Maximizing the information learned from finite data selects a simple model

Henry H. Mattingly[a,b,1], Mark K. Transtrum[c], Michael C. Abbott[d,2], and Benjamin B. Machta[b,e,2,3]

[a]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; [b]Lewis-Sigler Institute, Princeton University, Princeton, NJ 08544; [c]Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602; [d]Marian Smoluchowski Institute of Physics, Jagiellonian University, 30-348 Kraków, Poland; and [e]Department of Physics, Princeton University, Princeton, NJ 08544

Edited by Larry Wasserman, Carnegie Mellon University, Pittsburgh, PA, and approved January 9, 2018 (received for review September 1, 2017)

We use the language of uninformative Bayesian prior choice to study the selection of appropriately simple effective models. We advocate for the prior which maximizes the mutual information between parameters and predictions, learning as much as possible from limited data. When many parameters are poorly constrained by the available data, we find that this prior puts weight only on boundaries of the parameter space. Thus, it selects a lower-dimensional effective theory in a principled way, ignoring irrelevant parameter directions. In the limit where there are sufficient data to tightly constrain any number of parameters, this reduces to the Jeffreys prior. However, we argue that this limit is pathological when applied to the hyperribbon parameter manifolds generic in science, because it leads to dramatic dependence on effects invisible to experiment.

effective theory | model selection | renormalization group | Bayesian prior choice | information theory

Physicists prefer simple models not because nature is simple, but because most of its complication is usually irrelevant. Our most rigorous understanding of this idea comes from the Wilsonian renormalization group (1–3), which describes mathematically the process of zooming out and losing sight of microscopic details. These details influence the effective theory which describes macroscopic observables only through a few relevant parameter combinations, such as the critical temperature or the proton mass. The remaining irrelevant parameters can be ignored, as they are neither constrained by past data nor useful for predictions. Such models can now be understood as part of a large class called sloppy models (4–14), whose usefulness relies on a similar compression of a large microscopic parameter space down to just a few relevant directions.

This justification for model simplicity is different from the one more often discussed in statistics, motivated by the desire to avoid overfitting (15–21). Since irrelevant parameters have an almost invisible effect on predicted data, they cannot be excluded on these grounds. Here we motivate their exclusion differently: We show that simplifying a model can often allow it to extract more information from a limited dataset and that this offers a guide for choosing appropriate effective theories.

We phrase the question of model selection as part of the choice of a Bayesian prior on some high-dimensional parameter space. In a set of nested models, we can always move to a simpler model by using a prior which is nonzero only on some subspace. Recent work has suggested that interpretable effective models are typically obtained by taking some parameters to their limiting values, often 0 or ∞, thus restricting to lower-dimensional boundaries of the parameter manifold (22).

Our setup is that we wish to learn about a theory by performing some experiment which produces data $x \in X$. The theory and the experiment are together described by a probability distribution $p(x|\theta)$, for each value of the theory's parameters $\theta \in \Theta$. This function encodes both the quality and the quantity of data to be collected.

The mutual information (MI) between the parameters and their expected data is defined as $\mathrm{MI} = I(X;\Theta) = S(\Theta) - S(\Theta|X)$, where S is the Shannon entropy (23). The MI thus quantifies the information which can be learned about the parameters by measuring the data, or equivalently, the information about the data which can be encoded in the parameters (24, 25). Defining $p_\star(\theta)$ by maximizing this, we see the following:

i) The prior $p_\star(\theta)$ is almost always discrete (26–30), with weight only on a finite number K of points or atoms (Figs. 1 and 2): $p_\star(\theta) = \sum_{a=1}^{K} \lambda_a \delta(\theta - \theta_a)$.
ii) When data are abundant, $p_\star(\theta)$ approaches the Jeffreys prior $p_J(\theta)$ (31–33). As this continuum limit is approached, the proper spacing of the atoms shrinks as a power law (Fig. 3).
iii) When data are scarce, most atoms lie on boundaries of the parameter space, corresponding to effective models with fewer parameters (Fig. 4). The resulting distribution of weight along relevant directions is much more even than that given by the Jeffreys prior (Fig. 5).

After some preliminaries, we demonstrate these properties in three simple examples, each a stylized version of a realistic experiment. To see the origin of discreteness, we study the bias of an unfair coin and the value of a single variable corrupted with Gaussian noise. To see how models of lower dimension arise, we then study the problem of inferring decay rates in a sum of exponentials.

Significance

Most physical theories are effective theories, descriptions at the scale visible to our experiments which ignore microscopic details. Seeking general ways to motivate such theories, we find an information theory perspective: If we select the model which can learn as much information as possible from the data, then we are naturally led to a simpler model, by a path independent of concerns about overfitting. This is encoded as a Bayesian prior which is nonzero only on a subspace of the original parameter space. We differ from earlier prior selection work by not considering an infinite quantity of data. Having finite data is always a limit on the resolution of an experiment, and in our framework this selects how complicated a theory is appropriate.

Author contributions: H.H.M., M.K.T., M.C.A., and B.B.M. designed research; H.H.M., M.K.T., M.C.A., and B.B.M. performed research; H.H.M., M.K.T., M.C.A., and B.B.M. analyzed data; M.C.A. and B.B.M. wrote the paper; H.H.M. and M.C.A. performed the numerical experiments; and H.H.M. and M.K.T. contributed to writing.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Present address: Department of Molecular Cellular and Developmental Biology, Yale University, New Haven, CT 06520.

[2]To whom correspondence may be addressed. Email: abbott@th.if.uj.edu.pl or benjamin.machta@yale.edu.

[3]Present address: Department of Physics, Yale University, New Haven, CT, 06520; and Systems Biology Institute, Yale University, West Haven, CT, 06516.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1715306115/-/DCSupplemental.

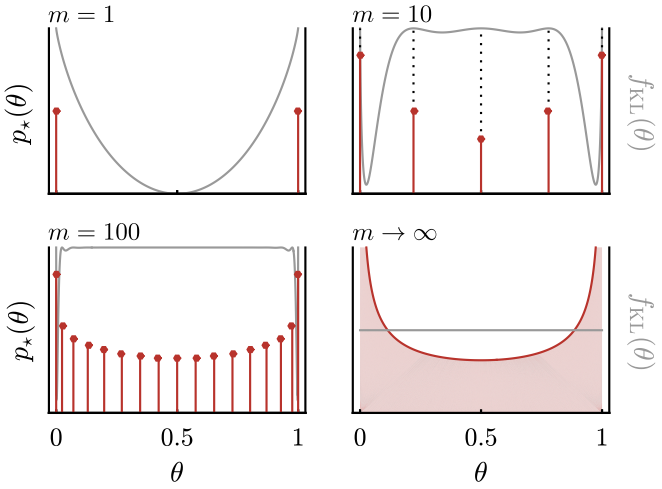

Fig. 1. Optimal priors for the Bernoulli model (Eq. **1**). Red lines indicate the positions of delta functions in $p_\star(\theta)$, which are at the maxima of $f_{\mathrm{KL}}(\theta)$, Eq. **3**. As $m \to \infty$ these coalesce into the Jeffreys prior $p_{\mathrm{J}}(\theta)$.

In *Supporting Information*, we discuss the algorithms used for finding $p_\star(\theta)$, and we apply some more traditional model selection tools to the sum of exponentials example.

Priors and Geometry

Bayes' theorem tells us how to update our knowledge of θ upon observing data x, from prior $p(\theta)$ to posterior $p(\theta|x) = p(x|\theta)\,p(\theta)/p(x)$, where $p(x) = \int d\theta\, p(\theta)\, p(x|\theta)$. In the absence of better knowledge we must pick an uninformative prior which codifies our ignorance. The naive choice of a flat prior $p(\theta) = \text{const.}$ has undesirable features, in particular making $p(x)$ depend on the choice of parameterization, through the measure $d\theta$.

The Jeffreys prior $p_{\mathrm{J}}(\theta)$ is invariant under changes of parameterization because it is constructed from some properties of the experiment (34). This $p_{\mathrm{J}}(\theta) \propto \sqrt{\det g_{\mu\nu}}$ is, up to normalization, the volume form arising from the Fisher information metric or matrix (FIM):

$$g_{\mu\nu}(\vec{\theta}) = \int dx\, p(x|\vec{\theta})\, \frac{\partial \log p(x|\vec{\theta})}{\partial \theta^\mu}\, \frac{\partial \log p(x|\vec{\theta})}{\partial \theta^\nu}.$$

This Riemannian metric defines a reparameterization-invariant distance between points, $ds^2 = \sum_{\mu,\nu=1}^{D} g_{\mu\nu}\, d\theta^\mu\, d\theta^\nu$. It measures the distinguishability of the data which θ and $\theta + d\theta$ are expected to produce, in units of standard deviations. Repeating an (identical and independently distributed) experiment m times means considering $p^m(\vec{x}|\theta) = \prod_{j=1}^{m} p(x_j|\theta)$, which leads to metric $g^m_{\mu\nu}(\theta) = m\, g_{\mu\nu}(\theta)$. However, the factor $m^{D/2}$ in the volume is lost by normalizing $p_{\mathrm{J}}(\theta)$. Thus, the Jeffreys prior depends on the type of experiment, but not the quantity of data.

Bernardo defined a prior $p_\star(\theta)$ by maximizing the MI between parameters Θ and the expected data X^m from m repetitions and then a reference prior by taking the limit $m \to \infty$ (29, 31). Under certain benign assumptions, this reference prior is exactly the Jeffreys prior (31–33), providing an alternative justification for $p_{\mathrm{J}}(\theta)$.

We differ in taking seriously that the amount of data collected is always finite.* Besides being physically unrealistic, the limit $m \to \infty$ is pathological both for model selection and for prior choice. In this limit any number of parameters can be perfectly inferred, justifying an arbitrarily complicated model. In addition, in this limit the posterior $p(\theta|x)$ becomes independent of any smooth prior.†

Geometrically, the defining feature of sloppy models is that they have a parameter manifold with hyperribbon structure (6–9): There are some long directions (corresponding to d relevant, or stiff, parameters) and many shorter directions ($D - d$ irrelevant, or sloppy, parameter combinations). These lengths are often estimated using the eigenvalues of $g_{\mu\nu}$ and have logarithms that are roughly evenly spaced over many orders of magnitude (4, 5). The effect of coarse graining is to shrink irrelevant directions (here using the technical meaning of irrelevant: a parameter which shrinks under renormalization group flow) while leaving relevant directions extended, producing a sloppy manifold (8, 14). By contrast, the limit $m \to \infty$ has the effect of expanding all directions, thus erasing the distinction between directions longer and shorter than the critical length scale of (approximately) 1 SD.

On such a hyperribbon, the Jeffreys prior has an undesirable feature: Since it is constructed from the D-dimensional notion of volume, its weight along the relevant directions always depends on the volume of the $D - d$ irrelevant directions. This gives it extreme dependence on which irrelevant parameters are included in the model.‡ The optimal prior $p_\star(\theta)$ avoids this dependence because it is almost always discrete, at finite m.§ It puts weight on a set of nearly distinguishable points, closely spaced along the relevant directions, but ignoring the irrelevant ones. Yet being the solution to a reparameterization-invariant optimization problem, the prior $p_\star(\theta)$ retains this good feature of $p_{\mathrm{J}}(\theta)$.

Maximizing the MI was originally done to calculate the capacity of a communication channel, and we can borrow techniques from rate-distortion theory here: The algorithms we use were developed there (37, 38), and the discreteness we exploit was discovered several times in engineering (26–28, 39). In statistics, this problem is more often discussed as an equivalent minimax problem (40). Discreteness was also observed in other minimax problems (41–43) and later in directly maximizing MI (29, 30, 33, 44). However, it does not seem to have been seen as useful, and none of these papers explicitly find discrete priors in dimension $D > 1$, which is where we see attractive properties. Discreteness has been useful, although for different reasons, in the idea of rational inattention in economics (45, 46). There, market actors have a finite bandwidth for news, and this drives them to make discrete choices despite all the dynamics being continuous. Rate-distortion theory has also been useful in several areas of biology (47–49), and discreteness emerges in a recent theoretical model of the immune system (50).

We view this procedure of constructing the optimal prior as a form of model selection, picking out the subspace of Θ on which $p_\star(\theta)$ has support. This depends on the likelihood function $p(x|\theta)$ and the data space X, but not on the observed data x. In this regard it is closer to Jeffreys' perspective on prior selection than to tools like the information criteria and Bayes factors,

*Interned for 5 y, John Kerrich flipped his coin only 10^4 times (35). With computers we can do better, but even the Large Hadron Collider generated only about 10^{18} bits of data (36).

†For simplicity we consider only regular models; i.e., we assume all parameters are structurally identifiable.

‡See Fig. 5 for a demonstration of this point. For another example, consider a parameter manifold Θ which is a cone, with Fisher metric $ds^2 = (50\, d\vartheta)^2 + \vartheta^2 d\Omega_n^2/4$: There is one relevant direction $\vartheta \in [0, 1]$ of length $L = 50$, and there are n irrelevant directions forming a sphere of diameter ϑ. Then the prior on ϑ alone implied by $p_{\mathrm{J}}(\vec{\theta})$ is $p(\vartheta) = (n+1)\vartheta^n$, putting most of the weight near $\vartheta = 1$, dramatically so if $n = D - d$ is large. But since only the relevant direction is visible to our experiment, the region $\vartheta \approx 0$ ought to be treated similarly to $\vartheta \approx 1$. The prior $p_\star(\vec{\theta})$ has this property.

§We offer both numerical and analytic arguments for discreteness below. The exception to discreteness is that if there is an exact continuous symmetry, $p_\star(\theta)$ will be constant along it. For example, if our Gaussian model Eq. **2** is placed on a circle (identifying both $\theta \sim \theta + 1$ and $x \sim x + 1$), then the optimum prior is a constant.

which are used at the stage of fitting to data. We discuss this difference at length in *Model Selection from Data*.

One-Parameter Examples

We begin with some problems with a single bounded parameter, of length L in the Fisher metric. These tractable cases illustrate the generic behavior along either short (irrelevant) or long (relevant, $L \gg 1$) parameter directions in higher-dimensional examples.

Our first example is the Bernoulli problem, in which we wish to determine the probability $\theta \in [0, 1]$ that an unfair coin gives heads, using the data from m trials. It is sufficient to record the total number of heads x, which occurs with probability

$$p(x|\theta) = \frac{m!}{x!(m-x)!}\,\theta^x(1-\theta)^{m-x}. \qquad [1]$$

This gives $g_{\theta\theta} = \frac{m}{\theta(1-\theta)}$, thus $p_{\mathrm{J}}(\theta) = [\pi\sqrt{\theta(1-\theta)}]^{-1}$, and proper parameter space length $L = \int \sqrt{ds^2} = \pi\sqrt{m}$.

In the extreme case $m = 1$, the optimal prior is two delta functions, $p_\star(\theta) = \frac{1}{2}\delta(\theta) + \frac{1}{2}\delta(\theta - 1)$, and $\mathrm{MI} = \log 2$, exactly one bit (29, 30, 33). Before an experiment that will run only once, this places equal weight on both outcomes; afterward it records the outcome. As m increases, weight is moved from the boundary onto interior points, which increase in number and ultimately approach the smooth $p_{\mathrm{J}}(\theta)$ (Figs. 1 and 3*A*).

Similar behavior is seen in a second example, in which we measure one real number x, normally distributed with known σ about the parameter $\theta \in [0, 1]$:

$$p(x|\theta) = \tfrac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\theta)^2/2\sigma^2}. \qquad [2]$$

Repeated measurements are equivalent to smaller σ (by $\sigma \to \sigma/\sqrt{m}$), so we fix $m = 1$ here. The Fisher metric is $g_{\theta\theta} = 1/\sigma^2$, and thus $L = 1/\sigma$. An optimal prior is shown in Fig. 2; in Fig. 5*A* it is shown along with its implied distribution of expected data. This $p(x)$ is similar to that implied by the Jeffreys prior, here $p_{\mathrm{J}}(\theta) = 1$.

We calculated $p_\star(\theta)$ numerically in two ways. After discretizing both θ and x, we can use the Blahut–Arimoto (BA) algorithm (37, 38). This converges to the global maximum, which is a discrete distribution (Fig. 2). Alternatively, using our knowledge that $p_\star(\theta)$ is discrete, we can instead adjust the positions θ_a and weights λ_a of a finite number of atoms. See *Algorithms* for more details.

To see analytically why discreteness arises, we write the MI as

$$\mathrm{MI} = I(X;\Theta) = \int d\theta\, p(\theta)\, f_{\mathrm{KL}}(\theta), \qquad [3]$$

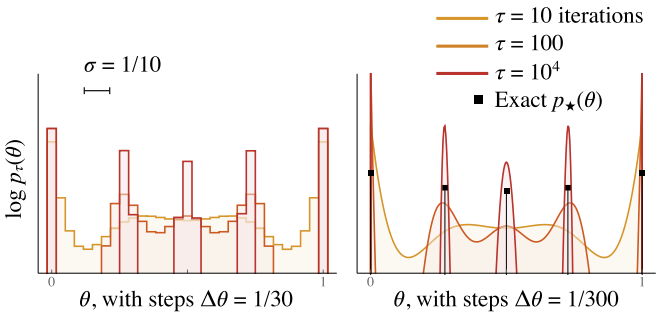

Fig. 2. Convergence of the BA algorithm. This is for the one-parameter Gaussian model Eq. **2** with $L = 10$ (comparable to $m = 10$ in Fig. 1). *Right* shows θ discretized into 10 times as many points, but $p_\tau(\theta)$ clearly converges to the same 5 delta functions.

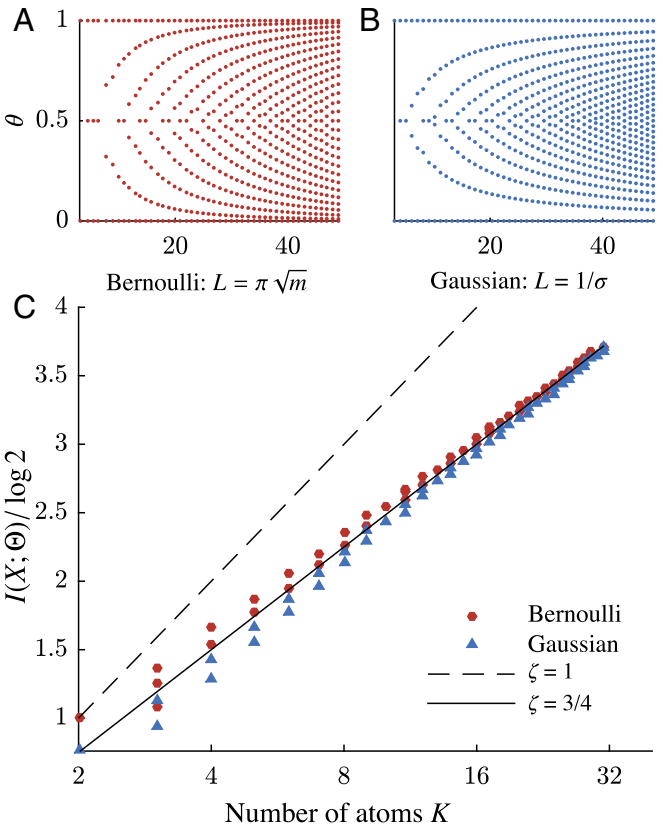

Fig. 3. Behavior of $p_\star(\theta)$ with increasing Fisher length. *A* and *B* show the atoms of $p_\star(\theta)$ for the two one-dimensional models as L is increased (i.e., we perform more repetitions m or have smaller noise σ). *C* shows the scaling of the MI (in bits) with the number of atoms K. The dashed line is the bound $\mathrm{MI} \le \log K$, and the solid line is the scaling law $\mathrm{MI} \sim 3/4 \log K$.

$$f_{\mathrm{KL}}(\theta) = D_{\mathrm{KL}}\left[p(x|\theta)\|p(x)\right] = \int dx\, p(x|\theta)\, \log \frac{p(x|\theta)}{p(x)},$$

where D_{KL} is the Kullback–Leibler divergence.[¶] Maximizing MI over all functions $p(\theta)$ with $\int d\theta\, p(\theta) = 1$ gives $f_{\mathrm{KL}}(\theta) = \text{const}$. But the maximizing function will not, in general, obey $p(\theta) \ge 0$. Subject to this inequality $p_\star(\theta)$ must satisfy

$$\{p_\star(\theta) > 0, f_{\mathrm{KL}}(\theta) = \mathrm{MI}\} \text{ or } \{p_\star(\theta) = 0, f_{\mathrm{KL}}(\theta) < \mathrm{MI}\}$$

at every θ. With finite data $f_{\mathrm{KL}}(\theta) - \mathrm{MI}$ must be an analytic function of θ and therefore must be smooth with a finite number of zeros, corresponding to the atoms of $p_\star(\theta)$ (Fig. 1). See refs. 28, 29, and 46 for related arguments for discreteness and refs. 41–43 for other approaches.

The number of atoms occurring in $p_\star(\theta)$ increases as the data improve. For K atoms there is an absolute bound $\mathrm{MI} \le \log K$, saturated if they are perfectly distinguishable. In Fig. 3*C* we

[¶]The function $f_{\mathrm{KL}}(\theta)$ is sometimes called the Bayes risk, as it quantifies how poorly the prior will perform if θ turns out to be correct. One of the problems equivalent to maximizing the MI (40) is the minimax problem for this (Fig. 1):

$$\max_{p(\theta)} I(X;\Theta) = \min_{p(\theta)} \max_{\theta} f_{\mathrm{KL}}(\theta) = \min_{q(x)} \max_{p(\theta)} \int d\theta\, p(\theta)\, D_{\mathrm{KL}}\left[p(x|\theta)\|q(x)\right].$$

The distributions we call expected data $p(x)$ are also known as Bayes strategies, i.e., distributions on X which are the convolution of the likelihood $p(x|\theta)$ with some prior $p(\theta)$. The optimal $q(x)$ from this third formulation (with $\min_{q(x)} \ldots$) can be shown to be such a distribution (40).

observe that the optimal priors instead approach a line $\mathrm{MI} \to \zeta \log K$, with slope $\zeta \approx 0.75$. At large L the length of parameter space is proportional to the number of distinguishable points, and hence $\mathrm{MI} \to \log L$. Together these imply $K \sim L^{1/\zeta}$, and so the average number density of atoms grows as

$$\rho_0 = K/L \sim L^{1/\zeta - 1} \approx L^{1/3}, \qquad L \gg 1. \qquad [4]$$

Thus, the proper spacing between atoms shrinks to zero in the limit of infinite data; i.e., neighboring atoms cease to be distinguishable.

To derive this scaling law analytically, in a related paper (51) we consider a field theory for the number density of atoms, in which the entropy density (omitting numerical factors) is $\mathcal{S} = \text{const.} - e^{-\rho^2}[\rho^4(\rho')^2 + 1]$. From this we find $\zeta = 3/4$, which is consistent with both examples presented above.

Multiparameter Example

In the examples above, $p_\star(\theta)$ concentrates weight on the edges of its allowed domain when data are scarce (i.e., when m is small or σ is large, and hence L is small). We next turn to a multiparameter model in which some parameter combinations are ill-constrained and where edges correspond to reduced models.

The physical picture is that we wish to determine the composition of an unknown radioactive source, from data of x_t Geiger counter clicks at some times t. As parameters we have the quantities A_μ and decay constants k_μ of isotopes μ. The probability of observing x_t should be a Poisson distribution (of mean y_t) at each time, but we approximate these by Gaussians of fixed σ to write[#]

$$p(\vec{x}|\vec{y}) \propto \prod_t e^{-(x_t - y_t)^2/2\sigma^2}, \qquad y_t = \sum_\mu A_\mu e^{-k_\mu t}. \qquad [5]$$

We can see the essential behavior with just two isotopes in fixed quantities: $A_\mu = \frac{1}{2}$, and thus $y_t = \frac{1}{2}(e^{-k_1 t} + e^{-k_2 t})$. Measuring at only two times t_1 and t_2, we almost have a 2D version of Eq. **2**, in which the center of the distribution $\vec{y} = (y_1, y_2)$ plays the role of θ above. The mapping between $\vec{k}$ and $\vec{y}$ is shown in Fig. 4*A*, fixing $t_2/t_1 = e$. The FIM is proportional to the ordinary Euclidean metric for $\vec{y}$, but not for $\vec{k}$:

$$g_{\mu\nu}(\vec{k}) = \frac{1}{\sigma^2}\sum_t \frac{\partial y_t}{\partial k_\mu}\frac{\partial y_t}{\partial k_\nu} \quad \Longleftrightarrow \quad g_{st}(\vec{y}) = \frac{1}{\sigma^2}\delta_{st}. \qquad [6]$$

Thus, the Jeffreys prior is a constant on the allowed region of the $\vec{y}$ plane.

Then we proceed to find the optimum $p_\star(\vec{y})$ for this model, shown in Fig. 4*B* for various values of σ. When σ is large, this has delta functions only in two of the corners, allowing only $k_1, k_2 = 0$ and $k_1, k_2 = \infty$. As σ is decreased, new atoms appear first along the lower boundary (corresponding to the one-dimensional model where $k_1 = k_2$) and then along the other boundaries. At sufficiently small σ, atoms start filling in the (2D) interior.

To show this progression in Fig. 4*C*, we define Ω_r as the total weight on all edges of dimension r and an effective dimensionality $d_{\text{eff}} = \sum_{r=1}^{D} r\,\Omega_r$. This increases smoothly from 0 toward $D = 2$ as the data improve.

At medium values of σ, the prior $p_\star(\vec{y})$ almost ignores the width of the parameter manifold and cares mostly about its length ($L_+ = \sqrt{2}/\sigma$ along the diagonal). This behavior is very different from that of the Jeffreys prior: In Fig. 5*B* we demonstrate

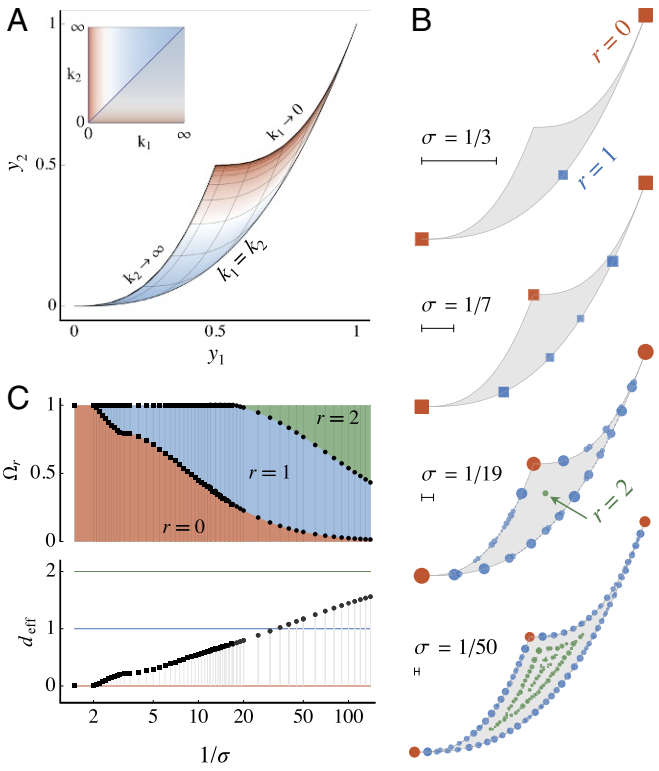

Fig. 4. Parameters and priors for the exponential model (Eq. 5). *A* shows the area of the $\vec{y}$ plane covered by all decay constants $k_1, k_2 \geq 0$. *B* shows the positions of the delta functions of the optimal prior $p_\star(\vec{y})$ for several values of σ, with colors indicating the dimensionality r at each point. *C* shows the proportion of weight on these dimensionalities.

this by plotting the distributions of data implied by these two priors. Jeffreys puts almost no weight near the ends of the long (i.e., stiff or relevant) parameter's range because the (sloppy or irrelevant) width happens to be even narrower there than in the middle. By contrast, our effective model puts significant weight on each end, much like the one-parameter model in Fig. 5*A*.

The difference between one and two parameters being relevant (in Fig. 4*B*) is very roughly $\sigma = 1/7$ to $\sigma = 1/50$, a factor 7 in Fisher length and thus a factor 50 in the number of repetitions m—perhaps the difference between a week's data and a year's. These numbers are artificially small to demonstrate the appearance of models away from the boundary: More realistic models often have manifold lengths spread over many orders of magnitude (5, 8) and thus have some parameters inaccessible even with centuries of data. To measure these we need a qualitatively different experiment, justifying a different effective theory.

The one-dimensional model along the lower edge of Fig. 4*A* is the effective theory with equal decay constants. This remains true if we allow more parameters $k_3, k_4, \ldots$ in Eq. **5**, and $p_\star(\vec{y})$ will still place a similar weight there.[‖] Measuring x_t also at later times $t_3, t_4, \ldots$ will add more thin directions to the manifold (7), but the one-dimensional boundary corresponding to equal decay constants will still have significant weight. The fact that such edges give human-readable simpler models (unlike arbitrary submanifolds) was the original motivation for preferring them in ref.

[#] Using a normal distribution of fixed σ here is what allows the metric in Eq. **6** to be so simple. However, the qualitative behavior from the Poisson distribution is very similar.

[‖] If we have more parameters than measurements, then the model must be singular. In fact the exponential model of Fig. 4 is already slightly singular, since $k_1 \leftrightarrow k_2$ does not change the data; we could cure this by restricting to $k_2 \geq k_1$, or by working with $\vec{y}$, to obtain a regular model.

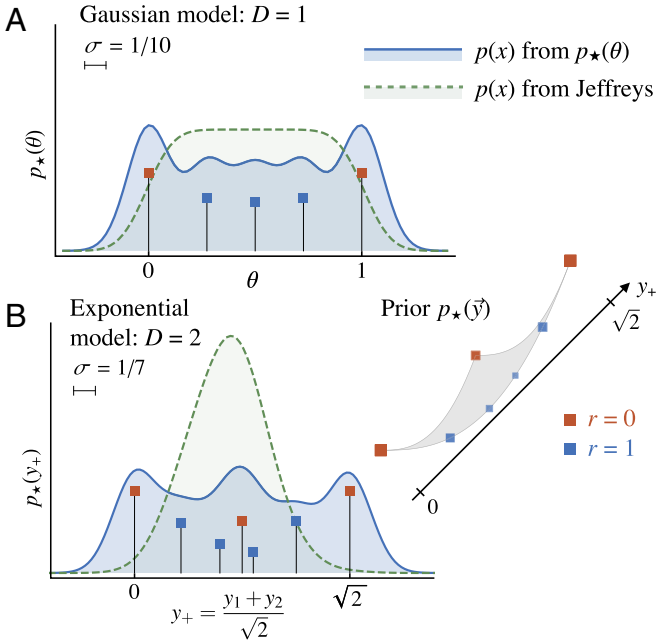

Fig. 5. Distributions of expected data $p(x)$ from different priors. *A* is the one-parameter Gaussian model, with $L = 10$. *B* projects the two-parameter exponential model onto the $y_1 + y_2$ direction, for $\sigma = 1/7$ where the perpendicular direction should be irrelevant. The length of the relevant direction is about the same as the one-parameter case: $L_+ = 7\sqrt{2}$. Note that the distribution of expected data $p(x_+)$ from the Jeffreys prior here is quite different, with almost no weight at the ends of the range (0 and $\sqrt{2}$), because this prior still weights the area and not the length.

22, and it is very interesting that our optimization procedure has the same preference.[**]

Discussion

While the three examples we have studied here are very simple, they demonstrate a principled way of selecting optimal effective theories, especially in high-dimensional settings. Following ref. 45, we may call this rational ignorance.

The prior $p_\star(\theta)$ which encodes this selection is the maximally uninformative prior, in the sense of leaving maximum headroom for learning from data. But its construction depends on the likelihood function $p(x|\theta)$, and thus it contains knowledge about the experiment through which we are probing nature. The Jeffreys prior $p_J(\theta)$ also depends on the experiment, but more weakly: It is independent of the number of repetitions m, precisely because it is the limit $m \to \infty$ of the optimal prior (32, 33).

Under either of these prescriptions, performing a second experiment may necessitate a change in the prior, leading to a change in the posterior not described by Bayes' theorem. If the second experiment is different from the first one, then changing to the Jeffreys prior for the combined experiment (and then applying Bayes' rule just once) will have this effect (55, 56).[††] Our prescription differs from that of Jeffreys in also regarding more repetitions of an identical experiment as being different. Many experiments would have much higher resolution if they could be repeated for all eternity. The fact that they cannot is an important limit on the accuracy of our knowledge, and our proposal treats this limitation on the same footing as the rest of the specification of the experiment.

Keeping m finite is where we differ from earlier work on prior selection. Bernardo's reference prior (31) maximizes the same MI, but always in the $m \to \infty$ limit where it gives a smooth analytically tractable function. Using $I(X;\Theta)$ to quantify what can be learned from an experiment goes back to Lindley (24). That finite information implies a discrete distribution was known at least since refs. 26 and 27. What has been overlooked is that this discreteness is useful for avoiding a problem with the Jeffreys prior on the hyperribbon parameter spaces generic in science (5): Because it weights the irrelevant parameter volume, the Jeffreys prior has strong dependence on microscopic effects invisible to experiment. The limit $m \to \infty$ has erased the divide between relevant and irrelevant parameters, by throwing away the natural length scale on the parameter manifold. By contrast, $p_\star(\theta)$ retains discreteness at roughly this scale, allowing it to ignore irrelevant directions. Along a relevant parameter direction this discreteness is no worse than rounding θ to as many digits as we can hope to measure, and we showed that in fact the spacing of atoms decreases faster than our accuracy improves.

Model selection is more often studied not as part of prior selection, but at the stage of fitting the parameters to data. From noisy data, one is tempted to fit a model which is more complicated than reality; avoiding such overfitting improves predictions. The Akaike information criterion (AIC), Bayesian information criterion (BIC) (15, 61), and related criteria (19, 20, 62–64) are subleading terms of various measures in the $m \to \infty$ limit, in which all (nonsingular) parameters of the true model can be accurately measured. Techniques like minimum description length (MDL), normalized maximum likelihood (NML), and cross-validation (62, 65, 66) need not take this limit, but all are applied after seeing the data. They favor minimally flexible models close to the data seen, while our procedure favors one answer which can distinguish as many different outcomes as possible. It is curious that both approaches can point toward simplicity. We explore this contrast in more detail in *Model Selection from Data*.[‡‡]

Being discrete, the prior $p_\star(\theta)$ is very likely to exclude the true value of the parameter, if such a $\theta_{\text{true}} \in \Theta$ exists. This is not a flaw: The spirit of effective theory is to focus on what is relevant for describing the data, deliberately ignoring microscopic effects which we know to exist (67). Thus, the same effective theory can emerge from different microscopic physics [as in the universality of critical points describing phase transitions (68)]. The relevant degrees of freedom are often quasiparticles [such as the Cooper pairs of superconductivity (69)] which do not exist in the microscopic theory, but give a natural and simple description at the scale being observed. We argued here for such simplicity not on the grounds of the difficulty of simulating 10^{23} electrons or of human limitations, but based on the natural measure of information learned.

There is similar simplicity to be found outside of physics. For example, the Michaelis–Menten law for enzyme kinetics (70) is derived as a limit in which only the ratios of some reaction rates matter and is useful regardless of the underlying system. In more complicated systems which we cannot solve by hand and, for which the symmetries and scaling arguments used in physics

[**] Edges of the parameter manifold give simpler models not only in the sense of having fewer parameters, but also in an algorithmic sense. For example, the Michaelis–Menten model is analytically solvable (52) in a limit which corresponds to a manifold boundary (53). Stable linear dynamical systems of order n are model boundaries of order $n+1$ systems (54). Taking some parameter combinations to the extreme can lock spins into Kadanoff blocks (53).

[††] This view is natural in the objective Bayesian tradition, but see refs. 57–60 for alternatives.

[‡‡] Model selection usually starts from a list of models to be compared, in our language a list of submanifolds of Θ. We can also consider maximizing mutual information in this setting, rather than with an unconstrained function $p(\theta)$, and unsurprisingly we observe a similar preference for highly flexible simpler models. This is also discussed in Eq. S3.

cannot be applied, we hope that our information approach may be useful for identifying the appropriately detailed theory.

ACKNOWLEDGMENTS. We thank Vijay Balasubramanian, William Bialek, Robert de Mello Koch, Peter Grünwald, Jon Machta, James Sethna, Paul Wiggins, and Ned Wingreen for discussion and comments. We thank International Centre for Theoretical Sciences Bangalore for hospitality. H.H.M. was supported by NIH Grant R01GM107103. M.K.T. was supported by National Science Foundation (NSF)-Energy, Power, and Control Networks 1710727. B.B.M. was supported by a Lewis-Sigler Fellowship and by NSF Division of Physics 0957573. M.C.A. was supported by Narodowe Centrum Nauki Grant 2012/06/A/ST2/00396.

1. Kadanoff LP (1966) Scaling laws for Ising models near T_c. *Physics* 2:263–272.
2. Wilson KG (1971) Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Phys Rev* B4:3174–3183.
3. Cardy JL (1996) *Scaling and Renormalization in Statistical Physics* (Cambridge Univ Press, Cambridge, UK).
4. Waterfall JJ, et al. (2006) Sloppy-model universality class and the Vandermonde matrix. *Phys Rev Lett* 97:150601–150604.
5. Gutenkunst RN, et al. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3:1871–1878.
6. Transtrum MK, Machta BB, Sethna JP (2010) Why are nonlinear fits to data so challenging? *Phys Rev Lett*, 104:060201.
7. Transtrum MK, Machta BB, Sethna JP (2011) Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys Rev E* 83:036701.
8. Machta BB, Chachra R, Transtrum MK, Sethna JP (2013) Parameter space compression underlies emergent theories and predictive models. *Science* 342:604–607.
9. Transtrum MK, et al. (2015) Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys* 143:010901.
10. O'Leary T, Sutton AC, Marder E (2015) Computational models in the age of large datasets. *Curr Opin Neurobiol* 32:87–94.
11. Niksic T, Vretenar D (2016) Sloppy nuclear energy density functionals: Effective model reduction. *Phys Rev C*:94:024333.
12. Dhruva VR, Anderson J, Papachristodoulou A (2017) Delineating parameter unidentifiabilities in complex models. *Phys Rev E* 95:032314.
13. Bohner G, Venkataraman G (2017) Identifiability, reducibility, and adaptability in allosteric macromolecules. *J Gen Physiol* 149:547–560.
14. Raju A, Machta BB, Sethna JP (2017) Information geometry and the renormalization group. *arXiv*:1710.05787.
15. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–772.
16. Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. *Commun Stat Theory Meth* 7:13–26.
17. Balasubramanian V (1997) Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comp* 9:349–368.
18. Myung IJ, Balasubramanian V, Pitt MA (2000) Counting probability distributions: Differential geometry and model selection. *Proc Natl Acad Sci USA* 97:11170–11175.
19. Spiegelhalter DJ, Best NG, Carlin BP, Linde AVD (2002) Bayesian measures of model complexity and fit. *J R Stat Soc B* 64:583–639.
20. Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *JMLR* 11:3571–3594.
21. LaMont CH, Wiggins PA (2017) Information-based inference for singular models and finite sample sizes. arXiv:1506.05855v4.
22. Transtrum MK, Qiu P (2014) Model reduction by manifold boundaries. *Phys Rev Lett* 113:098701.
23. Shannon CE (1948) A mathematical theory of communication. *Bell Sys Tech J* 27: 623–656.
24. Lindley DV (1956) On a measure of the information provided by an experiment. *Ann Math Stat* 27:986–100.
25. Rényi A (1967) On some basic problems of statistics from the point of view of information theory. *Proc 5th Berkeley Symp Math Stat Prob* 1:531–543.
26. Färber G (1967) Die Kanalkapazität allgemeiner Übertragunskanäle bei begrenztem Signalwertbereich beliebigen Signalübertragungszeiten sowie beliebiger Störung. *Arch Elektr Übertr* 21:565–574.
27. Smith JG (1971) The information capacity of amplitude-and variance-constrained scalar Gaussian channels. *Inf Control* 18:203–219.
28. Fix SL (1978) Rate distortion functions for squared error distortion measures. *Proc 16th Annu Allerton Conf Commun Control Comput*, 704–711.
29. Berger JO, Bernardo JM, Mendoza M (1988) On priors that maximize expected information. *Recent Developments in Statistics and Their Applications*, eds Klein J, Lee J (Freedom Academy, Seoul, Korea), pp 1–20.
30. Zhang Z (1994) Discrete noninformative priors. PhD thesis (Yale University, New Haven, CT).
31. Bernardo JM (1979) Reference posterior distributions for Bayesian inference. *J R Stat Soc B* 41:113–147.
32. Bertrand SC, Barron AR (1994) Jeffreys' prior is asymptotically least favorable under entropy risk. *J Stat Plan Infer* 41:37–60.
33. Scholl HR (1998) Shannon optimal priors on independent identically distributed statistical experiments converge weakly to Jeffreys' prior. *Test* 7:75–94.
34. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. *Proc R Soc A* 186:453–461.
35. Kerrich JE (1946) *An Experimental Introduction to the Theory of Probability* (E Munksgaard, Copenhagen).
36. O'Luanaigh C (2013) CERN data centre passes 100 petabytes. home.cern. Available at https://home.cern/about/updates/2013/02/cern-data-centre-passes-100-petabytes. Accessed May 1, 2017.
37. Arimoto S (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans Inf Theory* 18:14–20.
38. Blahut R (1972) Computation of channel capacity and rate-distortion functions. *IEEE Trans Inf Theory* 18:460–473.
39. Rose K (1994) A mapping approach to rate-distortion computation and analysis. *IEEE Trans Inf Theory* 40:1939–1952.
40. Haussler D (1997) A general minimax result for relative entropy. *IEEE Trans Inf Theory* 43:1276–1280.
41. Ghosh MN (1964) Uniform approximation of minimax point estimates. *Ann Math Stat* 35:1031–1047.
42. Casella G, Strawderman WE (1981) Estimating a bounded normal mean. *Ann Stat* 9: 870–878.
43. Feldman I (1991) Constrained minimax estimation of the mean of the normal distribution with known variance. *Ann Stat* 19:2259–2265.
44. Chen M, Dey D, Müller P, Sun D, Ye K (2010) *Frontiers of Statistical Decision Making and Bayesian Analysis* (Springer, New York).
45. Sims CA (2006) Rational inattention: Beyond the linear-quadratic case. *Am Econ Rev* 96:158–163.
46. Jung J, Kim J, Matějka F, Sims CA (2015) Discrete actions in information-constrained decision problems. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.696.267. Accessed May 1, 2017.
47. Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch C* 36:910–912.
48. Tkačik G, Callan CG, Bialek W (2008) Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA* 105:12265–12270.
49. Petkova MD, Tkačik G, Bialek W, Wieschaus EF, Gregor T (2016) Optimal decoding of information from a genetic network. arXiv:1612.08084.
50. Mayer A, Balasubramanian V, Mora T, Walczak AM (2015) How a well-adapted immune system is organized. *Proc Natl Acad Sci USA* 112:5950–5955.
51. Abbott MC, Machta BB (2017) An information scaling law $\zeta = 3/4$. arXiv:1710.09351.
52. Schnell S, Mendoza C (1997) Closed form solution for time-dependent enzyme kinetics. *J Theor Biol* 187:207–212.
53. Transtrum M, Hart G, Qiu P (2014) Information topology identifies emergent model classes. arXiv:1409.6203.
54. Paré PE, Wilson AT, Transtrum MK, Warnick SC (2015) A unified view of balanced truncation and singular perturbation approximations. *2015 American Control Conference*, 10.1109/ACC.2015.7171025.
55. Lewis N (2017) Combining independent Bayesian posteriors into a confidence distribution, with application to estimating climate sensitivity. *J Stat Plan Inference*, 10.1016/j.jspi.2017.09.013.
56. Lewis N (2013) Modification of Bayesian updating where continuous parameters have differing relationships with new and existing data. arXiv:1308.2791.
57. Poole D, Raftery AE (2000) Inference for deterministic simulation models: The Bayesian melding approach. *J Am Stat Assoc* 95:1244–1255.
58. Seidenfeld T (1979) Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. *Theory Decis* 11:413–440.
59. Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. *J Am Stat Assoc* 91:1343–1370.
60. Williamson J (2009) Objective Bayesianism, Bayesian conditionalisation and voluntarism. *Synthese* 178:67–85.
61. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
62. Rissanen J (1978) Modeling by shortest data description. *Automatica* 14:465–471.
63. Wallace CS, Boulton DM (1968) An information measure for classification. *Comput J* 11:185–194.
64. Watanabe S (2013) A widely applicable Bayesian information criterion. *J Mach Learn Res* 14:867–897.
65. Grünwald PD, Myung IJ, Pitt MA (2009) *Advances in Minimum Description Length: Theory and Applications* (MIT Press, Cambridge, MA).
66. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Stat Surv* 4:40–79.
67. Anderson PW (1972) More is different. *Science* 177:393–396.
68. Batterman RW (2017) Philosophical implications of Kadanoff's work on the renormalization group. *J Stat Phys* 167:559–574.
69. Bardeen J, Cooper LN, Schrieffer JR (1957) Theory of superconductivity. *Phys Rev* 108:1175–1204.
70. Michaelis L, Menten ML (2013) The kinetics of invertin action. *FEBS Lett* 587: 2712–2720.