MDPI

Article

Far from Asymptopia: Unbiased High-Dimensional Inference Cannot Assume Unlimited Data

Michael C. Abbott * and Benjamin B. Machta *

Department of Physics, Yale University, New Haven, CT 06520, USA
*   Correspondence: michael.abbott@yale.edu (M.C.A.); benjamin.machta@yale.edu (B.B.M.)

Abstract: Inference from limited data requires a notion of measure on parameter space, which is most explicit in the Bayesian framework as a prior distribution. Jeffreys prior is the best-known uninformative choice, the invariant volume element from information geometry, but we demonstrate here that this leads to enormous bias in typical high-dimensional models. This is because models found in science typically have an effective dimensionality of accessible behaviors much smaller than the number of microscopic parameters. Any measure which treats all of these parameters equally is far from uniform when projected onto the sub-space of relevant parameters, due to variations in the local co-volume of irrelevant directions. We present results on a principled choice of measure which avoids this issue and leads to unbiased posteriors by focusing on relevant parameters. This optimal prior depends on the quantity of data to be gathered, and approaches Jeffreys prior in the asymptotic limit. However, for typical models, this limit cannot be justified without an impossibly large increase in the quantity of data, exponential in the number of microscopic parameters.

Keywords: Bayesian inference; information geometry; posterior bias; model selection

1. Introduction

No experiment fixes a model's parameters perfectly. Every approach to propagating the resulting uncertainty must, explicitly or implicitly, assume a measure of the space of possible parameter values. A badly chosen measure can introduce bias, and we argue here that avoiding such bias is equivalent to the very natural goal of assigning equal weight to each distinguishable outcome. However, this goal is seldom reached, either because no attempt is made, or because the problem is simplified by prematurely assuming the asymptotic limit of nearly infinite data. We demonstrate here that this assumption can lead to a large bias in what we infer from the parameters, in models with features typical of many-parameter mechanistic models found in science. We propose a score for such bias, and advocate for using a measure that makes this zero. Such a measure allows for unbiased inference without the need to first simplify the model to just the right degree of complexity. Instead, weight is automatically spread according to a lower effective dimensionality, ignoring details irrelevant to visible outcomes.

We consider models which predict a probability distribution $p(x|\theta)$ for observing data x given parameters θ. The degree of overlap between two such distributions indicates how difficult it is to distinguish the two parameter points, which gives a notion of distance on parameter space. The simplifying idea of information geometry is to focus on infinitesimally close parameter points, for which there is a natural Riemannian metric, the Fisher information [1,2]. This may be thought of as having units of standard deviations, so that along a line of integrated length L there are about L distinguishable points, and thus any parameter which can be measured to a few digits of precision has length $L > 100$. It is a striking empirical feature of models in science that most have a few such long (or relevant) parameter directions, followed by many more short (or irrelevant) orthogonal directions [3–6]. The irrelevant lengths, all $L < 1$, show a characteristic spectrum of being

roughly evenly spaced on a log scale, often over many decades. As a result, much of the geometry of this Riemannian model manifold consists of features much smaller than 1, far too small to observe. However, the natural intrinsic volume measure, which follows from the Fisher metric, is sensitive to all of these unobservable dimensions, and as we demonstrate here, they cause this measure to introduce enormous bias.

To avoid this problem, we need a measure tied to the Fisher length scale $L \approx 1$, instead of one from the continuum. Locally, this length scale partitions dimensions into relevant and irrelevant, which in turn approximately factorizes the volume element into a relevant part and what we term the *irrelevant co-volume*. The wild variations of this co-volume are the source of the bias we describe, and it is rational to ignore them. As we illustrate in Figure 1 for a simple two-parameter model, equally distinguishable predictions do not correspond to equal intrinsic volumes, and this failure is detected by a score we call *bias pressure*. The measure $p_\star(\theta)$ for which this score is everywhere zero, by contrast, captures relevant distinguishability and ignores the very thin irrelevant direction. The same measure is also obtained by maximizing the information learned about parameter θ from seeing data x [7–9], or equivalently from a particular minimax game [10–12]. Since $p_\star(\theta)$ is usually discrete [9,13–18], it can be seen as implementing a length cutoff, replacing the smooth differential-geometric view of the model manifold with something quantized [19].
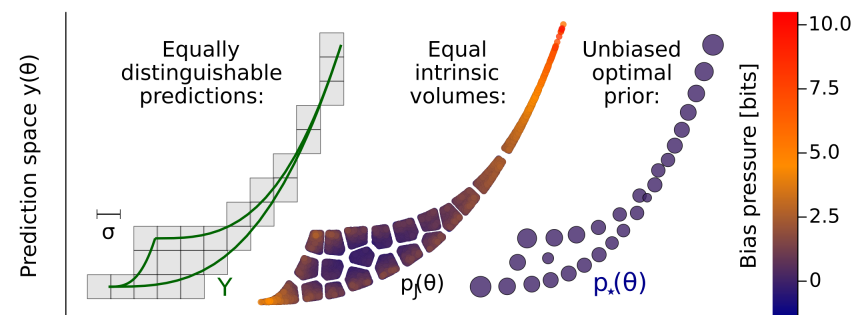

Figure 1. The natural volume is a biased measure for the space of distinguishable outcomes. The left panel outlines the space of possible predictions Y; the observed x is deterministic $y(\theta)$ plus measurement noise. With the scale of the noise σ as shown, the upper half is effectively one-dimensional. The center panel shows a sample from the volume measure $p_J(\theta)$, divided into blocks of equal weight. These are strongly influenced by the unobservable thickness of the upper portion. Points are colored by bias pressure $b(\theta)$, which we define in Equation (5). The right panel shows the explicitly unbiased optimal measure $p_\star(\theta)$, which gradually adjusts from two- to one-dimensional behavior (The model is Equation (6) with $a_1 = 0.8$, $a_2 = 0.2$, and $k_1 \geq k_2$, observed at times $t = 1, 3$ each with Gaussian noise $\sigma = 0.1$).

In the Bayesian framework, the natural continuous volume measure $p_J(\theta)$ is known as Jeffreys prior, and is the canonical example of an uninformative prior: a principled, ostensibly neutral choice. It was first derived based on invariance considerations [20], and can also be justified by information- or game-theoretic ideas, provided these are applied in the limit of infinitely many repetitions [7,8,17,21,22]. This asymptotic limit often looks like a technical trick to simplify derivations. However, in realistic models, this limit is very far from being justified, exponentially far in the number of parameters, often requiring an experiment to be repeated for longer than the age of the universe. We demonstrate here that using the prior derived in this limit introduces a large bias in such models. Furthermore, we argue that such bias, and not only computational difficulties, has prevented the wide use of uninformative priors.

The promise of principled ways of tracking uncertainty, Bayesian or otherwise, is to free us from the need to select a model with precisely the right degree of complexity. This idea is often encountered in the context of overfitting, where the maximum likelihood point of an overly complex model gives worse predictions. The bias discussed here is a distinct way for overly complex models to give bad predictions. We begin with toy models in which

the number of parameters can be easily adjusted. However, in the real models of interest, we cannot trivially tune the number of parameters. This is why we wish to find principled methods which are not fooled by the presence of many irrelevant parameters.

2. Results

We consider a model to be characterized by the likelihood $p(x|\theta)$ of observing data $x \in X$ when the parameters are $\theta \in \Theta$. In such a model, the Fisher information metric (FIM) measures the distinguishability of nearby points in parameter space as a distance $ds^2(\theta) = \sum_{\mu,\nu=1}^{d} g_{\mu\nu}(\theta) d\theta^\mu d\theta^\nu$, where

$$g_{\mu\nu}(\theta) = -\int dx\, p(x|\theta)\, \partial_\mu \partial_\nu \log p(x|\theta). \tag{1}$$

For definiteness, we may take points separated along a geodesic by a distance $L = \int \sqrt{ds^2(\theta)} > 1$ to be distinguishable. Intuitively, though incorrectly, the d-dimensional volume implied by the FIM might be thought to correspond to the total number of distinguishable parameter values inferable from an experiment:

$$Z = \int d\theta \sqrt{\det g(\theta)}.$$

However, this counting makes a subtle assumption that all structure in the model has a scale much larger than 1. When many dimensions are smaller than 1, their lengths weigh the effective volume along the larger dimensions, despite having no influence on distinguishability.

The same effect applies to the normalized measure, Jeffreys prior:

$$p_J(\theta) = \frac{1}{Z}\sqrt{\det g(\theta)}. \tag{2}$$

This measure's dependence on the *irrelevant co-volume* is an under-appreciated source of bias in posteriors derived from this prior. The effect is most clearly seen when the FIM is block-diagonal, $g = g_{\text{rel}} \oplus g_{\text{irrel}}$. Then the volume form factorizes exactly, and the relevant effective measure is the $\sqrt{\det g_{\text{rel}}(\theta_{\text{rel}})}$ factor times $V_{\text{irrel}}(\theta_{\text{rel}})$, an integral over the irrelevant dimensions.

A more principled measure of the (log of the) number of distinguishable outcomes is the mutual information between parameters and data, $I(X;\Theta)$:

$$I(X;\Theta) = \int d\theta\, p(\theta) D_{\text{KL}}\big[p(x|\theta)\big\|p(x)\big]$$

where D_{KL} is the Kullback–Leibler divergence between two probability distributions, which are not necessarily close: $p(x) = \int d\theta\, p(\theta)\, p(x|\theta)$ is typically much broader than $p(x|\theta)$. Unlike the volume Z, the mutual information depends on the prior $p(\theta)$. Past work, both by ourselves and others, has advocated for using the prior, which maximizes this mutual information, with [8] or without [9] taking the asymptotic limit:

$$p_\star(\theta) = \underset{p(\theta)}{\text{argmax}}\, I(X;\Theta). \tag{3}$$

The same prior arises from a minimax game in which you choose a prior, your opponent chooses the true θ, and you lose the (large) KL divergence [10–12]:

$$p_\star(\theta) = \underset{p(\theta)}{\text{argmin}}\, \underset{\theta}{\max}\, D_{\text{KL}}\big[p(x|\theta)\big\|p(x)\big]. \tag{4}$$

Here we stress a third perspective, defining a quantity we call *bias pressure* which captures how strongly the prior disfavors predictions from a given point:

$$b(\theta) = \left.\frac{\partial I(X;\Theta)}{\partial p(\theta)}\right|_{\int d\theta\, p(\theta)=1} = D_{\mathrm{KL}}\big[p(x|\theta)\|p(x)\big] - I(X;\Theta). \tag{5}$$

The optimal $p_\star(\theta)$ has $b(\theta) = 0$ on its support, and can be found by minimizing $B = \max_\theta b(\theta)$. Other priors have $b(\theta) > 0$ at some points, indicating that $I(X;\Theta)$ can be increased by moving weight there (and away from points where $b(\theta) < 0$). We demonstrate below that $b(\theta)$ deserves to be called a bias, as it relates to large deviations of the posterior center of mass. We enact this by presenting a number of toy models, chosen to have information geometry similar to that typically found in mechanistic models from many scientific fields [5].

2.1. Exponential Decay Models

The first model we study involves inferring rates of exponential decay. This may be motivated, for instance, by the problem of determining the composition of a radioactive source containing elements with different half-lives, using Geiger counter readings taken over some period of time. The mean count rate at time t is

$$y_t(\theta) = \sum_{\mu=1}^{d} a_\mu e^{-k_\mu t}, \quad k_\mu = e^{-\theta_\mu} > 0. \tag{6}$$

We take the decay rates as parameters, and fix the proportions a_μ, usually to $a_\mu = 1/d$, thus initial condition $y_0 = 1$. If we make observations at m distinct times t, then the prediction y is an m-vector, restricted to a compact region $Y \subset [0,1]^m$. For radioactivity, we would expect to observe y_t plus Poisson noise, but the qualitative features are the same if we simplify to Gaussian noise with constant width σ:

$$p(x|\theta) = e^{-|x-y(\theta)|^2/2\sigma^2}/(2\pi\sigma^2)^{m/2}. \tag{7}$$

The Fisher metric then simplifies to be the Euclidean metric in the space of predictions Y, pulled back to parameter space Θ:

$$g_{\mu\nu}(\theta) = \frac{1}{\sigma^2}\sum_{t,t'}^{m} \frac{\partial y_t}{\partial \theta^\mu}\frac{\partial y_{t'}}{\partial \theta^\nu}\delta_{tt'}$$

thus, plots of $p(y)$ in $\mathbb{R}^m$ will show Fisher distances accurately. This model is known to be ill conditioned, with many small manifold widths and many small FIM eigenvalues when d is large [23].

With just two dimensions, $d = m = 2$, Figure 1 shows the region $Y \subset \mathbb{R}^2$, Jeffreys prior $p_J(\theta)$, and the optimal prior $p_\star(\theta)$, projected to densities on Y. Jeffreys is uniform $p_J(y) \propto 1$ (since the metric is constant in y), and hence always weights a two-dimensional area, both where this is appropriate and where it's not. The upper portion of Y in the figure is thin compared to σ, so the points we can distinguish are those separated vertically: the model is effectively one-dimension there. Jeffreys does not handle this well, which we illustrate in two ways. First, the prior is drawn divided into 20 segments of equal weight (equal area), which roughly correspond to distinguishable differences where the model is two-dimensional, but not where it becomes one-dimensional. Second, the points are colored by $b(\theta)$, which detects this effect, and gives large values at the top (about 10 bits). The optimal prior avoids these flaws by smoothly adjusting from the one- to the two-dimensional part of the model [9].

The claim that some parts of the model are effectively one-dimensional depends on the amount of data gathered. Independent repetitions of the experiment have overall likelihood $p(x^M|\theta) = \prod_{i=1}^{M} p(x^{(i)}|\theta)$, which will always scale the FIM by M, hence all

distances by a factor $\sqrt{M}$. This scaling is exactly equivalent to smaller Gaussian noise σ. Increasing M increases the number of distinguishable points, and large enough M (or small enough σ) can eventually make any nonzero length larger than 1. Thus, the amount of data gathered affects which parameters are relevant. However, notice that such repetition has no effect at all on $p_J(\theta)$, since the scale of $g_{\mu\nu}(\theta)$ in Equation (2) is canceled by Z. In this sense, it is already clear that Jeffreys prior belongs to the fixed point of repetition, i.e., to the asymptotic limit $M \to \infty$.

Figure 2 shows a more complicated version of the model (6), with $d = 4$ parameters, and looks at the effect of varying the noise level σ. Jeffreys prior always fills the 4-dimensional bulk, but at moderate σ, most of the distinguishable outcomes are located far from this mass. At large σ, equivalent to few repetitions, all the weight of the optimal prior is on zero- and one-dimensional edges. As more data are gathered, it gradually fills in the bulk, until, in the asymptotic limit $\sigma \to 0$, it approaches Jeffreys prior [12,17,21,24]. However, while $p_\star(\theta)$ approaches a continuum at any interior point [25], it remains discrete at Fisher distances ~ 1 from the boundary. The worst-case bias pressure detects this; hence, the maximum for Jeffreys prior does not approach that for the optimal prior: $B_J \not\to 0$. However, since mutual information is dominated by the interior in this limit, we expect the values for $p_J(\theta)$ and $p_\star(\theta)$ to agree in the limit: $I_J - I_\star \to 0$.
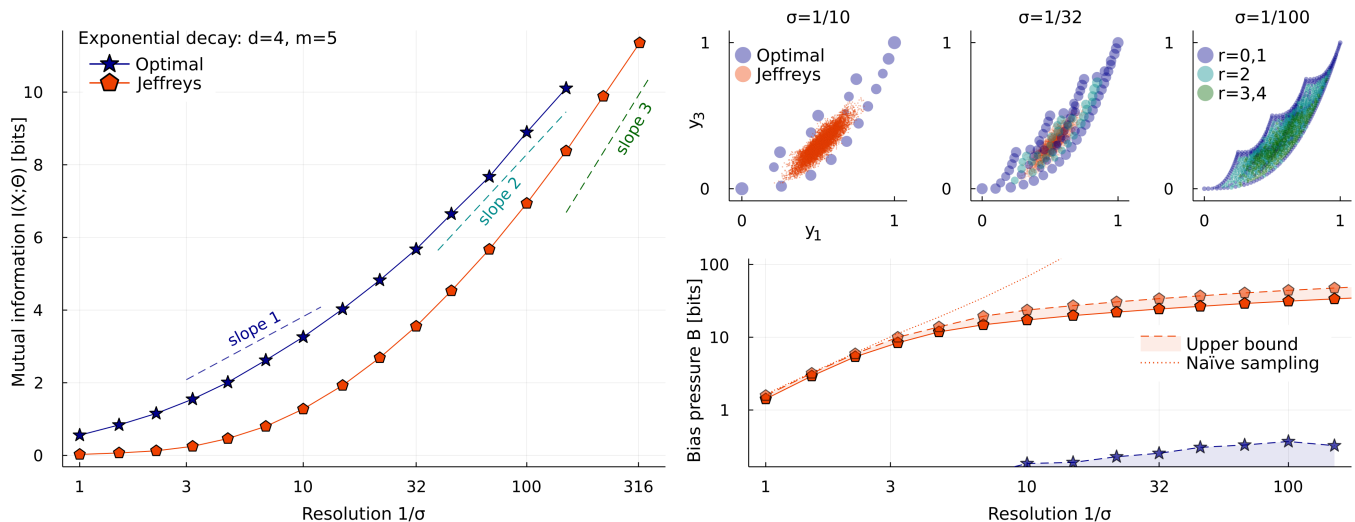

Figure 2. The effect of varying the noise level σ on a fixed model. The model is (6) with $d = 4$ parameters, observed at $m = 5$ times $t = 1, 2, \ldots 5$. Top right, the optimal prior $p_\star(\theta)$ has all of its weight on 0- and 1-dimensional edges at large σ, but adjusts to fill in the bulk at small σ (colors indicate the dimension r of the 4-dimensional shape's edge on which a point is located, the rank of the FIM there). Jeffreys prior $p_J(\theta)$ is independent of σ, and has nonzero density everywhere, but a sample of 10^6 points is largely located near the middle of the shape. Left, the slope of $I_\star(X;\Theta) \sim d_{\text{eff}} \log 1/\sigma$ gives a notion of effective dimensionality; in the asymptotic limit $\sigma \to 0$, we expect $d_{\text{eff}} = d = 4$. Bottom right, the worst-case bias pressure $B = \max_\theta b(\theta)$ is always zero for $p_\star(\theta)$, up to the numerical error, but remains nonzero for $p_J(\theta)$ even in the asymptotic limit. The Appendix A describes how upper and lower bounds for B are calculated.

One way to quantify the effective dimensionality is to look at the rate of increase in mutual information under repetition, or decreasing noise σ. Along a dimension with Fisher length $L \gg 1$, the number of distinguishable points is proportional to L, and thus a cube with d_{eff} large dimensions will have $\propto L^{d_{\text{eff}}}$ such points. This motivates defining d_{eff} by

$$I_\star(X;\Theta) \sim d_{\text{eff}} \log L, \qquad L = \int \sqrt{ds^2} \propto 1/\sigma. \qquad (8)$$

Figure 2 shows lines for slope $d_{\text{eff}} = 1, 2, 3$, and we expect $d_{\text{eff}} \to d$ in the limit $\sigma \to 0$.

2.2. The Costs of High Dimensionality

The problems of uneven measure grow more severe with more dimensions. To explore this, Figures 3–5 show a sequence of models with 1 to 26 parameters. All describe the same data: observations at the same list of $m = 26$ times in $1 \le t \le 5$ with the same noise $\sigma = 0.1$. While Jeffreys prior is nonzero everywhere, its weight is concentrated where the many irrelevant dimensions are largest. With a Monte Carlo sample of a million points, all are found within the small orange area on the right of Figure 3. For a particular observation x, we plot also the posterior $p(\theta|x)$ for each prior. The extreme concentration of weight in $p_J(\theta)$ in $d = 26$ pulls this some 20 standard deviations away from the maximum likelihood point $y(\hat{\theta}_x)$. We call this distance the posterior deviation Δ; it is the most literal kind of bias in results.
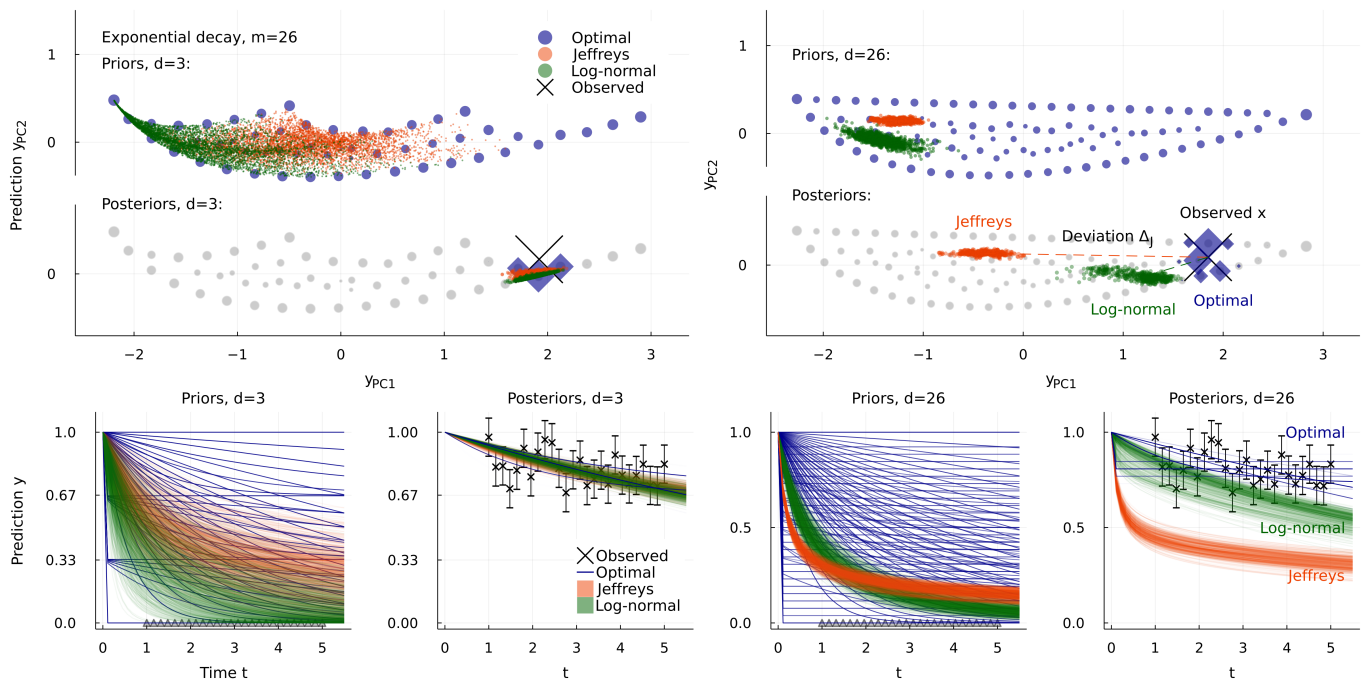

Figure 3. The effect of changing model dimension for fixed data and noise level. Left half, the exponential decay model of Equation (6) with $d = 3$ parameters, observed with noise $\sigma = 0.1$ at $m = 26$ times in $1 \le t \le 5$. Three priors are shown, drawn above by projecting onto the first two principal components of vector y, and below-left as a time course y_t (each point on the upper plot is a line on the lower one). The corresponding posteriors are shown for a particular fixed x, which is the large cross in the upper plot (where the prior is shown again in light gray as a visual guide) and the series of points in the lower plot. In the $d = 3$ model, all three posteriors are reasonable fits to the data. Right half, the similar model with $d = 26$ parameters, for the same observations with the same noise. Here Jeffreys prior is much more strongly concentrated, favoring the part of the manifold where the irrelevant dimensions are largest. This has the effect of biasing the posterior far from the data, more than 20 standard deviations away. Figures 4 and 5 explore the same setup further, including intermediate dimensions d. The log-normal prior is introduced in Equation (10).

Figure 4 compares the posterior deviation Δ to the bias pressure $b(\theta)$ defined in Equation (5). For each of many observations x, we find the maximum likelihood point $\hat{\theta}_x = \arg\max_\theta p(x|\theta)$, and calculate the distance from this point to the posterior expectation value of y:

$$\Delta(x) = \frac{1}{\sigma} \left| y(\hat{\theta}_x) - \int d\theta \, p(\theta|x) \, y(\theta) \right|. \qquad (9)$$

Then, using the same prior, we evaluate the corresponding bias pressure, $b(\hat{\theta}_x)$. The figure shows 100 observations x drawn from $p_\star(x) = \int d\theta \, p_\star(\theta) \, p(x|\theta)$, and we believe this

justifies the use of the word "bias" to describe $b(\theta)$. The figure is for $d = 11$, but a similar relationship is seen in other dimensionalities.
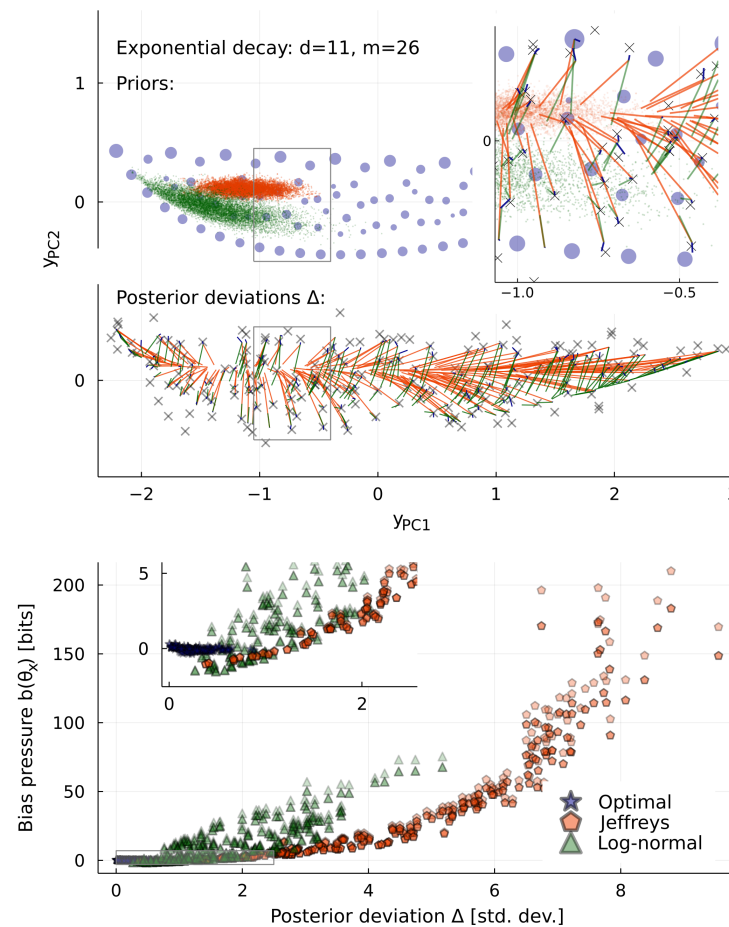

Figure 4. Posterior bias due to concentration of measure. Above, priors for the $d = 11$ case of the model in Figure 3. We calculate the posterior for each at 100 points x (marked), and draw a line connecting the maximum likelihood point $y(\hat{\theta}_x)$ to the posterior center of mass $\langle y(\theta) \rangle_x$. Inset enlarges to show that there are blue lines too, for the optimal prior, most much shorter than the spacing of its atoms. Below, we compare the length of such lines (divided by $\sigma = 0.1$) to the bias pressure $b(\hat{\theta}_x)$. Notice that $b(\theta)$ is sometimes negative (it has zero expectation value: $\int d\theta\, p(\theta)\, b(\theta) = 0$), although the worst-case $B = \max_\theta b(\theta)$ is non-negative. Each pair of darker and lighter points are a lower and an upper bound, explained in the Appendix A.

Instead of looking at particular observations x, Figure 5 shows global criteria $I(X; \Theta)$ and $B = \max_\theta b(\theta)$. The optimal prior is largely unaffected by the addition of many irrelevant dimensions. Once $d > 3$, it captures essentially the same information in any higher dimension and has zero bias (or near-zero bias, in our numerical approximation). We may think of this as a new invariance principle, that predictions should be independent of unobservable model details. This replaces one of the invariances of Jeffreys, that repetition of the experiment does not change the prior. Repetition invariance guarantees poor performance when we are far from the asymptotic limit, as we see here from the rapidly declining performance of Jeffreys prior with increasing dimension, capturing less than one bit in $d = 26$. This decline in information is mirrored by a rise in the worst-case bias B.
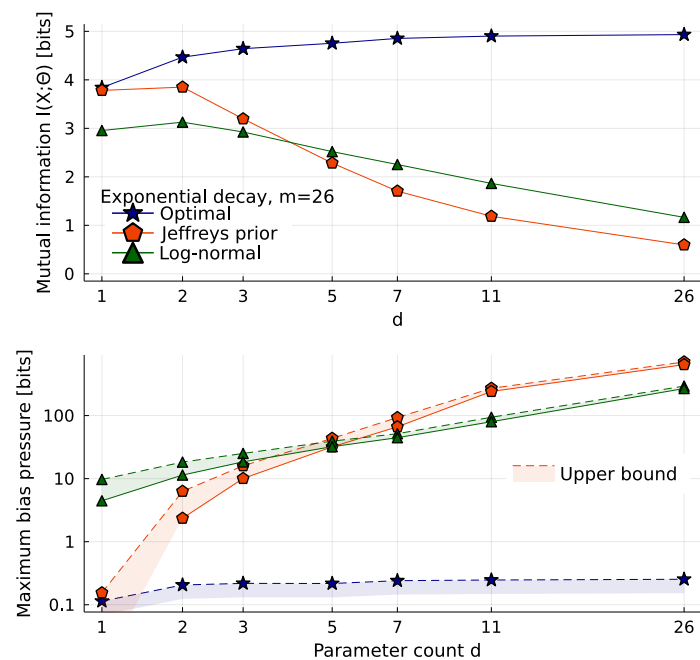
Figure 5. Information-theoretic scores for priors, as a function of dimensionality d. Like Figure 3, these models all describe the same data, with the same noise. Above, mutual information $I(X;\Theta)/\log 2$ (all plots are scaled thus to have units of bits). The optimal prior ignores the addition of more irrelevant parameters, but Jeffreys prior is badly affected, and ends up capturing less than 1 bit. Below, worst-case bias pressure $\max_\theta b(\theta)/\log 2$. This should be zero for the optimal prior, but our numerical solution has small errors. For the other priors, we plot lower and upper bounds, calculated using Bennett's method [26], as described in the Appendix A. The bias of Jeffreys prior increases strongly with the increasing concentration of its weight in higher dimensions.

Figures 3–5 also show a third prior, which is log-normal in each decay rate $k_\mu = e^{\theta_\mu} > 0$, that is, normal in terms of $\theta \in \mathbb{R}^d$:

$$p_{\mathrm{LN}}(\theta) \propto \prod_{\mu=1}^{d} e^{-(\theta_\mu - \bar\theta)^2/2\bar\sigma^2}, \qquad \bar\theta = 0, \ \bar\sigma = 1 \ \forall\mu. \tag{10}$$

This is not a strongly principled choice, but something like this is commonly used for parameters known to be positive. Here it produces better results than Jeffreys prior in high dimensions. We observe that it also suffers a decline in performance with increasing d, despite making no attempt to deliberately adapt to the high-dimensional geometry. The details of how well it works will, of course, depend on the values chosen for $\bar\theta, \bar\sigma$, and more complicated priors of this sort can be invented. With enough free "meta-parameters" such as $\bar\theta, \bar\sigma$, we can surely adjust such a prior to approximate the optimal prior, and in practice, such a variational approach might be more useful than solving for the optimal prior directly. We believe that worst-case bias $B = \max_\theta b(\theta)$ is a good score for this purpose, partly because its zero point is meaningful.

2.3. Inequivalent Parameters

Compared to these toy models, more realistic models often still have many parameter combinations poorly fixed by data, but seldom come in families that allow us to easily tune the number of dimensions. Instead of having many interchangeable parameters, each will often describe a different microscopic effect that we know to exist, even if we are not sure which combination of them will matter in a given regime [27]. To illustrate this, we now examine some models of enzyme kinetics, starting with the famous reaction:

$$E + S \underset{k_r}{\overset{k_f}{\rightleftharpoons}} ES \overset{k_p}{\rightarrow} E + P \tag{11}$$

This summarises differential equations for the concentrations, such as $\partial_t[P] = k_p[ES]$ for the final product P, and $\partial_t[E] = -k_f[E][S] + k_r[ES] + k_p[ES]$ for the enzyme, which combines with the substrate to form a bound complex.

   If the concentration of product $[P]$ is observed at some number times, with some noise, and starting from fixed initial conditions, then this model is not unlike the toy model above. Figure 6 shows the resulting priors for the rate constants appearing in Equation (11). The shape of the model manifold is similar, and the optimal prior again places most of its weight along two one-dimensional edges, while Jeffreys prior places it in the bulk, favoring the region where all three rate constants come closest to having independently visible effects on the data. However, the resulting bias is not extreme in three dimensions.
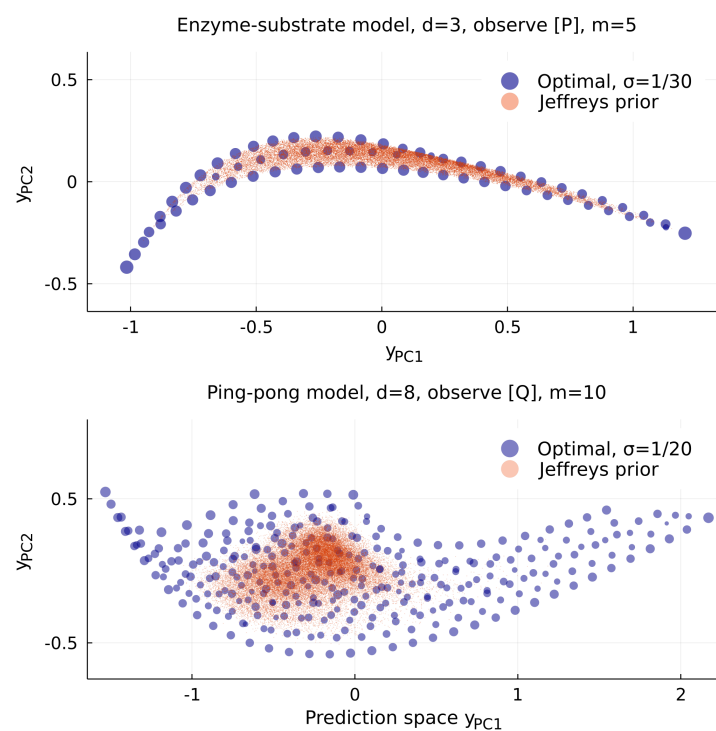

Figure 6. Priors for two models of enzyme kinetics. Above, the 3-parameter model from Equation (11), observing only the concentration of product $[P]$ at times $t = 1, 2, \ldots 5$. Below, the 8-parameter model from Equation (12), observing only the final product $[Q]$ at times $t = 1, 2, \ldots 10$. Here Jeffreys prior has worst-case bias $B \approx 28$ bits, comparable to the models in Figure 5 at similar dimension, while the optimal prior for the $d = 3$ model has its weight on well-known 2-parameter approximations, including that of Michaelis and Menten, the edge structure of the $d = 8$ model is much more complicated (for suitable initial conditions, it will include the $d = 3$ model as an edge).

   The edges of this model are known approximations, in which certain rate constants become infinite (or equal), which we discuss in the Appendix A [28]. These approximations are useful in practice since each spans the full length of the most relevant parameter. However, the more difficult situation is when many different processes of comparable speed are unavoidably involved. The model manifold may still have many short directions, but the simpler description selected by $p_\star(\theta)$ will tend to have weight on many different processes. In other words, the simpler model, according to information theory, is not necessarily one simpler model obtained by taking a limit, but instead, a mixture of many different analytic limits.

To see this, we consider a slightly more complicated enzyme kinetics model, the ping-pong mechanism with $d = 8$ rate constants:

$$E + A \rightarrow EA \rightleftharpoons E^*P \rightarrow E^* + P \tag{12}$$
$$E^* + B \rightarrow E^*B \rightleftharpoons EQ \rightarrow E + Q.$$
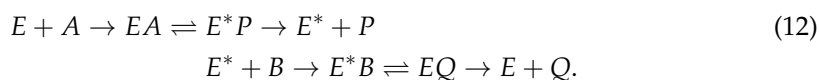
Here E^* is a deformed version of the enzyme E, which is produced in the reaction from A to P, and reverted in the reaction from B to final product Q. There are clearly many more possible limits in which some combination of the rate constants become large or small. Figure 6 shows that the optimal prior has weight on at least five different 1-edges, none of which is a good description by itself.

The concentration of weight seen in Jeffreys prior for these enzyme models is comparable to what we had before, with worst-case bias pressure $B \approx 14$ bits in $d = 3$ and 28 bits in $d = 8$. These examples share geometric features with many real models in science [5], and thus we believe the problems described here are generic.

3. Discussion

Before fitting a model to data, there is often a selection step to choose a model which is complex enough to fit the true pattern, but not so complex as to fit the noise. The motivation for this is clear in maximum likelihood estimation, where only one $\hat{\theta}_x$ is kept, and there are various criteria for making the trade-off [29–31]. The motivation is less clear in Bayesian analysis, where slightly different criteria can be derived by approximating $p(x)$ [32,33]. We might hope that if many different points θ are consistent with the noisy data x, then the posterior $p(\theta|x)$ should simply have weight on all of them, encoding our uncertainty about θ.

Why, then, is model selection needed at all in Bayesian inference? Our answer here is that this is performed to avoid measure-induced bias, not overfitting. When using a sub-optimal prior, models with too much complexity do indeed perform badly. This problem is seen in Figure 5, in the rapid decline of scores $I(X; \Theta)$ or B with increasing d, and would also be seen in the more traditional model evidence $p(x)$—all of these scores prefer models with $d \leq 3$. However, the problem is not overfitting, since the extra parameters being added are irrelevant, i.e., they can have very little effect on the predictions $y_t(\theta)$. Instead, the problem is concentration of measure. In models with tens of parameters, this effect can be enormous: It leads to posterior expectation values $\Delta > 20$ standard deviations away from ideal, for the $d = 26$ model with Jeffreys prior, and mutual information $I < 1$ bit learned, and $B > 500$ bits of bias. This problem is completely avoided by the optimal prior $p_\star(\theta)$, which suffers no decline in performance with increasing parameter count d.

Geometrically, we can view traditional model selection as adjusting d to ensure that the model manifold only has dimensions of length $L > 1$. This ensures that most of the posterior weight is in the interior of the manifold; hence, ignoring model edges is justified. By contrast, when there are dimensions of length $L < 1$, the optimal posterior will usually have its weight at their extreme values, on several manifold edges, which are themselves simpler models [9]. Fisher lengths L depend on the quantity of data to be gathered, and repeating an experiment M times enlarges all by a factor $\sqrt{M}$. Large enough M can eventually make any dimension larger than 1, and thus repetition alters what d traditional model selection prefers. Similarly, repetition alters the effective dimensionality of $p_\star(\theta)$. Some earlier work on model geometry studies a series in $1/M$ [22,33,34]; this expansion around $L = \infty$ captures some features beyond the volume but is not suitable for models with dimensions $L \ll 1$.

Real models in science typically have many irrelevant parameters [5,35–38]. It is common to have parameter directions 10^{-10} times as important as the most relevant one, but impossible to repeat an experiment the $M = 10^{20}$ times needed to bridge this gap. Sometimes it is possible to remove the irrelevant parameters and derive a simpler effective theory. This is what happens in physics, where a large separation of scales allows great

simplicity and high accuracy [39,40]. However, many other systems we would like to model cannot, or cannot yet, be so simplified. For complicated biological reactions, climate models, or neural networks, it is unclear which of the microscopic details can be safely ignored, or what the right effective variable is. Unlike our toy models, we cannot easily adjust d, since every parameter has a different meaning. This is why we seek statistical methods which do not require us to find the right effective theory. Furthermore, in particular, here we study priors almost invariant to complexity.

The optimal prior is discrete, which makes it difficult to find, and this difficulty appears to be why its good properties have been overlooked. It is known analytically only for extremely simple models such as $M = 1$ Bernoulli, and previous numerical work only treated slightly more complicated models, with $d \leq 2$ parameters [9]. While our concern here is with the ideal properties, for practical use nearly optimal approximations may be required. One possibility is the adaptive slab-and-spike prior introduced in [6]. Another would be to use some variational family $p_\lambda(\theta)$ with adjustable meta-parameters λ [41].

Discreteness is also how the exactly optimal $p_\star(\theta)$ encodes a length scale $L \approx 1$ in the model geometry, which is the divide between relevant and irrelevant parameters, between parameters that are constrained by data and those which are not. Making this distinction in some way is essential for good behavior, and it implies a dependence on the quantity of data. An effective model appropriate for much fewer data than observed will be too simple: the atoms of $p_\star(\theta)$ will be too far apart (much like recording too few significant figures), or else selecting a small d means picking just one edge (fixing some parameters which may, in fact, be relevant). On the other hand, what we have demonstrated here is that a model appropriate for much more data—infinitely more in the case of $p_J(\theta)$—will instead introduce enormous bias into our inference about θ.

Appendix A

For brevity, the main text omits some standard definitions. The KL divergence (or relative entropy) is defined

$$D_{\mathrm{KL}}\big[p(x)\big\|q(x)\big] = \int dx\, p(x)\, \log\frac{p(x)}{q(x)}.$$

With a conditional probability, in our notation $D_{\mathrm{KL}}\big[p(x|\theta)\big\|q(x)\big]$ integrates x but remains a function of θ. The Fisher information metric $g_{\mu\nu}(\theta)$ is the quadratic term from expanding $D_{\mathrm{KL}}\big[p(x|\theta + d\theta)\big\|p(x|\theta)\big]$.

The mutual information is

$$\begin{aligned}
I(X;\Theta) &= D_{\mathrm{KL}}\big[p(x,\theta)\|p(x)p(\theta)\big] \\
&= S(\Theta) - S(\Theta|X) \\
&= \iint dx\, d\theta\; p(x|\theta)p(\theta)\log\frac{p(x|\theta)}{p(x)}
\end{aligned}$$

where we use Bayes theorem,

$$p(\theta|x) = p(x|\theta)p(\theta)\big/p(x)$$

with $p(x) = \int d\theta\, p(x|\theta)p(\theta)$, and entropy

$$S(X) = -\int dx\, p(x)\log p(x).$$

Conditional entropy is $S(X|\theta) = -\int dx\, p(x|\theta)\log p(x|\theta)$ for one value θ, or $S(X|\Theta) = \int d\theta\, p(\theta)\, S(X|\theta)$. With a Gaussian likelihood, Equation (7), and $X = \mathbb{R}^m$, this is a constant:

$$S(X|\Theta) = \frac{m}{2}(1 + \log 2\pi\sigma^2).$$

Many of these quantities depend on the choice of prior, such as the posterior $p(\theta|x)$ and the mutual information $I(X;\Theta)$. This is also true of our bias pressure $b(\theta)$ and worst-case $B = \max_\theta b(\theta)$. When we need to refer to those for a specific prior, such as $p_\star(\theta)$, we use the same subscript, writing $I_\star$ and $B_\star$.

All probability distributions are normalized. In particular, the gradient in (5) is taken with the constraint of normalization—varying the density at each point independently would give a different constant.

Appendix A.1. Square Hypercone

Here we consider an even simpler toy model, in which we can more rigorously define what we mean by co-volume, and analytically calculate the posterior deviation.

Consider a d-dimensional cone consisting of a line of length L thickened to have a square cross-section. This is $y(\theta) = (\theta_1, r\theta_2, r\theta_3, \ldots)$ with scale $r(\theta_1) = \theta_1/L$, and co-ordinate ranges:

$$\begin{aligned}
0 &\le \theta_1 \le L \\
0 &\le \theta_\mu \le 1, \quad \mu = 2, 3, \ldots, d.
\end{aligned}$$

Fixing noise $\sigma = 1$, this has one relevant dimension, length L, and $d-1$ irrelevant dimensions whose lengths are always ≤ 1. The FIM is then

$$g(\theta) = \begin{bmatrix}
1 + \sum_{\mu=2}^{d} r^2 & \frac{\theta_1\theta_2}{L^2} & \frac{\theta_1\theta_3}{L^2} & \cdots \\
\frac{\theta_1\theta_2}{L^2} & r^2 & 0 & \\
\frac{\theta_1\theta_3}{L^2} & 0 & r^2 & \\
\vdots & & & \ddots
\end{bmatrix}.$$

Thus the volume element is

$$\sqrt{\det g(\theta)} = 1\, r(\theta_1)^{(d-1)}.$$

Regarding the first factor as $\sqrt{\det g_{\text{rel}}} + \mathcal{O}(1/L^2)$, it is trivial to integrate the second factor over θ_μ for all $\mu \geq 2$, and this factor $\int_0^1 d\theta_2 \cdots \int_0^1 d\theta_d \sqrt{\det g_{\text{irrel}}} = (r(\theta_1))^{(d-1)}$ is the irrelevant co-volume. The effective Jeffreys prior along the one relevant dimension is thus

$$p_J(\theta_1) \propto (\theta_1)^{(d-1)}$$

which clearly has much more weight at large θ_1, at the thick end of the cone.

Now observe some x, giving $p(\theta_1|x) \propto e^{-(x-\theta_1)^2/2}(\theta_1)^{(d-1)}$. With a few lines of algebra, we can derive, assuming $1 \ll x \ll L$, that the posterior deviation (9) is

$$\Delta = x - \langle \theta_1 \rangle_{p(\theta_1|x)} = \frac{d-1}{x} + \mathcal{O}\left(\frac{1}{x^3}\right).$$

Choosing $L = 50$ and $d = 26$ to roughly match Figure 3, at $x \approx 10$ the deviation is $\Delta \approx 2.5$. This is smaller than what is seen for the exponential decay model, whose geometry is, of course, more complicated. This difference is also detected by bias pressure. The maximum $b(\theta)$ for this cone is about 55 bits, which is close to the $d = 5$ model in Figure 5.

While this example takes all irrelevant dimensions to be of equal Fisher length, it would be more realistic to have a series $L, 1, L^{-1}, L^{-2}, \ldots$ equally spaced on a log scale. This makes no difference to the effective $p_J(\theta_1) \propto (\theta_1)^{(d-1)/2}$ and hence to the posterior deviation Δ. Figure A1 draws instead a cone with a round cross-section, which also makes no difference. It compares this to a shape of constant cross-section, for which $p_J(\theta_1) \propto 1$; hence, there is no such bias.
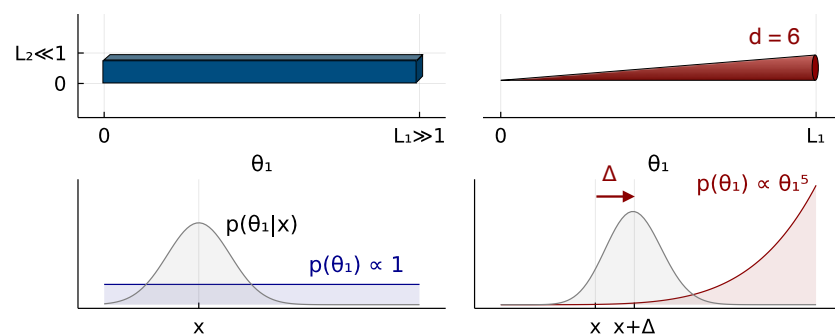

Figure A1. The simplest geometry in which to see measure-induced posterior bias. We compare two $d = 6$ model manifolds, both with θ_1 relevant, Fisher length $L_1 \gg 1$, and five irrelevant dimensions ($L_\mu \ll 1$, $\mu = 2, 3, \ldots 6$) which are either of constant size (**left**) or taper linearly (**right**). The distinction between these two situations is, by assumption, unobservable, but using the d-dimensional notion of volume as a prior gives an effective $p(\theta_1)$ which is either flat or $\propto (\theta_1)^5$. This can induce substantial bias in the posterior $p(\theta_1|x)$.

Appendix A.2. Estimating $I(X; \Theta)$ and Its Gradient

The mutual information can be estimated for a discrete prior

$$p(\theta) = \sum_{a=1}^K \lambda_a \delta(\theta - \theta_a)$$

and a Gaussian likelihood (7) by replacing the integral over x with normally distributed samples:

$$I(X; \Theta) = \sum_a \lambda_a \underbrace{\int dx\, p(x|\theta_a)}_{\text{sample } x \sim \mathcal{N}(\theta_a, \sigma^2)} \log \frac{p(x|\theta_a)}{\sum_b \lambda_b p(x|\theta_b)}.$$

This gives an unbiased estimate, and what we used for plotting $I(X; \Theta)$ in Figures 2 and 5.

However, for finding $p_\star(\theta)$, what we need is the very small gradients of I with respect to each θ_a: near to the optimum, the function is very close to flat. We find that, instead of Monte Carlo, the following kernel density approximation works well:

$$
\begin{aligned}
S(X) &= -\sum_a \lambda_a \int dx'\, p(x'|\theta_a) \log \sum_b \lambda_b\, p(x'|\theta_b) \\
&= -\sum_a \lambda_a \int dx'\, p(x'|\theta_a) \Big[\log \sum_b \lambda_b\, p(x = \theta_a|\theta_b) \\
&\qquad\qquad\qquad + \mathcal{O}(x' - \theta_a)^2 \Big] \\
&\approx -\sum_a \lambda_a \log \sum_b \lambda_b\, e^{[y(\theta_a) - y(\theta_b)]^2/2\sigma'^2} + \text{const.}
\end{aligned}
\tag{A1}
$$

Here we Taylor expand the $\log p(x)$ about $x = \theta_a$ for each atom [43]. For the purpose of finding $p_\star(\theta)$ we may ignore the constant and the conditional entropy in $I(X;\Theta) = S(X) - S(X|\Theta)$. Further, this is a better estimate used at $\sigma' = \sqrt{2}\sigma$.

Before maximizing $I(X;\Theta)$ using L-BFGS [44,45] to adjust all θ_a and λ_a together, we find it useful to sample initial points using Mitchell's best-candidate algorithm [46].

Appendix A.3. Other Methods

Our focus here is on the properties of the optimal prior $p_\star(\theta)$, but better ways to find nearly optimal solutions may be needed in order to use these ideas on larger problems. Some ideas have been explored in the literature:

- The famous algorithm for finding $p_\star(\theta)$ is due to Blahut and Arimoto [47,48], but it needs a discrete Θ, which limits d. This was adapted to use MCMC sampling instead by [49], although their work appears to need discrete X instead. Perhaps it can be generalized.
- More recently, the following lower bound for $I(X;\Theta)$ was used by [41] to find approximations to $p_\star(\theta)$:

$$
I(X;\Theta) \geq I_{\text{NS}} = -S(X|\Theta) - \int dx\, p(x) \log \max_\theta p(x|\theta).
$$

This bound is too crude to see the features of interest here: For all models in this paper, it favors a prior $p_2(\theta) = \text{argmax}_{p(\theta)} I_{\text{NS}}$ with just two delta functions, for any noise level σ.
- We mentioned above that adjusting some "meta-parameters" of some distribution $p_{\bar{\theta},\bar{\sigma}}(\theta)$ would be one way to handle near-optimal priors. This is the approach of [41], and of many papers maximizing other scores, often described as "variational".
- Another prior that typically has large $I(X;\Theta)$ was introduced in [6] under the name "adaptive slab-and-spike prior". It pulls every point x in a distribution

$$
p_{\text{NML}}(x) = \frac{\max_{\hat{\theta}} p(x|\hat{\theta})}{Z}, \quad Z = \int dx\, \max_{\hat{\theta}} p(x|\hat{\theta})
$$

back to its maximum likelihood point $\hat{\theta}$:

$$
p_{\text{proj}}(\theta) = \int dx\, p_{\text{NML}}(x)\, \delta\big(\theta - \text{argmax}_{\hat{\theta}} p(x|\hat{\theta})\big).
$$

The result has weight everywhere in the model manifold, but has extra weight on the edges. Because the amount of weight on edges is controlled by σ, it adopts an appropriate effective dimensionality (8), and has a low bias (5).

Appendix A.4. Bias Alla Bennett

The KL divergence integral needed for $b(\theta)$ is somewhat badly behaved when the prior's weight is far from point θ. To describe how we handle this, we begin with the naïve Monte Carlo calculation of $p(x) = \int d\theta\, p(\theta)\, p(x|\theta)$, which involves sampling from the prior. If x is very far from where the prior has most of its weight, then we will never obtain any samples where $p(x|\theta)$ is not exponentially small, so we will miss the leading contribution to $p(x)$.

Sampling from the posterior $p(\theta|x) \propto p(\theta)\, p(x|\theta)$ instead, we will obtain points in the right area, but the wrong answer. The following identity due to Bennett [26] lets us walk from the prior to the posterior and obtain the right answer, sampling from distributions $\propto p(\theta)\, p(x|\theta)^\alpha$ for several powers α. Defining $\Delta_\alpha^\delta(x)$:

$$0 = \alpha_0 < \alpha_1 < \ldots < \alpha_n = 1$$

$$e^{\Delta_\alpha^\delta(x)} = \left\langle p(x|\theta)^\delta \right\rangle_{x,\alpha} = \frac{\int d\theta\, p(\theta)\, p(x|\theta)^\alpha\, p(x|\theta)^\delta}{\int d\theta\, p(\theta)\, p(x|\theta)^\alpha}$$

the result is

$$\log p(x) = \sum_{i=0}^{n-1} \Delta_{\alpha_i}^{(\alpha_{i+1}-\alpha_i)}(x) = -\sum_{i=1}^{n} \Delta_{\alpha_i}^{(\alpha_{i-1}-\alpha_i)}(x)$$

i.e.,

$$p(x) = \prod_{\alpha \neq 1} \left\langle p(x|\theta)^{\alpha_{\text{next}}-\alpha} \right\rangle_{x,\alpha} = 1 \Big/ \prod_{\alpha \neq 0} \left\langle 1/p(x|\theta)^{\alpha - \alpha_{\text{prev}}} \right\rangle_{x,\alpha}.$$

For $n = 1$ the first is trivial, and the second reads $1/p(x) = \langle 1/p(x|\theta)\rangle_{\theta \sim p(\theta|x)}$.

Next, we want $D_{\text{KL}}[p(x|\varphi)\|p(x)]$ which involves $-\int dx\, p(x|\varphi)\, \log p(x)$. To plug this in, we would need to average $\langle \ldots \rangle_{x,\alpha}$ at every x in the integral. Rather than sample from a fresh set of α-distributions for each x, we can take one set at some x_0, and correct using importance sampling to write:

$$\langle \ldots \rangle_{x,\alpha} = \left\langle \ldots \frac{p(x|\theta)}{p(x_0|\theta)} \right\rangle_{x_0,\alpha}$$

All of this can be performed without knowing the normalization of the prior, Z in (2).

The same set of samples from α-distributions can be used to calculate either forward or backward. These give upper and lower bounds on the true value. When they are sufficiently far apart to be visible, the plots show both of them. Figure 2 uses, as the α-distributions, all larger σ-values on the plot, and also shows (dotted line) the result of naïve sampling from $p_J(\theta)$. Figure 5 uses about 30 steps.

Appendix A.5. Jeffreys and Vandermonde

To find Jeffreys prior, let us parameterize the exponential decay model (6) by $\phi_\mu = e^{-\exp(\theta_\mu)}$ which lives in the unit interval:

$$y_t(\theta) = \sum_{\mu=1}^{d} \frac{e^{-k_\mu t}}{d} = \sum_{\mu=1}^{d} \frac{(\phi_\mu)^t}{d}$$

where $\theta_\mu = \log k_\mu \in \mathbb{R}$, $0 \leq \phi_\mu \leq 1$. The Fisher information metric (1) reads $g_{\mu\nu} = \frac{1}{\sigma^2} \sum_t^m J_{\mu t} J_{\nu t}$ in terms of a Jacobian which, in these co-ordinates, takes the simple form:

$$J_{\mu t}(\phi) = \frac{\partial y_t}{\partial \phi_\mu} = \frac{1}{d} t\, \phi^{t-1}.$$

In high dimensionality, $g_{\mu\nu}$ is very badly conditioned, and thus it is difficult to find the determinant with sufficient numerical accuracy (or at least, it is slow, as high-precision

numbers are required). However, in the case $d = m$, where the Jacobian is a square matrix, it is of Vandermonde form. Hence, its determinant is known exactly, and we can simply write:

$$p_J(\phi) = \sqrt{\det g_{\mu\nu}(\phi)} = \left|\det J_{\mu t}(\phi)\right|$$
$$= \prod_{1 \leq \mu < \nu \leq d} |\phi_\mu - \phi_\nu| \prod_{t=1}^{m} \frac{t}{d}.$$

For $d \neq m$, more complicated formulae (involving a sum over Schur polynomials) are known, but the points in Figure 5 are chosen not to need them.

To sample from Jeffreys prior, or posterior, we use the affine-invariant "emcee" sampler [50]. This adapts well to badly conditioned geometries. Because the Vandermonde formula lets us work in machine precision, we can sample 10^6 points in a few minutes, which is sufficient for Figure 3.

In the enzyme kinetics models of Figure 6, finding $y_t(\theta)$ involves solving a differential equation. The gradient $\partial y_t / \partial \theta_\mu$ is needed both for Jeffreys density and for maximizing $S(X)$ via the above KDE Formula (A1). This can be handled efficiently by passing dual numbers through the solver [51].

Appendix A.6. Michaelis–Menten et al.

The arrows in (11) summarise the following differential equations for the concentrations of the four chemicals involved:

$$\partial_t[E] = -k_f[E][S] + k_r[ES] + k_p[ES]$$
$$\partial_t[S] = -k_f[E][S] + k_r[ES]$$
$$\partial_t[ES] = +k_f[E][S] - k_r[ES] - k_p[ES]$$
$$\partial_t[P] = k_p[ES].$$

These equations conserve $E_0 = [E] + [ES]$ (the enzyme is recycled) and $S_0 = [S] + [ES] + [P]$ (the substrate is converted to product), leaving two dynamical quantities, $[S]$ and $[P]$. The plot takes them to have initial values $[S] = 1$, $[P] = 0$ at $t = 0$, and we fix $E_0 = 1/4$, $S_0 = 1$.

The original analysis of Michaelis and Menten [52] takes the first two reactions to be in equilibrium. This can be viewed as taking the limit $k_r, k_f \to \infty$ holding fixed $K_D = k_r/k_f$, which picks a 2-parameter subspace of Θ, an edge of the manifold. Then, $[ES] = [E][S]/K_D$ becomes constant, leaving their equation

$$\partial_t[P] = \frac{k_p E_0[S]}{K_D + [S]}. \tag{A2}$$

If we do not observe $[E]$, then this is almost identical to the quasi-static limit of Briggs and Haldane [53], who take $k_r, E_0 \to 0$, and $k_f, k_p \to \infty$ holding fixed $K_M = k_p/f_f$ and $V_{\max} = k_p E_0$, which gives

$$\partial_t[P] = \frac{V_{\max}[S]}{K_M + [S]}$$

which was much later shown to be analytically tractable [54].

In Figure 6, most of the points of weight in the optimal prior lie on the intersection of these two 2-parameter models, that is, on a pair of one-parameter models. These and other limits were discussed geometrically in [28].

For the $d = 8$ ping-pong model (12), we take initial conditions of $[A] = [B] = 1$, $[E] = [E*] = 0.5$, $[EA] = [E * P] = [E * B] = [EQ] = 0.1$, $[P] = [Q] = 0$.

Appendix A.7. Ever since the Big Bang

The claim that the age of the universe constrains us from taking the asymptotic limit in realistic multi-parameter models deserves a brief calculation. The conventional age is 13.8×10^8 years, about 4.4×10^{17} s [55].

For the model (6), manifold widths are shown to scale such as $L \propto \sqrt{\mathrm{eig}\, g}$ in [23], and [5,23] shows FIM eigenvalues $\propto 10^{-d}$. Repeating an experiment M times scales lengths by $\sqrt{M}$, so we need roughly 10^d repetitions to make the smallest manifold width larger than 1. If the initial experiment took 1 s, then it is impossible to perform enough repetitions to make all dimensions of the $d = 26$ model relevant.

Other models studied in [5,23] have different slopes $\log(\mathrm{eig}\, g)$ vs. d, so the precise cutoff will vary. However, models with hundreds, or thousands, of parameters are also routine. It seems safe to claim that for most of these, the asymptotic limit is never justified.

Appendix A.8. Terminology

What we call Jeffreys prior, $p_J(\theta)$, is (apart from variations in apostrophes and articles) sometimes called "Jeffreys's nonlocation rule" [56] in order to stress that it is not quite what Jeffreys favored. He argued for separating "location" parameters (such as an unconstrained position $\phi \in \mathbb{R}$) and excluding them from the determinant. The models we consider here have no such perfect symmetries.

What we call the optimal prior $p_\star(\theta)$ is sometimes called "Shannon optimal" [17], and sometimes called a "reference prior" after Bernardo [8]. The latter is misleading, as definition 1 in [8] explicitly takes the asymptotic limit $M \to \infty$. This is then Jeffreys prior, until some ways to handle nuisance parameters are appended. The idea of considering $\mathrm{argmax}_{p(\theta)}\, I(X; \Theta)$ in the limit is older, for instance Lindley [7] considers it, and also notes that it leads to Jeffreys prior (which he does not like for multi-parameter models, but not for our reasons). In [15], the discrete prior for k repetitions is called the "k-th reference prior", and is understood to be discrete, but is of interest only as a tool for showing that the limit exists. We stress that the asymptotic limit removes what are (for this paper) the interesting features of this prior.

The same $p_\star(\theta)$ can also be obtained by an equivalent minimax game, for which Kashyap uses the term "optimal prior" [10]. However, he, too, takes the asymptotic limit.

References

1. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
2. Amari, S.I. A foundation of information geometry. *Electron. Commun. Jpn.* **1983**, *66*, 1–10. [CrossRef]
3. Brown, K.S.; Sethna, J.P. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E* **2003**, *68*, 021904. [CrossRef]
4. Daniels, B.C.; Chen, Y.J.; Sethna, J.P.; Gutenkunst, R.N.; Myers, C.R. Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotechnol.* **2008**, *19*, 389–395. [CrossRef]
5. Machta, B.B.; Chachra, R.; Transtrum, M.K.; Sethna, J.P. Parameter space compression underlies emergent theories and predictive models. *Science* **2013**, *342*, 604–607. [CrossRef]
6. Quinn, K.N.; Abbott, M.C.; Transtrum, M.K.; Machta, B.B.; Sethna, J.P. Information geometry for multiparameter models: New perspectives on the origin of simplicity. *Rep. Prog. Phys.* **2023**, *86*, 035901. [CrossRef]
7. Lindley, D.V. The use of prior probability distributions in statistical inference and decisions. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 30 June–30 July 1961.
8. Bernardo, J.M. Reference posterior distributions for Bayesian inference. *J. Roy. Stat. Soc. B* **1979**, *41*, 113–128. [CrossRef]
9. Mattingly, H.H.; Transtrum, M.K.; Abbott, M.C.; Machta, B.B. Maximizing the information learned from finite data selects a simple model. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1760–1765. [CrossRef]
10. Kashyap, R. Prior probability and uncertainty. *IEEE Trans. Inform. Theory* **1971**, *17*, 641–650. [CrossRef]
11. Haussler, D. A general minimax result for relative entropy. *IEEE Trans. Inform. Theory* **1997**, *43*, 1276–1280. [CrossRef]
12. Krob, J.; Scholl, H.R. A minimax result for the Kullback Leibler Bayes risk. *Econ. Qual. Control* **1997**, *12*, 147–157.
13. Färber, G. Die Kanalkapazität allgemeiner Übertragunskanäle bei begrenztem Signalwertbereich beliebigen Signalübertragungszeiten sowie beliebiger Störung. *Arch. Elektr. Übertr.* **1967**, *21*, 565–574.
14. Smith, J.G. The information capacity of amplitude-and variance-constrained scalar gaussian channels. *Inf. Control.* **1971**, *18*, 203–219. [CrossRef]

15. Berger, J.O.; Bernardo, J.M.; Mendoza, M. On priors that maximize expected information. In *Recent Developments in Statistics and Their Applications*; Seoul Freedom Academy Publishing: Seoul, Republic of Korea, 1989.

16. Zhang, Z. Discrete Noninformative Priors. Ph.D. Thesis, Yale University, New Haven, CT, USA, 1994.

17. Scholl, H.R. Shannon optimal priors on independent identically distributed statistical experiments converge weakly to Jeffreys' prior. *Test* **1998**, *7*, 75–94. [CrossRef]

18. Sims, C.A. Rational inattention: Beyond the linear-quadratic case. *Am. Econ. Rev.* **2006**, *96*, 158–163. [CrossRef]

19. Connes, A. *Noncommutative Geometry*; Academic Press: San Diego, CA, USA, 1994.

20. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A* **1946**, *186*, 453–461. [CrossRef]

21. Clarke, B.S.; Barron, A.R. Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stat. Plan. Inference* **1994**, *41*, 37–60. [CrossRef]

22. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **1997**, *9*, 349–368. [CrossRef]

23. Transtrum, M.K.; Machta, B.B.; Sethna, J.P. Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.* **2010**, *104*, 060201. [CrossRef]

24. Clarke, B.; Barron, A. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **1990**, *36*, 453–471. [CrossRef]

25. Abbott, M.C.; Machta, B.B. A scaling law from discrete to continuous solutions of channel capacity problems in the low-noise limit. *J. Stat. Phys.* **2019**, *176*, 214–227. [CrossRef]

26. Bennett, C.H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys* **1976**, *22*, 245–268. [CrossRef]

27. Hines, K.E.; Middendorf, T.R.; Aldrich, R.W. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *J. Gen. Physiol.* **2014**, *143*, 401–416. [CrossRef] [PubMed]

28. Transtrum, M.K.; Qiu, P. Bridging mechanistic and phenomenological models of complex biological systems. *PLoS Comput. Biol.* **2016**, *12*, e1004915. [CrossRef]

29. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **1974**, *19*, 716–723. [CrossRef]

30. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]

31. Grünwald, P.; Roos, T. Minimum description length revisited. *Int. J. Math. Ind.* **2019**, *11*, 1930001. [CrossRef]

32. Schwarz, G. Estimating the dimension of a model. *Ann. Statist.* **1978**, *6*, 461–464. [CrossRef]

33. Myung, I.J.; Balasubramanian, V.; Pitt, M.A. Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11170–11175. [CrossRef]

34. Piasini, E.; Balasubramanian, V.; Gold, J.I. Effect of geometric complexity on intuitive model selection. In *Machine Learning, Optimization, and Data Science*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–24. [CrossRef]

35. O'Leary, T.; Williams, A.H.; Caplan, J.S.; Marder, E. Correlations in ion channel expression emerge from homeostatic tuning rules. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E2645–E2654. [CrossRef]

36. Wen, M.; Shirodkar, S.N.; Plecháč, P.; Kaxiras, E.; Elliott, R.S.; Tadmor, E.B. A force-matching Stillinger-Weber potential for MoS_2: Parameterization and Fisher information theory based sensitivity analysis. *J. Appl. Phys.* **2017**, *122*, 244301. [CrossRef]

37. Marschmann, G.L.; Pagel, H.; Kügler, P.; Streck, T. Equifinality, sloppiness, and emergent structures of mechanistic soil biogeochemical models. *Environ. Model. Softw.* **2019**, *122*, 104518. [CrossRef]

38. Karakida, R.; Akaho, S.; Amari, S.i. Pathological spectra of the Fisher information metric and its variants in deep neural networks. *Neural Comput.* **2021**, *33*, 2274–2307. [CrossRef] [PubMed]

39. Kadanoff, L.P. Scaling laws for Ising models near T_c. *Physics* **1966**, *2*, 263–272. [CrossRef]

40. Wilson, K.G. Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Phys. Rev.* **1971**, *B4*, 3174–3183. [CrossRef]

41. Nalisnick, E.; Smyth, P. Learning approximately objective priors. *arXiv* **2017**, arXiv:1704.01168.

42. Bezanson, J.; Karpinski, S.; Shah, V.B.; Edelman, A. Julia: A fast dynamic language for technical computing. *arXiv* **2012**, arXiv:1209.5145.

43. Huber, M.F.; Bailey, T.; Durrant-Whyte, H.; Hanebeck, U.D. On entropy approximation for Gaussian mixture random vectors. In Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, Seoul, Republic of Korea, 20–22 August 2008; pp. 181–188. [CrossRef]

44. Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comp.* **1980**, *35*, 773–782. [CrossRef]

45. Johnson, S.G. The NLopt Nonlinear-Optimization Package. Available online: http://github.com/stevengj/nlopt (accessed on 6 May 2022).

46. Mitchell, D.P. Spectrally optimal sampling for distribution ray tracing. *SIGGRAPH Comput. Graph.* **1991**, *25*, 157–164. [CrossRef]

47. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory* **1972**, *18*, 460–473. [CrossRef]

48. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inform. Theory* **1972**, *18*, 14–20. [CrossRef]

49. Lafferty, J.; Wasserman, L.A. Iterative Markov chain Monte Carlo computation of reference priors and minimax risk. *Uncertain. AI* **2001**, *17*, 293–300.

50. Goodman, J.; Weare, J. Ensemble samplers with affine invariance. *CAMCoS* **2010**, *5*, 65–80. [CrossRef]

51. Ma, Y.; Dixit, V.; Innes, M.; Guo, X.; Rackauckas, C. A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. *arXiv* **2021**, arXiv:1812.01892.
52. Michaelis, L.; Menten, M.L. Die Kinetik der Invertinwirkung. *Biochem. Z* **1913**, *49*, 333–369.
53. Briggs, G.E.; Haldane, J.B.S. A note on the kinetics of enzyme action. *Biochem. J.* **1925**, *19*, 338–339. [CrossRef]
54. Schnell, S.; Mendoza, C. Closed Form Solution for Time-dependent Enzyme Kinetics. *J. Theor. Biol.* **1997**, *187*, 207–212. [CrossRef]
55. Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **2021**, *641*, A6. [CrossRef]
56. Kass, R.E.; Wasserman, L. The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.* **1996**, *91*, 1343–1370. [CrossRef]