lie on hyperribbons in the behavior space, naturally forming an isometric, low-dimensional embedding that we can visualize using, for example, principal component analysis (as in Exercise S3.1). Can we find a way to do this for probabilistic models?

The key problem is that systems like large Ising models, or experiments measuring the cosmic microwave background radiation, have too much precise data. Quinn [13] pointed out that, once parameters have shifted even a tiny amount, the behavior is obviously distinguishable with simple measurements: the probability of a snapshot of one Ising model (or one Universe's cosmic microwave background radiation) being reproduced at the other temperature and pressure (or Hubble constant and baryon density) becomes near zero.

Consider the sphere distance (eqn S6.1), whose square

$$d^2_{\text{sphere}}(\rho_1, \rho_2) = \sum_{\mathbf{x}} \left( \sqrt{\rho_1(\mathbf{x})} - \sqrt{\rho_2(\mathbf{x})} \right)^2 = 2(1 - \sqrt{\rho_1} \cdot \sqrt{\rho_2}), \tag{S6.5}$$

can be written in terms of a dot product between points on the sqrt probability sphere. (The Hellinger distance and the natural FIM local distances agree with the sphere distance up to constants.)

(a) *Note that the two forms of $d^2_{\text{sphere}}$ in eqn S6.5 are equal, because the densities are normalized. Suppose there is no overlap between two Ising model ensembles. That is, for all spin configurations $\mathbf{s}$ of an Ising model with significant probability $\rho_{T,H}(\mathbf{s}) > 0$, the probability $\rho_{T',H'} = 0$. What is the sphere distance between the two?*

[1] This exercise is based on Quinn et al. [12].

(b) *Suppose there are many ensembles at different temperatures and fields, with mutually orthogonal probability distributions. What geometrical figure will they form on the hypersphere? Can this be viewed as a hyperribbon?*

Quinn et al. used the *replica trick* to take the limit of zero data. Suppose we take $n$ snapshots of our Ising model, $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. We can view this as one snapshot of $n$ uncoupled replicas of the Ising model, all with the same parameters, with a probability distribution

$$\rho_{T,H}^{[n]}(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \rho_{T,H}(\mathbf{s}_1)\rho_{T,H}(\mathbf{s}_2)\ldots\rho_{T,H}(\mathbf{s}_n). \tag{S6.6}$$

(c) *Note that the replicated dot product* $\sqrt{\rho_1^{[n]}}\cdot\sqrt{\rho_2^{[n]}} = \sum_{\mathbf{s}_1}\cdots\sum_{\mathbf{s}_n}\prod_{i=1}^{n}\sqrt{\rho_1(\mathbf{s}_i)}\sqrt{\rho_2(\mathbf{s}_i)}$. *Show that it can be written as* $\left(\sum_{\mathbf{s}}\sqrt{\rho_1(\mathbf{s})}\sqrt{\rho_2(\mathbf{s})}\right)^n = (\sqrt{\rho_1}\cdot\sqrt{\rho_2})^n$. *Have we made the orthogonality problem better or worse?*

Thus you have shown that the distance per replica

$$(d_{\text{replicated}}^{[n]})^2(\rho_1,\rho_2) = d_{\text{replicated}}^2(\rho_1^{[n]},\rho_2^{[n]})/n = 2(1-(\sqrt{\rho_1}\cdot\sqrt{\rho_2})^n)/n. \tag{S6.7}$$

The replica trick is to boldly use the formula $\lim_{n\to 0}(x^n-1)/n = \log(x)$. Especially regarding the replicated partition function of disordered systems like spin glasses, using this formula is mathematically dubious but physically extremely useful. Here it yields the Bhattacharyya divergence between the two probability distributions.

(d) *Show that* $\lim_{n\to 0} d^{[n]}(\rho_1,\rho_2)^2 = -2\log(\sqrt{\rho_1}\cdot\sqrt{\rho_2})$. Up to the multiplicative factor separating our sphere distance from that of the FIM, you can check on your favorite Web search or AI that this yields the Bhattacharyya divergence.

Using this metric, along with a version of principal component analysis called MDS[2] we succeed in averting the curse of dimensionality, as shown in Figs 6.5 and 6.6.

S6.3 **Kullback–Leibler and isKLe.**[3] (Information geometry, Statistics) @

To visualize the model manifold in information geometry, we want to preserve some measure of the distance between the predictions. For models whose output is a probability defined over a large number of states, it is usually not possible to directly calculate that distance. Of course, a direct measure of the magnetization of an $N$-spin Ising model would also demand an infeasible $2^N$ computations of the energy: perhaps one could develop a Monte-Carlo method for calculating the Bhattacharyya divergence of inPCA (Exercise S6.2) or the Hellinger distance (Exercise S6.1).

In Exercises S2.6 and S2.7 we found a deep relation between the Fisher Information Metric (measuring local distances between probability distributions) and second derivatives of the free energy like the specific heat and thermal expansion coefficient. Here

[2] We rederived multidimensional scaling by taking the zero-replica limit of principal component analysis. It can be used for any distance measure, not just the Bhattacharyya divergence. So, for example, our isKLe embeddings [19] use MDS with the symmetrized Kullback–Leibler embedding (see Exercise S6.3 and Figs 6.1, 6.7, and 6.8).

[3] This exercise is based on Teoh et al. [19].

we discover a deep relation between a global measure of distance between distributions and first derivatives of the free energy. You shall show that the Kullback–Leibler divergence between two distributions, once symmetrized, can be written in terms of the magnetization and the energy for Ising models (uncovered by Teoh et al. [19], see Fig. 6.1).

The Ising model is a particular example of what probabilists call an *exponential family*. A probability distribution $\rho_{\boldsymbol{\theta}}(\mathbf{x})$ is an exponential family if it can be written in the form

$$\rho_{\boldsymbol{\theta}}(\mathbf{x}) = f(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{\gamma} \eta_{\gamma}(\boldsymbol{\theta}) T_{\gamma}(\mathbf{x})\right). \tag{S6.8}$$

The $T_\gamma$ are called the *sufficient statistics* for the distribution; they hold all the information about the configuration $\mathbf{x}$ needed to determine the probability; The $\eta_\gamma$ are called the *natural parameters* (see Exercise S1.2).

Note that eqn S6.8 has a form similar to a Boltzmann distribution: for the Ising model $\rho_{T,H}(\mathbf{s}) = \exp\left(-(J\sum_{<ij>} s_i s_j - H\sum_i s_i)/k_B T\right)/Z(T,H)$.

(a) *What are the two natural parameters and two sufficient statistics for our Ising model? What is $g(\boldsymbol{\theta})$? Show that $f(\mathbf{x}) = 1$.* It will be convenient to work in these natural parameters (as we did in Exercises S2.6 and S2.7). Let us call $\boldsymbol{\eta} = \{-\beta, h\}$ and $\mathbf{T}(\mathbf{s}) = \{e(\mathbf{s}), m(\mathbf{s})\}$.

The Kullback–Leibler divergence between two probability distributions $\rho(\mathbf{x})$ and $\sigma(\mathbf{x})$ is

$$KL(\rho||\sigma) = \sum_{\mathbf{x}} \rho(\mathbf{x}) \log\left(\rho(\mathbf{x})/\sigma(\mathbf{x})\right). \tag{S6.9}$$

Notice that the formula reminds us of the formula for the Shannon entropy $S = -k\sum(\rho\log(\rho))$. It has many physical interpretations and uses in statistics (expected surprise from using the wrong model, extra bits using the wrong coding algorithm, and so on.)

(b) *Calculate the divergence $KL(\rho_1(\beta_1,h_1)||\rho_2(\beta_2,h_2))$ in terms of $Z_1$, $Z_2$, the four natural parameters, $e_1 = \langle e(\mathbf{s})\rangle_{\rho_1}$ and $m_1 = \langle m(\mathbf{s})\rangle_{\rho_1}$.*

The KL divergence is not symmetric (as a distance should be). But we can use the symmetrized KL divergence (sometimes called the Jeffrey's divergence)

$$sKL(\rho,\sigma) = KL(\rho||\sigma) + KL(\sigma||\rho) \tag{S6.10}$$

as a kind of distance.[4]

(c) *Show that the sKL divergence is $-(\beta_1-\beta_2)(e_1-e_2)+(h_1-h_2)(m_1-m_2)$.*

[4]It is zero if $\rho = \sigma$ and symmetric, and one can check that it agrees with the FIM when the deviation between $\rho$ and $\sigma$ goes to zero. It does not satisfy the triangle inequality, but this is precisely why it is valuable to us, since the triangle inequality dooms low dimensional embeddings (see Exercise S6.1).

This is a known result, for a general exponential family. But what was not realized (to our knowledge) is that this can be used to find coordinates for the two points! Although, as Exercise S6.2 and more generally for multidimensional scaling embeddings, some of the coordinates are "timelike", with negative squared contributions to the distance. The space-like coordinates are the averages[5] of the natural parameters and their statistics, $(-\beta + e)/2$ and $(h + m)/2$, and the time-like coordinates are half the differences $(-\beta - e)/2$ and $(h - m)/2$,

$$\begin{aligned} T_\beta^\pm &= \tfrac{1}{2}(-\beta \pm e) \\ T_h^\pm &= \tfrac{1}{2}(h \pm m) \end{aligned} \tag{S6.11}$$

(d) *Show that the sKL divergence between $\rho_1$ and $\rho_2$ indeed is given by the sum of the squares of the space-like coordinate differences minus the sum of the squares of the time-like coordinate differences.*

Note the relation between our model manifold embedding in a Minkowski space and the *model graph* formed by plotting $(e(\beta, h), m(\beta, h))$ in four dimensions. Rotating each conjugate pair $(\beta, e)$ and $(h, m)$ of the model graph by 45 degrees generates the isometric embedding. Note that space-time rotations are *not* isometric, however. Indeed, the fully magetized states in Fig. 6.1 at low temperatures and large fields are at zero distance from one another (they are the same state), but arrange in a 45 degree diagonal with light-like separations.

The explicit formulas make generating model manifolds in most statistical mechanics problems completely straightforward. Fig. 6.1, showing the two-dimensional Ising model, uses a standard Monte-Carlo evaluation of the field and temperature-dependent energy and magnetization. (The Wolff algorithm can be generalized to work in an external field [8], making simulations like these fast even near the critical point.) One imagines using simulations like these to

Figs 6.8 and 6.7 show other illustrations of applications of the isKLe embedding.

S6.4 **Distances in probability space.**[6] (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 S2.4 introduces four problems (loaded dice, statistical mechanics, the height distribution of women, and least-squares fits to data), each of which have parameters $\boldsymbol{\theta}$

[5] Note that the natural parameter for the energy is $-\beta$. It may seem weird to add and subtract fields and magnetizations! One can find other coordinate sets (e.g., $(\lambda h \pm 1/\lambda m)/2$) which can be used to fix the units. These correspond to Lorentz boosts in the Minkowski-like embedding space.

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

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 S2.21) 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 S2.4 also introduced the Fisher information metric (FIM) in eqn 3.2:

$$g_{\mu\nu}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \log(P(\mathbf{x}))}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\mathbf{x}} \tag{S6.12}$$

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

$$d^2(P(\boldsymbol{\theta}), P(\boldsymbol{\theta} + \epsilon\boldsymbol{\Delta})) = \epsilon^2 \sum_{\mu\nu} \Delta_\mu g_{\mu\nu} \Delta_\nu. \tag{S6.13}$$

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 [13].

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 S2.4, 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:[7]

$$\begin{aligned} d^2_{\mathrm{Hel}}(P,Q) &= (P-Q)^2 = 2 - 2P\cdot Q \\ &= \int \mathrm{d}\mathbf{x} \left(\sqrt{P(\mathbf{x})} - \sqrt{Q(\mathbf{x})}\right)^2. \end{aligned} \tag{S6.14}$$

(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*

$$P_N \cdot Q_N = (P \cdot Q)^N, \tag{S6.15}$$

*and hence*

$$d^2_{\mathrm{Hel}}(P_N, Q_N) = 1 - (P \cdot Q)^N \tag{S6.16}$$

*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.[8]

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.

[7] 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 $\sqrt{2}$.

[8] 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.

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:

$$\begin{aligned} d^2_{\text{Bhatt}}(P,Q) &= \lim_{N\to 0} \tfrac{1}{2} d^2_{\text{Hel}}(P_N, Q_N)/N \\ &= -\log(P \cdot Q) \\ &= -\log\left(\sum_{\mathbf{x}} \sqrt{P(\mathbf{x})}\sqrt{Q(\mathbf{x})}\right). \end{aligned} \tag{S6.17}$$

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 S6.17.* (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 $P$.

$$d_{\mathbf{KL}}(Q|P) = -\int \mathrm{d}\mathbf{x} P(\mathbf{x}) \log(Q(\mathbf{x})/P(\mathbf{x})). \tag{S6.18}$$

(c) *Show that the Kullback–Liebler divergence is positive, zero only if $P = Q$, but is not symmetric. Show that, to quadratic order in $\epsilon$ in eqn S6.13, that the Kullback–Liebler divergence does lead to the FIM.*

The Kullback–Liebler divergence is sometimes symmetrized:

$$\begin{aligned} & d_{\mathbf{sKL}}(Q,P) \\ &\quad = \tfrac{1}{2}(d_{\mathbf{KL}}(Q|P) + d_{\mathbf{KL}}(P|Q)) \\ &\quad = \int \mathrm{d}\mathbf{x}(P(\mathbf{x}) - Q(\mathbf{x})) \log(P(\mathbf{x})/Q(\mathbf{x})). \end{aligned} \tag{S6.19}$$

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

(d) *Show that they are* intensive *[13]—that the distance grows linearly with repeated measurements*[9] *(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 S2.21, that they equal the distance in data space between the two predictions.*

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

S6.5 **Can we burn information?**[10] (Mathematics, Thermodynamics, Information geometry) ④

The use of entropy to measure information content has been remarkably fruitful in computer science, communications, and even in studying the efficiency of signaling and sensing in biological systems. The Szilard engine (Exercise 5.2) was a key argument that thermodynamic entropy and information entropy could be exchanged for one another—that one could burn information. We ask here—can they be exchanged? Are information and entropy fungible?

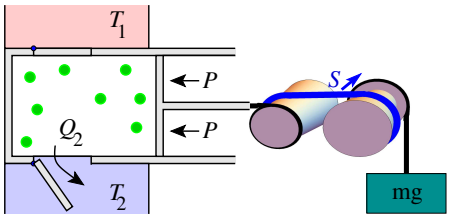



**Fig. S6.9 Piston control**. Machta [9] studies a piston plus a control system to extract work during expansion. To change the pressure, a continuously variable transmission, controlled by a gradient of the entropy $S$, connects the piston to a mass under a gravitational force. Minimizing the control cost plus the entropy cost due to fluctuations in the gear ratio lead to a minimum entropy cost for control.

Szilard stores a bit of information as an atom on one side of a piston, and extracts $P\mathrm{d}V$ work $k_B T \log 2$ as the piston expands—the same work needed to store a bit. Machta [9] argues that there is a fundamental bound on the entropy cost for extracting this work. He considers a system consisting of the piston plus a control mechanism to slowly decrease the pressure and extract the work, Fig. S6.9. (See Feynman's *Ratchet and pawl* discussion [5, I.46], discussing fluctuations in a similar system.)

Machta argues that this cost is given by a *path length in parameter space*. To be specific, Machta argues that to guide a system through a change in pressure from $P_i$ to $P_f$ should cost an entropy[11]

$$\langle \Delta S_{\text{control}} \rangle = 2 \int_{P_i}^{P_f} \sqrt{g_{PP}} |\mathrm{d}P|. \tag{S6.20}$$

The metric in this space, as discussed in Exercise S2.4, is

$$g_{PP} = -\langle \partial^2 \log(\rho)/\partial P^2 \rangle, \tag{S6.21}$$

the *Fisher information metric*, giving the natural distance between two nearby probability distributions.

[10] This exercise was developed in collaboration with Ben Machta, Archishman Raju, Colin Clement, and Katherine Quinn

[11] We shall follow Machta and set $k_B = 1$ in this exercise, writing it explicitly only when convenient.

For example, in a Gibbs ensemble at constant pressure $P = \theta_1$, the squared distance between two nearby pressures at the same temperature is

$$d^2(\rho(\mathbf{X}|P), \rho(\mathbf{X}|P + \mathrm{d}P)) = g_{PP}(\mathrm{d}P)^2, \tag{S6.22}$$

leading directly to eqn S6.20.

Let us compute the metric $g_{PP}$ in the coordinates for the ideal gas in a piston (Exercise S2.5), and then analyze the cost for thermodynamic control for Szilard's burning information engine in Exercise 5.2.

(a) *Using eqns S6.21 and S2.26, show that* $g_{PP} = (1 + N)/P^2$.

(b) *What is the entropy cost to expand a piston containing a single atom at constant temperature by a factor of two? What is the work done by the piston? How does this affect Szilard's argument about burning information in Exercise 5.2?*

Machta's result thus challenges Szilard's argument that information entropy and thermodynamic entropy can be exchanged. It also gives a (subextensive) cost for the Carnot cycle (see Exercise S2.6).

### S6.6 Averaging over disorder.[12] ③

A two-state spin takes values $S = \pm 1$. It is in an external field $h$, so that its Hamiltonian is

$$\mathcal{H} = -hS. \tag{S6.23}$$

It is connected to a heat bath at temperature $T$.

(a) *Compute its partition function $Z$, its Helmholtz free energy $A$, the entropy $S$, and the specific heat*[13] *$c$ as a function of $h$ and $T$. What is the entropy at $T = 0, h > 0$ and at $T = \infty$?* (The $T \to 0$ limit is tricky: a graphical solution is fine.) *Is the difference as expected from our understanding of information entropy?*

To model a system with dirt – a disordered system – one often adds a random term to the Hamiltonian (like a random field for each spin). One then averages the answer over the probability distribution of the disorder to predict the behavior of a large system. This turns out to be trickier than it seems.

Let us calculate the average properties of our spin in a random field $h$, averaged over a Gaussian probability distribution $\rho(h) = \exp(-h^2/2\sigma^2)/(\sqrt{2\pi}\sigma)$.

(b) *Write in integral form the average of each of the quantities $\overline{Z}$, $\overline{A}$, $\overline{S}$, and $\overline{c}$ over the probability density $\rho(h)$.* All but one of these will be infeasible to evaluate in closed form. *Evaluate the integral for $\overline{Z}$.*

In interacting systems like spin glasses, it is much easier to calculate the average of $Z$ than the average of $\log Z$ or $A$. But we run into trouble.

[12] This problem was developed in collaboration with Stephen Thornton.

[13] Section 6.1 discusses the specific heat at constant volume $c_v$, but the formulas are the same because here there is no volume to be fixed.

(c) *Define $Z_a = \overline{Z}$, and calculate the corresponding quantities $A_a$ and $S_a$. Show that $S_a$ goes negative at low temperatures.*

The entropy for each disorder you calculated in part (a) never goes negative. So its average cannot be negative! We seem to be stuck with the integrals we cannot do in closed form.

(d) *Define $A_q = \overline{A}$. Argue that $S_q$, defined as the appropriate derivative of $A_q$, is equal to $\overline{S}$ from part (b).*

Let us briefly consider a simpler scenario, where $h$ can take only the three values 0 or $\pm h_0$ (with $h_0 > 0$), each with probability 1/3.

(e) *Write $A_q$ and $A_a$ exactly for this case, and evaluate them in the limit $T \to 0$. Using $\overline{A} = \overline{\langle E \rangle} - T\overline{S}$, what value should you expect for the average free energy at $T = 0$? Does $A_a$ appear to be giving unfair weights to disorder configurations with lower-energy states?*

Thus $A_a$ gives an unfairly large weight to members of the disordered ensemble that have unusually low energy configurations. For spin glasses, $A_a$ gives unfair weights to systems like the non-disordered Ising model, where a single spin configuration can make all the bonds happy. This leads to an unphysical ferromagnetic–like transition.

Why the choice of subscripts? When we want to freeze our dirt into a particular configuration, we quench the system quickly to a low temperature. (The blacksmith pounding the red-hot horseshoe, after they get it into shape, quenches it in a bucket of water.) $A_q$ is the *quenched* free energy. We anneal a defective crystal by heating it up to a large temperature $T_0$ where its defects have enough energy to rearrange and come to equilibrium. $A_a = -k_B T \log(\overline{Z})$ is called the *annealed* free energy. But why does our $A_a$ correspond to an annealed free energy, where the "defects" come to equilibrium?

(f) *Show that $Z_a(T_0)$ from part (c) at a particular temperature $T_0$ is the true partition function for a Hamiltonian*

$$\mathcal{H}_a = h^2 k_B T_0 / 2\sigma^2 - hS + C, \tag{S6.24}$$

*where the constant $C = \frac{1}{2} k_B T_0 \log(2\pi\sigma^2)$. Thus $Z_a$ discusses a system where $h$ and $S$ are both weighted according to the Boltzmann distribution (so the field fluctuates to equilibrate with the spin).* In systems like spin glasses, one can calculate annealed averages because they are, in disguise, the correct partition function for an undisordered equilibrium system.

We must end with the *replica trick* that people use to bypass the infeasible integrals we get from trying to average the $\log(Z)$, as in $A_q = -k_B T \overline{\log Z}$. One can often calculate $\overline{Z^n}$, the annealed disorder average of $n$ replicas of a system. (Again, it is feasible because it is in disguise an equilibrium physical system, whose dirt equilibrates with the spins.) We then can find the average $\overline{\log(Z)}$ and hence $A_q$:

(g) *Show that $\log x = \lim_{n \to 0}(x^n - 1)/n$ by writing $x^n = \exp(n \log x)$.*

We can then take the average of both sides and write $\overline{\log(Z)} = \lim_{n\to 0}(\overline{Z^n} - 1)/n$. Finding the right way of taking the limit $n \to 0$ is harder than we are suggesting. The original researchers used a "replica symmetric" method that works for many systems, and works well in spin glasses for temperatures above the glass transition. Below the glass transition, one must do something more exotic. Georgio Parisi received the Nobel Prize in Physics in 2021 for showing certain disordered systems undergo a "replica symmetry breaking" transition as the temperature is lowered, where certain correlations within the system change dramatically in the spin glass phase. These methods have been shown by Parisi and others to be powerful tools for solving models of ordinary glass, analyzing deep neural network models in machine learning, and providing the fastest algorithms for challenging "NP complete" models in computer science (see Exercise 8.15).

# Bibliography


[1] (2019). The NIST reference on constants, units, and uncertainty, from CODATA internationally recommended 2018 values of the fundamental physical constants. http://physics.nist.gov/cuu/Constants/.

[2] Anderson, P. W. (1972). More is different. *Science*, **177**(4047), 393–6.

[3] Brown, K. S., Hill, C. C., Calero, G. A., Lee, K. H., Cerione, R. A., and Sethna, J. P. (2002). Integrated approaches to signal transduction: PC12 differentiation. *Biophysical Journal*, **82**, 1067.

[4] Brown, K. S. and Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, **68**, 021904.

[5] Feynman, R. P., Leighton, R. B., and Sands, M. (1963). *The Feynman lectures on physics*. Addison-Wesley, Menlo Park, CA.

[6] Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, **3**, 1871–1878.

[7] Hayden, Lorien X., Chachra, Ricky, Alemi, Alexander A., Ginsparg, Paul H., and Sethna, James P. (2018). Canonical sectors and evolution of firms in the us stock markets. *Quantitative Finance*, **18**(10), 1619–1634.

[8] Kent-Dobias, J. and Sethna, J. P. (2018). Cluster representations and the Wolff algorithm in arbitrary external fields. *Physical Review E*, **98**, 063306.

[9] Machta, B. B. (2015). Dissipation bound for thermodynamic control. *Physical Review Letters*, **115**, 260603.

[10] Mao, Jialin, Griniasty, Itay, Teoh, Han Kheng, Ramesh, Rahul, Yang, Rubing, Transtrum, Mark K, Sethna, James P, and Chaudhari, Pratik (2023). The training process of many deep networks explores the same low-dimensional manifold. *arXiv preprint arXiv:2305.01604*.

[11] Quinn, Katherine N, Abbott, Michael C, Transtrum, Mark K, Machta, Benjamin B, and Sethna, James P (2022, dec). Information geometry for multiparameter models: new perspectives on the origin of simplicity. *Reports on Progress in Physics*, **86**(3), 035901.

[12] Quinn, Katherine N., Bernardis, Francesco De, Niemack, Michael D., and Sethna, James P. (2017). Visualizing theory space: Isometric embedding of probabilistic predictions, from the Ising model to the cosmic microwave background. https://arxiv.org/abs/1709.02000.

[13] Quinn, K. N., Clement, C. B., De Bernardis, F., Niemack, M. D., and Sethna, J. P. (2019). Visualizing probabilistic models and data with intensive principal component analysis. *Proceedings of the National Academy of Sciences*, **116**(28), 13762–7.

[14] Quinn, Katherine N., Wilber, Heather, Townsend, Alex, and Sethna, James P. (2019, Apr). Chebyshev approximation and the global geometry of model predictions. *Phys. Rev. Lett.*, **122**, 158302.

[15] Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York.

[16] Sethna, J. P., Chachra, R., Machta, B. B., and Transtrum, M. K. (2013). Why is science possible? http://sethna.lassp.cornell.edu/Sloppy/WhyIsSciencePossible.html.

[17] Sethna, J. P. and Myers, C. R. (2020). *Entropy, order parameters, and complexity*, computer exercises and course materials. http://sethna.lassp.cornell.edu/StatMech/EOPCHintsAndMaterials.html.

[18] Taimina, Daina (2011). Daina Taimina—Fiber Sculptures, Hyperbolic plane, from the textile collection at the Cooper-Hewitt National Design museum. https://dainataimina.blogspot.com/2011/12/hyperbolic-planes.html.

[19] Teoh, Han Kheng, Quinn, Katherine N., Kent-Dobias, Jaron, Clement, Colin B., Xu, Qingyang, and Sethna, James P. (2020, Aug). Visualizing probabilistic models in Minkowski space with intensive symmetrized Kullback-Leibler embedding. *Phys. Rev. Research*, **2**, 033221.

[20] Transtrum, Mark K., Machta, Benjamin B., Brown, Kevin S., Daniels, Bryan C., Myers, Christopher R., and Sethna, James P. (2015). Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of Chemical Physics*, **143**(1), 010901.

[21] Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2010). Why are nonlinear fits to data so challenging? *Physical Review Letters*, **104**, 060201.

[22] Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2011). Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, **83**, 036701.

[23] Wikipedia contributors (2024). Singular value decomposition — Wikipedia, the free encyclopedia. [Online; accessed 9-January-2024].