Unification of Field Theory and Maximum Entropy Methods for Learning Probability Densities

Justin B. Kinney
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Bayesian field theory and maximum entropy are two methods for learning smooth probability distributions (a.k.a. probability densities) from finite sampled data. Both methods were inspired by statistical physics, but the relationship between them has remained unclear. Here I show that Bayesian field theory subsumes maximum entropy density estimation. In particular, the most common maximum entropy methods are shown to be limiting cases of Bayesian inference using field theory priors that impose no boundary conditions on candidate densities. This unification provides a natural way to test the validity of the maximum entropy assumption on one’s data. It also provides a closer-fitting nonparametric density estimate when the maximum entropy assumption is rejected.

Maximum entropy (MaxEnt) density estimation is carried out as follows. One first uses the sampled data to estimate the values for a chosen set of moments, e.g., mean and variance. Typically, all moments up to some chosen order are selected. The probability density that matches these moments while having the maximum possible entropy is then adopted as one’s estimate. All other information in the data is discarded. One can therefore think of the MaxEnt estimate as a null hypothesis reflecting the assumption that there is no useful information in the data beyond the values of the specified moments.

In the Bayesian field theory approach, one first defines a prior on the space of continuous probability densities. This prior is formulated using a scalar field theory that favors smooth probability densities over rugged ones. The data are then used to compute a Bayesian posterior, from which a maximum a posteriori (MAP) density estimate is obtained. Although such field theory priors invoke an explicit length scale \( \ell \) when quantifying smoothness, the optimal value for \( \ell \) can be learned from the data in a natural way.

Its scale-free nonparametric Bayesian formulation arguably makes this field theory approach the most principled way to estimate probability densities from sampled data. However, the field theory priors considered in imposed periodic boundary conditions on candidate densities. These boundary conditions limit the types of data sets for which such priors are appropriate. MaxEnt, by contrast, does not impose any boundary conditions on estimated densities.

Here I describe a class of Bayesian field theory priors that have no boundary conditions. These yield density estimates that, like standard MaxEnt estimates, match the first few moments of the data. The MaxEnt estimate itself is recovered when the smoothness length scale \( \ell \) is infinite.

More generally, any MaxEnt density estimate can be recovered from Bayesian field theory in the infinite smoothness limit. One need only choose a field theory prior that defines “smoothness” appropriately. This connection provides a natural way to test the validity of any MaxEnt hypothesis. If field theory density estimation identifies \( \ell = \infty \) as being optimal for one’s data set, the corresponding MaxEnt hypothesis is validated. If instead the optimal \( \ell \) is finite, the MaxEnt hypothesis is rejected and a nonparametric density estimate having lower entropy is obtained.

A predictor-corrector algorithm allows the optimal value for \( \ell \), as well as the corresponding density estimate, to be rapidly computed. Moreover, perturbation theory in the \( \ell \to \infty \) limit reveals an analytic criterion for rejecting the MaxEnt hypothesis. Software for carrying out this analysis in one dimension is available at [https://github.com/jbkinney/14_maxent](https://github.com/jbkinney/14_maxent).

Bayesian field theory – These results are elaborated in the context of one-dimensional density estimation. Suppose we are given \( N \) data points \( x_1, x_2, \ldots, x_N \) sampled from a smooth probability density \( Q_{\text{true}}(x) \) that is confined to an interval of length \( L \). We wish to estimate \( Q_{\text{true}} \) from these data.

Following [7], we represent each density \( Q(x) \) in terms of a real field \( \phi(x) \) via

\[
Q(x) = \frac{e^{-\phi(x)}}{\int dx' e^{-\phi(x')}}.
\]

This parametrization ensures that \( Q \) is positive and normalized. We then adopt a field theory prior on \( \phi \). Specifically, we consider priors of the form

\[
p(\phi|\ell) = \frac{e^{-S_\ell^{0}[\phi]}}{Z_\ell^{0}} \quad \text{where} \quad S_\ell^{0}[\phi] = \int dx \frac{\ell}{2} \left( \partial^2 \phi \right)^2.
\]

PACS numbers: 02.50.-r, 11.10.Lm, 89.70.Cf, 02.60.-x
Here, $S'_0$ plays the role of an action in statistical field theory, $Z'_0 = \int D\phi \, e^{-S'_0[\phi]}$ is the corresponding partition function, $\ell$ is a length scale below which fluctuations in $\phi$ are strongly damped, and $\alpha$ is a positive integer of our choosing. After some calculation [17], the resulting Bayesian posterior on $Q$ is found to be consistent with a posterior on $\phi$ of the form

$$p(\phi|\text{data}, \ell) = \frac{e^{-S_i[\phi]}}{Z_\ell},$$

(3)

where

$$S_i[\phi] = \int dx \left\{ \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2 + NLR\phi + N e^{-\phi} \right\}$$

(4)

is a nonlinear action, $Z_\ell$ is the corresponding partition function, and $R(x) = N^{-1} \sum_{n=1}^N \delta(x-x_n)$ is the raw data density.

Eliminating boundary conditions – The MAP field, here denoted by $\phi_\ell$, minimizes the action $S_i[\phi]$. To obtain a differential equation for $\phi_\ell$, previous work [5–7] imposed periodic boundary conditions on $\phi$ and used integration by parts to derive

$$\ell^{2\alpha} (\partial^\alpha \phi_\ell + NLR - N e^{-\phi_\ell} = 0.$$  

(5)

With the assumed boundary conditions in place, this differential equation has a unique solution. However, imposing these boundary conditions amounts to assuming that $Q_{\text{true}}(x)$ is the same at both ends of the $x$-interval. It is not hard to imagine data sets for which this assumption would be problematic.

This imposition of boundary conditions, however, is unnecessary. From Eq. (4) alone we can derive a differential equation for $\phi_\ell$ that has a unique solution.

To do so, we first define a differential operator $\Delta^\alpha$ by the requirement that

$$\phi \Delta^\alpha \psi \equiv (\partial^\alpha \phi)(\partial^\alpha \psi)$$

(6)

for any two fields $\phi$ and $\psi$. This operator is Hermitian [18] even without boundary conditions on $\phi$ and $\psi$. We refer to $\Delta^\alpha$ as the “bilateral Laplacian of order $\alpha$.”

Setting $\delta S_i/\delta \phi = 0$ then gives the differential equation

$$\ell^{2\alpha} \Delta^\alpha \phi_\ell + NLR - N e^{-\phi_\ell} = 0.$$  

(7)

Any solution of Eq. (7) also satisfies Eq. (5) in the interior of the $x$-interval. However, Eq. (7) also specifies the behavior of $\phi_\ell$ at the boundary in a way that Eq. (5) by itself does not. These extra constraints render the solution of Eq. (7) unique, whereas Eq. (5) is degenerate without the additional assumption of boundary conditions on $\phi_\ell$.

The reason for this difference between the standard Laplacian and the bilateral Laplacian is perhaps clearest in the discrete representation. Let us restrict our attention to $G$ grid points evenly spaced throughout the $x$-interval. The field $\phi_\ell$ becomes a $G$-dimensional vector, and the derivative operator $\partial_G$ becomes a $(G-1) \times G$ matrix having elements $(\partial_G)_{ij} = \frac{\ell^{2\alpha}}{2} (-\delta_{ij} + \delta_{i+1,j})$. Similarly, the standard $\alpha$-order Laplacian $(-1)^\alpha \partial_G^2 \phi$ is given by $(-1)^\alpha \partial_G \cdot 2a \cdot \cdots \partial_G \cdot \partial_G \partial_G \phi$, a $(G - 2\alpha) \times G$ matrix. Eq. (5) therefore provides only $G - 2\alpha$ equations for the $G$ unknown values of $\phi_\ell$ at the grid points. $2\alpha$ boundary conditions are thus needed to obtain a unique solution. By contrast, the $\alpha$-order bilateral Laplacian is represented by the $G \times G$ matrix $\Delta^\alpha_\ell = (\partial_G^2)^{\alpha} \partial_G^2$, where $\partial_G^2 = \partial_G \cdot \cdots \partial_G \partial_G \partial_G \partial_G \partial_G \partial_G \partial_G$. Eq. (7) therefore provides $G$ equations for the $G$ unknown elements of $\phi_\ell$, and has a unique solution without additional constraints.

Connection to maximum entropy – Integrating Eq. (7) gives $\int dx \, e^{-\phi_\ell} = L$, and so the MAP density is $Q_\ell = e^{-\phi_\ell}/L$. Multiplying Eq. (4) on the left by polynomials of order $\alpha - 1$ and integrating reveals that $Q_\ell$ matches the first $\alpha$ moments of the data, i.e.,

$$\int dx Q_\ell x^k = \int dx R x^k, \quad k = 0, 1, \ldots, \alpha - 1.$$  

(8)

This moment matching behavior results from the kernel of $\Delta^\alpha$ begin spanned by polynomials of order $\alpha - 1$. Eq. (8) holds exactly in the discrete representation.

At $\ell = \infty$, the MAP field $\phi_\infty$ is restricted to the kernel of the bilateral Laplacian. The corresponding density thus has the form

$$Q_\infty(x) = \frac{1}{L} \exp \left( -\sum_{k=0}^{\alpha-1} a_k x^k \right),$$  

(9)

where the values of the coefficients $a_k$ are determined by Eq. (5). $Q_\infty$ is therefore identical to the MaxEnt density that matches the first $\alpha$ moments of the data [3]. This equivalence again holds in the discrete representation.

Choosing the length scale – To determine the optimal value for $\ell$, we compute the evidence $p(\text{data}|\ell) = \int D\phi p(\text{data}|\phi) p(\phi|\ell)$ [5, 6, 20, 21]. This quantity is infinitesimal when $\alpha > 1$, due to $p(Q|\ell)$ being an improper prior [22]. However, the evidence ratio $E = p(\text{data}|\ell)/p(\text{data}|\infty)$ is finite for all $\ell > 0$. Using a Laplace approximation, which is valid for large $N$, we find

$$E = e^{S_\infty[\phi_\infty]} - S_i[\phi_\ell] \frac{\text{det}_{\text{ker}}[\mathbf{e}^{-\phi_\infty}]}{\text{det}_{\text{ker}}[L^{2\alpha} \Delta^\alpha]} \frac{\text{det}_{\text{ker}}[L^{2\alpha} \Delta^\alpha + \mathbf{e}^{-\phi_\infty}]}{\eta^{-\alpha}} \frac{\text{det}_{\text{ker}}[L^{2\alpha} \Delta^\alpha]}{\eta^{-\alpha}},$$  

(10)

where $\eta = N(L/\ell)^{2\alpha}$.

By construction, the evidence ratio $E$ approaches unity in the large $\ell$ limit. Whether this limiting value is approached from above or below is relevant to the question of whether the MaxEnt hypothesis is optimal. Using perturbation theory about $\eta = 0$ ($\ell = \infty$), we find that

$$\ln E = K\eta + O(\eta^2),$$  

(11)

where the coefficient $K$ is [24]

$$K = \sum_{i,j} \frac{N\eta_i^2 - z_i^2}{2\lambda_i} + \sum_{i,j} \frac{z_i^2}{2\lambda_i} + \sum_{i,j} \frac{v_i z_{ij} z_{j,k}}{2\lambda_i \zeta_j \zeta_k}.$$  

(12)
Numerical examples – In practice this density estimation procedure is carried out on a grid spanning the interval of interest. First, a predictor-corrector homotopy algorithm [25] is used to compute the value of $\phi_\ell$ over all length scales $\ell$. As in [7], this computation can be performed rapidly and deterministically due to the sparsity of $\Delta_0^\ell$ and the convexity of $S_\ell$. The result is a one-parameter family of moment-matching densities that smoothly interpolates between the MaxEnt density at $\ell = \infty$ and the data histogram at $\ell = 0$ (Fig. 1b). The value of the evidence ratio is then computed over all length scales using Eq. (10). The optimal length scale, which we denote $\ell^*$, is thereby identified (Fig. 1c).

The validity of the MaxEnt assumption depends on the data in hand. Sometimes $\ell^* = \infty$, in which case the MaxEnt hypothesis is deemed optimal. When $\ell^*$ is finite, however, the estimated density $Q_{\ell^*}$ fits the data more closely and has lower entropy than $Q_\infty$ (Figs. 1a,d). This improved fit and reduced entropy reflects the use of additional information in the data beyond the first $\alpha$ moments.

Bayesian field theory can therefore be used to test whether $Q_{\text{true}}$ has a hypothesized functional form. For example, the value of $\ell^*$ that results from using $\alpha = 3$ provides a test of whether $Q_{\text{true}}$ is Gaussian. Indeed, when using $\alpha = 3$ to analyze simulated data drawn from a Gaussian density, $\ell^* = \infty$ was obtained most of the time (Figs. 2a,d). By contrast, when data was drawn from a mixture of two Gaussians, the fraction of data sets yielding $\ell^* = \infty$ decreased sharply as the separation between the two Gaussians was increased (Figs. 2b,c,e,f). Other choices for $\alpha$ can be used to test other functional forms for $Q_{\text{true}}$.

The “empirical Bayes” method used to select $\ell^*$ both here and in previous work [5] is arguably inferior to the fully Bayesian approach of [7]. However, the fully Bayesian approach requires a length scale prior $p(\ell)$ that obscures the nontrivial and potentially useful large $\ell$ behavior of the evidence ratio $E$. In particular, the $K$ coefficient in Eq. (12) might provide a useful way to test the functional form of $Q_{\text{true}}$ when the direct computation of $E$ over all length scales is not feasible. In the simulations performed for Fig. 2, the sign of $K$ performed well as a proxy for the finiteness of $\ell^*$: for Fig. 2d (2e), $K < 0$ was found for $100\%$ ($95\%$) of the $\ell^* = \infty$ data sets, whereas $K > 0$ was found for $96\%$ ($69\%$) of the finite $\ell^*$ data sets; $K > 0$ and finite $\ell^*$ were found for all of the data sets analyzed in Fig. 2f.

Summary and discussion – Bialek et al. [5] showed that probability density estimation can be formulated as a Bayesian inference problem using field theory priors. Among other results, [5] derived the action in Eq. (4) and showed how to use a Laplace approximation of the evidence to select the optimal smoothness length scale [27]. However, periodic boundary conditions were imposed on candidate densities in order to transform the standard Laplacian into a Hermitian operator. The MaxEnt den-
sity typically violates these boundary conditions, and as a result the ability of Bayesian field theory to subsume MaxEnt density estimation went unrecognized in [5] and in follow-up studies [6] [7].

Here we have seen that boundary conditions on candidate densities are unnecessary. The bilateral Laplacian, defined in Eq. (1), is a Hermitian operator that imposes no boundary conditions on functions in its domain, yet is equivalent to the standard Laplacian in the interior of the interval of interest. Using the bilateral Laplacian of various orders to define field theory priors, we recovered standard MaxEnt density estimates in cases where the smoothness length scale was infinite. We also obtained criteria for judging the appropriateness of the MaxEnt hypothesis on individual data sets.

More generally, Bayesian field theories can be constructed for any set of moment-matching constraints. One need only replace the bilateral Laplacian in the above equations with a differential operator that has a kernel spanned by the functions whose mean values one wishes to match to the data. The resulting field theory will subsume the corresponding MaxEnt hypothesis, and thereby allow one to judge the validity of that hypothesis. Analogous approaches for estimating discrete probability distributions can be formulated by replacing integrals over $x$ with sums over states.

The elimination of boundary conditions removes a considerable point of concern with using Bayesian field theory for estimating probability densities. As demonstrated here and in [7], the necessary computations are readily carried out in one dimension. One issue not investigated here – the large $N$ assumption used to compute the evidence ratio – can likely be addressed by using Feynman diagrams in the manner of [8].

How to choose an appropriate prior thus appears to be the primary issue standing in the way of a definitive practical solution to the density estimation problem in one dimension. In the author’s opinion, this issue likely reflects a lingering ambiguity in what “density estimation” means. Different situations may ultimately call for the use of different priors, but understanding which situations call for which priors will require further work.

This field theory approach to density estimation readily generalizes to higher dimensions – at least in principle. Additional care is required in order to construct field theories that do not produce ultraviolet divergences [5], and the best way to do this remains unclear. The need for a very large number of grid points also presents a substantial practical challenge. Grid-free methods, such as those used by [10] [28], may provide a way forward.

I thank Gurinder Atwal, Curtis Callan, William Bialek, and Vijay Kumar for helpful discussions. Support for this work was provided by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

FIG. 2. The optimal estimated density may or may not have maximum entropy. Panels (a-c) show three difference choices for $Q_{\text{true}}$ (black), along with a histogram (gray) of $N = 100$ data points binned at $G = 100$ grid points. $Q_{\text{true}}$ was chosen to be the sum of two equally weighted normal distributions separated by a distance of (a) 0, (b) 2.5, or (c) 5. Panels (d-f) show the evidence ratio curves computed for 100 data sets respectively drawn from the $Q_{\text{true}}$ densities in (a-c). Blue curves indicate $\ell^* = \infty$; orange curves indicate finite $\ell^*$. Titles in (d-f) give the percentage of data sets for which $\ell^* = \infty$ was found.

* Email correspondence to jkinney@cshl.edu

[1] B. Silverman, Density Estimation for Statistics and Data Analysis (Chapman and Hall, 1986).
[2] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley, 1992).
[3] E. T. Jaynes, Phys. Rev. 106, 620 (1957).
[4] L. R. Mead and N. Papanicolaou, J. Math. Phys. 25, 2404 (1984).
[5] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. 77, 4693 (1996).
[6] I. Nemenman and W. Bialek, Phys. Rev. E 65, 026137 (2002).
[7] J. B. Kinney, Phys. Rev. E 90, 011301(R) (2014).
[8] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D 80, 105005 (2009).
[9] J. C. Lemm, *Bayesian Field Theory* (Johns Hopkins, 2003).
[10] T. E. Holy, Phys. Rev. Lett. **79**, 3545 (1997).
[11] V. Periwal, Phys. Rev. Lett. **78**, 4671 (1997).
[12] T. Aida, Phys. Rev. Lett. **83**, 3554 (1999).
[13] D. M. Schmidt, Phys. Rev. E **61**, 1052 (2000).
[14] D. Ormoneit and H. White, Economet. Rev. **18**, 127 (1999).
[15] I. J. Good, Ann. Math. Stat. **34**, 911 (1963).
[16] Integrals over $x$ are restricted to the interval of length $L$.
[17] See Supplemental Material (SM) for a derivation of Eq. (1).
[18] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists: A comprehensive guide*, 7th ed. (Academic Press, 2011).
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, 2006).
[20] D. J. C. MacKay, “Information theory, inference, and learning algorithms,” (Cambridge University Press, 2003) Chap. 28.
[21] V. Balasubramanian, Neural Comput. **9**, 349 (1997).
[22] See SM for a derivation of this result.
[23] The subscripts “row” and “ker” respectively indicate restriction to the row space and kernel of $A^\alpha$; see SM for a derivation of Eq. (10).
[24] See SM for a derivation of Eq. (12).
[25] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction* (Springer, 1990).
[26] This predictor-corrector algorithm is more robust and precise than the integration-based algorithm of [7]. See SM for algorithm details.
[27] See SM for a discussion of related work by Silverman [29].
[28] I. J. Good and R. A. Gaskins, Biometrika **58**, 255 (1971).
[29] B. W. Silverman, Ann. Stat. **10**, 795 (1982).
Supplemental Material:
Unification of Field Theory and Maximum Entropy Methods for
Learning Probability Densities

Justin B. Kinney
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

CONTENTS

Derivation of the action $S_\ell[\phi]$ (Eq. 4) 1

Derivation of the evidence ratio $E$ (Eq. 10) 3

Derivation of the $K$ coefficient (Eq. 12) 4

Step 1: Expansion of $\phi_\ell$ to first order in $\eta$. 5

Step 2: Expansion of $S_\ell[\phi_\ell]$ to first order in $\eta$. 6

Step 3: Expansion of $\ln \det[L^{2\alpha} \Delta^\alpha + \eta e^{-\phi_\ell}]$ to first order in $\eta$. 7

Putting it all together 9

Homotopy algorithm 9

Computing the MaxEnt density 10

Predictor step 10

Corrector step 11

Discussion of Silverman (1982) 11

References 12

DERIVATION OF THE ACTION $S_\ell[\phi]$ (EQ. 4)

Our derivation of the action $S_\ell[\phi]$ in Eq. 4 of the main text is essentially that used in [1]. This derivation is not entirely straight-forward, however, and the details of it have yet to be published.

The prior we consider is

$$p(\phi/\ell) = \frac{1}{Z_\ell^0} e^{-S_\ell^0[\phi]}, \quad S_\ell^0 = \int \frac{dx}{L} \frac{\ell^{2\alpha}}{2} \phi \Delta^\alpha \phi, \quad Z_\ell^0 = \int \mathcal{D}\phi e^{-S_\ell^0[\phi]}. \quad (1)$$

As written, this prior is improper due to the $\alpha$-dimensional kernel of $\Delta^\alpha$. To avoid unnecessary mathematical difficulties, we can render $p(\phi/\ell)$ proper by considering a regularized form of the action

$$S_\ell^0 = \int \frac{dx}{L} \frac{1}{2} \phi \left[ \ell^{2\alpha} \Delta^\alpha + \epsilon \right] \phi, \quad (2)$$

where $\epsilon$ is an infinitesimal positive number. All quantities of interest, of course, should be evaluated in the $\epsilon \to 0$ limit.

The prior over $Q$ that results from Eq. (1) is

$$p(Q/\ell) = \int_{-\infty}^{\infty} d\phi_c p(\phi/\ell)$$

$$= \frac{1}{Z_\ell^0} \int_{-\infty}^{\infty} d\phi_c \exp \left( -\frac{1}{2} \int \frac{dx}{L} \phi \left[ \ell^{2\alpha} \Delta^\alpha + \epsilon \right] \phi \right)$$

$$= \frac{1}{Z_\ell^0} \left\{ \int_{-\infty}^{\infty} d\phi_c \exp \left( -\frac{\epsilon}{2} \frac{d\phi_c^2}{\phi_{nc}} \right) \right\} \exp \left( -\frac{1}{2} \int \frac{dx}{L} \phi_{nc} \left[ \ell^{2\alpha} \Delta^\alpha + \epsilon \right] \phi_{nc} \right)$$

$$= \frac{1}{Z_\ell^0} \sqrt{2\pi} \frac{\epsilon}{\phi_{nc}} \exp \left( -\frac{1}{2} \int \frac{dx}{L} \phi_{nc} \left[ \ell^{2\alpha} \Delta^\alpha + \epsilon \right] \phi_{nc} \right). \quad (6)$$
where $\phi_c$ is the constant Fourier component of $\phi$ and $\phi_{nc}(x)$ is the non-constant component.

The likelihood of $Q$ given the data,

$$p(\text{data}|Q) = \prod_i Q(x_i),$$

(7)

can be expressed in a somewhat non-obvious but nevertheless useful form. Using the identity

$$\frac{1}{a^N} = \frac{N^N}{\Gamma(N)} \int_{-\infty}^{\infty} du e^{-N(u+a-\nu)},$$

(8)

which holds for any positive number $a$ and positive integer $N$, we find that

$$p(\text{data}|Q) = \exp\left( -\sum_i \phi_{nc}(x_i) \right) \frac{1}{L^N} \left[ \int dx e^{-\phi_{nc}(x)} \right]^{-N}$$

$$= \exp\left( -\int dx N R \phi_{nc} \right) \frac{N^N}{L^N \Gamma(N)} \int_{-\infty}^{\infty} d\phi_c \exp\left( -N \left[ \phi_c + \int dx \frac{1}{L} e^{-\phi_{nc}(x)-\phi_c} \right] \right)$$

$$= \frac{N^N}{L^N \Gamma(N)} \int_{-\infty}^{\infty} d\phi_c \exp\left( -\int dx \frac{1}{L} \left\{ N L R \phi + N e^{-\phi} \right\} \right).$$

(9)

(10)

(11)

The prior probability of $Q$ and the data together is therefore given by

$$p(\text{data}, Q|\ell) = p(\text{data}|Q) p(Q|\ell)$$

$$= \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \int_{-\infty}^{\infty} d\phi_c \exp\left( -\int dx \frac{1}{L} \left( N L R \phi - N e^{-\phi} \right) \right) \frac{1}{Z_{\ell}^{2\alpha}} \exp\left( -\int dx \frac{\ell^{2\alpha}}{2} (\partial^a \phi_{nc})^2 \right)$$

$$= \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \int_{-\infty}^{\infty} d\phi_c \exp\left( -\int dx \frac{1}{L} \left\{ \frac{\ell^{2\alpha}}{2} (\partial^a \phi)^2 \right. \right.$$

$$+ N L R \phi + N e^{-\phi} \right)$$

$$= \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \int_{-\infty}^{\infty} d\phi_c e^{-S_{\ell}[\phi]},$$

(12)

(13)

(14)

(15)

where

$$S_{\ell}[\phi] = \int dx \frac{\ell^{2\alpha}}{2} (\partial^a \phi)^2 + N L R \phi + N e^{-\phi}$$

is the action in Eq. 4 of the main text.

The “evidence” for $\ell$ – i.e., the probability of the data given $\ell$ – is therefore given by,

$$p(\text{data}|\ell) = \int Q p(\text{data}, Q|\ell)$$

$$= \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \int Z_{\ell}^{\alpha} Q e^{-S_{\ell}[\phi]}$$

$$= \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} Z_{\ell},$$

(16)

(17)

(18)

(19)

where

$$Z_{\ell} = \int Q e^{-S_{\ell}[\phi]}.$$

(20)

is the partition function corresponding to the action $S_{\ell}$. The posterior distribution over $Q$ is then given by Bayes’s rule:

$$p(Q|\text{data}, \ell) = \frac{p(\text{data}, Q|\ell)}{p(\text{data}|\ell)} = \int_{-\infty}^{\infty} d\phi_c \frac{e^{-S_{\ell}[\phi]}}{Z_{\ell}}.$$ 

(21)

This motivates us to define the posterior distribution over $\phi$ as

$$p(\phi|\text{data}, \ell) \equiv \frac{e^{-S_{\ell}[\phi]}}{Z_{\ell}}.$$ 

(22)
This definition of \( p(\phi|\text{data}, \ell) \) is consistent with the formula for \( p(Q|\text{data}, \ell) \) obtained in Eq. (21), in that

\[
p(Q|\text{data}, \ell) = \int_{-\infty}^{\infty} d\phi_e \, p(\phi|\text{data}, \ell).
\]

(23)

However, Eq. (22) violates Bayes’s rule, since

\[
p(\phi|\text{data}, \ell) \neq \frac{p(\text{data} \mid \phi)}{p(\text{data} \mid \ell)}.
\]

(24)

This is not a problem, however, since \( \phi \) itself is not directly observable; only \( Q \) is observable, and so only Eq. (21) is required for our mathematics to make sense. Put another way, Eq. (22) violates Bayes’s rule only in the way that it specifies the behavior of \( \phi_c \). This constant component of \( \phi \), however, has no effect on \( Q \).

Note that all of the above calculations have been exact; no large \( N \) approximation was used. This contrasts with prior work [2, 3], which used a large \( N \) Laplace approximation to derive the formula for \( S_l[\phi] \). Also note that the regularization parameter \( \epsilon \) has vanished in the formulas for the posterior distributions \( p(Q|\text{data}, \ell) \) and \( p(\phi|\text{data}, \ell) \). However, this parameter still appears in the formula for the evidence \( p(\text{data} \mid \ell) \), both explicitly and implicitly through the value of \( Z \).

**DERIVATION OF THE EVIDENCE RATIO \( E \) (EQ. 10)**

In Eq. (19), we found that the evidence for a length scale \( \ell \) is given by

\[
p(\text{data} \mid \ell) = \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \frac{Z_\ell}{Z_0}.
\]

(25)

We now turn to the task of evaluating the partition functions \( Z_\ell \) and \( Z_0 \), so that we can compute this quantity. Defining \( \Lambda = L^{2\alpha} \Delta^\alpha \) and \( \eta = N(L/\ell)^{2\alpha} \), and using a Laplace approximation to compute the path integral, we get,

\[
Z_\ell = \int \mathcal{D}\phi e^{-S_l[\phi]} \quad \text{(26)}
\]

\[
= e^{-S_l[\phi_e]} \int \mathcal{D}\phi \, \exp\left( -\frac{1}{2} \int \delta\phi(x) \delta\phi(x') \frac{\delta^2 S_\ell}{\partial x \partial x'} \delta\phi(x) \delta\phi(x') \right) \quad \text{(27)}
\]

\[
= e^{-S_l[\phi_e]} \int \mathcal{D}\phi \, \exp\left( -\frac{1}{2} \int \delta\phi \left[ \frac{\ell^{2\alpha} \Delta^{\alpha}}{L^{2\alpha}} + N \epsilon e^{-\phi_i} \right] \delta\phi \right) \quad \text{(28)}
\]

\[
= e^{-S_l[\phi_e]} \int \mathcal{D}\phi \, \exp\left( -\frac{1}{2} \sum_{i,j} \delta\phi_i \left[ \frac{\ell^{2\alpha} \Delta^{\alpha}}{G L^{2\alpha}} \right] \delta\phi_j \right) \quad \text{(29)}
\]

\[
= e^{-S_l[\phi_e]} \left( \frac{\ell^{2\alpha}}{2\pi GL^{2\alpha}} \right)^G \det \left[ \Lambda + \eta e^{-\phi_i} \right]^{-1/2} \quad \text{(30)}
\]

In Eq. (30) we switched to the grid representation, in which \( \delta\phi_i = \delta\phi(x_i), \Lambda_{ij} = L^{2\alpha} \left[ \partial_i^\alpha \partial_j^\alpha \right] G_{ij}, \left[ e^{-\phi_i} \right]_{ij} = \delta_{ij} e^{-\phi_i(x_i)} \), etc. Note that the operator \( \Lambda \) is unitless and has eigenvalues that, if ranked in ascending order, approach limiting values as \( G \to \infty \). Also note that \( \eta \) is unitless. For these reasons, \( \eta \) will emerge as a natural perturbation parameter in the \( \ell \to \infty \) limit.

Evaluating the partition function \( Z_0 \) requires more care because the regularized form of the action \( S_0^\ell \), given in Eq.
(2), has to be used in order for the equations we derive to make sense. We get
\[ Z_\ell^0 = \int D\phi e^{-S_0^\ell[\phi]} \]
(32)
\[ = \int D\phi \exp \left( -\frac{1}{2} \int \frac{dx}{L} \phi \left( \ell^{2\alpha} \Delta^\alpha + \epsilon \right) \phi \right) \]
(33)
\[ = \int D\phi \exp \left( -\frac{1}{2} \int \frac{dx}{L} \phi \left( \ell^{2\alpha} \right) \left[ \Lambda + \frac{\eta}{N} \right] \phi \right) \]
(34)
\[ = \int D\phi \exp \left( -\frac{1}{2} \sum_{i,j} \phi_i \left( \ell^{2\alpha} \right) \left[ \Lambda + \frac{\eta}{N} \right] \phi_j \right) \]
(35)
\[ = \left\{ \frac{\ell^{2\alpha}}{2\pi G L^{2\alpha}} \right\}^G \det \left[ \Lambda + \frac{\eta}{N} \right]^{-1/2} \]
(36)
\[ = \left\{ \frac{\ell^{2\alpha}}{2\pi G L^{2\alpha}} \right\}^G \left( N^{-\alpha} \eta^\alpha e^{\alpha \det \text{row}[\Lambda]} \right)^{-1/2} \]
(37)

As in the main text, the subscript “row” on the determinant denotes restriction to the row space of \( \Lambda \). Putting these values for \( Z_\ell \) and \( Z_\ell^0 \) back into Eq. (25), we get
\[ p(\text{data}|\ell) = \frac{N^N}{L^N \Gamma(N)} e^{-S_\ell[\phi_\ell]} \sqrt{\frac{2\pi}{\epsilon}} \sqrt{\frac{N^{-\alpha} \eta^\alpha \det \text{row}[\Lambda]}{\det [\Lambda + \epsilon \phi_\ell]}} \]
(38)
\[ = \epsilon^{\frac{N-2}{4}} \sqrt{\frac{2\pi N^N}{L^N \Gamma(N)}} e^{-S_\ell[\phi_\ell]} \sqrt{\frac{\det \text{row}[\Lambda]}{\eta^{-\alpha} \det [\Lambda + \epsilon \phi_\ell]}} \]
(39)

Although both \( Z_\ell \) and \( Z_\ell^0 \) depend strongly on the number of grid points \( G \), the value for the evidence does not. The evidence does, however, depend on the regularization parameter \( \epsilon \); specifically, it is proportional to \( \epsilon^{\frac{N-2}{2}} \). This is the basis for the claim in the main text that the evidence vanishes for \( \alpha > 1 \).

In the large \( \ell \) limit, \( \eta \to 0 \), and so
\[ \det [\Lambda + \epsilon \phi_\ell] \to \eta^\alpha \det \text{ker} \left[ e^{-\phi_\ell} \right] \]
(40)
where “ker” denotes restriction to the kernel of \( \Lambda \). As a result,
\[ p(\text{data}|\infty) = \epsilon^{\frac{N-2}{4}} \sqrt{\frac{2\pi N^N}{L^N \Gamma(N)}} e^{-S_\infty[\phi_\infty]} \sqrt{\frac{\det \text{ker}[e^{-\phi_\infty}]}{\eta^{-\alpha} \det [\Lambda + \epsilon \phi_\ell]}} \]
(41)

The evidence ratio is therefore given by
\[ E = \frac{p(\text{data}|\ell)}{p(\text{data}|\infty)} = e^{S_\infty[\phi_\infty] - S_\ell[\phi_\ell]} \sqrt{\frac{\det \text{ker}[e^{-\phi_\ell}]}{\eta^{-\alpha} \det [\Lambda + \epsilon \phi_\ell]}} \]
(42)

Note that \( E \), unlike the evidence itself, does not depend on the regularization parameter \( \epsilon \).

**DERIVATION OF THE K COEFFICIENT (EQ. 12)**

The goal of this section is to clarify the large length scale behavior of
\[ \ln E = S_\infty[\phi_\infty] - S_\ell[\phi_\ell] + \frac{1}{2} \ln \left\{ \frac{\det \text{ker}[e^{-\phi_\ell}]}{\eta^{-\alpha} \det [\Lambda + \epsilon \phi_\ell]} \right\} \]
(43)
To do this we expand \( \ln E \) as a power series in \( \eta \) about \( \eta = 0 \). We will find that
\[ \ln E = K \eta + O(\eta^2) \]
(44)
where \( K \) is a nontrivial coefficient that can be either positive or negative, depending on the data. We carry out this expansion in three steps:
1. Expand $\phi_\ell$ to first order in $\eta$.
2. Expand $S_\ell[\phi_\ell]$ to first order in $\eta$.
3. Expand $\ln \det[\Lambda + \eta e^{-\phi_\ell}]$ to first order in $\eta$.

In what follows we will sometimes use the ket notion of quantum mechanics, primarily as a notational convenience. For any two functions $f$ and $g$ and any Hermitian operator $H$, we define

$$\langle f | H | g \rangle = \int \frac{dx}{L} f^* H g.$$  \hspace{1cm} (45)

Note the convention of dividing by $L$ in the inner product integral. This allows us to take inner products without altering units. The eigenvalues $\lambda_i$ and corresponding eigenfunctions $\psi_i(x)$ defined in the main text satisfy

$$\langle \psi_i | \psi_j \rangle = \delta_{ij} \text{ for all } i,j.$$  \hspace{1cm} (46)

and

$$\langle \psi_i | \Lambda | \psi_j \rangle = \delta_{ij} \zeta_j \text{ for } i,j \leq \alpha.$$  \hspace{1cm} (47)

for some positive numbers $\zeta_j$. We will also make use of the following indexed quantities,

$$v_i = L \langle \psi_i | Q_\infty - R \rangle = \int dx (Q_\infty - R) \psi_i \hspace{1cm} (48)$$

$$z_{ij} = L \langle \psi_i | Q_\infty | \psi_j \rangle = \int dx Q_\infty \psi_i \psi_j \hspace{1cm} (49)$$

$$z_{ijk} = L \langle \psi_i | Q_\infty \psi_j | \psi_k \rangle = \int dx Q_\infty \psi_i \psi_j \psi_k.$$ \hspace{1cm} (50)

**Step 1: Expansion of $\phi_\ell$ to first order in $\eta$.**

We begin by representing $\phi_\ell$ as a power series in $\eta$. Abusing notation somewhat, we write

$$\phi_\ell = \phi_\infty + \eta \phi_1 + \eta^2 \phi_2 + \cdots.$$ \hspace{1cm} (51)

Plugging this expansion into the equation of motion,

$$\Lambda \phi_\ell + \eta (LR - e^{-\phi_\ell}) = 0,$$ \hspace{1cm} (52)

then collecting terms of equal order in $\eta$, we get,

$$\Lambda \phi_\infty + \eta (\phi_1 + LR - e^{-\phi_\infty}) + \eta^2 (\Lambda \phi_2 + e^{-\phi_\infty} \phi_1) + \cdots = 0.$$ \hspace{1cm} (53)

At lowest order in $\eta$, we recover

$$\Lambda \phi_\infty = 0.$$ \hspace{1cm} (54)

This just reflects the restriction of $\phi_\infty$ to the kernel of $\Lambda$. At first order we get

$$\Lambda \phi_1 + LR - e^{-\phi_\infty} = 0,$$ \hspace{1cm} (55)

which we will proceed to investigate. To compute $\phi_1$, we first write it in terms of the eigenfunctions of $\Lambda$.

$$\phi_1(x) = \sum_i A_i \psi_i(x).$$ \hspace{1cm} (56)

Taking the inner product of Eq. (55) with $\psi_i$, we get

$$\lambda_i A_i + L \langle \psi_i | R - Q_\infty \rangle = 0.$$ \hspace{1cm} (57)

Since $\lambda_i > 0$ for $i > \alpha$,

$$A_i = \frac{\psi_i}{\lambda_i}, \hspace{1cm} i > \alpha.$$ \hspace{1cm} (58)
As yet we have no information about the value of $A_i$ for $i \leq \alpha$. For this we need to consider the second order terms in Eq. (53). At second order in $\eta$, we obtain,

$$\Lambda \phi_2 + e^{-\phi_0} \phi_1 = 0. \quad (59)$$

Dotting this with $\psi_j$ for $j \leq \alpha$, we obtain

$$0 = \langle \psi_j | e^{-\phi_0} | \phi_1 \rangle = \sum_i A_i \langle \psi_i | e^{-\phi_0} | \psi_j \rangle = A_j \zeta_j + \sum_{i > \alpha}^{N \nu_2} \frac{\nu_i \zeta_j}{\lambda_i}. \quad (60)$$

The coefficients left unspecified in Eq. (58) are therefore given by,

$$A_j = - \sum_{i > \alpha}^{N \nu_2} \frac{\nu_i \zeta_j}{\lambda_i}, \quad j \leq \alpha. \quad (61)$$

This completes our computation of $\phi_\ell$ to first order in $\eta$:

$$\phi_\ell = \phi_\infty + \eta \phi_1 + O(\eta^2) \quad \text{where} \quad \phi_1(x) = \sum_{i \leq \alpha}^{N \nu_2} \frac{\nu_i}{\lambda_i} \psi_i(x) - \sum_{j \leq \alpha}^{N \nu_2} \frac{\nu_i \zeta_j}{\lambda_i} \psi_j(x). \quad (62)$$

**Step 2: Expansion of $S_\ell[\phi_\ell]$ to first order in $\eta$.**

The value of the action $S_\ell$, evaluated at its minimum $\phi_\ell$, is

$$S_\ell[\phi_\ell] = \int \frac{dx}{L} \left\{ \frac{\ell^2}{2} \phi_\ell \Delta^\alpha \phi_\ell + N \Lambda R \phi_\ell + N e^{-\phi_\ell} \right\} \quad (63)$$

$$= N \int \frac{dx}{L} \left\{ \frac{1}{2 \eta} \phi_\ell \Lambda \phi_\ell + L \Lambda \phi_\ell - e^{-\phi_\ell} \right\} \quad (64)$$

$$= N \int \frac{dx}{L} \left\{ \frac{1}{2 \eta} (\phi_\infty + \eta \phi_1) \Lambda (\phi_\infty + \eta \phi_1) + L \Lambda (\phi_\infty + \eta \phi_1) + L Q_\infty - \eta L Q_\infty \phi_1 \right\} + O(\eta^2) \quad (65)$$

$$= S_\infty[\phi_\infty] + N \eta \left\{ \frac{1}{2} \langle \phi_1 | \Lambda | \phi_1 \rangle - \langle \phi_1 | Q_\infty - R \rangle \right\} + O(\eta^2). \quad (66)$$

The first inner product term in Eq. (66) is

$$\frac{1}{2} \langle \phi_1 | \Lambda | \phi_1 \rangle = \frac{1}{2} \sum_{i,j}^{N \nu_2} A_i A_j \lambda_j \langle \psi_i | \psi_j \rangle \quad (67)$$

$$= \frac{1}{2} \sum_{i,j \geq \alpha}^{N \nu_2} \frac{\nu_i \nu_j}{\lambda_i \lambda_j} \delta_{ij} \quad (68)$$

$$= \frac{1}{2} \sum_{i \geq \alpha}^{N \nu_2} \frac{\nu_i^2}{\lambda_i}. \quad (69)$$

Here we have used the fact that $\lambda_k = 0$ for $k \leq \alpha$ to restrict the sum to $i,j > \alpha$ in Eq. (68). The second inner product term in Eq. (66) is

$$- \langle \phi_1 | Q_\infty - R \rangle = - \sum_i^{N \nu_2} A_i \langle \psi_i | Q_\infty - R \rangle \quad (70)$$

$$= - \sum_{i \geq \alpha}^{N \nu_2} \frac{\nu_i}{\lambda_i} \quad (71)$$

$$= - \sum_{i \geq \alpha}^{N \nu_2} \frac{\nu_i^2}{2 \lambda_i}. \quad (72)$$

Here, the sum over $i$ is restricted to $i > \alpha$ due to the moment-matching condition $\langle \psi_i | Q_\infty - R \rangle = 0$ for $k \leq \alpha$. To first order in $\eta$, we therefore find the rather simple result,

$$S_\ell[\phi_\ell] - S_\infty[\phi_\infty] = - \eta \sum_{i \geq \alpha}^{N \nu_2} \frac{\nu_i^2}{2 \lambda_i}. \quad (73)$$
We pause to note that this moment matching condition has interesting implications in the context of perturbation theory. For $k \leq \alpha$,

$$0 = \langle \psi_k | R - Q \rangle = \left\{ \psi_k | R - Q_{\infty} + \eta Q_{\infty} \phi_1 + O(\eta^2) \right\}. \quad (74)$$

This expression must vanish at each order in $\eta$, and so

$$0 = \langle \psi_k | Q_{\infty} \phi_1 \rangle = \sum_{\ell, j \alpha} \frac{\nu_i \nu_{\ell j}}{\lambda_i \zeta_{j}}. \quad (75)$$

In a similar manner, considering the higher-order terms of Eq. (74) leads to nontrivial constraints on the higher-order corrections to $\phi_\ell$. We don’t make use of this finding in what follows, but it seems worth noting.

**Step 3: Expansion of $\ln \det \left[ L^{2\alpha} \Delta^\alpha + \eta e^{-\phi_\ell} \right]$ to first order in $\eta$.**

We first outline how we will go about computing

$$\ln \det \Gamma \quad \text{where} \quad \Gamma = \Delta + \eta e^{-\phi_\ell}. \quad (76)$$

Computing this quantity requires computing the eigenvalues of $\Gamma$. We will use $\gamma_i$ and $\rho_i$ to denote the eigenvalues and corresponding orthonormal eigenfunctions of $\Gamma$, i.e.,

$$\Gamma \rho_i = \gamma_i \rho_i. \quad (77)$$

Our primary task is to compute each eigenvalue $\gamma_i$ as a power series in $\eta$:

$$\gamma_i = \lambda_i + \eta \gamma_i^1 + \eta^2 \gamma_i^2 + \cdots. \quad (78)$$

This task, as we will see, also requires computing the eigenfunctions $\rho_i$ as power series in $\eta$:

$$\rho_i = \psi_i + \eta \rho_i^1 + \eta^2 \rho_i^2 + \cdots. \quad (79)$$

As usual, it will help to expand $\rho_i^1$ in the eigenfunctions of $\Lambda$:

$$\rho_i^1(x) = \sum_j B_{ij} \psi_j(x). \quad (80)$$

Keeping in mind that $\lambda_i > 0$ for $i > \alpha$, and $\lambda_j = 0$ for $j \leq \alpha$, we see that

$$\ln \det \Gamma = \ln \prod_i \gamma_i \quad (81)$$

$$= \ln \left\{ \prod_{j=\alpha} \gamma_j \left( 1 + \eta \gamma_j^1 + O(\eta^2) \right) \right\} + \ln \left\{ \prod_{i=\alpha} \lambda_i \left( 1 + \eta \gamma_i^1 + O(\eta^2) \right) \right\} \quad (82)$$

$$= \ln \left\{ \eta^{\alpha} \left( \prod_{j=\alpha} \gamma_j \right) \left( \prod_{i=\alpha} \lambda_i \right) \right\} + \eta \left\{ \sum_{i=\alpha} \gamma_i^1 \lambda_i + \sum_{j=\alpha} \gamma_j^2 \right\} + O(\eta^2). \quad (83)$$

So while the larger eigenvalues of $\Gamma$ need only be computed to first order in $\eta$, the $\alpha$ lowest eigenvalues must actually be computed to second order in $\eta$. Performing this second order calculation will require that we also compute the eigenfunctions $\rho_i$ to first order in $\eta$.

Plugging Eq. (78) and Eq. (79) into Eq. (77) and collecting terms by order in $\eta$, we get

$$0 = (\Lambda + \eta e^{-\phi_\infty} - \eta^2 e^{-\phi_\infty} \phi_1 + \cdots) (\psi_i + \eta \rho_i^1 + \eta^2 \rho_i^2 + \cdots) - (\lambda_i + \eta \gamma_i^1 + \eta^2 \gamma_i^2 + \cdots) (\psi_i + \eta \rho_i^1 + \eta^2 \rho_i^2 + \cdots) \quad (84)$$

$$= (\Lambda \psi_i - \lambda_i \psi_i) + \eta (\Lambda \rho_i^1 + e^{-\phi_\infty} \psi_i - \lambda_i \rho_i^1 - \gamma_i^1 \psi_i) + \eta^2 \left( \Lambda \rho_i^2 + e^{-\phi_\infty} \rho_i^1 - e^{-\phi_\infty} \phi_1 \psi_i - \lambda_i \rho_i^2 - \gamma_i^1 \rho_i^1 - \gamma_i^2 \psi_i \right) + O(\eta^3). \quad (85)$$

From the zeroth order term, we recover the eigenvalue equation $\Lambda \psi_i = \lambda_i \psi_i$, which we already knew. From the first order term, we get,

$$0 = \Lambda \rho_i^1 + e^{-\phi_\infty} \psi_i - \lambda_i \rho_i^1 - \gamma_i^1 \psi_i. \quad (86)$$
Dotting this with $\psi_k$ gives

$$0 = \{\psi_k|A|\rho_i^1\} + \{\psi_k|e^{-\phi_{\infty}}|\psi_i\} - \lambda_i \{\psi_k|\rho_i^1\} - \gamma_i^1 \{\psi_k|\psi_i\}$$  \hspace{1cm} (87)

$$= \sum_j \lambda_k B_j^i \{\psi_k|\psi_j\} + z_{ki} - \lambda_i \sum_j B_j^i \{\psi_k|\psi_j\} - \gamma_i^1 \delta_{ik}$$  \hspace{1cm} (88)

$$= \lambda_k B_k^i + z_{ik} - \lambda_i B_i^k - \gamma_i^1 \delta_{ik}$$  \hspace{1cm} (89)

$$= (\lambda_k - \lambda_i) B_k^i + z_{ik} - \gamma_i^1 \delta_{ik}. \hspace{1cm} (90)$$

Choosing $k = i$, we recover the standard first order correction to the eigenvalues,

$$\gamma_i^1 = z_{ii} \text{ for all } i, \hspace{1cm} (91)$$

in particular,

$$\gamma_j^1 = \zeta_j, \text{ } j \leq \alpha. \hspace{1cm} (92)$$

Moreover, for $i \leq \alpha$, Eq. (90) gives

$$\rho_i^1(x) = \sum_{j > \alpha} B_j^i \psi_j(x) = - \sum_{j > \alpha} \frac{z_{ij}}{\lambda_j} \psi_j(x). \hspace{1cm} (93)$$

From the second order term in Eq. (85), we get,

$$0 = \Lambda \rho_i^2 + e^{-\phi_{\infty}} \rho_i^1 - e^{-\phi_{\infty}} \phi_1 \psi_i - \lambda_i \rho_i^2 - \gamma_i^1 \rho_i^1 - \gamma_i^2 \psi_i. \hspace{1cm} (94)$$

Focusing on $i \leq \alpha$ and dotting with $\psi_i$,

$$0 = \{\psi_i|A|\rho_i^2\} + \{\psi_i|e^{-\phi_{\infty}}|\rho_i^1\} - \{\psi_i|e^{-\phi_{\infty}} \phi_1|\psi_i\} - \lambda_i \{\psi_i|\rho_i^1\} - \gamma_i^1 \{\psi_i|\rho_i^1\} - \gamma_i^2 \{\psi_i|\psi_i\}$$

$$= \{\psi_i|e^{-\phi_{\infty}}|\rho_i^1\} - \{\psi_i|e^{-\phi_{\infty}} \phi_1|\psi_i\} - \gamma_i^1 \{\psi_i|\rho_i^1\} - \gamma_i^2. \hspace{1cm} (95)$$

Using Eq. (93) to evaluate this expression gives

$$\gamma_i^2 = \{\psi_i|e^{-\phi_{\infty}}|\rho_i^1\} - \{\psi_i|e^{-\phi_{\infty}} \phi_1|\psi_i\}$$

$$= - \sum_{j > \alpha} \frac{z_{ij}}{\lambda_j} \{\psi_i|e^{-\phi_{\infty}}|\psi_i\} - \sum_{j > \alpha} \frac{v_j}{\lambda_j} \{\psi_i|e^{-\phi_{\infty}} \psi_j|\psi_i\} + \sum_{j > \alpha \leq k \leq \alpha} \frac{v_j z_{jk}}{\lambda_j \gamma_k} \{\psi_i|e^{-\phi_{\infty}} \psi_k|\psi_i\}$$

$$= - \sum_{j > \alpha} \frac{z_{ij}}{\lambda_j} - \sum_{j > \alpha} \frac{v_j z_{ij}}{\lambda_j} + \sum_{j > \alpha \leq k \leq \alpha} \frac{v_j z_{jk} z_{ii}}{\lambda_j \gamma_k}. \hspace{1cm} (96)$$

Switching indices from $i \leftrightarrow j$, so that $i > \alpha$ and $j, k \leq \alpha$, we thus find the required second order correction to $\gamma_j$:

$$\gamma_j^2 = - \sum_{i > \alpha} \frac{z_{ii}^2 + v_i z_{ij}^j}{\lambda_i} + \sum_{j > \alpha \leq k \leq \alpha} \frac{v_i z_{ik}^i z_{jj}^k}{\lambda_i \gamma_k}, \text{ } j \leq \alpha. \hspace{1cm} (100)$$

We can now compute $\ln \det \Gamma$. Plugging the values for $\gamma_i^1$ and $\gamma_i^2$ into Eq. (83), and using

$$\prod_{j \leq \alpha} \zeta_j = \det_{\text{ker}} \{e^{-\phi_{\infty}}\}, \hspace{1cm} \prod_{i > \alpha} \lambda_i = \det_{\text{row}}\{A\}, \hspace{1cm} (101)$$

we get what we set out to find:

$$\ln \det\{A + \eta^{-\phi_{\infty}}\} = \ln \left\{\eta^\alpha \det_{\text{ker}}\{e^{-\phi_{\infty}}\} \det\{A\}\right\} + \eta \left\{\sum_{i > \alpha} \frac{z_{ii}}{\lambda_i} - \sum_{j > \alpha \leq k \leq \alpha} \frac{z_{ij}^2 + v_i z_{ij}^j}{\lambda_i \gamma_j} + \sum_{j > \alpha \leq k \leq \alpha} \frac{v_i z_{ik}^i z_{jj}^k}{\lambda_i \gamma_k}\right\} + O(\eta^2). \hspace{1cm} (102)$$
FIG. S1. Illustration of the predictor-corrector homotopy algorithm used for the computations reported in the main text. (a) The MAP curve (brown) is approximated using finite set of densities \{R, Q_ℓ, \ldots, Q_ℓ_0, Q_ℓ_1, Q_ℓ_2, \ldots, Q_ℓ_m, Q_∞\}. First the MAP density at an intermediate length scale \(ℓ_0\) is computed. A predictor-corrector algorithm is then used to extend the set of MAP densities outward to larger and to smaller values of \(ℓ\). These \(ℓ\) values are chosen so that neighboring MAP densities lie within a geodesic distance of \(\leq ϵ\) of each other. (b) Each step \(Q_{ℓ_n} \to Q_{ℓ_n'}\) has two parts. First, a predictor step (magenta) computes a new length scale \(ℓ_n\) and an approximation \(Q_{ℓ_n}^{(0)}\) of \(Q_{ℓ_n}\). A series of corrector steps \(Q_{ℓ_n}^{(0)} \to Q_{ℓ_n}^{(1)} \to Q_{ℓ_n}^{(2)} \cdots\) (orange) then converges to \(Q_{ℓ_n}\).

Putting it all together

Putting together our results from Eq. (73) and Eq. (102), and switching indices \(j \leftrightarrow k\) in the last term of Eq. (102) just to smooth out the notation, we get

\[
\ln E = - (S_ι(ℓ) - S_∞(ϕ_∞)) - \frac{1}{2} \ln \left\{ \frac{\det[Λ + ηe^{-ϕ_∞}]}{η^α \det[\ker(e^{-ϕ_∞})] \det[\text{row}(Λ)]} \right\} \tag{103}
\]

\[
= η \left\{ \sum_{i>α} \frac{Nv_i^2 - z_{ii}}{2\lambda_i} + \sum_{i>α, j≤α} \frac{z_{ij}^2 + v_jz_{jj}}{2λ_jζ_j} - \sum_{i>α, j,k≤α} \frac{v_iz_{ij}z_{jj}}{2λ_iζ_jζ_k} \right\} + O(η^2) \tag{104}
\]

\[
= Kη + O(η^2), \tag{105}
\]

where the \(K\) coefficient is

\[
K = \sum_{i>α} \frac{Nv_i^2 - z_{ii}}{2\lambda_i} + \sum_{i>α, j≤α} \frac{z_{ij}^2 + v_jz_{jj}}{2λ_iζ_j} - \sum_{i>α, j,k≤α} \frac{v_iz_{ij}z_{jj}}{2λ_iζ_jζ_k}. \tag{106}
\]

**HOMOTOPY ALGORITHM**

In the space of probability densities, the set of MAP densities forms a one-parameter curve extending from the MaxEnt density at \(ℓ = \infty\) to the data histogram at \(ℓ = 0\). The algorithm used to perform the computations shown in Figs. 1 and 2 of the main text approximates this “MAP curve” by computing the MAP densities at a finite set of length scales. This set of length scales extends from \(ℓ = 0\) to \(ℓ = \infty\) (inclusively), and are chosen so that, at any two neighboring length scales \(ℓ\) and \(ℓ'\), the MAP densities \(Q_ℓ\) and \(Q_{ℓ'}\) are sufficiently similar. By computing this “string of beads” approximation to the MAP curve, the homotopy algorithm allows one to compute the evidence ratio \(E\) over all length scales.

This algorithm proceeds as follows. First, the MaxEnt density \(Q_∞\) is computed essentially as described by Ormoneit and White [4]. The details of this computation are reported below.

Next, the MAP density \(Q_{ℓ_0}\) is computed at an intermediate length scale \(ℓ_0 = L/√G\). This density, \(Q_{ℓ_0}\), serves as the starting point for a predictor-corrector homotopy algorithm [5] that traces the MAP curve from \(ℓ_0\) towards both larger and smaller length scales (Fig. S1a). The result is a finite set of length scales \(\{0, ℓ_0, ℓ_1, ℓ_2, \ldots, ℓ_m, ℓ_∞\}\) chosen so that MAP densities at any two neighboring length scales \(ℓ\) and \(ℓ'\) satisfy

\[
D_{geo}(Q_ℓ, Q_{ℓ'}) \leq ϵ, \tag{107}
\]
where $D_{geo}$ is the geodesic distance measure [1, 6] and $\epsilon$ is a small user-specified distance. The value $\epsilon = 10^{-2}$ was used for the computations reported in the main text.

Each step $Q_{\ell_n'} \rightarrow Q_{\ell_n}$ is accomplished in two parts (Fig. S1b). First, a “predictor step” is used to compute the new length scale $\ell_n$ as well as an approximation $Q_{\ell_n}^{(0)}$ to the corresponding MAP density $Q_{\ell_n}$. A series of “corrector steps” is then used to compute a corresponding series of densities $Q_{\ell_n}^{(1)}, Q_{\ell_n}^{(2)}$, . . . that converges to $Q_{\ell_n}$. These predictor and corrector steps are described in more detail below.

**Computing the MaxEnt density**

We saw in the main text that $\phi_\infty$ is a polynomial of order $\alpha - 1$, i.e.,

$$\phi_\infty(x) = \sum_{i=0}^{\alpha-1} a_i x^i.$$  \hfill (108)

The problem of computing the MaxEnt density $Q_\infty$ therefore reduces to finding the vector of coefficients $a = (a_0, a_1, \ldots, a_{n-1})$ that minimizes the action

$$S_\infty(a) = N \int \frac{dx}{L} \left( LR \sum_{i=0}^{\alpha-1} a_i x^i + \exp \left[ - \sum_{i=0}^{\alpha-1} a_i x^i \right] \right).$$  \hfill (109)

As suggested by Ormoneit and White [4], we solve this optimization problem using the Newton-Raphson algorithm with backtracking. Starting at $a_0 = 0$, we iterate

$$a^n \rightarrow a^{n+1} = a^n + \gamma_n \delta a^n$$  \hfill (110)

where $0 < \gamma_n \leq 1$ and the vector $\delta a^n$ is the solution to

$$\sum_{j=0}^{\alpha-1} \frac{\partial^2 S}{\partial a_i \partial a_j} \left|_{a^n} \right. \delta a^n_j = - \frac{\partial S}{\partial a_i} \left|_{a^n} .$$  \hfill (111)

From Eq. (109),

$$\frac{\partial S}{\partial a_i} = N \mu_i - N \int \frac{dx}{L} x^i \exp \left[ - \sum_{k=1}^{\alpha-1} a_k x^k \right] \quad \text{and} \quad \frac{\partial^2 S}{\partial a_i \partial a_j} = N \int \frac{dx}{L} x^{i+j} \exp \left[ - \sum_{k=1}^{\alpha-1} a_k x^k \right],$$  \hfill (112)

where $\mu_i = \int dx R x^i$ is the $i$’th moment of the data. The Hessian matrix $\frac{\partial^2 S}{\partial a_i \partial a_j}$ is positive definite at all $a$, so a unique solution for $\delta a^n$ can always be found [7]. The scalar $\gamma_n$ is then chosen so that

$$S_\infty(a^n + \gamma_n \delta a^n) - S_\infty(a^n) < \gamma_n \left( \frac{\alpha}{2} \sum_{i=0}^{\alpha-1} \frac{\partial S}{\partial a_i} \left|_{a^n} \right. \right) \delta a^n.$$  \hfill (113)

Specifically, $\gamma_n$ is first set to 1, then is reduced by factors of $\frac{1}{2}$ until Eq. (113) is satisfied. This “dampening” of the Newton-Raphson method is sufficient to guarantee convergence [4, 8]. This algorithm is terminated when the magnitude of the change in the action $|S_\infty(a^{n+1}) - S_\infty(a^n)|$ falls below a specified tolerance.

**Predictor step**

The predictor step computes a change $\ell \rightarrow \ell'$ in the length scale, as well as an approximation to the corresponding change $\phi_\ell \rightarrow \phi_{\ell'}$ in the MAP field. Specifically, the predictor step returns a scalar $\delta t$ and a function $\rho(x)$ such that,  

$$t' = t + \delta t \quad \text{and} \quad \phi_{\ell'}(x) \approx \phi_\ell(x) + \rho(x) \delta t,$$  \hfill (114)

where $t = \ln \eta$ is a numerically convenient reparametrization of the length scale $\ell$. To determine the function $\rho$, we examine the equation of motion at $\ell'$:

$$0 = \Delta \phi_{\ell'} + \eta \left( (LR - e^{-\phi_{\ell'}}) \right) \right)$$  \hfill (115)

$$= \Delta (\phi_\ell + \rho \delta t) + \eta(1 + \delta t)(LR - e^{-\phi_\ell} + e^{-\phi_\ell'} + O((\delta t)^2))$$  \hfill (116)

$$= \Delta (\phi_\ell + \eta(LR - e^{-\phi_\ell}) + \delta t \left[ \left[ \Lambda + \eta e^{-\phi_\ell} \right] + \eta(LR - e^{-\phi_\ell'}) \right] + O((\delta t)^2).$$  \hfill (117)
The \( O(\delta t) \) term must vanish, and thus we obtain a linear equation for \( \rho \):

\[
\left[ \Lambda + \eta e^{-\phi_\ell} \right] \rho = \eta \left( e^{-\phi_\ell} - LR \right).
\] (118)

The scalar \( \delta t \) is then chosen to satisfy the distance criterion,

\[
\epsilon^2 \approx D^2_\text{geo}(Q_\ell, Q_{\ell'}) \approx \int dx \frac{\delta Q^2}{Q_\ell} \approx (\delta t)^2 \int dx Q_\ell \rho^2.
\] (119)

We therefore adopt

\[
\delta t = \pm \frac{\epsilon}{\sqrt{\int dx Q_\ell \rho^2}},
\] (120)

with the sign of \( \delta t \) chosen according to the direction we wish to traverse the MAP curve.

**Corrector step**

The purpose of the corrector step is to accurately solve the equation of motion,

\[
\Lambda \phi_\ell + \eta (LR - e^{-\phi_\ell}) = 0,
\] (121)

at fixed length scale \( \ell \). This step is used initially to compute \( Q_{\ell_0} \) at the starting length scale \( \ell_0 \). It is also used to hone in on the MAP density at each new length scale chosen by the predictor step of the homotopy algorithm.

As with the computation of the MaxEnt density, this corrector step uses the Newton-Raphson algorithm with backtracking. Starting from a hypothesized field \( \phi^{(0)} \) (e.g., returned by the predictor algorithm), the iteration

\[
\phi^{(n)} \to \phi^{(n+1)} = \phi^{(n)} + \gamma_n \delta \phi^{(n)}
\] (122)

is repeatedly performed. The function \( \delta \phi^{(n)} \) and scalar \( \gamma_n \) are chosen so that this iteration converges to the true field \( \phi_\ell \). To derive the field perturbation \( \delta \phi^{(n)} \), we plug this formula for \( \phi^{(n+1)} \) into the equation of motion. Keeping only terms of order \( \delta \phi^{(n)} \) or less, we see that

\[
\Lambda \phi^{(n)} + \Lambda \delta \phi^{(n)} + \eta (LR - e^{-\phi^{(n)}} + e^{-\phi^{(n)}} \delta \phi^{(n)}) = 0.
\] (123)

The field perturbation \( \delta \phi^{(n)} \) is therefore the solution to the linear equation

\[
\left[ \Lambda + \eta e^{-\phi^{(n)}} \right] \delta \phi^{(n)} = \eta \left( e^{-\phi^{(n)}} - LR \right) - \Lambda \phi^{(n)}.
\] (124)

As before, \( \gamma_n \) is chosen so that so that the action decreases by a respectable amount compared to what is expected from the linear approximation, i.e.,

\[
S_{\ell} \left[ \phi^{(n)} + \gamma_n \delta \phi^{(n)} \right] - S_{\ell} \left[ \phi^{(n)} \right] < \frac{\gamma_n}{2} \int dx \left. \frac{\delta S_{\ell}}{\delta \phi(x)} \right|_{\phi^{(n)}} \delta \phi^{(n)}.
\] (125)

This iterative process is terminated when the magnitude of the change in the action, \( |S_{\ell} \left[ \phi^{(n+1)} \right] - S_{\ell} \left[ \phi^{(n)} \right]| \) falls below a specified tolerance.

**DISCUSSION OF SILVERMAN (1982)**

In statistics there is a class of nonparametric techniques called “maximum penalized likelihood” estimation [9]. The central idea behind these methods for estimating smooth functions is to maximize the likelihood function modified by a heuristic roughness penalty. In this context, Silverman (1982) proposed using \(-S_{\ell} \phi \) as the penalized likelihood function for probability density estimation [10]. This choice was motivated by the observation that, when \( \ell = \infty \), one recovers a moment-matching distribution having a familiar parametric form. This early work by Silverman is therefore relevant to the results reported here.
However, the results reported here move beyond [10] in a number of ways. First, the connection with MaxEnt estimation was not reported in [10]. The fact that the MAP density $Q_\ell$ matches the same moments as $Q_\infty$ even at finite values of $\ell$ was also not observed. Moreover, periodic boundary conditions on $Q_\ell$ were assumed in the subsequent analysis described in [10], and the contradiction between these boundary conditions and the results for $\ell = \infty$ was not discussed.

Perhaps most importantly, the non-Bayesian interpretation of the maximum penalized likelihood approach illustrates, by deficiency, the importance of adopting a Bayesian interpretation. Although the penalized likelihood context may be sufficient to motivate the formula for $S_\ell[\phi]$, it provides no motivation for computing the evidence $p(\text{data}|\ell)$. Without this notion of “evidence,” it is unclear how to determine the optimal smoothness length scale $\ell^*$ without resorting to empirical methods, such as cross-validation. By contrast, the Bayesian interpretation introduced by Bialek et al. [2] and built upon here transparently motivates the Laplace approximation of $p(\text{data}|\ell)$. This results in an explicit mathematical criterion for choosing $\ell^*$. In particular, this Bayesian interpretation is essential for our derivation of the $K$ coefficient in Eq. 12 of the main text.

* Email correspondence to jkinney@cshl.edu

[1] J. B. Kinney, Phys. Rev. E 90, 011301(R) (2014).
[2] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. 77, 4693 (1996).
[3] I. Nemenman and W. Bialek, Phys. Rev. E 65, 026137 (2002).
[4] D. Ormoneit and H. White, Econom. Rev. 18, 127 (1999).
[5] E. L. Allgower and K. Georg, Numerical Continuation Methods: An Introduction (Springer, 1990).
[6] J. Skilling, AIP Conf. Proc. 954, 39 (2007).
[7] L. R. Mead and N. Papanicolaou, J. Math. Phys. 25, 2404 (1984).
[8] S. Boyd and L. Vandenberghe, Convex Optimization (Cambridge University Press, 2009).
[9] P. P. B. Eggermont and V. N. LaRiccia, Maximum Penalized Likelihood Estimation: Volume 1: Density Estimation (Springer, 2001).
[10] B. W. Silverman, Ann. Stat. 10, 795 (1982).
