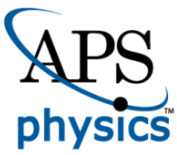

Correspondence between thermodynamics and inference

Colin H. LaMont and Paul A. Wiggins

Phys. Rev. E **99**, 052140 — Published 30 May 2019

DOI: 10.1103/PhysRevE.99.052140

On the correspondence between thermodynamics and inference

Colin H. LaMont and Paul A. Wiggins
*Departments of Physics, Bioengineering and Microbiology, University of Washington, Box 351560. 3910 15th Avenue Northeast, Seattle, WA 98195, USA**

We expand upon a natural analogy between Bayesian statistics and statistical physics in which sample size corresponds to inverse temperature. This analogy motivates the definition of two novel statistical quantities: a learning capacity and a Gibbs entropy. The analysis of the learning capacity, corresponding to the heat capacity in thermal physics, leads to new insight into the mechanism of learning and explains why some models have anomalously-high learning performance. We explore the properties of the learning capacity in a number of examples, including a sloppy model. Next, we propose that the Gibbs entropy provides a natural device for counting distinguishable distributions in the context of Bayesian inference. We use this device to define a *generalized principle of indifference* (GPI) in which every distinguishable model is assigned equal *a priori* probability. This principle results in a new solution to a long-standing problem in Bayesian inference: the definition of an objective or uninformative prior. A key characteristic of this new approach is that it can be applied to analyses where the model dimension is unknown and circumvents the automatic rejection of higher-dimensional models in Bayesian inference.

I. INTRODUCTION

Despite an intensifying interest in applications of machine learning to the analysis of big data, fundamental questions remain about the mechanism of learning and the development of efficient learning algorithms. In this paper, we explore the phenomenology of learning by exploiting a correspondence between Bayesian inference and statistical mechanics. This correspondence has been previously described by Jaynes, Balasubramanian, and many others [1–6] and methods from statistical physics have been adapted to statistical calculations [7–17]. Motivated by the success of this previous work, we propose to exploit the correspondence at a more conceptual level. By using the canonical bridge between statistical mechanics and thermodynamics, we define statistical analogues to the standard thermodynamic potentials and properties of a system. We then explore the statistical properties of these new analogues. The correspondence identifies two novel statistical quantities, a *learning capacity* and the *Gibbs entropy* which give new physical insight into the mechanism of learning and defines a novel Bayesian learning algorithm, respectively.

A. Model complexity

How does model complexity affect learning and why do some models have anomalously good learning performance? These are questions of great topical interest due to the increasing application of machine learning algorithms and the analysis of extremely complex models in the context of systems biology and other fields (*e.g.* [18]). The analysis of a novel *learning capacity* (corresponding to the heat capacity) reveals a natural connection between the equipartition theorem in statistical mechanics [19], which predicts the thermal energy as a function of the *number of degrees of freedom of a physical system*, and the information loss in prediction as a function of the *number of degrees of freedom of a model*. This connection provides physical insight into why there are universal properties of learning systems that are independent of the detailed functional dependence of the likelihood functions or the learning algorithm.

In spite of these universal limits, it has long been known that some high-dimensional models learn anomalously well. These models have been termed *sloppy* [18]. We demonstrate that the learning capacity both provides a natural definition for the sloppiness phenomenon and identifies a mechanism for the anomalously-high predictive performance, a statistical analogue of the well-known freeze-out mechanism of statistical mechanics. We speculate that this mechanism is responsible for the anomalously-high predictive performance of high-dimensional models more generally.

B. Prior selection

The correspondence also provides new insight into another important question: How do you define an objective or uninformative prior? Prior selection has been a subject of debate in Bayesian inference from its inception to the current day [20]. A key application of the correspondence is to translate insights into improved learning algorithms. In this light, we propose that the Gibbs entropy provides a natural device for determining model multiplicity, *i.e.* counting indistinguishable distributions in the context of statistical inference. This interpretation allows us to define a *generalized principle of indifference* (GPI) for selecting a prior in the absence of *a priori* information on the parameters or models. The GPI unifies a number of known, but seemingly unconnected objective Bayesian methods [21], while also pro-

* pwiggins@uw.edu; http://mtshasta.phys.washington.edu/

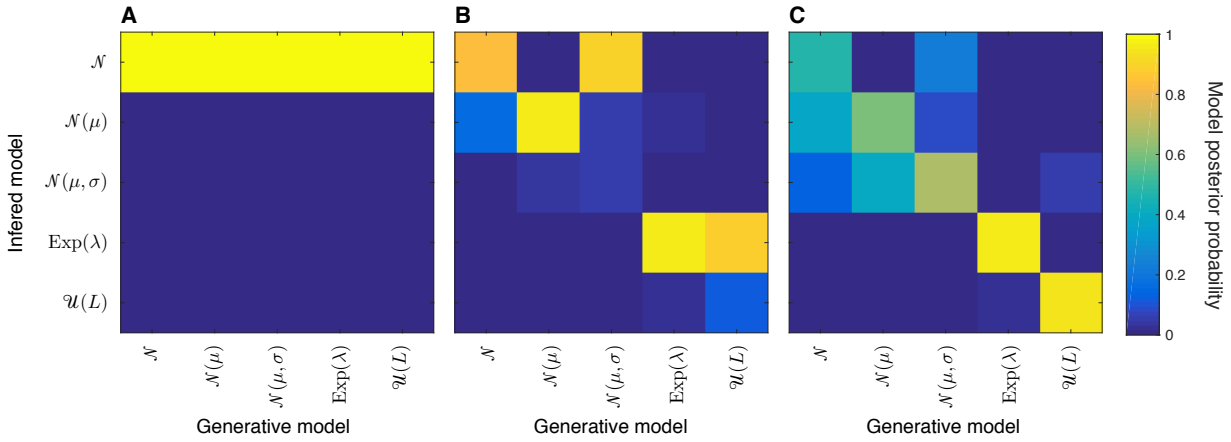

FIG. 1. **Bayesian inference on model identity.** The posterior of model identity (y axis) was computed for datasets generated by each model (x axis). **Panel A: Objective prior.** The non-compactness of the parameter manifolds implies automatic rejection of all the higher-dimensional models. Since the model $\mathcal{N}$ is parameter-free, it has posterior probability of 1 for all datasets, regardless of the fit. **Panel B: Revised objective prior.** To avoid this undesirable anomaly, we tune the prior parameter support to result in a reasonable posterior model probability. (See Tab. IV.) Inference is no longer objective, as the posterior probabilities depend on how this tuning is performed. One representative plot of posterior probabilities to shown. In general, inference cannot be expect to identify the generative model unless the KL divergence is so large as to make the prior irrelevant. **Panel C: GPI prior.** Using the GPI prior, the generative model had the highest posterior probability as expected. When $\mathcal{N}$ is the generative distribution, the data is realizable in both $\mathcal{N}(\mu)$ and $\mathcal{N}(\mu,\sigma)$ as well. The fact that these model have lower posterior probability reveals that there remains a natural mechanism for the selection of parsimonious models using the GPI prior.

viding an algorithm applicable in contexts where existing approaches fail [22].

C. The Lindley-Bartlett Paradox

The most important advantage of the GPI approach over existing approaches is in the context of models with unknown dimension (*i.e.* model selection [21, 23, 24]). Here, the use of uninformative priors can often result in the automatic rejection of higher-dimensional models, a phenomenon called the Lindley-Bartlett paradox [25, 26]. To demonstrate this phenomenon, we analyze simulated datasets where the generative distribution is known and test the performance of inference. As we will demonstrate, the canonical Bayesian approach to inference leads to difficulties in generating a meaningful posterior probability on model identity.

Numerical experiment: We consider five competing models: three realizations of the normal model (known mean and variance, unknown mean and known variance, and unknown mean and variance), the exponential model with unknown rate and the uniform distribution model with unknown end-point. We generate datasets from all fives models. We then perform inference on the model identity for each dataset using an uninformative prior. A detailed definition of the likelihood functions, priors, parameter support and generative model parameters are provided in the Appendix (A 1) and Tab. IV.

If the canonical Bayesian approach is interpreted literally, the posterior probability of the parameter-free normal model is one, regardless of which distribution was used to generate the data! (See Fig. 1A.) This approach ensures that the smallest model is always selected, however much data is accumulated and however poorly this model fits the data. Although this scenario would appear paradoxical, it perfectly logical from a Bayesian perspective: Due to the non-compact manifold, a measure zero fraction of the parameter space is consistent with the data. The relative merits of Bayesian and Frequentist inference in this context were first debated by Lindley and Bartlett in the 1950s [25, 26].

A number of *ad hoc* modifications to this naive Bayesian procedure are possible to avoid the automatic rejection of larger models. These include some formal methods: *e.g.* the use of empirical or variational Bayes, pseudo-Bayes factors, partitioning of the data to train the prior *etc.* Many practical-minded Bayesians will revise their prior beliefs, presumably remembering *a posteri* that they knew *a priori*. We use this approach in Fig. 1B. In this context not only does inference fail to identify the generative distribution, but it is also depends sensitively on *ad hoc* decisions made in the analysis, which we would argue is not acceptable in the context of objective scientific analysis.

How can priors be defined over models to give sensible and robust results? The use of the proposed GPI prior circumvents the paradox by naturally assigning a mutually-consistent weighting, not only between parameter values, but between model families, irrespective of model dimension. (The GPI approach is shown

in Fig. 1C.) Our approach is asymptotically consistent with a powerful but non-Bayesian approach to inference on model identity: the use of the pseudo-Bayes factor [21, 27–30]. Thus the GPI approach provides a novel solution to one of the oldest problems in statistics: the specification of an objective prior.

D. Outline

The paper is organized as follows: In Sec. II A-III B, we define the correspondence and compute the thermodynamic potentials and properties of inference in the analytically tractable large-sample-size limit and use these results to deduce the statistical meaning of each quantity. In Sec. III C, we explore the statistical properties of the learning capacity. In Sec. III D, we use the Gibbs entropy to define a generalized principle of indifference and an objective prior (the GPI prior). In Sec. III E, we compute the GPI prior in a number of examples. In Sec. III F, we discuss how the GPI prior circumvents the Lindley-Bartlett paradox.

II. METHODS

A. Defining the correspondence

We assume that a true parameter value $\boldsymbol{\theta}_0$ is drawn from a known prior distribution $\varpi(\boldsymbol{\theta})$. We observe N samples $x^N \equiv \{x_1, ..., x_N\}$ which are distributed like $q(x|\boldsymbol{\theta}_0)$:

$$X_i \sim q(\cdot|\boldsymbol{\theta}_0), \tag{1}$$

where we use capital X to denote random variables and the symbol $\sim$ to denote *distributed like*. For simplicity, we will assume that the observations are independent and identically distributed, but the approach can be generalized.

The correspondence between statistical physics and Bayesian inference is clearest when expressions are written in terms of the empirical estimator of the cross-entropy:

$$\hat{H}(\boldsymbol{\theta}) \equiv -\left\langle \log q(X|\boldsymbol{\theta}) \right\rangle_{X\sim x^N}, \tag{2}$$

$$\equiv -N^{-1}\sum_{i=1}^{N} \log q(x_i|\boldsymbol{\theta}), \tag{3}$$

where the angle brackets represent an expectation over the variable in the subscript, which in this case is an empirical expectation over the observed data x^N (Eq. 3). The marginal likelihood (*i.e.* evidence) can be written [4]:

$$Z(x^N) \equiv \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\, \varpi(\boldsymbol{\theta})\, e^{-N\hat{H}}, \tag{4}$$

which can be directly compared to the partition function in the canonical ensemble[31] [1–6]. The model parameters $\boldsymbol{\theta}$ correspond to the variables that define the physical state vector, the cross entropy estimator $\hat{H}(\boldsymbol{\theta})$ corresponds to the energy $E(\boldsymbol{\theta})$, the prior $\varpi(\boldsymbol{\theta})$ corresponds to the density of states $\rho(\boldsymbol{\theta})$. The data x^N is quenched disorder[32] in the physical system [4]. The sample size N is identified with the inverse temperature $\beta \equiv (k_B T)^{-1}$ [4]. (Henceforth, we will set $k_B \equiv 1$.) This assignment is natural in the following sense: At small sample size N, many parameter values are consistent with the data, in analogy with the large range of states $\boldsymbol{\theta}$ occupied at high temperature. In contrast, at large sample size N the parameter values consistent with the data are tightly localized around the true value, in analogy to a statistical system at low temperature where only states $\boldsymbol{\theta}$ in very close proximity to the ground state are occupied. We note that choosing $T^{-1} \leftrightarrow N$ is only one of at least two proposals for the identification of the temperature. See the Appendix (A 3).

B. Application of thermodynamic identities

To extend the previously proposed correspondence, we follow the standard prescriptions from statistical mechanics to compute thermodynamic potentials, properties, and variables for the system [19, 33]. These are shown in the lower half of Tab. I. The thermodynamic quantities depend on the particular realization of the data X^N. In the current context we are interested in the expectation over this *quenched disorder* (*i.e.* data). We define the disorder average with an overbar:

$$\overline{f}(N) \equiv \left\langle\, f(X^N, \boldsymbol{\theta}_0)\, \right\rangle_{X,\boldsymbol{\theta}_0}, \tag{5}$$

where $X \sim q(\cdot|\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_0 \sim \varpi$.

III. RESULTS

A. Models

Our immediate interest in the next few sections is not performing inference but rather exploring the properties of the statistical counterparts of well-understood thermodynamic quantities. The similarity between the statistical and thermodynamic quantities suggests that these novel statistical quantities may have an analogous interpretation to their thermodynamic counterparts. We will motivate this hypothesis by comparing the properties of statistical models in the the large-sample-size limit to the properties of a free particle (*i.e.* a gas).

Thermodynamics			**Statistics**	
Quantity:	**Interpretation:**		**Quantity:**	**Interpretation:**
$\beta = T^{-1}$	Inverse temperature	$\leftrightarrow$	N	Sample size
$\boldsymbol{\theta}$	State variables/vector	$\leftrightarrow$	$\boldsymbol{\theta}$	Model parameters
X^N	Quenched disorder	$\leftrightarrow$	X^N	Observations
$E_X(\boldsymbol{\theta})$	State energy	$\leftrightarrow$	$\hat{H}_X(\boldsymbol{\theta})$	Cross entropy estiamtor
E_0	Disorder-averaged ground state energy	$\leftrightarrow$	H_0	Shannon entropy
$\rho(\boldsymbol{\theta})$	Density of states	$\leftrightarrow$	$\varpi(\boldsymbol{\theta})$	Prior
Z	Partition function	$\leftrightarrow$	Z	Evidence
$Z^{-1}\rho\ \exp -\beta E_X$	Normalized Boltzmann weight	$\leftrightarrow$	$\varpi(\boldsymbol{\theta}\|X^N)$	Posterior
$F = -\beta^{-1}\log Z$	Free energy	$\leftrightarrow$	$F = -N^{-1}\log Z$	Minus-log-evidence
$U = \partial_\beta \beta F$	Average energy	$\leftrightarrow$	$U = \partial_N N F$	Minus-log-prediction
$C = -\beta^2\partial_\beta^2 \beta F$	Heat capacity	$\leftrightarrow$	$C = -N^2\partial_N^2 N F$	*Learning capacity*
$S = \beta^2 \partial_\beta F$	Gibbs entropy	$\leftrightarrow$	$S = N^2 \partial_N F$	*Statistical Gibbs entropy*

TABLE I. **Thermodynamic-Bayesian correspondence.** The top half of the table lists the correspondences that can be determined directly from the definition of the marginal likelihood as the partition function. The lower half of the table lists the implied thermodynamic expressions and their existing or proposed statistical interpretation.

1. *Free particle*

In the context of statistical mechanics (and thermodynamics), we will analyze a classical free particles in K dimensions in the canonical ensemble. The particle has internal energy E_0 and a phase-space density of states h^{-K}, where h is the Planck constant. The particle is confined to a K-cube with side-length L. We define a critical temperature T_0 at which the de Broglie wavelength is equal to L:

$$T_0 \equiv \beta_0^{-1} \equiv h^2/2\pi m L^2. \tag{6}$$

The thermodynamics quantities are straightforward to compute and are summarized in Tab. II. The Free energy is:

$$F = E_0 + \tfrac{K}{2\beta}\log\tfrac{\beta}{\beta_0}, \tag{7}$$

which is written in terms of the critical inverse temperature β_0, a parameter which holds all the information about both the density of states as well as the geometry of the system.

2. *Large-sample-size limit in regular and singular models*

In the context of statistics, we will compare and contrast two classes of models: *regular* and *singular*. Models are called *singular* when parameters are *structurally unidentifiable*, defined as the absence of a one-to-one map between the space of candidate distribution functions (q) and the parameter manifold. In other words there exists some parameter $\boldsymbol{\theta}_1$ such that

$$q(x|\boldsymbol{\theta}_1) = q(x|\boldsymbol{\theta}_2) \quad \text{for} \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2. \tag{8}$$

A model is singular when the unidentifiability cannot be removed by re-parameterization [34]. In this case, the Fisher Information Matrix (Eq. A6) contains at least one zero eigenvalue at $\boldsymbol{\theta}_1$ [34]. In contrast, in a *regular model*, all parameters are identifiable, the parameter manifold is continuous, and the Fisher information matrix is therefore positive definite in a suitable parameter coordinate system.

B. Definition of thermodynamic analogues

There are known asymptotic results for the evidence (partition function) in large-sample-size limit of both regular and singular models on continuous parameter manifolds [34]. It is therefore straightforward to compute the thermodynamic quantities in this limit once Z is known. These results are shown in Tab. II. A comparisons between the properties of a free particle and a regular model reveal an identical structure to leading order in N (or β).

1. *Free energy*

The model that maximizes the evidence and therefore minimizes F is selected in the canonical approach to Bayesian model selection [21]. Since, a relation between the partition function and Bayesian evidence has long been discussed [1–6], the definition of F is not particularly novel.

For a regular model, F breaks up into two parts:

$$\overline{F} = H_0 + \tfrac{K}{2N}\log\tfrac{N}{N_0} + \mathcal{O}(N^{-1}), \tag{9}$$

		K-Dimensional Free Particle	K-Dimensional Regular Model	K-Dimensional Singular Model
Free energy	$\overline{F}$	$E_0 + \frac{K}{2\beta}\log\frac{\beta}{\beta_0}$	$H_0 + \frac{K}{2N}\log\frac{N}{N_0} + \mathcal{O}(N^{-1})$	$H_0 + \frac{\gamma}{2N}\log N + \mathcal{O}(N^{-1})$
Average energy	$\overline{U}$	$E_0 + \frac{K}{2\beta}$	$H_0 + \frac{K}{2N} + \mathcal{O}(N^{-2})$	$H_0 + \frac{\gamma}{2N} + \mathcal{O}(N^{-2})$
Heat capacity	$\overline{C}$	$\frac{K}{2}$	$\frac{K}{2} + \mathcal{O}(N^{-1})$	$\frac{\gamma}{2} + \mathcal{O}(N^{-1})$
Gibbs entropy	$\overline{S}$	$\frac{K}{2}\left(1 - \log\frac{\beta}{\beta_0}\right)$	$\frac{K}{2}\left(1 - \log\frac{N}{N_0}\right) + \mathcal{O}(N^{-1})$	$-\frac{\gamma}{2}\log N + \mathcal{O}(N^0)$

TABLE II. **Thermodynamic-Bayesian correspondence.** The thermodynamic quantities of a K-dimensional free particle with ground-state energy E_0 are compared to a K-dimensional regular model. Inspection reveals that the β/N dependence of free particle is identical to a regular model to order N^{-1}. For the singular model, the learning coefficient is $\gamma \leq K$. The special case of $\gamma = K$ is a regular model.

to order N^{-1} where the dependence in the prior is absorbed into a *critical sample-size* N_0. We call N_0 the *critical sample size* because $N = N_0$ corresponds to the sample size at which the thermodynamic properties of inference change, as we will shall discuss shortly (Sec. III C 2). Due to the dependence of F on the critical sample size N_0, the evidence is clearly dependent on the choice of prior, if only logarithmically.

From the perspective of statistical mechanics, a direct comparison between this expression (Eq. 9) and the free energy of a free particle (Eq. 7) allows the reader intuitively understand the motivation for the correspondence defined in Sec. II A: H_0 corresponds to the ground-state energy and the second term in Eq. 9 is an entropic contribution to the free energy.

2. *Average energy*

The thermodynamic prescription for computing the average energy involves a derivative with respect to temperature (Tab. I):

$$U \equiv \partial_N NF. \tag{10}$$

In the context of a discrete temperature, we will formally interpret this derivative using a finite difference definition:

$$\partial_N f(N) \to f(N) - f(N-1). \tag{11}$$

Such finite-difference approximations are already implicit to statistical mechanics, where we take derivatives with respect to many variables which are in fact discrete (*e.g.* particle number, energy, *etc.*). In the context of statistics where there are N independent choices for reducing the sample size by one sample, it is convenient to define the finite difference derivative by averaging over the choices:

$$U(x^N) \equiv -\left\langle \log q(x_i|x^{\neq i})\right\rangle_{i=1..N}, \tag{12}$$

where $q(x_i|x^{\neq i}) \equiv Z(x^N)/Z(x^{\neq i})$ is known as the posterior-predictive distribution [21]. The RHS is a well-known statistical object: the Leave-One-Out-Cross-Validation (LOOCV) estimator of model performance. See the Appendix (A 4). The statistical interpretation of average energy U is therefore the minus-expected-predictive-performance of the model (*e.g.* [35]).

To explore the correspondence to the free particle, we compute the averaged energy for the regular model. U can be written as the sum of two contributions:

$$\overline{U} = H_0 + \tfrac{K}{2N} + \mathcal{O}(N^{-2}), \tag{13}$$

to order N^{-2}. The first term H_0 corresponds to a ground-state energy. The second term corresponds to the thermal energy in a physical system. From a statistical perspective, the term represents the information loss associated with predicting a new observation X using parameters estimated from the training set x^N rather than the true parameter $\boldsymbol{\theta}_0$. This predictive information loss is often called the *Generalization Error*, defined (*e.g.* [35]):

$$\mathrm{GE} \equiv H_0 - \overline{U}. \tag{14}$$

In a regular model, GE is [34]:

$$\mathrm{GE} = -\tfrac{K}{2N} + \mathcal{O}(N^{-2}), \tag{15}$$

which depends only on the model dimension K and sample size N but it is independent of the detailed structure of the model (*i.e.* independent of the likelihood function q).

This universal generalization error has a well-known thermodynamic analogue in the equipartition theorem: *There is a half k_BT of thermal energy per harmonic degree of freedom* [19]. In a statistical context, there is a universal generalization error of $\frac{1}{2N}$ per degree of freedom in the model. This universal property of the generalization error is known in statistics (*e.g.* [34]) and can be interpreted as the mechanism by which the Akaike Information Criterion (AIC) estimates the predictive performance [23, 24]. However, the connection between this result and the equipartition theorem had not yet been described.

3. *Learning capacity*

To study the generalization error associated with learning from a finite-sized sample, it is natural to study the statistical quantity corresponding to the heat capacity. The heat capacity measures the rate of increase in thermal energy with temperature ($\overline{C}$ in Tab. I). The statistical analogue of the heat capacity, a *learning capacity*:

$$C \equiv -N^2 \partial_N U, \tag{16}$$

is a measure of the rate of increase in predictive performance with sample size.

To explore the correspondence to the free particle, we compute the learning capacity for a regular model:

$$\overline{C} = \tfrac{1}{2} K + \mathcal{O}(N^{-1}), \tag{17}$$

as implied by the equipartition theorem. To learn how this analogy generalizes to a generic statistical model, we use the large-sample-size limit asymptotic expression for the Bayesian evidence for a singular model from Ref. 34 to compute the learning capacity. (See Tab. II.) Like the normal model, the learning capacity for a singular model has the equipartition form but with an effective complexity:

$$\overline{C} = \tfrac{1}{2} K_{\text{eff}} + \mathcal{O}(N^{-1}), \tag{18}$$

where $K_{\text{eff}} = \gamma$ is the learning coefficient defined by Watanabe [34]. A regular model is a special case of this expression where $\gamma = K$, the dimension of the parameter manifold. The learning capacity is a novel statistical object defined by the correspondence. We will explore its properties in detail in Sec. III C.

.

4. *The Gibbs entropy*

In physics, the Gibbs entropy generalizes the Boltzmann formula: $S = \log \Omega$ where Ω is the number of accessible states. We propose that the Gibbs entropy has the analogous meaning in the context of Bayesian statistics: The Gibbs entropy is the log-number of models consistent with the data. The correspondence implies that the statistical analogue to the Gibbs entropy is defined:

$$S(x^N) \equiv N(U - F), \tag{19}$$

in analogy to statistical mechanics.

To explore the correspondence to the free particle, we compute the Gibbs entropy for a regular model:

$$\overline{S} = \tfrac{K}{2} \left(1 - \log \tfrac{N}{N_0}\right) + \mathcal{O}(N^0). \tag{20}$$

When the data is informative to the parameter values, the number of models consistent with the data is reduced as the sample size grows. As a result the Gibbs entropy is always negative for a normalized prior and becomes increasingly negative as the sample size N grows.

Like the learning capacity, the Gibbs entropy is a novel statistical object defined by the correspondence which we shall argue provides a natural mechanism for defining an objective prior in which all models are assigned equal *a priori* weight. We will explore its properties in detail in Sec. III D.

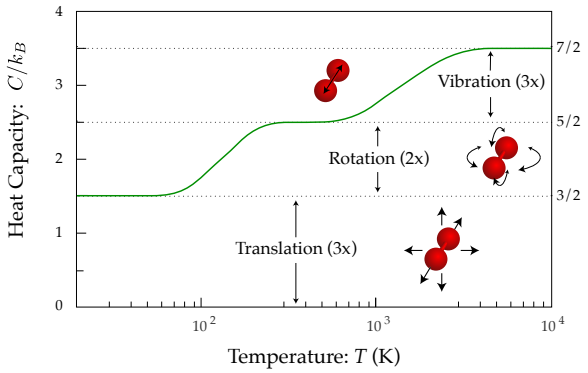

FIG. 2. **The failure of equipartition. Low-temperature freeze-out in a quantum system.** The heat capacity is plotted as a function of temperature. Equipartition predicts that the reduced heat capacity should be constant, equal to half the degrees of freedom in the system. Plateaus can clearly be observed at half-integer values, but the number of degrees of freedom is temperature dependent due to the discrete nature of quantum energy levels. At low temperature, some degrees of freedom are frozen out since the first excited state is thermally inaccessible. This discrete topology of the energy levels implies anharmonicity in the potential and therefore failure of the equipartition theorem.

C. Examples of the Learning Capacity

In this section, we investigate the phenomenology of the learning capacity in a series of simple examples. We shall demonstrate that the learning capacity can show large deviation from the equipartition limit. From a physical perspective, this is no surprise. The failure of the equipartition theorem is a well understood phenomenon in physics where degrees of freedom can become anharmonic at both high and low temperature, altering their contribution to the heat capacity [19]. See Fig. 2. We will analyze statistical models where analogous transitions occur as a function of sample size.

1. *High-temperature freeze out*

It is well known in statistical physics that degrees of freedom can become irrelevant at high temperature. For instance, the position degrees of freedom of a gas do not contribute to the heat capacity [36]. Exactly this

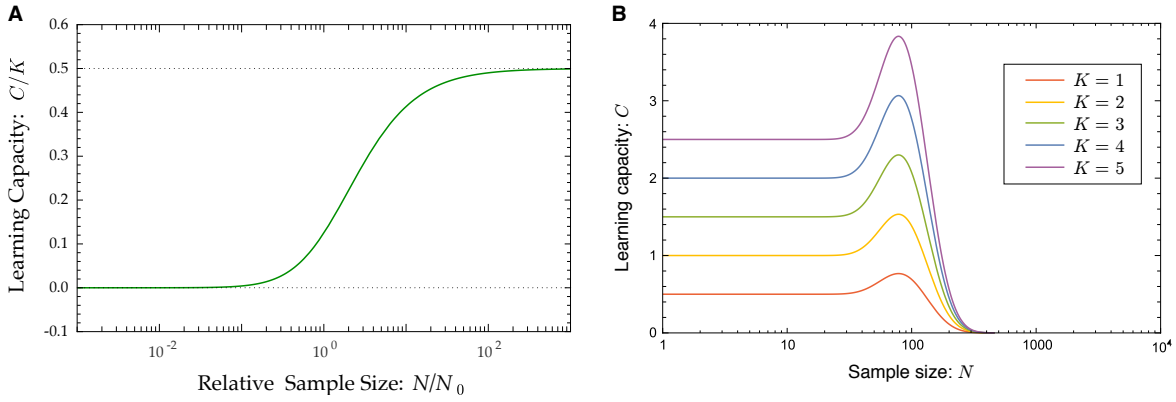

FIG. 3. **The failure of equipartition. Panel A: "High-temperature freeze-out" in the Learning Capacity.** Analogous to the statistical mechanics system, the statistical learning capacity transitions between half integer plateaus, reflecting a temperature-dependent number of degrees of freedom. At low sample size N (high temperature), the parameters are completely specified by model constraints (the prior) and therefore the parameters do not contribute to the learning capacity. At large sample size N, the parameters become data dominated and therefore the learning capacity is predicted by equipartition ($\frac{1}{2}K$). **Panel B: Low-temperature freeze out in the Learning Capacity.** The learning capacity of a normal model with an unknown D-dimensional mean $\vec{\mu} \in \mathbb{Z}^D$ and variance $\sigma^2 = 15$. For statistical uncertainty $\delta\mu \gg 1$, the learning capacity is predicted by equipartition since the discrete nature of the parameter manifold cannot be statistically resolved. For $\delta\mu \ll 1$, there is no statistical uncertainty in the parameter value (due to the discreteness of μ) and the degrees of freedom freeze out, giving a learning capacity of zero.

type of phenomenon occurs in inference as well where a significant fraction of the degrees of freedom of a model can become anharmonic and are not data dominated. We present two simple examples of this fairly general phenomenon.

2. *A normal model with an informative prior*

The canonical model in a statistical context is the normal model (a Gaussian distribution), many of whose properties generalize to more generic models in the large-sample-size limit.

Model: We define a normal model on a D-dimensional observational space with *unknown mean* and *known variance* σ^2. The likelihood function is:

$$q(\vec{x}|\boldsymbol{\theta}) \equiv (2\pi\sigma^2)^{-D/2} \exp[-\tfrac{1}{2\sigma^2}(\vec{x} - \vec{\mu})^2], \qquad (21)$$

with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$. The true parameter $\boldsymbol{\theta}_0$ is drawn from a K-dimensional normal distribution:

$$\varpi(\boldsymbol{\theta}) \equiv (2\pi\sigma_\varpi^2)^{-D/2} \exp[-\tfrac{1}{2\sigma_\varpi^2}(\vec{\mu} - \vec{\mu}_\varpi)^2], \qquad (22)$$

where μ_ϖ and σ_ϖ are assumed to be known hyperparameters. In close analogy to the inverse critical temperature of the free particle, it will be convenient to define a critical sample size:

$$N_0 \equiv \sigma^2/\sigma_\varpi^2, \qquad (23)$$

where N_0 can be understood as the number of previous observations x^{N_0} required to determine the prior in the absence of other information.

Analysis: The learning capacity can be computed analytically:

$$\overline{C} = \tfrac{K}{2(1+N_0/N)^2}, \qquad (24)$$

as shown in the Appendix (C 1 a). At large sample size, the learning capacity is equal to the equipartition expression (Eq. 18). At small sample size (high temperature), the prior determines the parametrization, $\overline{C} \to 0$ and the model appears anomalously predictive. (See Figs. 2 and 3A.)

3. *Exponential mixture models.*

In the previous example, there are fairly transparent constraints applied to the parameters in the form of a prior which reduce the learning capacity. Our next example illustrates another *higher-temperature* freeze-out phenomenon, but here the mechanism is model singularity: A zero mode appears in the Fisher information matrix.

Model: We analyze the exponential mixture model which has previously been identified as *sloppy model* by Transtrum, Machta and coworkers using a criterion defined by the distribution of the eigenvalues of the Fisher information matrix [18]. Consider a model for the lifetime of a mixed population consisting of several different chemical species I with different transition rates. Both the transition rates (k_I) and the relative abundance of the species (p_I) are unknown. For an m species model, the likelihood function for the lifetime t is:

$$q(t|\boldsymbol{\theta}) \equiv \sum_{I=1}^{m} p_I \, k_I \, \mathrm{e}^{-k_I t}, \qquad (25)$$

with parameters:

$$\boldsymbol{\theta} \equiv \begin{pmatrix} p_1 & \dots & p_m \\ k_1 & \dots & k_m \end{pmatrix}, \tag{26}$$

subject to the constraint: $\sum_I p_I = 1$ and we apply improper prior $\varpi(\boldsymbol{\theta}) = 1$. For simplicity, we analyze the smallest model with a singularity ($m = 2$) to facilitate the numerical Bayesian marginalization. The exponential mixture model is singular since parameter k_I is unidentifiable for $p_I = 0$ and p_1 is unidentifiable for $k_1 = k_2$. (See Eq. 8.)

Analysis: We compute the learning capacity at two locations in parameter manifold, at the singularity ($\boldsymbol{\theta}_S$) and far from it ($\boldsymbol{\theta}_R$):

$$\boldsymbol{\theta}_S = \begin{pmatrix} 1 & 0 \\ 1 & 10 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta}_R = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 10 \end{pmatrix}. \tag{27}$$

The learning capacity is computed numerically for $N = 100$ observations with distribution $T^N \sim q(\cdot|\boldsymbol{\theta})$:

$$\overline{C}(\boldsymbol{\theta}) = \begin{cases} 0.61, & \boldsymbol{\theta} = \boldsymbol{\theta}_S \\ 1.5, & \boldsymbol{\theta} = \boldsymbol{\theta}_R \end{cases}. \tag{28}$$

Far from the singularity ($\boldsymbol{\theta}_R$), the equipartition theorem predicts the learning capacity ($\dim/2$) whereas close to the singularity ($\boldsymbol{\theta}_S$), where the model is effectively described by only a single parameter (k_1), the learning capacity reflects this smaller effective model dimension.

4. *Low-temperature freeze out*

The freeze out phenomenon owes its name to the loss of harmonic degrees of freedom at low temperature in physical systems. Here the discrete nature of the quantum energy states plays an essential role in the physics. As the temperature becomes too low to populate the lowest-energy excited state, this degree of freedom no longer contributes to the heat capacity. Again, the analogous phenomenon is found in inference. In this context, the discrete nature of space is typically the result of a discrete parameter.

Model: To explore the low-temperature freeze-out phenomenon, consider a normal model with an unknown discrete mean. The likelihood for the D-dimensional normal model is defined in Eq. 21. As before, the parameters are $\boldsymbol{\theta} = (\vec{\mu})$, but with the mean constrained to have an integer values: $\vec{\mu} \in \mathbb{Z}^D$.

Analysis: The learning capacity can be computed analytically. See the Appendix (C 1 c). The learning capacity is plotted in Figs. 2 and 3B.

To discuss the phenomenology, it is useful to define a frequentist *statistical resolution* with respect to parameter coordinate θ^i:

$$\delta\theta^i(N) \equiv N^{-\frac{1}{2}} \sqrt{[\boldsymbol{I}^{-1}]^{ii}}, \tag{29}$$

in terms of the Fisher information matrix $\boldsymbol{I}$ (Eq. A6), which is a naturally covariant symmetric tensor on the continuous parameter manifold [37]. $\delta\theta^i(N)$ is the width of the posterior in the large-sample-size limit. For the normal model, $\delta\mu = \sigma/\sqrt{N}$. For a regular model with discrete parameters in the large sample size limit, the learning capacity is:

$$\overline{C} = \begin{cases} \frac{1}{2}K, & \Delta\theta^i \ll \delta\theta^i \text{ for all } i \\ 0, & \Delta\theta^i \gg \delta\theta^i \text{ for all } i \end{cases} \tag{30}$$

where $\Delta\theta^i$ is the lattice spacing for parameter coordinate θ^i. The physical interpretation is clear: At large sample size, the system condenses into a single state. Therefore the corresponding degrees of freedom freeze out, and no longer contribute to the learning capacity. At small sample size, the discrete nature of the parameter manifold cannot be resolved, and the parameter manifold is effectively continuous. The learning capacity therefore assumes the equipartition value (provided that the sample size is large enough such that the information is effectively harmonic yet small enough so the discrete nature of the parameter manifold cannot be resolved).

5. *Learning capacity for additional models*

In the Appendix, we provide a series of other examples of learning capacity analyses to further explore its phenomenology. In Appendix Secs. C 1 and C 3, we analyze a range of exactly tractable (but non-Gaussian) models. These analyses reveal that the learning capacity can also be larger than the equipartition limit at finite sample size. In the Appendix (C 2), we work an exactly tractable but non-regular model where the learning capacity is larger than the equipartition value for all sample sizes.

D. The Gibbs entropy and prior selection

In statistical physics, the density of states is known (*i.e.* measured) but in Bayesian inference the selection of a prior is often subjective. The construction of an objective or uninformative prior is a long-standing problem in Bayesian statistics. What insight does the proposed correspondence provide for prior choice?

Prior construction since Bayes and Laplace has often attempted to apply a *Principle of Indifference*: All *mutually exclusive* and *exhaustive* possibilities should be assigned equal prior probability [38, 39]. One interpretation of this prescription is that it maximizes entropy [1, 40]. However, the principle of indifference is difficult to interpret in the context of continuous parameters, or across models of different dimension. For example, are normal models with means μ and $\mu + \mathrm{d}\mu$ mutually exclusive (distinguishable)? Even if the mean

were constrained to be an integer ($\mu \in \mathbb{Z}$), which would define *mutually exclusive*, the exhaustive condition is also problematic. Exhaustive would correspond to a prior with uniform weighting over all integers. This vanishing prior weight ($1/\infty$) on the non-compact set $\mathbb{Z}$ results in a paradoxical value for the evidence $Z \to 0$ and the rejection of the model irrespective of the data, as described in the Lindley-Bartlett Paradox (Sec. I C).

1. A generalized principle of indifference

To define *mutually exclusive* in a statistical context, we look for natural analogues to this problem in statistical physics. A surprising result from the perspective of classical physics is that Nature makes no distinction between states with identical particles exchanged (*e.g.* electrons) and counts only distinguishable states (the Gibbs paradox). Following Balasubramanian [4], we proposed that the concept of indistinguishability must be applied to objective Bayesian inference. We take the *mutually exclusive* criteria in the principle of indifference to refer to distributions which are mutually distinguishable at the experimental resolution available at sample size N. We propose a *generalized principle of indifference*: sets of indistinguishable models are each collectively assigned the weight of a single distinguishable model.

To study the weighting of each model, *we must prepare the data using a different procedure*. We distribute X^N according to an assumed true parameter $\boldsymbol{\theta}$: $X^N \sim q(\cdot|\boldsymbol{\theta})$, omitting the expectation over $\boldsymbol{\theta}$:

$$\overline{f}(\boldsymbol{\theta}, N) \equiv \langle\, f(X^N, \boldsymbol{\theta}) \,\rangle_X \,, \tag{31}$$

where $X \sim q(\cdot|\boldsymbol{\theta})$. A generalized principle of indifference states that the prior ϖ should be chosen such that:

$$\overline{S}(\boldsymbol{\theta}; N, \varpi) \approx \text{const} \quad \forall \quad \boldsymbol{\theta} \in \Theta, \tag{32}$$

at sample size N, where the Gibbs entropy is now a function of $\boldsymbol{\theta}$. Qualitatively, Eq. 32 realizes the condition of equal weighting on *mutually exclusive models* since the Gibbs entropy is understood as the log-number of accessible models and constant entropy implies equal weighting between models.

The correspondence *also* offers a natural mechanism for resolving statistical anomalies arising from the *exhaustive* condition in the principle of indifference. In statistical mechanics, the partition function Z is not a probability by construction since the density of states ρ is a density but not a probability density. Therefore, a natural solution to statistical anomalies arising from the exhaustive condition is to re-interpret the objective inference prior as a *density of models*.

To circumvent the Lindley-Bartlett Paradox, we must specify a consistent density of models between different parameter values and model families. We replace the prior $\varpi(\boldsymbol{\theta})$ with a model density $w(\boldsymbol{\theta})$ such that:

$$\overline{S}(\boldsymbol{\theta}; N, w) \approx 0, \tag{33}$$

assigning unit multiplicity to all parameters $\boldsymbol{\theta}$ and model families I. Eq. 33 is reparametrization invariant. See the Appendix (A 6). The prior w will be improper, but none-the-less the normalization is well defined. We shall refer to Eq. 33 as the *Generalized Principle of Indifference* which realizes both the mutually-exclusive and exhaustive conditions using a principled statistical approach, regardless of the nature of the parameter manifold. We will call the prior w that satisfies Eq. 33 the *GPI prior*.

2. Technical note

Eqs. 32 and 33 cannot be defined as equalities since the condition is typically not exactly realizable for all $\boldsymbol{\theta}$ at finite sample size N. To define the GPI prior precisely, we minimize the largest violation of the GPI condition (Eq. 33), using a mini-max approach analogous to that of Kashyap [41]: We choose the prior ϖ normalization such that

$$\max_{\boldsymbol{\theta}} \overline{S}(\boldsymbol{\theta}; N, \varpi) = 0, \tag{34}$$

and then the GPI prior is the prior maximizes the minimum Gibbs entropy:

$$w = \arg\max_{\varpi} \min_{\boldsymbol{\theta}} \overline{S}(\boldsymbol{\theta}; N, \varpi). \tag{35}$$

Qualitatively, this procedure enforces Eq. 33 as precisely as possible.

E. Examples of the GPI prior

GPI has properties that rectify significant shortcomings with other approaches to prior selection. Our first aim in this section is to compute the GPI prior for a series of models to show that the calculation is tractable in many applications. In Sec. III E 1, we compute the GPI prior for regular models in the large-sample-size limit. This analysis reveals a connection between the GPI prior and the Jeffreys prior. In Sec. III E 2, we demonstrate an exact computation of the GPI prior for a number of non-harmonic models.

1. Approximate GPI prior for regular models

We will first explore the properties of the generalized principle of indifference by computing the GPI prior in the large-sample-size limit of a regular model. To define the GPI prior, it is first useful to define the scaled-Jeffreys prior:

$$\rho(\boldsymbol{\theta}; N) \equiv \left(\tfrac{N}{2\pi}\right)^{K/2} I^{1/2}, \tag{36}$$

where I is the determinant of the Fisher information matrix defined for a single sample (Eq. A6), K is the dimension of the continuous parameter manifold Θ. The

Model name	Likelihood $q(x\|\boldsymbol{\theta})$	Parameters $\boldsymbol{\theta}$	Support $\boldsymbol{\Theta}$	GPI prior $w(\boldsymbol{\theta})$	Effective complexity: $\mathcal{K}$
Normal	$(2\pi\sigma^2)^{-D/2}\ \exp[-\frac{1}{2\sigma^2}(\vec{x}-\vec{\mu})^2]$	$(\vec{\mu})$	$\vec{\mu}\in\mathbb{R}^D$	$(\frac{N}{2\pi\sigma^2})^{D/2}e^{-\mathcal{K}}$	$\frac{D}{2}\left[1+N\log(1+N^{-1})\right]$
		(σ)	$\sigma\in\mathbb{R}_+$	$(\frac{N}{\pi\sigma^2})^{1/2}e^{-\mathcal{K}}$	Eq. C55, $\alpha=\frac{1}{2}$
		$(\vec{\mu},\sigma)$	$\vec{\mu}\in\mathbb{R}^D, \sigma\in\mathbb{R}_+$	$\sqrt{2}(\frac{N}{2\pi\sigma^2})^{(D+1)/2}e^{-\mathcal{K}}$	Eq. C38
Exponential	$\lambda e^{-\lambda x}$	(λ)	$\lambda\in\mathbb{R}_+$	$(\frac{N}{2\pi\lambda^2})^{1/2}e^{-\mathcal{K}}$	Eq. C55, $\alpha=1$
Uniform	$H_{\rm SF}(x)H_{\rm SF}(L-x)$	(L)	$L\in\mathbb{R}_+$	$\frac{N}{L}e^{-\mathcal{K}}$	$N\log(1+N^{-1})+1$
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	(β)	$\beta\in\mathbb{R}_+$	$(\frac{\alpha N}{2\pi\beta^2})^{1/2}e^{-\mathcal{K}}$	Eq. C55

TABLE III. A summary of exact GPI priors for a collection of standard models. $H_{\rm SF}$ is the Heaviside step function and we write the set of positive real numbers: $\mathbb{R}_{>0}\equiv\{y\in\mathbb{R}|y>0\}$. Many of the effective complexities are defined in the Appendix.

prior ρ is a density on the parameter manifold with the qualitative meaning of the inverse volume of indistinguishable models at sample size N. The GPI prior is

$$\log w(\boldsymbol{\theta};N)=\log\rho\ -K+\mathcal{O}(N^{-1}),\qquad(37)$$

where $K=\dim\Theta$, as shown in the Appendix (A 5).

In the large-sample-size limit, the parameter dependence of the GPI prior is identical to the Jeffreys prior, which has enjoyed a long and successful history [20]. The Jeffreys prior was initially proposed because it was reparametrization invariant [42]. More recently the same prior has been motivated by numerous other arguments (*e.g.* [4, 43]). From the perspective of parameter inference the GPI approach simply recapitulates a widely-applied method in the large-sample-size limit of a regular model.

2. *Exact GPI prior for models with symmetry*

For simple models, symmetry and dimensional analysis often imply that w must still be proportional to the Jeffreys prior even at small sample size. We compute the exact GPI prior analytically for the normal model with unknown mean and variance, the exponential model, the uniform model and the Gamma model in the Appendix (C). In the Appendix (C 2), we demonstrate an exact computation of the GPI prior for a non-regular model. A summary of the exact GPI priors is shown in Tab. III.

All these models have a log-likelihood that is anharmonic in the parameters and therefore are expected to have non-trivial high-temperature behavior. The calculation reveals that the asymptotic form of the GPI prior (Eq. 37) closely approximates the exact prior. In many models it is convenient to define the finite sample-size correction as an effective complexity $\mathcal{K}$ that replaces the model dimension K in Eq. 37:

$$w(\boldsymbol{\theta};N)=\rho\,e^{-\mathcal{K}},\qquad(38)$$

$\mathcal{K}$ is plotted as a function of sample size (N) for a number of different models in Fig. 4. On an empirical basis, it is clear that Eq. 37 is typically an excellent approximation for w even small to intermediate sample sizes.

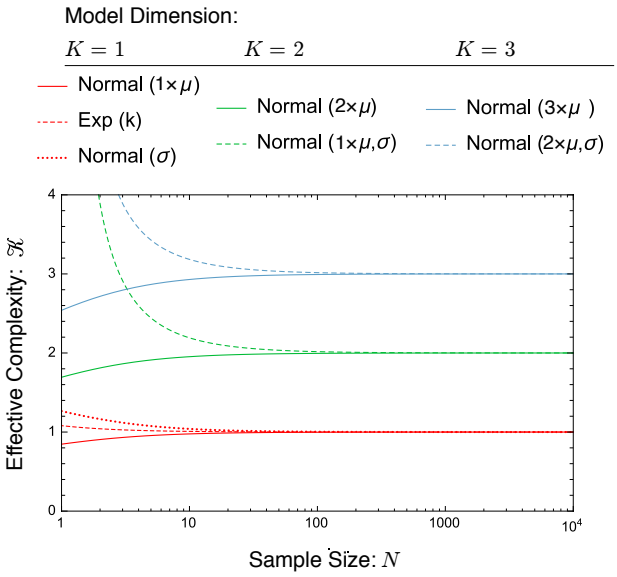

FIG. 4. **Effective complexity of models at finite sample size.** We computed the exact GPI prior for a series of models of different dimension. At large sample size, the dimension determines the effective complexity: $\mathcal{K}=\dim\Theta/2$. At finite sample size there are significant corrections. The effective complexity divergences for the normal model (dashed curves) with unknown mean and variance at $N=1$.

F. GPI circumvents the Lindley-Bartlett paradox

To demonstrate that the GPI prior automatically leads to non-anomalous inference (*i.e.* free from infinite normalization factors) and is also free from *ad hoc* parameters, we return to the example we used to introduce the Lindley-Bartlett paradox in the introduction in Sec. I C: We generate data from five competing models, then perform inference on the model identity on each of the datasets using the five models as candidates. A detailed description of the generative parameters is provided in the Appendix (A 1).

1. GPI approach

We have computed the GPI prior for each of the proposed models. Inference on parameter values follows the standard Bayesian framework using the GPI prior w. The GPI prior includes the model prior (Appendix Sec. A 8) and therefore the posterior probability of model I is:

$$\varpi(I|x^N) = Z_I / \sum_J Z_J, \tag{39}$$

where the model index J runs over the five competing models. The model posteriors for the five sets of simulated data for a sample size of $N = 20$ are shown in Fig. 1C. The results show a number of important characteristics of the GPI prior: (i) There is an unambiguous Bayesian procedure for computing inference on both parameters and models. The approach is automatic or free from *ad hoc* or subjective choices. (ii) Inference on both parameters and models leads to non-anomalous results in which the generative distribution has non-zero posterior probability. In our example, the highest posterior model is the generative distribution in each example. (iii) For the normal models, the higher-dimensional models have lower posterior probability for the data generated by model $\mathcal{N}$, even though the generative distribution is realizable in $\mathcal{N}(\mu)$ and $\mathcal{N}(\mu, \sigma)$. This shows that the GPI prior contains an endogenous model selection mechanism favoring model parsimony, and we will discuss this in detail in Sec. IV B 1. (iv) Finally, we note that the Jeffreys prior approach is not even possible in the context of the uniform model since the Fisher Information Matrix is undefined. However, we demonstrate in the Appendix (C) that the GPI prior can be computed analytically.

IV. DISCUSSION

A. Learning capacity

One valuable feature of the proposed correspondence is the potential to gain new insights into statistical phenomenology using physical insights into the thermodynamic properties of physical systems. Artificial Neural Networks (ANN) and systems-biology models are two examples of systems with a large number of poorly-specified parameters that none-the-less prove qualitatively predictive. This phenomenon has been termed *model sloppiness* [18, 44]. These models often have a logarithmic distribution of Fisher information matrix eigenvalues and this characteristic has been used as a definition of sloppiness [18]. But, this definition is unsatisfactory since it is not reparametrization invariant. It is easy to construct counter examples for this definition: For instance, in a K-dimensional normal model where the variance for each dimension is logarithmically distributed, the Fisher information eigenvalues are likewise logarithmically distributed, but the model none-the-less behaves like a normal regular model from the standpoint of prediction and statistical analyses.

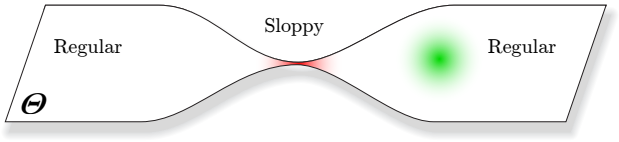

FIG. 5. **Sloppiness is determined by parameter manifold geometry and posterior width.** Parameters are defined on a compact manifold $\boldsymbol{\Theta}$. In sloppy regions of parameter manifold, the parameters are model-structure dominated (red posterior) whereas in regular regions of parameter manifold parameters are data dominated (green posterior). From the perspective of the learning capacity, the model is effectively one dimensional in proximity to the red posterior and two dimensional in proximity to the green posterior.

The correspondence we describe suggests a definition directly written in terms of the predictive performance of the model and the equipartition theorem. We propose that *predictive sloppiness* be defined as models that have a smaller learning capacity than estimated from the model dimension:

$$\overline{C} < \tfrac{1}{2} \dim \Theta. \tag{40}$$

This definition (i) would exclude all regular models in the large sample-size limit, (ii) is reparametrization invariant and (iii) can be generalized to other non-Bayesian frameworks by expressing the learning capacity in terms of the predictive performance.

The mechanism of sloppiness is model parameters being determined by model structure rather than the data. This scenario is drawn schematically in Fig. 5. We sketched a compact parameter manifold after reparameterizing the model so that the Fisher-Rao metric is the identity matrix. At a regular point (green), all parameter coordinates are regular and data dominated, the effective dimension of the model is $K_{\text{eff}} = 2$. At the sloppy point (red), the manifold is not rigorously singular but the manifold constraints determine the parameter value in the vertical coordinate direction. Therefore the effective dimension is $K_{\text{eff}} \approx 1$. In summary, when model structure not data determines the parameter values, the learning capacity will be anomalously small and the model will be anomalously predictive.

B. The generalized principle of indifference

We argue that a natural approach to objective Bayesian inference is to choose a prior such that the number of indistinguishable distributions is one for all parameter values. (See Eq. 33.) Schematically, this procedure assigns equal prior weighting to all models that

can be distinguished at finite sample size N. As the sample size increases, the prior must be modified to accommodate the increased resolution (Eq. 29) due to shrinking of the posterior support. (See Fig. 6.) For a regular model in the large-sample-size limit, no calculation is required and GPI prior is equal to the scaled-Jeffreys prior (Eq. 36.) At small sample size or in singular models, the GPI prior must be computed explicitly.

It is important to stress that GPI gives rise to a *sample-size-dependent prior* and therefore this inference is *not* Bayesian in a classical sense: (i) It violates Lindley's dictum: *today's posterior is tomorrow's prior*. (ii) Furthermore, the evidence and prior are no longer interpretable as probabilities but rather statistical weightings. On-the-other-hand, the method codes parameter uncertainty in terms of a posterior probability distribution and facilitates Bayesian parameter and model averaging. Therefore, we would argue the approach maintains all of the attractive features of the Bayesian framework while avoiding problematic aspects.

1. Model selection

The normalization of the GPI prior has significant consequences for inference on model identity (*i.e.* model selection). Returning to the regular model, it is straight forward to apply the Laplace approximation to compute the minus-log evidence using the GPI prior:

$$-\log Z(x^N; w) \approx -\log q(x^N|\hat{\boldsymbol{\theta}}) + K, \tag{41}$$

where $K = \dim \Theta$. The scaled-Jeffreys prior cancels the Occam factor from the integration. The two remaining contributions each have clear qualitative interpretations: the MLE estimate of the information ($-\log q$) is a measure of the *goodness-of-fit* and a penalty for model complexity (K). Eq. 41 is already well known as the Akaike Information Criterion (AIC)[45]:

$$\mathrm{AIC}(x^N) \equiv -\log q(x^N|\hat{\boldsymbol{\theta}}) + K. \tag{42}$$

Information-based inference is performed by selecting the model which minimizes AIC, maximizing the estimated predictive performance [24]. The reason why AIC and GPI Bayesian inference are equivalent is most easily understood by rewriting Eq. 33:

$$-N\overline{F} \approx -N\overline{U}, \tag{43}$$

which in statistical language corresponds to using a prior that makes the log partition function (LHS) an unbiased estimator of the log predictive performance (RHS). Since the Akaike Information Criterion (AIC) is an unbiased estimator of RHS at large sample size N, the generalized principle of indifference encodes an AIC-like model selection [46] and an information-based (AIC) realization of Occam's razor: *parsimony increases predictivity* [24]. The log-predictive performance

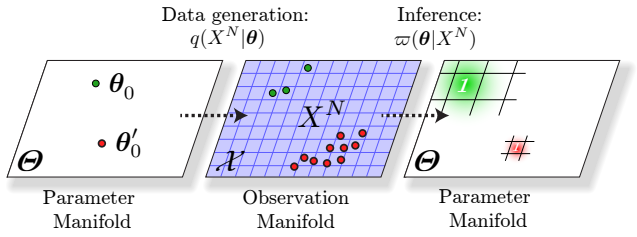

FIG. 6. **Generalized principle of indifference.** The posterior distribution $\varpi(\boldsymbol{\theta}_0|X^N)$ is shown schematically for two different sample sizes N. The resolution increases with sample size as the posterior shrinks. In the GPI prior (Eq. 33), all parameter values consistent with the posterior are assigned unit prior weight collectively.

(RHS Eq. 43) has already been advocated in the context of Bayesian model selection through the use of pseudo-Bayes factors by Gelman and coworkers [21, 27–30].

2. Posterior impropriety

The use of GPI prior often, but not always, gives non-zero evidence for all models under consideration. One such exception is shown in Fig. 4 which reveals that the normal model with unknown mean and variance has a divergent effective complexity at a sample size of $N = 1$. The effect of this divergence is to give these models zero statistical weight. Although this may initially appear problematic, it is an important feature of the generalized principle of indifference. A mean and variance cannot be estimated from a single observation and as a result the model parameter posterior would be improper and the predictive loss would be infinite. The generalized principle of indifference automatically removes these models from consideration by assigning these models zero statistical weight. The inability to automatically handle posterior impropriety is recognized as a significant shortcoming of these competing approaches [20].

C. Comparison with existing approaches

The generalized principle of indifference subsumes a patchwork of conflicting methods for prior and model selection, resolving many conflicting approaches and generating a single, generally-applicable and self-consistent framework. The GPI approach subsumes the following approaches: (i) For discrete parameter manifolds in the large sample size limit, the GPI gives equal weight to all mutually exclusive models, consistent with the original formulation of the principle of indifference by Bayes and Laplace [38, 39]. (See Eq. C56.) (ii) In the large-sample-size limit, GPI generates a GPI prior proportional to the well-known Jeffreys prior. In this sense, the approach is closely related to the reference prior approach of Bernardo and

Berger [43, 47]. (iii) With respect to model selection (inference on model identity), the GPI evidence behaves like pseudo-Bayes factors (or AIC) and therefore circumvents the Lindley-Bartlett paradox. (See Sec. IV B 1.) To date, the pseudo-Bayes approach has always been un-Bayesian in the sense that the pseudo-Bayes method consists of the *ad hoc* combination of a canonical Bayesian prior for inference on parameters but a cross-validation-based weighting for inference on models. The GPI provides a self-consistent approach to inference on both parameters and models.

The GPI addresses a number of problems with existing approaches to objective Bayesian inference. (iv) *Lindley-Bartlett paradox:* As already discussed above, an important shortcoming with existing objective Bayesian approaches relates to the compactness of the parameter manifold and the automatic rejection of higher-dimensional models in model selection (the Bartlett-Lindley paradox [25, 26]). More generally, the evidence of the canonical objective Bayesian approach depends on *ad hoc* modeling decisions, like the range of allowed parameter values. The GPI-Bayes evidence circumvents these anomalies by generating a consistent distribution density w over competing models. As a result the GPI evidence is independent of *ad hoc* modeling decisions. (v) *Unification of statistical paradigms:* The absence of the Lindley-Bartlett paradox implies coherent inference between paradigms [48, 49] and therefore the generalized principle of indifference naturally unifies objective Bayesian inference with information-based inference. (vi) *Prior and posterior impropriety:* Another important flaw identified in other objective Bayesian approaches is the inability to handle impropriety. In many cases where parameters are defined on non-compact manifolds, the prior (and sometimes the posterior) cannot be normalized. The redefinition of the prior as a density of models introduces a well-defined and consistent method for defining prior normalization, regardless of the global structure of the manifold. Furthermore, the approach automatically assigns zero statistical weight to models that suffer from posterior impropriety. (See Sec. IV B 2.) (vii) *Singularity and sloppiness:* Finally, the GPI-Bayes approach does not assume model regularity. It treats singularity and the sloppiness phenomenon in a natural way. (See Sec. III E 2.)

D. Conclusion

Nature reveals an elegant formulation of statistics in the thermal properties of physical systems. Measurements of the heat capacity, compressibility or susceptibility reveal unambiguously how Nature enumerates states and defines entropy. These physical insights provide clues to the definition of novel statistical quantities and the resolutions of ambiguities in the formulation of objective Bayesian statistics. We have refined a previously proposed correspondence between the Bayesian marginal likelihood and the partition function of statistical physics. We demonstrate a novel and substantive mapping between the average energy, heat capacity, entropy and other statistical quantities. The newly-defined learning capacity is a natural quantity for characterizing and understanding learning algorithms and generates new physical insight into the mechanism of model sloppiness through a correspondence with the equipartition theorem and the freeze-out phenomenon. A key motivation for exploring the phenomenology of learning is to apply these insights to develop new learning algorithms. We provide one example in the paper: We use the Gibbs entropy to define a generalized principle of indifference and an objective Bayesian GPI prior with the property that all distributions have equal prior weight. This approach subsumes many seemingly inconsistent and disparate methods into a single, coherent statistical approach. For the first time, we demonstrate a self-consistent Bayesian approach to performing inference on models of unknown dimensional with uninformative priors.

Acknowledgements: PAW and CHL acknowledge helpful discussions with M. Linden, J. Kinney, D. Mayo, C. Heilig, M. Abbott, B. Machta and M. Transtrum. This work was support by NSF grants NSF-PHY-084845 and NSF-MCB-1151043-CAREER.

Model		Initial parameter support: $\boldsymbol{\theta}$	Revised parameter support: $\boldsymbol{\theta}$	Generative parameters $\boldsymbol{\theta}_0$	Likelihood $q(x\|\boldsymbol{\theta})$	Prior $\varpi(\boldsymbol{\theta})$
Normal	$\mathcal{N}$			$\mu_0 = 5, \sigma_0 = 1$	$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$	1
	$\mathcal{N}(\mu)$	$\mu \in \mathbb{R}$	$\mu \in [0, 10]$	$\mu_0 = 6, \sigma_0 = 1$		C_0
	$\mathcal{N}(\mu, \sigma)$	$\mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}$	$\mu \in [0, 10], \sigma \in [0.1, 10]$	$\mu_0 = 5, \sigma_0 = 0.75$		C_0/σ^2
Exponential	$\text{Exp}(\lambda)$	$\lambda \in \mathbb{R}_{>0}$	$\lambda \in [0.1, 10]$	$\lambda_0 = 2$	$\lambda e^{-\lambda x}$	C_0/λ
Uniform	$\mathcal{U}(L)$	$L \in \mathbb{R}_{>0}$	$L \in [0, 10]$	$L_0 = 10$	$H_{\text{SF}}(x) H_{\text{SF}}(L - x)/L$	C_0

TABLE IV. **Models for inference on simulated data.** Five datasets were generated, one for each model, using the generative parameters: $X^N \sim q(\cdot|\boldsymbol{\theta}_0)$. In the canonical objective Bayesian approach, inference was performed on the simulated data using the the objective prior shown in the final column. The normalization constant C_0 was chosen in each case to make the prior proper over the original or revised parameter support. In the GPI approach, the relevant GPI prior w was used, as computed in this appendix, for each model. See Tab. III. $H_{\text{SF}}(x)$ is the Heaviside step function.

Appendix A: Supplemental results

1. Details on the Lindley-Bartlett Paradox example

a. *Canonical objective Bayesian approach*

Analysis: For the canonical objective Bayesian approach, we will use a proper objective prior. We attempt to use the Jeffreys prior for each model. A problem immediately presents itself in the context of the Uniform model where the Jeffreys prior is undefined. (See Sec. C 2.) We must therefore deviate from our protocol and apply some other prior. We set a flat prior on the parameter L motivated by the principle of indifference. The priors for the four models with parameters cannot be normalized due to their non-compact parameter manifolds. Formally, we can work in a finite interval, then consider the limit as the limits of the intervals approach infinity.

Parameters: Parameter posteriors for each model can be computed using this procedure and the results are identical to the GPI approach, except for the uniform model where no Jeffreys prior exists. These are clearly acceptable results.

Models: Unlike the parameter posteriors, inference on the model leads to anomalous results. We find that the model posterior for the parameter-free normal model is one, regardless of which distribution was used to generate the data, due to the prior impropriety of the other models. (See Fig. 1A.)

b. *Revised uninformative Bayesian approach*

After having seen the data, a Bayesian will often reconsider the prior and localize it around the values favored by the data. Here we normalize the priors on the revised finite intervals defined in Tab. IV. It is important to stress that the boundary of each interval is *ad hoc* and investigators will make different choices. Sensible choices typically do not strongly affect the parameter posteriors, but they do affect model posteriors as shown in Fig. 1B. This approach is formalized in variational or empirical Bayesian methods. In this case, the prior is no longer determined *a priori* and this double-use of data can lead to difficulties due to the potential for overfitting.

2. Definitions of information, cross entropy, Fisher information matrix

The Shannon Information is defined:

$$h(x|\boldsymbol{\theta}) \equiv -\log q(x|\boldsymbol{\theta}). \tag{A1}$$

Let X be distributed with a true distribution with parameter $\boldsymbol{\theta}_0$: $X \sim q(\cdot|\boldsymbol{\theta}_0)$. The cross entropy is defined:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \equiv \langle h(X|\boldsymbol{\theta}) \rangle_{X \sim q(\cdot|\boldsymbol{\theta}_0)}, \tag{A2}$$

and which has a minimum at the true entropy:

$$H_0(\boldsymbol{\theta}_0) \equiv H(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0). \tag{A3}$$

The empirical estimator of the cross entropy is defined:

$$\hat{H}(\boldsymbol{\theta}) \equiv N^{-1} \sum_{i=1}^{N} h(x_i|\boldsymbol{\theta}), \tag{A4}$$

which is independent of N to leading order, in spite of the prefactor. The KL-Divergence:

$$D_{\mathrm{KL}}(\boldsymbol{\theta}_0||\boldsymbol{\theta}) = H(\boldsymbol{\theta};\boldsymbol{\theta}_0) - H_0(\boldsymbol{\theta}_0), \tag{A5}$$

is the natural distance-like measure on the parameter manifold. The Fisher information matrix is defined:

$$I_{ij} = \left[\tfrac{\partial}{\partial\theta^i}\tfrac{\partial}{\partial\theta^j} H(\boldsymbol{\theta};\boldsymbol{\theta}_0)\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{A6}$$

which is a rank-two symmetric covariant tensor known as the Fisher-Rao metric [37].

3. An alternate correspondence

We do not believe that statistical mechanics prescribes a unique procedure for objective Bayesian inference. In establishing the correspondence between inference and statistical mechanics, we identify the partition function Z as the marginal likelihood and $N \leftrightarrow \beta$ in agreement with V. Balasubramanian [4]. However, this is not the only choice that has been proposed. For instance, Watanabe [34] instead chooses to define the inverse-temperature β^* so that the likelihood is raised to an arbitrary power β^*:

$$q(X^N|\boldsymbol{\theta}) \to q^{\beta^*}(X^N|\boldsymbol{\theta}). \tag{A7}$$

We have explored both possibilities in some detail. In the large-sample-size limit, these two correspondences have many similar properties. However, we present only the analysis of the $N \leftrightarrow \beta$ correspondence in this papers because we feel that it identifies statistical objects with the most desirable properties.

It is important to note that the β^* correspondence does have a number of convenient computational properties: (i) It is continuous and therefore no finite difference definition need be introduced. (ii) It allows one to interpolate between a Bayesian posterior (given by $\beta^* = 1$) and the point estimates of the MLE's (given by $\beta^* \to \infty$). This temperature has also been applied in tempering schemes in MCMC methods, and simulated annealing—increasing the temperature promotes a better exploration of the sample space (chain-mixing) that can be used to better sample multimodal distributions, or find the minima in a rough function in close analogy to analogous methods in physics.

However, the β^* correspondence also has a number of features we find undesirable: First, β^* is not a preexisting statistical parameter within the Bayesian framework. Only $\beta^* = 1$ corresponds to a Bayesian statistics. High and low β^* correspond to non-Bayesian statistics whereas the sample size interpretation of β has a sensible and natural Bayesian interpretation in terms of small and large sample sizes. Second, the internal energy under the choice of β^* is not the predictive performance U. Consequently, the principle of indifference, which results from a likelihood-power β^*, does not generate the Akaike weights as the model averaging procedure. It is the reproduction of AIC, with its proven asymptotic efficiency [50], that is an important motivation of the proposed correspondence.

4. Finite difference is equivalent to cross validation

The log-predictive distribution can be written as a finite difference:

$$\log q(X_i|X^{\neq i}) = \log Z(X^N) - \log Z(X^{\neq i}). \tag{A8}$$

We can interpret the $-\log q(X_i|X^{\neq i})$ as a finite difference estimate of the the sample size derivative of the free energy. We take the mean over all permutations of the data so that this estimate is symmetric with respect to all data points. Under expectation, analytically continuing sample size, the LOOCV relationship to the internal energy is clear:

$$\langle \log q(X_i|X^{\neq i})\rangle \approx \frac{\partial}{\partial N}\langle \log Z(X^N)\rangle + O(N^{-1}). \tag{A9}$$

This identity is crucial in establishing the thermodynamic interpretation in terms of predictive performance.

5. Jeffreys prior is proportional to GPI prior in the large-sample-size limit

In the large-sample-size limit, the partition function can be evaluated using the Lapalce (saddle-point) approximation and the resulting prior is proportional to the Jeffreys prior. The integral is evaluated by expanding around the minimum of $\hat{H}_X(\boldsymbol{\theta})$, the maximum likelihood estimator: $\hat{\boldsymbol{\theta}}_X$. The partitition function $Z(X^N) = \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\, \varpi(\boldsymbol{\theta})\, \exp[-N\hat{H}_X(\boldsymbol{\theta})]$, becomes

$$Z(X^N) \approx e^{-N\hat{H}_X(\hat{\boldsymbol{\theta}}_X)} \left(\frac{2\pi}{N(\det I)^{1/K}}\right)^{K/2} \varpi(\boldsymbol{\theta}_X) \tag{A10}$$

By the standard χ^2_K representation of the overfitting error, $\langle \hat{H}(\hat{\theta}_X)\rangle_X = H_0 - \frac{K}{2N}$. Therefore the disorder average becomes

$$\overline{F}(\boldsymbol{\theta}_0, \varpi, N) = H_0 - \frac{K}{2N} - \frac{K}{2N}\log\frac{2\pi}{N(\det I)^{1/K}} - \frac{1}{N}\log\varpi(\boldsymbol{\theta}_0) + O(N^{-2}) \tag{A11}$$

We can then calculate the Gibbs entropy $N^2\partial_N F$,

$$\overline{S}(\boldsymbol{\theta}_0, \varpi, N) = \frac{K}{2}\log\frac{2\pi}{N(\det I)^{1/K}} + K + \log\varpi(\boldsymbol{\theta}_0) + O(N^{-1}) \tag{A12}$$

If we enforce the generalized principle of indifference, ignoring order N^{-1},

$$0 = S(\boldsymbol{\theta}_0, w, N) \tag{A13}$$

and substituting the w for ϖ in the entropy expression then gives us the condition

$$w(\boldsymbol{\theta}_0) = (\det I)^{1/2}\left(\frac{N}{2\pi}\right)^{K/2} e^{-K}. \tag{A14}$$

Thus the GPI condition is satisfied by the Jeffries prior in the large-sample-size limit. The constant weighting factor is important in model selection as e^{-K} encodes the Akaike weight [24]. The GPI prior has sample-size dependence. This sample size dependence will break the de-Finetti likelihood principle: that the prior should not depend on the nature of the data-generating procedure (including the sample size) [51]. The departure from the likelihood principle is the origin of the departure from the conventional Bayesian model selection behavior.

6. Reparametrization invariance of the GPI approach

An important property of an objective prior is *reparameterization invariance*. Is the GPI approach reparameterization invariant? First consider the properties of the partition function:

$$Z(X^N) = \int \mathrm{d}\boldsymbol{\theta}\, \varpi(\boldsymbol{\theta})\, q(X^N|\boldsymbol{\theta}), \tag{A15}$$

which is invariant under reparameterization $\boldsymbol{\theta} \to \boldsymbol{\theta}'$ if the prior ϖ transforms as a density, *i.e.*:

$$\varpi'(\boldsymbol{\theta}') = J^{-1}\varpi(\boldsymbol{\theta}), \tag{A16}$$

where J is the determinant of the Jacobian of the coordinate transformation. Since the GPI condition is written in terms of the partition function and the sample size N, the GPI condition itself is also reparameterization invariant. Therefore if $w(\boldsymbol{\theta})$ is the GPI prior for $\boldsymbol{\theta}$ coordinates,

$$w'(\boldsymbol{\theta}') = J^{-1}w(\boldsymbol{\theta}), \tag{A17}$$

will be the GPI prior in the $\boldsymbol{\theta}'$ coordinates. Therefore the GPI prior will transform like a density under reparameterization.

7. Effective temperature of confinement

To calculate the free energy F of a classical free particle confined to a volume $V = L^3$, we calculate the partition function by integrating over available phase space:

$$Z(\beta) = \int \frac{d^K \boldsymbol{p}\, d^K \boldsymbol{x}}{(2\pi\hbar)^K} e^{-\beta H(\boldsymbol{p},\boldsymbol{x})} \tag{A18}$$

$$= \frac{e^{-\beta E_0} L^K}{(2\pi\hbar)^K} \left(\int dp\, e^{\frac{-\beta p^2}{2m}} \right)^K = \left(\frac{mL^2}{2\pi\hbar^2\beta} \right)^{K/2} e^{-\beta E_0}. \tag{A19}$$

The Free energy is then

$$F(\beta) = E_0 + \frac{K}{2\beta} \log \frac{mL^2}{2\pi\hbar^2\beta} = E_0 + \frac{K}{2\beta} \log \frac{\beta_0}{\beta} \tag{A20}$$

where we have made the identifications

$$\beta_0 = \frac{mL^2}{2\pi\hbar^2} \quad \text{and} \quad K = 3. \tag{A21}$$

β_0 can be interpreted as the inverse of the (typically negligibly small) temperature at which the thermal de Broglie wavelength of the confined particle is on the order of the width of the confining box.

8. A Bayesian re-interpretation

The replacement of the prior (a probability density) with an unnormalized density of states may make a Bayesian reader uncomfortable since the evidence (Z) no longer has the meaning of a probability. But there is a natural Bayesian interpretation in terms of the model prior.

Typically, when models are compared in a Bayesian context, all mutually exclusive models are assigned equal *a priori* probabilities (*i.e.* the principle of indifference). But, we have now proposed a new concept of model enumeration by introducing a density of models. We can compute the total number of distinguishable distributions in model I at sample size N by integrating the GPI prior (density of states) over the parameter manifold:

$$\mathcal{N}_I(N) \equiv \int_\Theta \mathrm{d}\boldsymbol{\theta}\; w_I(\boldsymbol{\theta}; N). \tag{A22}$$

Since models I and J contain different numbers of distinguishable distributions, we reason that the principle of indifference should be interpreted to apply at the distinguishable distribution level rather than the model level. Therefore the *a prior* model probabilities should be:

$$\varpi_I \equiv \mathcal{N}_I / \textstyle\sum_J \mathcal{N}_J. \tag{A23}$$

and the proper parameter prior is

$$\varpi(\boldsymbol{\theta}|I) \equiv w_I(\boldsymbol{\theta}; N)/\mathcal{N}_I. \tag{A24}$$

Inference with the improper GPI prior is equivalent to assuming proper prior ϖ_I on models and proper prior $\varpi(\boldsymbol{\theta}|I)$ on parameters. The numerator in RHS of Eq. A23 will cancel the denominator in the RHS of Eq. A24 when the model posterior is computed and the normalization $\mathcal{N}_I$ divides out of parameter posterior distributions.

Appendix B: Methods

1. Computation of the free energy using a sufficient statistic

It is often convenient to work in terms of sufficient statistics because (i) all the data dependence of the posterior enters through the sufficient statistic and (ii) the statistics have well known statistical distributions that significantly simplify many calculations. We define a sufficient statistic $\boldsymbol{t} = \boldsymbol{T}(X^N)$ such that

$$\Pr(\boldsymbol{\theta}|X^N) = \Pr(\boldsymbol{\theta}|\boldsymbol{t}), \tag{B1}$$

or all the information about the parameters is encoded in $\boldsymbol{t}$. We can therefore write:

$$q(X^N|\boldsymbol{\theta}) = q(X^N|\boldsymbol{t})\; q(\boldsymbol{t}|\boldsymbol{\theta}), \tag{B2}$$

and we can define a Statistic Shanon entropy:

$$H_{\boldsymbol{t}}(\boldsymbol{\theta}) = -\overline{\log q(\boldsymbol{t}|\boldsymbol{\theta})}. \tag{B3}$$

In terms of the sufficient statistic, the partition function factors:

$$Z(X^N; \varpi) = q(X^N|\boldsymbol{t})\; z(\boldsymbol{t}; N, \varpi), \tag{B4}$$

where the statistic partition function is

$$z(\boldsymbol{t}; N, \varpi) \equiv \int_\Theta \mathrm{d}\boldsymbol{\theta}\; \varpi(\boldsymbol{\theta})\; q(\boldsymbol{t}|\boldsymbol{\theta}). \tag{B5}$$

The expected free energy can be written:

$$\overline{F}(\boldsymbol{\theta}; N, \varpi) = -N^{-1}\overline{\log z(\boldsymbol{t}; N, \varpi)} + H_0(\boldsymbol{\theta}) - N^{-1}H_{\boldsymbol{t}}(\boldsymbol{\theta}), \tag{B6}$$

where H_0 is the entropy.

2. Computation of learning capacity

To compute the learning capacity, we will use the definition from Tab. I:

$$\overline{C}(\boldsymbol{\theta}; N, \varpi) = N^2\partial_N^2\overline{\log Z}(\boldsymbol{\theta}; N, \varpi), \tag{B7}$$

where $X \sim q(\cdot|\boldsymbol{\theta})$. We will ignore the finite-difference definition in these computations for simplicity.

3. Direct computation of GPI prior

We will use the finite-difference definition of the entropy (Eqs. 12 and 19) to enforce the generalized principle of indifference (Eq. 33). The relation for the GPI prior can be written:

$$(N+1)\,\overline{\log Z}(\boldsymbol{\theta}; N, w) = N\,\overline{\log Z}(\boldsymbol{\theta}; N+1, w), \tag{B8}$$

in terms of the partition function. We will use Eq. B8 explicitly to solve for the GPI prior w. For the models we work analytically, we will be be able to use the asymptotic form of w (Eq. 37) to define an effective model dimension $\mathcal{K}$ (Eq. 38). The general strategy will be:

1. Use symmetry and dimensional analysis to deduce the scaling of w with respect to the parameters $\boldsymbol{\theta}$.
2. Compute $\log Z(X^N; w)$ and re-express in terms of canonical random variables.
3. Compute $\overline{\log Z(X^N; w)}$.
4. Solve for the unknown normalization c of w using the GPI condition (Eq. B8).

4. Computation of the GPI prior using a recursive approximation

The Gibbs entropy has the property that it is linear in the prior such that:

$$\overline{S}(\theta_0, N, e^\alpha\varpi) = \alpha + \overline{S}(\theta_0, N, \varpi) \tag{B9}$$

If the prior and entropy are flat, then setting $\alpha = -\overline{S}(\theta_0, N, \varpi)$ will result in $\overline{S}(\theta_0, N, e^\alpha\varpi) = 0$; the GPI prior condition. This suggests the following simple recursive scheme for a successive approximation for the GPI prior:

procedure RECURSIVEW(ϖ)
    repeat

3: $\varpi(\theta) \leftarrow \varpi(\theta) e^{-\overline{S}(\theta,\varpi,N)}$
4: **until** $\overline{S}(\theta;\varpi,N) \approx 0$
5: **end procedure**

To the extent the entropy is slowly varying and only locally dependent on the prior, this algorithm will very quickly converge to an exact GPI prior. However, effects due to manifold boundaries and model singularities may create artifacts that lead to unstable updates. Empirical evidence suggests that the algorithm should be terminated before the exact GPI prior condition is met. Typically very few iterations are required.

Appendix C: Detailed analysis of the learning capacity and GPI prior of selected models

1. Normal models

a. *Normal model with unknown mean, known variance and an informative prior*

The likelihood for the normal model is defined by Eq. 21 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\mu \in \mathbb{R}^K$ for a normal model with unknown mean and known variance σ^2. In this example, we assume a conjugate prior:

$$\varpi(\boldsymbol{\theta}) = (2\pi\sigma_\varpi^2)^{-K/2} \exp[-\tfrac{1}{2\sigma_\varpi^2}(\vec{\mu} - \vec{\mu}_\varpi)^2], \tag{C1}$$

where we define the critical sample size:

$$N_0 \equiv \sigma^2/\sigma_\varpi^2. \tag{C2}$$

The partition function is computed by completing the square in the exponential. If $X^N \sim q(\cdot|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \varpi$, the log partition function can be expressed in terms of three chi-squared random variables:

$$\sigma^{-2} \sum_{i=1}^{N} (\vec{X}_i - \hat{\vec{\mu}}_X)^2 \sim \chi^2_{K(N-1)}, \tag{C3}$$

$$\sigma^{-2} N (\vec{\mu} - \hat{\vec{\mu}}_X)^2 \sim \chi^2_K, \tag{C4}$$

$$\sigma^{-2} N_0 (\vec{\mu} - \vec{\mu}_\varpi)^2 \sim \chi^2_K. \tag{C5}$$

The log partition function is therefore distributed:

$$\log Z(X^N;\varpi) \sim -\tfrac{KN}{2} \log 2\pi\sigma^2 - \tfrac{K}{2} \log \tfrac{N+N_0}{N_0} - \tfrac{1}{2}\chi^2_{K(N-1)} - \tfrac{1}{2}\tfrac{N_0 N}{N+N_0}(N^{-1}\chi^2_K + N_0^{-1}\chi^2_K{}'), \tag{C6}$$

where χ^2_j is a chi-squared random variable dimension j and the expect-log partition function is

$$\overline{\log Z}(N,\varpi) = -NH_0 - \tfrac{K}{2} \log \tfrac{N+N_0}{N_0}, \tag{C7}$$

where H_0 is the entropy and the free energy is:

$$\overline{F}(N,\varpi) = H_0 + \tfrac{K}{2N} \log \tfrac{N+N_0}{N_0}. \tag{C8}$$

The other results in the Tab. II are generated by apply the definitions of the correspondence in Tab. I.

b. *Normal model with unknown mean and known variance*

The likelihood for the normal model is defined by Eq. 21 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\vec{\mu} \in \mathbb{R}^D$ for a normal model with unknown mean and known variance σ^2. The cross entropy is

$$H(\boldsymbol{\theta};\boldsymbol{\theta}_0) \equiv \tfrac{D}{2}\left[\log 2\pi\sigma^2 + 1\right] + \tfrac{1}{2\sigma^2}(\vec{\mu} - \vec{\mu}_0)^2, \tag{C9}$$

where the true distribution is $X \sim q(\vec{x}|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = \sigma^{-2D}. \tag{C10}$$

The scaled Jeffreys prior (Eq. 36) is therefore:

$$\rho = (\tfrac{N}{2\pi\sigma^2})^{D/2}. \tag{C11}$$

We will assume w matches the asymptotic form:

$$w = c\sigma^{-D}, \tag{C12}$$

and solve for the unknown constant $c(N, D)$. The partition function is

$$\log Z(X^N; w) \sim \log c - \tfrac{DN}{2} \log 2\pi\sigma^2 + \tfrac{D}{2} \log \tfrac{2\pi}{N} - \tfrac{1}{2}\chi^2_{D(N-1)} \tag{C13}$$

where $\chi^2_{D(N-1)}$ is a $D(N-1)$-dimensional chi-squared random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N, w) = -NH_0(\boldsymbol{\theta}) + \log c + \tfrac{D}{2}\left[\log \tfrac{2\pi}{N} + 1\right], \tag{C14}$$

where H_0 is the true entropy. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = \tfrac{D}{2}, \tag{C15}$$

where D is both the dimension of the model. The unknown normalization of w is:

$$\log c = \tfrac{D}{2} \log \tfrac{N}{2\pi} - \tfrac{D}{2}\left[1 + N \log(1 + N^{-1})\right] \tag{C16}$$

which can be re-written as an effective complexity:

$$\mathcal{K} = \tfrac{D}{2}\left[1 + N \log(1 + N^{-1})\right], \tag{C17}$$

to define the GPI prior w using Eq. 38.

c. *Normal model with unknown discrete mean*

The likelihood for the normal model is defined by Eq. 21 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\vec{\mu} \in \mathbb{Z}^D$ for a normal model with unknown mean and known variance σ^2. We use Eq. B6 to treat the problem in terms of sufficient statitics. The statistic partition function breaks up by dimension: For each dimension with the flat improper prior:

$$\varpi(\mu) = \sum_m \delta(\mu - m), \tag{C18}$$

and sufficient statistic $t = N^{-1}\sum_i^N X_i$ the statistic partition function becomes the sum over discrete prior values:

$$z(t; N, \varpi) = \sum_{m=-\infty}^{\infty} q(t|m) = \left(\frac{N}{2\pi}\right)^{1/2} \sum_{m=-\infty}^{\infty} e^{-N(t-m)^2} \tag{C19}$$

$$= \vartheta\left(t; r = e^{-\frac{2\pi^2}{N}}\right) = \sum_{m=-\infty}^{\infty} r^{m^2} e^{2\pi t} \tag{C20}$$

Where ϑ is the Jacobi theta function with nome r. We can use the Jacobi triple product formula to write down an epxression for the log partition function:

$$\log z(t; N, \varpi) = \sum_{m=1}^{\infty} \log(1 - r^{2m}) + \log\left(1 + r^{2m-1}e^{-i2\pi t}\right) + \log\left(1 + r^{2m-1}e^{i2\pi t}\right). \tag{C21}$$

Assume (without loss of generality) that $m_0 = 0$, then, because $|r| < 1$, we can Taylor expand the logarithm:

$$\mathop{\mathbb{E}}_{t|m_0} \sum_{m=1}^{\infty} \log\left(1 + r^{2m-1}e^{-i2\pi t}\right) = -\sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^k}{k} r^{(2m-1)k} \mathop{\mathbb{E}}_{t|m_0} e^{-i2\pi kt}, \tag{C22}$$

then compute the expectation term by term:

$$= -\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}r^{(2m-1)k+k^2}, \tag{C23}$$

$$= \sum_{k=1}^{\infty}\frac{(-1)^k}{k}\frac{r^{k^2+k}}{r^{2k}-1}. \tag{C24}$$

This series is convergent, but converges slowly for very large N. We therefore must also develop a series for when $N \gg 1$. We can use the Poisson resummation formula to convert the partition function into a sum over reciprocal space. First we have to subtract off the singularity at zero, by adding a piece to the summand that can be explicitly summed. Then we can extend this function to both positive and negative integers:

$$\mathop{\mathbb{E}}_{t|m_0}\sum_{m=1}^{\infty}\log\left(1+r^{2m-1}e^{-i2\pi t}\right) = \frac{1}{2}\sum_{k=1}^{\infty}\frac{\cos(\pi k)}{k^2}\frac{k}{\sinh\left(\frac{2\pi^2 k}{N}\right)}e^{-\frac{2\pi^2k^2}{N}} \tag{C25}$$

$$= \frac{N}{48}+\frac{N}{4\pi^2}\sum_{k=1}^{\infty}\frac{\cos(\pi k)}{k^2}\left(\frac{2\pi^2 k}{n\sinh\left(\frac{2\pi^2 k}{N}\right)}e^{-\frac{2\pi^2k^2}{N}}-1\right) \tag{C26}$$

$$= \frac{N}{48}+\frac{1}{4}-\frac{\pi^2}{12N}+\frac{N}{8\pi^2}\sum_{k=-\infty}^{\infty}\frac{\cos(\pi k)}{k^2}\left(\frac{2\pi^2 k}{n\sinh\left(\frac{2\pi^2 k}{N}\right)}e^{-\frac{2\pi^2k^2}{N}}-1\right). \tag{C27}$$

The sum can now be represented as the sum of the Fourier transform of the summand. At large N, even the first term is exponentially small, and the whole sum can be ignored, leaving:

$$\mathop{\mathbb{E}}_{t|m_0}\sum_{m=1}^{\infty}\log\left(1+r^{2m-1}e^{-i2\pi t}\right) = \begin{cases}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\frac{r^{k^2+k}}{r^{2k}-1} & \text{for all } N\\ \frac{N}{48}+\frac{1}{4}-\frac{\pi^2}{12N} & N>10^2\end{cases} \tag{C28}$$

Similarly we have for the Euler function piece of the triple product

$$\sum_{m=1}^{\infty}\log(1-r^{2m}) = \begin{cases}\sum_{k=1}^{\infty}\frac{1}{k}\frac{r^{2k}}{1-r^{2k}} & \text{for all } N\\ -\frac{N}{24}+\frac{\pi^2}{6N}+\frac{1}{2}\log\left(\frac{N}{2\pi}\right) & N>50.\end{cases} \tag{C29}$$

The other term in Eq. C21 can be computed in the same way. The free energy can then be computed using Eq. B6 and the computation of other derived quantities (learning capacity, GPI prior *etc*) is left to the reader.

d. *Normal model unknown mean and variance*

The likelihood for the normal model is defined by Eq. 21 with parameters $\boldsymbol{\theta} = (\vec{\mu}, \sigma)$ with support $\vec{\mu} \in \mathbb{R}^D$ and $\sigma \in \mathbb{R}_{>0}$. The cross entropy is:

$$H(\boldsymbol{\theta};\boldsymbol{\theta}_0) \equiv \tfrac{D}{2}\left[\log 2\pi\sigma^2 + \tfrac{\sigma_0^2}{\sigma^2}\right] + \tfrac{1}{2\sigma^2}(\vec{\mu}-\vec{\mu}_0)^2, \tag{C30}$$

where the true distribution is $X \sim q(\vec{x}|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = 2\sigma^{-2(D+1)}. \tag{C31}$$

The scaled Jeffreys prior (Eq. 36) is therefore:

$$\rho = \sqrt{2}(\tfrac{N}{2\pi\sigma^2})^{(D+1)/2}. \tag{C32}$$

We will assume w matches the asymptotic form:

$$w = c\,\sigma^{-D-1}. \tag{C33}$$

Note that w must have units of inverse length to the $D+1$ power in order to give the evidence the correct units. Due to translation symmetry in μ, w must be a function of σ only. The partition function is

$$\log Z(X^N) \sim \log c - \tfrac{DN}{2}\log 2\pi\sigma^2 + \tfrac{D}{2}\log\tfrac{2\pi}{N} - \log 2 + \log\Gamma(\tfrac{DN}{2}) - \tfrac{DN}{2}\log\tfrac{\chi^2_{D(N-1)}}{2} \tag{C34}$$

where $\chi^2_{D(N-1)}$ is a $D(N-1)$-dimensional chi-squared random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -NH_0(\boldsymbol{\theta}) + \log c + \tfrac{DN}{2} + \tfrac{D}{2}\log\tfrac{2\pi}{N} - \log 2 + \log\Gamma(\tfrac{DN}{2}) - \tfrac{DN}{2}\psi(\tfrac{D(N-1)}{2}) \tag{C35}$$

where H_0 is the true entropy and ψ is the polygamma function. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N, w) = \tfrac{D}{2} + N^2(\tfrac{D}{2})^2\psi^{(1)}(\tfrac{DN}{2}) - 2N^2(\tfrac{D}{2})^2\psi^{(1)}(\tfrac{D(N-1)}{2}) - N^3(\tfrac{D}{2})^3\psi^{(2)}(\tfrac{D(N-1)}{2}), \tag{C36}$$

where D is both the dimension of mean parameter and the model. The unknown normalization of w is:

$$\log c = \tfrac{D+1}{2}\log\tfrac{N}{2\pi} + \tfrac{1}{2}\log 2 - \mathscr{K} \tag{C37}$$

written in terms of the effective complexity:

$$\mathscr{K} = \tfrac{1}{2}\log\tfrac{N}{2\pi} - \tfrac{1}{2}\log 2 - \tfrac{DN}{2}\log\tfrac{N}{N+1} - N\log\Gamma[\tfrac{D(N+1)}{2}] + (N+1)\log\Gamma[\tfrac{DN}{2}] + \tfrac{D(N+1)N}{2}\left[\psi(\tfrac{DN}{2} - \psi(\tfrac{D(N-1)}{2}))\right], \tag{C38}$$

which is used to define the GPI prior w using Eq. 38.

2. Uniform distribution

In the exponential mixture model example, the non-regular model showed reduced learning capacity at the singularity but non-regular models can also have increased learning capacity as well. To illustrate this phenomenon, consider a continuous version of the German Tank problem, estimation of the end point of a uniform distribution [52].

The likelihood for the normal model is defined:

$$q(x|\boldsymbol{\theta}) \equiv \begin{cases} L^{-1}, & 0 \le x \le L \\ 0, & \text{otherwise} \end{cases}, \tag{C39}$$

where the parameter $\boldsymbol{\theta} \equiv (L)$ with support $L \in \mathbb{R}_{>0}$. The cross entropy is:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = \begin{cases} \log L, & L_0 \le L \\ \infty, & \text{otherwise} \end{cases}, \tag{C40}$$

which is minimized at $L = L_0$ but neither the first nor second derivative is defined at this point and therefore the Fisher information matrix cannot be defined. We can still infer the dependence of the w by symmetry and dimensional analysis:

$$w = c\,L^{-1}. \tag{C41}$$

The partition function is

$$\log Z(X^N) \sim \log c - N\log L - \log N - N\log Y, \tag{C42}$$

where Y is the maximum of N uniformly-distributed random variables on the interval $[0, 1]$. The CDF for Y is the Nth power of the CDF for a single uniformly-distributed random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -NH_0(\boldsymbol{\theta}) + \log c - \log N + 1. \tag{C43}$$

The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = 1. \tag{C44}$$

The unknown normalization of w is:

$$\log c = \log N - N\log(1 + N^{-1}) - 1, \tag{C45}$$

which can be plugged into Eq. C41 to calculation the GPI prior w.

3. Gamma model

The likelihood for the Gamma model is defined:

$$q(x|\boldsymbol{\theta}) \equiv \tfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{C46}$$

where the parametersare $\boldsymbol{\theta} \equiv (\beta)$ with support $\beta \in \mathbb{R}_{>0}$. Note that the expontial model corresponds to $\alpha = 1$, the Uniform model corresponds to $\alpha \to 0$ and the normal model with unknown variance known mean corresponds to $\alpha = \frac{1}{2}$, after the transformation $x \to x^{1/\alpha}$.

The cross entropy is:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = -\alpha \log \beta + \log \Gamma(\alpha) - (\alpha - 1)\left[\psi(\alpha) - \log \beta\right] + \tfrac{\beta\alpha}{\beta_0}, \tag{C47}$$

where the true distribution is $X \sim q(x|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = \frac{\alpha}{\beta^2}. \tag{C48}$$

The scaled Jeffreys prior (Eq. 36) is therefore:

$$\rho = (\tfrac{\alpha N}{2\pi\beta^2})^{1/2}. \tag{C49}$$

We will assume w matches the asymptotic form:

$$w = c\, \beta^{-1}. \tag{C50}$$

The partition function is

$$\log Z(X^N) \sim -\alpha N \log Y^{\alpha N} + (\alpha - 1) \sum_{i=1}^{N} \log Y_i^\alpha - N \log \beta_0 + \log \Gamma(\alpha N) - N \log \Gamma(\alpha), \tag{C51}$$

where Y^I is a Gamma-distributed random variable with unit scale and shape I. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -N H_0(\boldsymbol{\theta}) + \log \Gamma(N\alpha) - \alpha N \psi(N\alpha) + N\alpha, \tag{C52}$$

where H_0 is the true entropy and ψ is the polygamma function. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = -(\alpha N)^2 [\psi^{(1)}(\alpha N) + \alpha N \psi^{(2)}(\alpha N)], \tag{C53}$$

where ψ is the polygamma function. The unknown normalization of w is:

$$\log c = \tfrac{1}{2} \log \tfrac{\alpha N}{2\pi} - \mathscr{K} \tag{C54}$$

which can be re-written as an effective complexity:

$$\mathscr{K} = \tfrac{1}{2} \log \tfrac{\alpha N}{2\pi} + (1 + N) \log \Gamma(\alpha N) + -N \log \Gamma(\alpha(1 + N)) - \alpha N (1 + N)(\psi[\alpha N] - \psi[\alpha(1 + N)]), \tag{C55}$$

to define the GPI prior w using Eq. 38.

4. GPI prior for discrete parameter manifolds

For discrete parameter manifolds two competing and well-established methods exist for choosing a prior: (i) A literal interpretation of the principle of indifference would seem to imply that all parameter values are given equal weight. (ii) Alternatively, we can consider the continuous parameter limit where the prior can be chosen to give consistent results with the Jeffreys prior. Both approaches have desirable properties in different contexts [22]. GPI provides an elegant resolution to this conflict: When the discrete nature of the parameter manifold can be statistically resolved, the GPI prior prior assigns equal weight to all discrete parameter values (i), whereas, if the discreteness of the space cannot be statistically resolved, the large N limit gives rise to a Jeffreys prior (ii):

$$w = \begin{cases} \rho\, e^{-K} \prod_i \Delta\theta^i, & \Delta\theta^i \ll \delta\theta^i \\ 1, & \Delta\theta^i \gg \delta\theta^i \end{cases} \tag{C56}$$

where $\Delta\theta^i$ is the lattice spacing and the statistical resolution $\delta\theta^i$ is defined in Eq. 29. The GPI prior for a normal model with a discrete mean can be computed exactly and is described in the appendix (C 1 c).

---

[1] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[2] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press., 2003.

[3] Ole E Barndorff-Nielsen and Peter E Jupp. Statistics, yokes and symplectic geometry. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 6, pages 389–427, 1997.

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

[5] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[6] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[7] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[8] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[9] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[10] Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[11] Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc., 2002.

[12] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[13] Teppo Mikael Niinimäki and Mikko Koivisto. Annealed importance sampling for structure learning in bayesian networks. In *IJCAI*, pages 1579–1585, 2013.

[14] Faming Liang and Wing Hung Wong. Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, 2001.

[15] Jun-ichi Inoue. Application of the quantum spin glass theory to image restoration. *Physical Review E*, 63(4):046114, 2001.

[16] Marc Mézard and Andrea Montanari. Reconstruction on trees and spin glass transition. *Journal of statistical physics*, 124(6):1317–1350, 2006.

[17] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[18] Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–7, Nov 2013.

[19] R.K. Pathria and P.D. Beale. *Statistical Mechanics*. Elsevier Science, 1996.

[20] Robert E Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 1996.

[21] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[22] James O. Berger, Jose M. Bernardo, and Dongchu Sun. Objective priors for discrete parameter spaces. *Journal Of The American Statistical Association*, 107(498):636–648, 2012.

[23] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov B. N. and E. Csaki, editors, *2nd International Symposium of Information Theory.*, pages 267–281. Akademiai Kiado, Budapest., 1973.

[24] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference.* Springer-Verlag New York, Inc., 2nd. edition, 1998.

[25] M. S. Bartlett. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(3/4):533–534, 1957.

[26] D. V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[27] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514, 1994.

[28] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, B64:583–639, 2002.

[29] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[30] Kenneth P. Burnham and David R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[31] In the canonical ensemble, the probability of the occupancy of a microstate i is $p_i = Z^{-1} g_i e^{-\beta E_i}$ where Z is the partition function, E_i is the energy, g_i is the degeneracy and $\beta^{-1} \equiv k_B T$.

[32] Quenched disorder refers to variables in the system that are randomly generated when the system is assembled but remain constant in time. These variables remain fixed in expectations over thermal fluctuations.

[33] Josiah Willard Gibbs. Elementary principles of statistical mechanics. *Compare*, 289:314, 1902.

[34] S. Watanabe. *Algerbraic geometry and statistical learning theory.* Cambridge Univeristy Press, 2009.

[35] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[36] F. Reif. *Statistical Physics.* McGraw-Hill (New York), 1967.

[37] S. I. Amari. *Differential Geometrical Methods in Statistics.* Springer-Verlag, Berlin., 1985.

[38] Marquis de. Laplace, Pierre Simon. *Théorie Analytique des Probabilités.* Courcier, Paris., 1812.

[39] J. M. Keynes. *A Treatise on Probability.* Macmillan Limited, London, 1921.

[40] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory, IEEE Transactions on*, 26(1):26–37, January 1980.

[41] R. L. Kashyap. Prior probability and uncertainty. *IEEE Transactions on information theory*, 17(6):641–650, 1971.

[42] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[43] J. M. Bernardo. *Bayesian statistics 6: Nested Hypothesis Testing: The Bayesian Reference Criterion*. Oxford University Press, 1999.

[44] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys*, 143(1):010901, Jul 2015.

[45] Where AIC is defined in nats, rather than the more common demi-nat expression which is twice Eq. 41.

[46] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47., 1977.

[47] J. O. Berger and J.-M. Bernardo. On the development of the reference prior method. Techincal Report 91-15C, Purdue University, 1991.

[48] Robert D. Cousins. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, pages 1–38, 2014.

[49] Colin H. LaMont and Paul A. Wiggins. The lindley paradox: The loss of resolution in bayesian inference. *Unbder review. (arXiv:1610.09433)*, 2017.

[50] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[51] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Chichester: Wiley., 1994.

[52] Leo A Goodman. Some practical techniques in serial number analysis. *Journal of the American Statistical Association*, 49(265):97–112, 1954.