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

The need to estimate smooth probability distributions (a.k.a. probability densities) from finite sampled data is ubiquitous in science. Many approaches to this problem have been described, but none is yet regarded as providing a definitive solution. Maximum entropy estimation and Bayesian field theory are two such approaches. Both have origins in statistical physics, but the relationship between them has remained unclear. Here I unify these two methods by showing that every maximum entropy density estimate can be recovered in the infinite smoothness limit of an appropriate Bayesian field theory. I also show that Bayesian field theory estimation can be performed without imposing any boundary conditions on candidate densities, and that the infinite smoothness limit of these theories recovers the most common types of maximum entropy estimates. Bayesian field theory thus provides a natural test of the maximum entropy null hypothesis and, furthermore, returns an alternative (lower entropy) density estimate when the maximum entropy hypothesis is falsified. The computations necessary for this approach can be performed rapidly for one-dimensional data, and software for doing this is provided.

PACS number(s): 02.50.Tt, 02.60.−x, 11.10.Lm, 89.70.Cf

I. INTRODUCTION

Research in nearly all fields of science routinely calls for the estimation of smooth probability densities from finite sampled data [1,2]. Indeed, the presence of histograms in a large fraction of the scientific literature attests to this need. But the problem of how to go beyond a histogram and recover a smooth probability distribution has yet to find a definitive solution, even in one dimension.

The reader might find this state of affairs surprising. Many different methods for estimating smooth probability densities are well known and commonly used. One of the most popular methods is kernel density estimation [1,2]. Kernel density estimation is easy to carry out, but this approach has little theoretical justification and there is no consensus on certain basic aspects of its use, such as how to choose a kernel width or how to treat data points near a boundary [3]. Bayesian inference of a Gaussian mixture model is another common method. This approach, however, requires that one assume an explicit functional form for the density that one wishes to learn.

Concepts from statistical physics have given rise to two alternative approaches to the density estimation problem: maximum entropy (MaxEnt) [4,5] and Bayesian field theory [6–14]. Each of these approaches has a firm but distinct theoretical basis. MaxEnt derives from the principle of maximum entropy as described by Jaynes in 1957 [4]. Bayesian field theory, which is also referred to as "information field theory" in some of the literature [9], instead uses the standard methods of Bayesian inference together with priors that weight possible densities according to an explicit measure of smoothness without requiring that these densities have a particular functional form. Perhaps because the principles

underlying these two methods are different, the relationship between these approaches has remained unclear.

MaxEnt density estimation is carried out as follows. One first uses sampled data to estimate values for a chosen set of moments, e.g., mean and variance. Typically, all moments up to some specified order are selected [5,15]. 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 [16].

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, and from this posterior one identifies the maximum *a posteriori* (MAP) density estimate. Simple field theory priors require that one assume an explicit smoothness length scale ℓ. However, an optimal value for ℓ can be learned from the data in a natural way if one instead adopts a prior formed from a scale-free mixture of these simple field theories [6–8]. Scale-free Bayesian field theories thus provide a way to estimate probability densities without having to specify any tunable parameters.

One problem with the field theory priors that have been considered for this purpose thus far [6–8] is that they impose boundary conditions on candidate densities. This assumption of boundary conditions is standard practice in physics; it greatly aids analytic calculations and is often well motivated by physical reasoning. In the density estimation context, however, boundary conditions limit the types of data sets for which such field theory priors would be appropriate. MaxEnt, by contrast, does not impose any boundary conditions on the density estimates it provides.

Here I describe a class of Bayesian field theory priors that have no boundary conditions. These priors yield MAP density estimates that exactly match the first few moments of the data. In the $\ell \to \infty$ limit, such MAP estimates become identical to MaxEnt estimates constrained by these same moments. More

---

[*]jkinney@cshl.edu

generally, I show that a MaxEnt density estimate matched to any moments of the data 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 unification of Bayesian field theory and MaxEnt density estimation further suggests a natural way to test the validity of the MaxEnt hypothesis against one's data. If Bayesian field theory identifies $\ell = \infty$ as being optimal for one's data set, the MaxEnt hypothesis is validated. If instead the optimal ℓ is finite, the MaxEnt hypothesis is rejected in favor of a nonparametric density estimate that matches the same moments of the data but has lower entropy.

This paper is structured as follows. Section II describes the derivation of an action, S_ℓ, that governs the posterior probability of densities under a specific class of Bayesian field theories. Section III describes how the MAP density, which minimizes this action, can be uniquely derived without assuming any boundary conditions. A differential operator I call the "bilateral Laplacian" plays a central role in eliminating the need for boundary conditions.

Section IV shows that such MAP density estimates reduce to MaxEnt estimates in the $\ell \to \infty$ limit. Section V derives an expression for a quantity, the "evidence ratio" $E(\ell)$, that allows one to select the optimal value for ℓ given the data. The large ℓ behavior of this evidence ratio is shown to be characterized by a "K coefficient," the sign of which provides a novel analytic test of the MaxEnt assumption.

Section VI formalizes a discrete-space representation of this Bayesian field theory inference procedure. In addition to being essential for the computational implementation of this method, this discrete representation greatly clarifies why no boundary conditions are required to derive the MAP density when one makes use of the bilateral Laplacian. Section VII describes how to compute the MAP density (to a specified precision) at all length scales ℓ. Section VIII illustrates this density estimation approach on simulated data sets. A summary and discussion are provided in Sec. IX.

Detailed derivations of various results from Secs. II through VI are provided in Appendices A–D. Appendix E presents details of a predictor-corrector homotopy algorithm that allows the density estimation computations described in this paper to be carried out. An open source software implementation of this algorithm for one-dimensional density estimation is provided [17]. Finally, Appendix F gives an expanded discussion of how Bayesian field theory relates to earlier work in statistics on "maximum penalized likelihood" [3,18,19].

II. BAYESIAN FIELD THEORY

The main results of this paper are elaborated in the context of one-dimensional density estimation. Many of these results, however, are readily extended to higher dimensions. This issue is discussed in more detail later.

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. Our goal is to estimate Q_{true} from these data. Following Ref. [8], we first represent each candidate density $Q(x)$ in terms of a real field $\phi(x)$ via

$$Q(x) = \frac{e^{-\phi(x)}}{\int dx' e^{-\phi(x')}}. \tag{1}$$

This parametrization ensures that Q is positive and normalized [20]. Next we adopt a field theory prior on ϕ. Specifically we consider priors of the form

$$p(\phi|\ell) = \frac{e^{-S_\ell^0[\phi]}}{Z_\ell^0}, \tag{2}$$

where

$$S_\ell^0[\phi] = \int \frac{dx}{L} \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2 \tag{3}$$

is the "action" corresponding to this prior and

$$Z_\ell^0 = \int \mathcal{D}\phi \, e^{-S_\ell^0[\phi]} \tag{4}$$

is the associated "partition function." The real parameter ℓ is a length scale below which fluctuations in ϕ are strongly damped. The parameter α, on the other hand, reflects a fundamental choice in how we define "smoothness." In this paper we consider arbitrary positive integer values of α, for reasons that will become clear. It should be noted, however, that previous work has explored the consequences of using noninteger values of α [7].

As shown in Appendix A, this choice of prior allows us to compute an exact posterior probability $p(\phi|\text{data},\ell)$ over candidate densities. We find that

$$p(\phi|\text{data},\ell) = \frac{e^{-S_\ell[\phi]}}{Z_\ell}, \tag{5}$$

where

$$S_\ell[\phi] = \int \frac{dx}{L} \left\{ \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2 + NLR\phi + Ne^{-\phi} \right\} \tag{6}$$

is a nonlinear action,

$$Z_\ell = \int \mathcal{D}\phi \, e^{-S_\ell[\phi]} \tag{7}$$

is the corresponding partition function, and

$$R(x) = N^{-1} \sum_{n=1}^{N} \delta(x - x_n) \tag{8}$$

is the raw data density.

The derivation of Eq. (6) is somewhat subtle. In particular, the action $S_\ell[\phi]$ gives a posterior probability $p(\phi|\text{data},\ell)$ that is not related to $p(\phi|\ell)$ via Bayes's rule. However, upon marginalizing over the constant component of ϕ, one finds that $p(Q|\text{data},\ell)$ is indeed related to $p(Q|\ell)$ via Bayes's rule. This latter fact is sufficient to justify the use of Eq. (6) in what follows. See Appendix A for details.

III. ELIMINATING BOUNDARY CONDITIONS

We define the MAP field ϕ_ℓ as the field that minimizes the action S_ℓ. To obtain a differential equation for ϕ_ℓ, previous work [6–8] imposed periodic boundary conditions on ϕ and used integration by parts to derive

$$\ell^{2\alpha}(-1)^\alpha \partial^{2\alpha} \phi_\ell + NLR - Ne^{-\phi_\ell} = 0. \tag{9}$$

With the periodic 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.

It is true, of course, that Eq. (9) requires boundary conditions in order to have a unique solution. The reason boundary conditions are needed is the appearance of the standard α-order Laplacian operator, $(-1)^{\alpha}\partial^{2\alpha}$. However, we assumed boundary conditions on ϕ in order to derive Eq. (9) in the first place. It therefore has not been established that boundary conditions are required for $S_{\ell}[\phi]$ to have a unique minimum.

In fact, $S_{\ell}[\phi]$ has a unique minimum without the imposition of any boundary conditions on ϕ. The boundary conditions on ϕ assumed in previous work [6–8] are therefore unnecessary. Indeed, from Eq. (6) alone we can derive a differential equation that uniquely specifies the MAP field ϕ_{ℓ}.

We start by rewriting the action as

$$S_{\ell}[\phi] = \int \frac{dx}{L}\left\{\frac{\ell^{2\alpha}}{2}\phi\Delta^{\alpha}\phi + NLR\phi + Ne^{-\phi}\right\}, \quad (10)$$

where the differential operator Δ^{α} is defined by the requirement that

$$\varphi\Delta^{\alpha}\phi = (\partial^{\alpha}\varphi)(\partial^{\alpha}\phi) \quad (11)$$

for any two fields φ and ϕ. In what follows we refer to Δ^{α} as the "bilateral Laplacian of order α." Note that Δ^{α} is a positive semidefinite operator, since

$$\int dx\,\phi\Delta^{\alpha}\phi = \int dx\,(\partial^{\alpha}\phi)^2 \geqslant 0 \quad (12)$$

for every real field ϕ.

We now prove that ϕ_{ℓ} is unique by showing that $S_{\ell}[\phi]$ is a strictly convex function of ϕ when $N > 0$. Consider the change in $S_{\ell}[\phi]$ upon the perturbation $\phi \to \phi + \epsilon\psi$, where ϕ and ψ are two real fields, ϵ is an infinitesimal number, and the field ψ is normalized so that $L^{-1}\int dx\,\psi^2 = 1$. The action will change by an amount

$$S_{\ell}[\phi + \epsilon\psi] = S_{\ell}[\phi] + \epsilon\int dx\,\psi\left.\frac{\delta S}{\delta\phi}\right|_{\phi}$$
$$+ \frac{\epsilon^2}{2}\int dx\left\{\frac{\ell^{2\alpha}}{L}\psi\Delta^{\alpha}\psi + \frac{N}{L}e^{-\phi}\psi^2\right\} + \cdots. \quad (13)$$

Because Δ^{α} is positive semidefinite, the $O(\epsilon^2)$ term will be bounded from below by $\epsilon^2 N\exp[-\max(\phi)]$ and must therefore be positive. The Hessian of S_{ℓ} is therefore positive definite at every ϕ, establishing the strict convexity of S_{ℓ} and thus the uniqueness of ϕ_{ℓ}.

The requirement that $\delta S_{\ell}/\delta\phi = 0$ gives the following differential equation for ϕ_{ℓ}:

$$0 = \ell^{2\alpha}\Delta^{\alpha}\phi_{\ell} + NLR - Ne^{-\phi_{\ell}}. \quad (14)$$

From the argument above we see that this differential equation, unlike Eq. (9), has a unique solution without the imposition of any boundary conditions on ϕ_{ℓ}.

This lack of a need for boundary conditions in Eq. (14), despite the need for boundary conditions in Eq. (9), is due to a fundamental difference between the standard Laplacian and the bilateral Laplacian. This difference occurs only at the boundaries of the x interval. Roughly speaking, $\Delta^{\alpha}\phi$ is well defined at both $x_{\min}$ and $x_{\max}$, whereas $(-1)^{\alpha}\partial^{2\alpha}\phi$ is not. This point will be clarified in Sec. VI, when we formulate our Bayesian field theory approach on a finite set of grid points.

In the interior of the x interval, however, the bilateral Laplacian is identical to the standard Laplacian. To see this, we integrate Eq. (11) over x and use integration by parts to derive

$$\int dx\,\varphi\Delta^{\alpha}\phi = \int dx\,\varphi[(-1)^{\alpha}\partial^{2\alpha}]\phi$$
$$+ \sum_{b=0}^{\alpha-1}[(-1)^b(\partial^{\alpha-b-1}\varphi)(\partial^{\alpha+b}\phi)]_{x_{\min}}^{x_{\max}}. \quad (15)$$

The second term on the right-hand side vanishes if the test function φ is chosen so that $\partial^b\varphi = 0$ at $x_{\min}$ and $x_{\max}$ for $b = 0, 1, \ldots, \alpha - 1$. The value of such test functions φ within the interior of the interval are unconstrained, and so

$$\Delta^{\alpha}\phi(x) = (-1)^{\alpha}\partial^{2\alpha}\phi(x), \text{ for all } x_{\min} < x < x_{\max}. \quad (16)$$

IV. CONNECTION TO MAXIMUM ENTROPY

From its definition in Eq. (11), we see that the bilateral Laplacian is symmetric and real. This operator is therefore Hermitian and possesses a complete set of orthonormal eigenvectors with corresponding real eigenvalues; see Appendix B for a discussion of this spectrum.

The kernel of Δ^{α} is particularly relevant to the density estimation problem. A field ϕ is in the kernel of Δ^{α} if and only if

$$\int dx\,\phi\Delta^{\alpha}\phi = \int dx\,(\partial^{\alpha}\phi)^2 = 0. \quad (17)$$

From this we see that the kernel of Δ^{α} is equal to the space of polynomials of order $\alpha - 1$.

In particular, $\phi = 1$ is in the kernel of Δ^{α} for all positive integers α. As a result, multiplying Eq. (14) on the left by unity and integrating gives $\int dx\,e^{-\phi_{\ell}} = L$. The MAP density Q_{ℓ}, which is defined in terms of ϕ_{ℓ} by Eq. (1), is thereby seen to have the simplified form,

$$Q_{\ell} = \frac{e^{-\phi_{\ell}}}{L}. \quad (18)$$

If we multiply Eq. (14) on the left by other polynomials of order $\alpha - 1$ and integrate, we further find that

$$\int dx\,Q_{\ell}\,x^k = \int dx\,R\,x^k, \ k = 1, \ldots, \alpha - 1. \quad (19)$$

Therefore, at every length scale ℓ, the first $\alpha - 1$ moments of the MAP density Q_{ℓ} exactly match those of the data.

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

$$Q_{\infty}(x) = \frac{1}{L}\exp\left(-\sum_{k=0}^{\alpha-1}a_k x^k\right), \quad (20)$$

where the values of the coefficients a_k are determined by Eqs. (18) and (19). Q_∞ is therefore identical to the MaxEnt density that matches the first $\alpha - 1$ moments of the data [5].

At $\ell = 0$, the kinetic term in Eq. (14) vanishes. As a result, setting $\delta S_0 / \delta \phi = 0$ gives

$$Q_0(x) = R(x). \quad (21)$$

We therefore see that the MAP density Q_0 is simply the "histogram" of the data, i.e., the normalized sum of delta functions centered at each data point. When we formulate our inference procedure on a grid in Sec. VI, we will see that $Q_0 = R$ indeed becomes a bona fide histogram with bins defined by our choice of grid.

The set of MAP densities Q_ℓ thus forms a one-parameter "MAP curve" in the space of probability densities. This MAP curve extends from the data histogram at $\ell = 0$ to the MaxEnt density at $\ell = \infty$. Every density Q_ℓ along this MAP curve exactly matches the first $\alpha - 1$ moments of the data.

More generally, the MaxEnt density estimate constrained to match any set of moments can be recovered in the infinite smoothness limit of an appropriate Bayesian field theory. To see this, consider a MaxEnt estimate Q_{ME} chosen to satisfy the generalized moment-matching criteria

$$\int dx \, Q_{\text{ME}} f_j = \int dx \, R f_j, \ \ j = 1, 2, \ldots, J \quad (22)$$

for some set of user-specified functions $f_1(x), f_2(x), \ldots, f_J(x)$. A Bayesian field theory that recovers this MaxEnt estimate in the infinite smoothness limit can be readily constructed by using a prior defined by the action

$$S_\xi^0[\phi] = \int \frac{dx}{L} \frac{\xi}{2} \phi \Delta \phi, \quad (23)$$

where ξ is a (positive) smoothness parameter and Δ is a positive semidefinite operator whose kernel is spanned by the specified functions $f_1, f_2, \ldots, f_J$ together with the constant function $f_0(x) = 1$. The posterior probability on ϕ will then be governed by the action

$$S_\xi[\phi] = \int \frac{dx}{L} \left\{ \frac{\xi}{2} \phi \Delta^\alpha \phi + NLR\phi + Ne^{-\phi} \right\}. \quad (24)$$

Following the same line of reasoning as above, we find that the MAP density Q_ξ, corresponding to the field ϕ_ξ that minimizes S_ξ, will satisfy

$$\int dx \, Q_\xi f_j = \int dx \, R f_j, \ \ j = 0, 1, \ldots, J \quad (25)$$

regardless of the value of ξ. In the infinite smoothness limit ($\xi \to \infty$), the MaxEnt density will be recovered, i.e.,

$$Q_\infty(x) = \frac{1}{L} \exp\left(-\sum_{j=0}^{J} a_j f_j(x) \right) = Q_{ME}(x), \quad (26)$$

where the coefficients $a_0, a_1, \ldots, a_j$ are determined by Eq. (25).

V. CHOOSING THE LENGTH SCALE

To determine the optimal value for ℓ, we compute $p(\text{data}|\ell) = \int \mathcal{D}\phi \, p(\text{data}|\phi) p(\phi|\ell)$. This quantity, commonly called the "evidence," forms the basis for Bayesian model selection [6,7,21,22].

For the problem in hand, the evidence vanishes when $\alpha > 1$ regardless of the data. The reason for this is that $p(Q|\ell)$ is an improper prior; see Appendix C. 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 that

$$E(\ell) = e^{S_\infty[\phi_\infty] - S_\ell[\phi_\ell]} \sqrt{\frac{\det_{\text{ker}}[e^{-\phi_\infty}] \det_{\text{row}}[L^{2\alpha} \Delta^\alpha]}{\eta^{-\alpha} \det[L^{2\alpha} \Delta^\alpha + \eta e^{-\phi_\ell}]}}, \quad (27)$$

where $\eta = N(L/\ell)^{2\alpha}$. Here the subscripts "row" and "ker" indicate restriction to the row space and kernel of Δ^α, respectively; the $e^{-\phi_\ell}$ terms inside the determinants stand for matrices that have the values $e^{-\phi_\ell(x)}$ (for all xs) arrayed along the main diagonal and zeros everywhere else. See Appendix C for the derivation of Eq. (27).

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

$$\ln E = K\eta + O(\eta^2), \quad (28)$$

where the coefficient K is [23]

$$K = \sum_{i > \alpha} \frac{N v_i^2 - z_{ii}}{2\lambda_i} + \sum_{\substack{i > \alpha \\ j \leqslant \alpha}} \frac{z_{ij}^2 + v_i z_{ijj}}{2\lambda_i \zeta_j} - \sum_{\substack{i > \alpha \\ j,k \leqslant \alpha}} \frac{v_i z_{ij} z_{jkk}}{2\lambda_i \zeta_j \zeta_k}. \quad (29)$$

Here λ_i and $\psi_i(x)$ ($i = 1, 2, \ldots$) denote the eigenvalues and eigenfunctions of $L^{2\alpha} \Delta^\alpha$ and are indexed so that $\lambda_i = 0$ for $i \leqslant \alpha$. The eigenfunctions are normalized so that $\int dx \, L^{-1} \psi_i \psi_j = \delta_{ij}$, and in the degenerate subspace ($i, j \leqslant \alpha$) they are oriented so that $\int dx \, Q_\infty \psi_i \psi_j = \delta_{ij} \zeta_j$ for some positive real numbers ζ_j. The other indexed quantities are $v_i = \int dx \, (Q_\infty - R) \psi_i$, $z_{ij} = \int dx \, Q_\infty \psi_i \psi_j$, and $z_{ijk} = \int dx \, Q_\infty \psi_i \psi_j \psi_k$.

Equation (29) provides a plug-in formula that can be used to assess the validity of the MaxEnt hypothesis. If $K > 0$, there is guaranteed to be a finite value of ℓ that has a larger evidence ratio than $\ell = \infty$. In this case the MaxEnt estimate is guaranteed to be nonoptimal. On the other hand, if $K < 0$, then $\ell = \infty$ is a local optimum that may or may not be a global optimum as well. Numerical computation of E over all values of ℓ is thus needed to resolve whether the MaxEnt hypothesis provides the best explanation of the data in hand.

VI. DISCRETE SPACE REPRESENTATION

In this section we retrace the entire analysis above in the discrete representation, i.e., where the continuous x interval is replaced by an evenly spaced set of G grid points. This discrete representation is necessary for the computational implementation of our field theoretic density estimation method. Happily, our main findings above are seen to hold exactly upon discretization. This discrete representation also sheds light on how the bilateral Laplacian differs from the standard Laplacian and why this difference eliminates the need for boundary conditions.

We consider G grid points evenly spaced throughout the interval $[x_{\min}, x_{\max}]$. Specifically, we place grid points at

$$x_i = x_{\min} + h\left(n - \frac{1}{2}\right), \quad n = 1, 2, \ldots, G, \quad (30)$$

where $h = L/G$ is the grid spacing. In moving to this discrete representation, functions of x become G-dimensional vectors with elements denoted by the subscript n. For instance, the field $\phi(x)$ becomes a vector with elements ϕ_n. Integrals become sums, i.e.,

$$\int dx\, f(x) \to h \sum_{n=1}^{G} f_n, \quad (31)$$

and path integrals over ϕ become G-dimensional integrals over the elements ϕ_n, i.e.,

$$\int \mathcal{D}\phi \to \int_{-\infty}^{\infty} d\phi_1 \int_{-\infty}^{\infty} d\phi_2 \cdots \int_{-\infty}^{\infty} d\phi_G. \quad (32)$$

We denote differential operators in this discrete representation with a subscript G. The derivative operator, ∂_G, becomes a $(G-1)$ by G matrix having elements $(\partial_G)_{nm} = h^{-1}(-\delta_{n,m} + \delta_{n+1,m})$. For instance, setting $G = 8$ gives the 7 by 8 matrix,

$$\partial_8 = \frac{1}{h}\begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}. \quad (33)$$

Similarly, the standard α-order Laplacian becomes a $(G - 2\alpha)$ by G matrix, given by $(-1)^\alpha \partial_{G-2\alpha+1} \cdots \partial_{G-1} \partial_G$. For example, choosing $\alpha = 3$ and $G = 8$ yields the a 2 by 8 Laplacian matrix

$$-\partial_8^6 = \frac{1}{h^6}\begin{pmatrix} -1 & 6 & -15 & 20 & -15 & 6 & -1 & 0 \\ 0 & -1 & 6 & -15 & 20 & -15 & 6 & -1 \end{pmatrix}. \quad (34)$$

Because 2α elements are eliminated from the vector ϕ_ℓ upon application of the standard Laplacian, the discrete version of Eq. (9) provides only $G - 2\alpha$ equations for the G unknown values of ϕ_ℓ. Thus 2α additional constraints, typically provided in the form of boundary conditions, are needed to obtain a unique solution.

By contrast, the α-order bilateral Laplacian is represented by the G by G matrix $\Delta_G^\alpha = (\partial_G^\alpha)^\top \partial_G^\alpha$, where $\partial_G^\alpha = \partial_{G-\alpha+1} \cdots \partial_{G-1} \partial_G$. Indeed, again choosing $\alpha = 3$ and $G = 8$, we recover an 8 by 8 bilateral Laplacian matrix

$$\Delta_8^3 = \frac{1}{h^6}\begin{pmatrix} 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 \\ -3 & 10 & -12 & 6 & 1 & 0 & 0 & 0 \\ 3 & -12 & 19 & -15 & 6 & 1 & 0 & 0 \\ -1 & 6 & -15 & 20 & -15 & 6 & -1 & 0 \\ 0 & -1 & 6 & -15 & 20 & -15 & 6 & -1 \\ 0 & 0 & -1 & 6 & -15 & 19 & -12 & 3 \\ 0 & 0 & 0 & -1 & 6 & -12 & 10 & -3 \\ 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{pmatrix}. \quad (35)$$

The middle two rows of Δ_8^3 match those of $-\partial_8^6$, reflecting the equivalence of bilateral Laplacians and standard Laplacians in the interior of the x interval. However, Δ_8^3 contains six additional rows, three at either end. These are sufficient to specify a unique solution for the eight elements of the ϕ_ℓ vector. More generally, the discrete version of Eq. (14) provides G equations for the G unknown elements of ϕ_ℓ and is therefore able to specify a unique solution without the imposition of any boundary conditions.

Using the bilateral Laplacian, we can readily define a discretized version of the field theory prior $p(\phi|\ell)$ by adopting the action

$$S_\ell^0[\phi] = \frac{\ell^{2\alpha}}{2G} \sum_{n,m} \Delta_{nm}^\alpha \phi_n \phi_m. \quad (36)$$

This leads to the posterior action

$$S_\ell[\phi] = \sum_{n,m} \left\{ \frac{\ell^{2\alpha}}{2G} \Delta_{nm}^\alpha \phi_n \phi_m \right. $$
$$\left. + \frac{NL}{G} R_n \phi_n \delta_{nm} + \frac{N}{G} e^{-\phi_n} \delta_{nm} \right\}, \quad (37)$$

where R_n is the value of the data histogram at grid point n, i.e., the fraction of data points discretized to grid point n, divided by bin width h.

The corresponding equation of motion is

$$\ell^{2\alpha} \sum_m \Delta_{nm}^\alpha \phi_{\ell m} + NLR_n - Ne^{-\phi_{\ell n}} = 0. \quad (38)$$
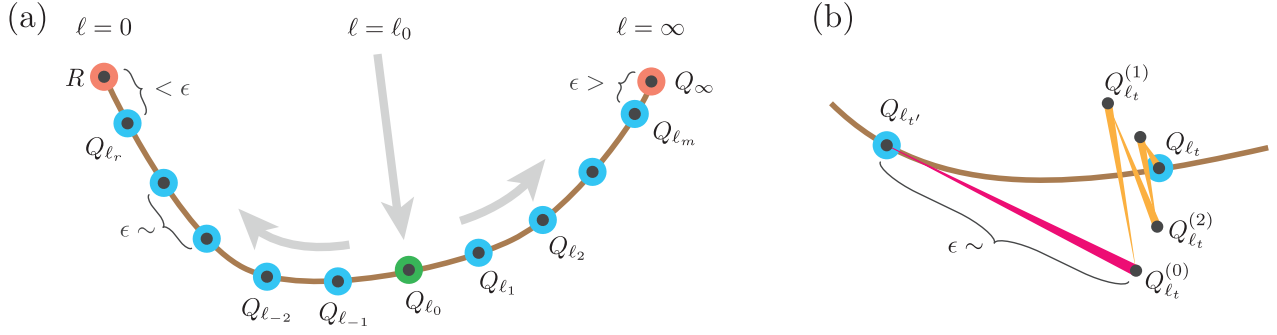
FIG. 1. (Color) Illustration of the predictor-corrector homotopy algorithm. (a) The MAP curve (brown) is approximated using a finite set of densities $\{R, Q_{\ell_r}, \ldots, Q_{\ell_{-2}}, Q_{\ell_{-1}}, Q_{\ell_0}, Q_{\ell_1}, Q_{\ell_2}, \ldots, Q_{\ell_m}, Q_\infty\}$. First the MAP density at an intermediate length scale $\ell_0 = L/\sqrt{G}$ 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 $\lesssim \epsilon$ of each other. (b) Each step $Q_{\ell_{t'}} \to Q_{\ell_t}$ has two parts. First, a predictor step (magenta) computes a new length scale ℓ_t and an approximation $Q_{\ell_t}^{(0)}$ of Q_{ℓ_t}. A series of corrector steps $Q_{\ell_t}^{(0)} \to Q_{\ell_t}^{(1)} \to Q_{\ell_t}^{(2)} \cdots$ (orange) then converges to Q_{ℓ_t}.

The kernel of Δ_G^α is spanned by vectors ϕ whose elements have the polynomial form $\phi_n = \sum_{b=0}^{\alpha-1} a_b x_n^b$. The analogous moment-matching behavior therefore holds exactly in the discrete representation, i.e.,

$$h \sum_{n=1}^G Q_{\ell n} x_n^k = h \sum_{n=1}^G R_n x_n^k, \qquad (39)$$

where Q_ℓ is related to ϕ_ℓ via Eq. (18). In the $\ell \to \infty$ limit, the MAP density Q_∞ again has the analogous form

$$Q_{\infty n} = \frac{1}{L} \exp\left(-\sum_{k=0}^{\alpha-1} a_k x_n^k\right), \qquad (40)$$

where the coefficients a_k are chosen to satisfy Eq. (39). Thus, the connection between Bayesian field theory and MaxEnt density estimation remains intact upon discretization.

The derivation of the evidence ratio in Eq. (27) follows through without modification. The only change is that the functional determinants now become determinants of finite-dimensional matrices. The derivation of the K coefficient in Eq. (29) also follows in a similar manner; the only change to Eq. (29) is that the the index i now ranges from 1 to G, not 1 to ∞.

VII. COMPUTING DENSITY ESTIMATES

To compute density estimates using this field theory approach, we work in the discrete representation described in the previous section. First, the user specifies the number of grid points G as well as a bounding box $[x_{\min}, x_{\max}]$ for the data. MAP densities Q_ℓ are then computed at a finite set of length scales $\{0, \ell_r, \cdots, \ell_{-2}, \ell_{-1}, \ell_0, \ell_1, \ell_2, \cdots, \ell_m, \infty\}$, as illustrated in Fig. 1(a). This "string of beads" approximation to the MAP curve allows us to evaluate the evidence ratio E over all length scales and, to finite precision, identify the length scale ℓ^* that maximizes E.

This approximation of the MAP curve is computed using a predictor-corrector homotopy algorithm [25]. An overview of this algorithm is now given. Please refer to Appendix E for algorithm details. I note that this algorithm provides more

transparent precision bounds on the computed Q_ℓ densities than does the previously reported algorithm of Ref. [8].

First, an intermediate length scale ℓ_0 is chosen and the corresponding MAP density Q_{ℓ_0} is computed. This density, Q_{ℓ_0}, serves as the starting point from which to trace the MAP curve towards both larger and smaller length scales [Fig. 1(a)]. The algorithm then proceeds in both directions, stepping from length scale to length scale and updating the MAP density at each step.

During each step, the subsequent length scale is chosen so that the corresponding MAP density is sufficiently similar to the MAP density at the preceding length scale. Specifically, in stepping from $\ell_{t'}$ to ℓ_t, the algorithm chooses ℓ_t so that the geodesic distance D_{geo} (see Refs. [8,26]) between $Q_{\ell_{t'}}$ and Q_{ℓ_t} matches a user-specified tolerance ϵ, i.e.,

$$D_{\text{geo}}[Q_{\ell_t}, Q_{\ell_{t'}}] \equiv 2 \cos^{-1}\left(\int dx\, \sqrt{Q_{\ell_t} Q_{\ell_{t'}}}\right) \lesssim \epsilon. \quad (41)$$

The value $\epsilon = 10^{-2}$ was used for the computations described below and shown in Figs. 2 and 3. Stepping in the decreasing ℓ direction is terminated at a length scale ℓ_r such that $D_{\text{geo}}[Q_{\ell_r}, R] < \epsilon$. Similarly, stepping in the increasing ℓ direction is terminated at a length scale ℓ_m such that $D_{\text{geo}}[Q_{\ell_m}, Q_\infty] < \epsilon$; the MaxEnt density Q_∞ is computed at the start of the algorithm essentially as described by Ormoneit and White [15].

Each step along the MAP curve is accomplished in two parts [Fig. 1(b)]. Given the MAP density $Q_{\ell_{t'}}$ at length scale $\ell_{t'}$, a "predictor step" is used to compute both the next length scale ℓ_t as well as an approximation $Q_{\ell_t}^{(0)}$ to the corresponding MAP density Q_{ℓ_t}. The repeated application of a "corrector step" is then used to compute a series of densities $Q_{\ell_t}^{(1)}, Q_{\ell_t}^{(2)}, \ldots$ that converges to Q_{ℓ_t}.

If the numerics are properly implemented, this predictor-corrector algorithm is guaranteed to identify the correct MAP density Q_ℓ at each of the chosen length scales ℓ. This is because the action $S_\ell[\phi]$ is strictly convex in ϕ and therefore has a unique minimum (as was shown in Sec. III). The distance criteria in Eq. (41) further ensures that the stepping procedure does not drastically overstep ℓ^*. It is worth noting that, because
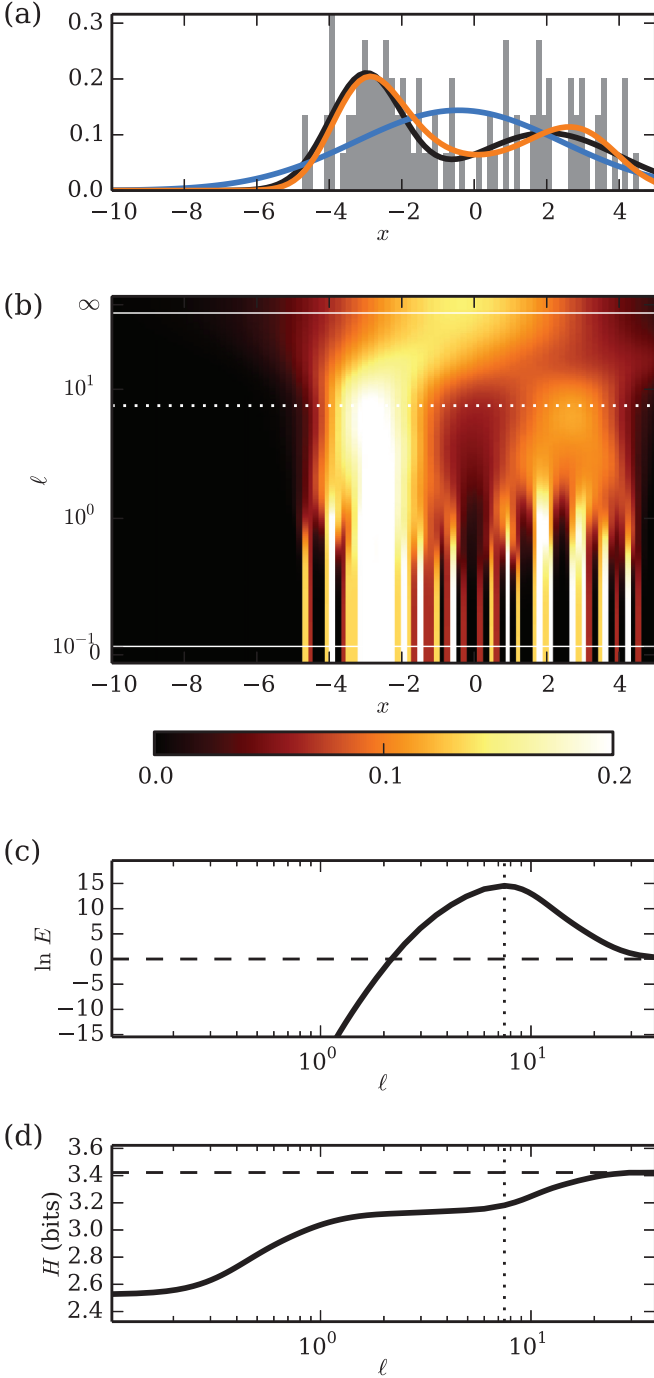
FIG. 2. (Color) Density estimation without boundary conditions using the $\alpha = 3$ field theory prior. (a) N = 100 data points were drawn from a simulated density Q_{true} (black) and binned at $G = 100$ grid points. The resulting histogram (gray) is shown along with the field theory density estimate $Q_{\ell*}$ (orange) and the corresponding MaxEnt estimate Q_∞ (blue). (b) The heat map shows the densities Q_ℓ interpolating between the MaxEnt density at $\ell = \infty$ and the data histogram at $\ell = 0$. (c) The log of the evidence ratio E is shown as a function of ℓ. (d) The differential entropy $H = -\int dx\, Q_\ell \ln Q_\ell$ [24] is shown as a function of ℓ; the entropy at $\ell = \infty$ is indicated by the dashed line. Dotted lines in (b–d) mark the optimal length scale ℓ^*.

Δ_G^α is sparse, these predictor and corrector steps can be sped up by using numerical sparse matrix methods.

VIII. EXAMPLE ANALYSES

Figure 2 provides an illustrated example of this density estimation procedure in action. Starting from a set of sampled data [Fig. 2(a), gray], the homotopy algorithm computes the MAP density Q_ℓ at a set of length scales spanning the interval $\ell \in [0, \infty]$ [Fig. 2(b)]. The evidence ratio E is then computed at each of these chosen length scales using Eq. (27), and the length scale ℓ^* that maximizes E is identified [Fig. 2(c)]. $Q_{\ell*}$ is then reported as the best estimate of the underlying density [Fig. 2(a), orange]. If one further wishes to report "error bars" on this estimate, other plausible densities Q can be drawn from the posterior $p(Q|\text{data})$ as described in Ref. [8].

The optimal length scale ℓ^* may or may not be infinite. If $\ell^* = \infty$, then $Q_{\ell*}$ is the MaxEnt estimate that matches the first $\alpha - 1$ moments of the data. On the other hand, if ℓ^* is finite as in Fig. 2, then $Q_{\ell*}$ will have lower entropy than the MaxEnt estimate [Fig. 2(d)] while still exactly matching the first $\alpha - 1$ moments of the data. This reduced entropy reflects the use of addition information in the data beyond the values of the first $\alpha - 1$ moments. It should be noted that ℓ^* is never zero due to a vanishing Occam factor in this limit.

The density estimation procedure proposed in this paper thus provides an automatic test of the MaxEnt hypothesis. It can therefore be used to test whether Q_{true} has a hypothesized functional form. For example, using $\alpha = 3$ we can test whether our data was drawn from a Gaussian distribution. This is demonstrated in Fig. 3. In these tests, when data were indeed drawn from a Gaussian density, $\ell^* = \infty$ was obtained about 64% of the time [Fig. 3(a) and 3(d)]. By contrast, when data were 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. 3(b), 3(c), 3(e), and 3(f)]. In a similar manner, this density estimation approach can be used to test other functional forms for Q_{true}, either by using the bilateral Laplacian of different order α, or by replacing the bilateral Laplacian with a differential operator having a kernel spanned by other functions whose expectation values one wishes to match to the data.

The method used to select ℓ^* both here and in previous work [6,7] is sometimes referred to as "empirical Bayes": we choose ℓ^* to be the value of ℓ that maximizes $p(\text{data}|\ell)$. By contrast, Ref. [8] used a fully Bayesian approach by positing a Jeffreys prior for $p(\ell)$ and then choosing the length scale ℓ that maximizes $p(\text{data}, \ell) \sim p(\text{data}|\ell)p(\ell)$. It can be reasonably argued that the empirical Bayes method adopted here is less sensible than the fully Bayesian approach. However, in the fully Bayesian approach, the assumed prior $p(\ell)$ obscures the large ℓ behavior of the evidence ratio E. This large ℓ behavior is nontrivial and potentially useful.

As shown in Sec. V, the behavior of E in the large ℓ limit is governed by the K coefficient defined in Eq. (29). The sign of the K coefficient can therefore be used to assess the MaxEnt hypothesis without having to compute E at every length scale. This suggestion is supported by the simulations shown in Fig. 3. Here the sign of K (positive or negative) performed well as a proxy for whether the MaxEnt estimate was recovered (no or yes, respectively) from a full computation
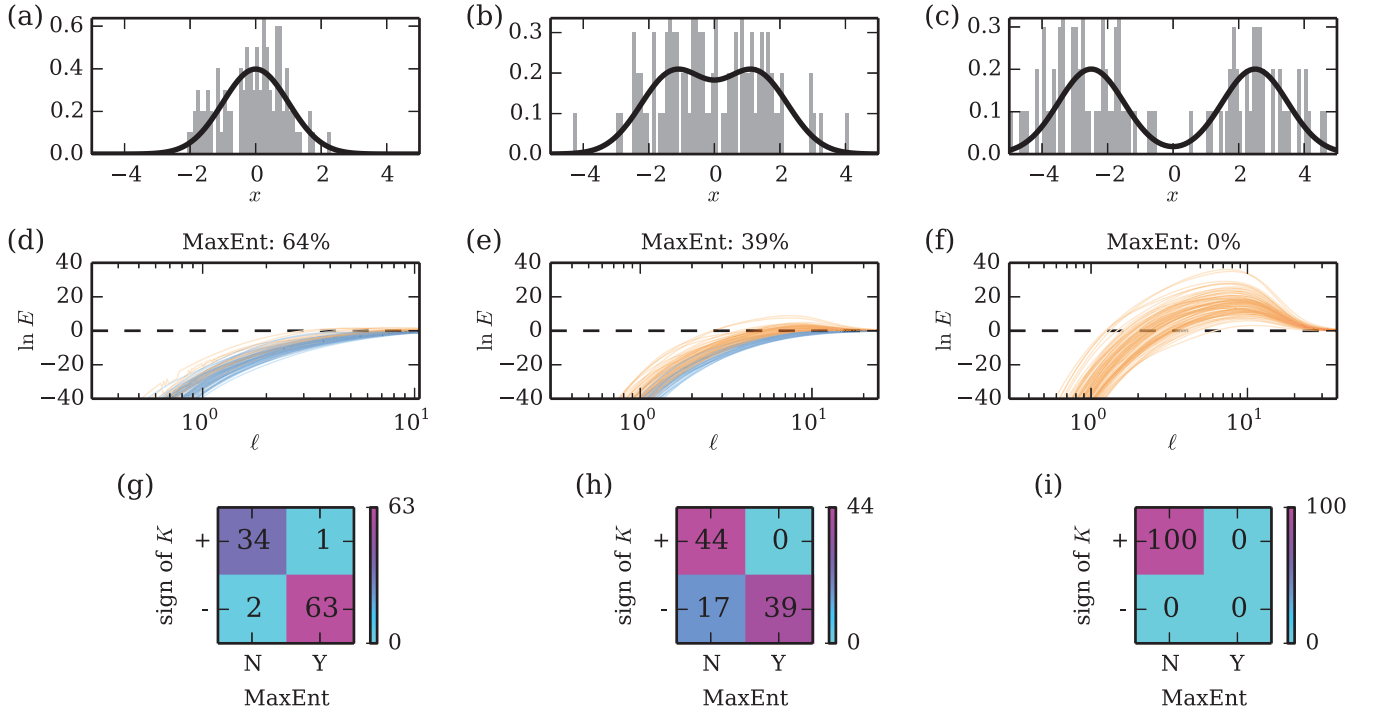
FIG. 3. (Color) The optimal estimated density for any particular data set might or might not have maximum entropy. Panels (a–c) show three different choices for Q_{true} (black), along with a histogram (gray) of $N = 100$ data points binned at $G = 100$ grid points. In each panel, Q_{true} was chosen 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_{true} densities in (a–c). Blue curves indicate $\ell^* = \infty$; orange curves indicate finite ℓ^*. Titles in (d–f) give the percentage of data sets for which $\ell^* = \infty$ was found. The heat maps shown in panels (g–i) report the number of simulated data sets for which the K coefficient was positive or negative and for which the MaxEnt density was or was not recovered.

of the MAP curve; see Figs. 3(g) [27], 3(h), and 3(i). These results suggest that the K coefficient, for which Eq. (29) provides an analytic expression, might allow an expedient test of the MaxEnt hypothesis when computation of the entire MAP curve is less feasible, e.g., in higher dimensions.

IX. SUMMARY AND DISCUSSION

Bialek *et al.* [6] showed that probability density estimation can be formulated as a Bayesian inference problem using field theory priors. Among other results, Ref. [6] derived the action in Eq. (6) and showed how to use a Laplace approximation of the evidence to select the optimal smoothness length scale [28]. However, periodic boundary conditions were imposed on candidate densities in order to transform the standard Laplacian into a Hermitian operator. The MaxEnt density estimate typically violates these boundary conditions, and as a result the ability of Bayesian field theory to subsume MaxEnt density estimation went unrecognized in Ref. [6] and in follow-up studies [7,8].

Here we have seen that boundary conditions on candidate densities are unnecessary. The bilateral Laplacian, defined in Eq. (11), 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 a novel criterion for judging the appropriateness of the MaxEnt hypothesis on individual data sets.

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 Ref. [8], the necessary Bayesian field theory 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 Ref. [9].

In the author's opinion, the problem of how to choose an appropriate prior appears to be the primary issue standing in the way of a definitive practical solution to the density estimation problem in one dimension. It is not hard to imagine different situations that would call for qualitatively different priors, but understanding which situations call for which priors will require further study. The author is optimistic, however, that the variety of priors needed to address most of the one-dimensional

density estimation problems typically encountered might not be that large.

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 [6], 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 Refs. [11,29], might provide a way forward.

APPENDIX A: DERIVATION OF THE ACTION

Our derivation of the action $S_\ell[\phi]$ in Eq. (6) is essentially that used in Ref. [8]. This derivation is not entirely straightforward, however, and the details of it have yet to be reported.

The prior $p(\phi|\ell)$, which is defined by the action S_ℓ^0 in Eq. (3), is improper due to the differential operator Δ^α having an α-dimensional kernel; see Sec. III and Appendix C. To avoid unnecessary mathematical difficulties, we can render $p(\phi|\ell)$ proper by considering a regularized form of the action,

$$S_\ell^0[\phi] = \int \frac{dx}{L} \frac{1}{2} \phi [\ell^{2\alpha} \Delta^\alpha + \epsilon] \phi, \qquad (A1)$$

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

The corresponding prior over Q is

$$p(Q|\ell) = \int_{-\infty}^{\infty} d\phi_c \, p(\phi|\ell) = \sqrt{\frac{2\pi}{\epsilon}} \frac{e^{-S_\ell^0[\phi_{nc}]}}{Z_\ell^0}. \qquad (A2)$$

Here we have decomposed the field ϕ using

$$\phi(x) = \phi_{nc}(x) + \phi_c, \qquad (A3)$$

where ϕ_c is the constant Fourier component of ϕ and $\phi_{nc}(x)$ is the nonconstant component of ϕ. The likelihood of Q given the data is given by

$$p(\text{data}|Q) = \prod_{n=1}^{N} Q(x_n). \qquad (A4)$$

Using the identity

$$a^{-N} = \frac{N^N}{\Gamma(N)} \int_{-\infty}^{\infty} du \, e^{-N(u+ae^{-u})}, \qquad (A5)$$

which holds for any positive numbers a and N, we find that the likelihood of Q can be expressed as

$$p(\text{data}|Q) = \frac{N^N}{L^N \Gamma(N)} \int_{-\infty}^{\infty} d\phi_c e^{-\int \frac{dx}{L}\{NLR\phi + Ne^{-\phi}\}}. \qquad (A6)$$

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

$$p(\text{data}, Q|\ell) = \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \frac{1}{Z_\ell^0} \int_{-\infty}^{\infty} d\phi_c \, e^{-S_\ell[\phi]}, \qquad (A7)$$

where $S_\ell[\phi]$ is the action from Eq. (6).

The "evidence" for ℓ, i.e., the probability of the data given ℓ, is therefore given by

$$p(\text{data}|\ell) = \frac{N^N}{L^N \Gamma(N)} \sqrt{\frac{2\pi}{\epsilon}} \frac{Z_\ell}{Z_\ell^0}, \qquad (A8)$$

where Z_ℓ is the partition function from Eq. (7). The posterior distribution over Q is then given by Bayes's rule:

$$p(Q|\text{data}, \ell) = \frac{p(\text{data}, Q|\ell)}{p(\text{data}|\ell)} \qquad (A9)$$

$$= \int_{-\infty}^{\infty} d\phi_c \, \frac{e^{-S_\ell[\phi]}}{Z_\ell}. \qquad (A10)$$

This motivates us to *define* the posterior distribution over ϕ as

$$p(\phi|\text{data}, \ell) \equiv \frac{e^{-S_\ell[\phi]}}{Z_\ell}. \qquad (A11)$$

This definition of $p(\phi|\text{data}, \ell)$ is consistent with the formula for $p(Q|\text{data}, \ell)$ obtained in Eq. (A10), in that

$$p(Q|\text{data}, \ell) = \int_{-\infty}^{\infty} d\phi_c \, p(\phi|\text{data}, \ell). \qquad (A12)$$

However, Eq. (A11) violates Bayes's rule, since

$$p(\phi|\text{data}, \ell) \neq \frac{p(\text{data}, \phi|\ell)}{p(\text{data}|\ell)}. \qquad (A13)$$

This violation of Bayes's rule is not a problem, however, since ϕ itself is not directly observable; only Q is observable. Put another way, Eq. (A11) violates Bayes's rule only in the way that it specifies the behavior of ϕ_c. This constant component of ϕ, 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 [6,7], which used a large N Laplace approximation to derive the formula for $S_\ell[\phi]$. Also note that the regularization parameter ϵ 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 $p(\text{data}|\ell)$, both explicitly and implicitly through the value of Z_ℓ^0.

APPENDIX B: SPECTRUM OF THE BILATERAL LAPLACIAN

In the continuum limit, $\Delta^\alpha = (\partial^\alpha)^\top \partial^\alpha$ remains manifestly Hermitian and therefore possesses a complete orthonormal basis of eigenfunctions. We now consider the spectrum of this operator. In what follows we will make use of 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

$$\langle f|H|g \rangle \equiv \int \frac{dx}{L} f^* H g. \qquad (B1)$$

Note the convention of dividing by L in the inner product integral. This allows us to take inner products without altering units.

From

$$\langle \phi | \Delta^\alpha | \phi \rangle = \int \frac{dx}{L} |\partial^\alpha \phi|^2 \geqslant 0, \tag{B2}$$

we see that Δ^α is positive semidefinite. Equality in Eq. (B2) obtains if and only if ϕ is a polynomial of order $\alpha - 1$; such polynomials therefore comprise the null space of Δ^α.

More generally, any solution to the eigenvalue equation $\Delta^\alpha \psi = \lambda \psi$ implies that $\lambda \langle \phi | \psi \rangle = \langle \phi | \Delta^\alpha | \psi \rangle$ for any test function ϕ. Integrating this by parts gives

$$\lambda \int_{x_{\min}}^{x_{\max}} dx \phi^* \psi = \sum_{b=0}^{\alpha-1} [(-1)^b (\partial^{\alpha-b-1} \phi^*)(\partial^{\alpha+b} \psi)]_{x_{\min}}^{x_{\max}}$$
$$+ (-1)^\alpha \int_{x_{\min}}^{x_{\max}} dx \phi^* \partial^{2\alpha} \psi. \tag{B3}$$

Considering test functions $\phi(x)$ whose first $\alpha - 1$ derivatives all vanish at the boundary, we see that in the bulk of the interval, $x_{\min} < x < x_{\max}$, the eigenfunction equation for the standard α-order Laplacian holds, i.e.,

$$\lambda \psi = (-1)^\alpha \partial^{2\alpha} \psi. \tag{B4}$$

Any function of Δ^α must therefore be an eigenfunction of the standard Laplacian as well. Moreover, all boundary terms in Eq. (B3) must vanish because the values of ϕ and its derivatives at the boundary are independent of its integral with any function in the interior. The eigenfunction ψ must therefore have the boundary behavior

$$0 = \partial^{\alpha+b} \psi |_{x_{\min}} = \partial^{\alpha+b} \psi |_{x_{\max}} \quad \text{for } 0 \leqslant b < \alpha. \tag{B5}$$

Note in particular that this behavior is exhibited by polynomials of order $\alpha - 1$, which comprise the kernel of Δ^α. If $\lambda > 0$, the corresponding eigenfunction ψ will instead consist of 2α terms of the form $\exp[i\kappa x]$, where $\kappa = \lambda^{1/2\alpha} e^{i\pi b/\alpha}$ for $b = 0, 1, \ldots, 2\alpha - 1$. The coefficients of these terms will be such that the boundary behavior in Eq. (B5) is satisfied. Typically such eigenfunctions will not be purely sinusoidal or purely exponential, but rather will exhibit a nontrivial combination of sinusoidal and exponential behavior [see Fig. 4(b)].

It should be emphasized that the boundary behavior of the eigenfunctions [Eq. (B5)] is not a boundary condition that all functions ϕ in the domain of Δ^α must satisfy. Specifically, although any well-behaved function ϕ can be expressed in the eigenbasis via

$$\phi = \sum_{k=0}^{\infty} c_k \psi_k \tag{B6}$$

for some set of coefficients c_k, one will typically find that

$$\partial^b \phi \neq \sum_{k=0}^{\infty} c_k \partial^b \psi_k \tag{B7}$$

because the sum on the right-hand side of Eq. (B7) will not be well defined. The reason for this seeming disagreement is that the convergence criterion for Eq. (B6) is weaker than that
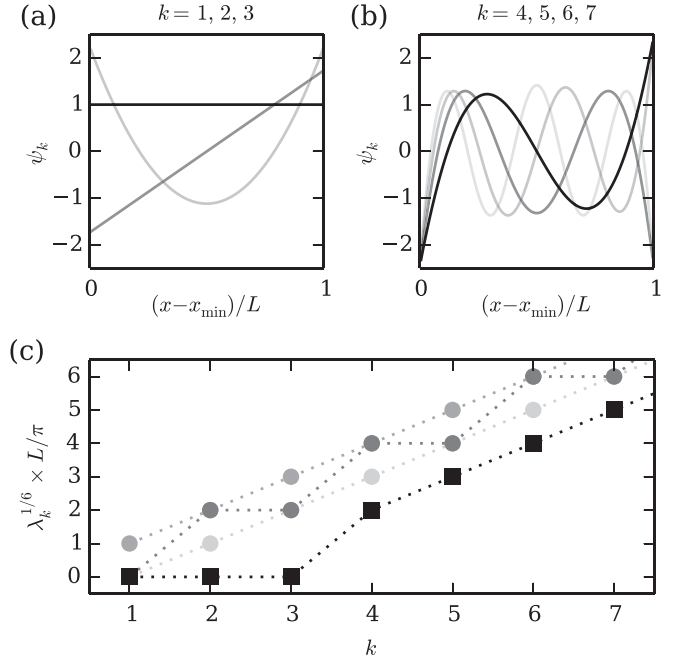

FIG. 4. Spectrum of the bilateral Laplacian of order $\alpha = 3$. (a) The first three Legendre polynomials provide an orthonormal basis for the kernel of Δ^3. These choices for ψ_1, ψ_2, and ψ_3 are plotted with decreasing opacity. (b) All other eigenfunctions are nontrivial linear combinations of factors of the form $\exp(i\kappa x)$. Their general behavior is illustrated by the basis functions ψ_4, ψ_5, ψ_6, and ψ_7, which are plotted with decreasing opacity. (c) The rank-ordered eigenvalues of Δ^3 lie at or below those of the standard Laplacian with any choice of boundary conditions. Shown are the eigenvalues of Δ^3 (black squares), together with the eigenvalues of $-\partial^6$ with periodic, Dirichlet, or Neumann boundary conditions (dark, medium, and light gray circles, respectively).

of Eq. (B7), due to the fact that $\psi_k \sim 1$ whereas $\partial^b \psi_k \sim k^b$. Therefore, even though the right-hand side of Eq. (B6) will converge for any particular ϕ, the right-hand side of Eq. (B7) typically will not.

The ordered eigenvalues of the bilateral Laplacian provide a lower bound for the eigenvalues of the standard Laplacian with any set of boundary conditions. To see this, note that we can define a positive semidefinite Hermitian operator H_{bc} whose kernel is spanned by all functions satisfying a set of specified boundary conditions. Let us denote the standard Laplacian with these boundary conditions as Δ_{bc}^α. Then we can express δ_{bc}^α in terms of the bilateral Laplacian as

$$\Delta_{bc}^\alpha = \lim_{A \to \infty} \Delta_A^\alpha, \tag{B8}$$

where

$$\Delta_A^\alpha = \Delta^\alpha + A H_{bc}. \tag{B9}$$

In the $A \to \infty$ limit, a prior defined using Δ_A^α in place of Δ^α will restrict candidate fields ϕ to those that satisfy the boundary condition specified by H_{bc}.

From first-order perturbation theory, we know that the kth eigenvalue of Δ_{A+dA}^α is related to that of Δ_A^α by

$$\lambda_k^{A+dA} = \lambda_k^A + dA \langle \psi^A | H_{bc} | \psi^A \rangle. \tag{B10}$$

Therefore, the kth eigenvalue of δ^α_{bc} is given by

$$\lambda^{bc}_k = \lambda_k + \int_0^\infty dA \langle \psi^A_k | H_{bc} | \psi^A_k \rangle, \qquad \text{(B11)}$$

where ψ^A_k is the (appropriately defined) kth eigenvector of the operator Δ^α_A. Because H_{bc} is positive semidefinite, the integral on the right-hand side is greater or equal to zero. We therefore see that $\lambda^{bc}_k \geqslant \lambda_k$ for all k, regardless of what the actual boundary conditions are.

This point is illustrated in Fig. 4(c), which plots the ordered eigenvalues for the $\alpha = 3$ bilateral Laplacian together with the ordered eigenvalues of the standard α-order Laplacian with three different types of boundary conditions: periodic boundary conditions,

$$\partial^b \psi|_{x_{\min}} = \partial^b \psi|_{x_{\max}}, \quad b = 0,1,\dots,2\alpha-1; \qquad \text{(B12)}$$

Dirichlet boundary conditions,

$$\partial^{2b}\psi|_{x_{\min}} = \partial^{2b}\psi|_{x_{\max}} = 0, \quad b = 0,1,\dots,\alpha-1; \qquad \text{(B13)}$$

and Neumann boundary conditions,

$$\partial^{2b+1}\psi|_{x_{\min}} = \partial^{2b+1}\psi|_{x_{\max}} = 0, \quad b = 0,1,\dots,\alpha-1. \qquad \text{(B14)}$$

APPENDIX C: DERIVATION OF THE EVIDENCE RATIO

We now turn to the task of evaluating the partition functions Z_ℓ and Z^0_ℓ so that we can compute the evidence $p(\text{data}|\ell)$ using Eq. (A8). Defining $\Lambda = L^{2\alpha}\Delta^\alpha$ and $\eta = N(L/\ell)^{2\alpha}$, and working in the grid representation, we find a Hessian of the form

$$\left.\frac{\partial^2 S}{\partial \phi_m \partial \phi_n}\right|_{\phi_\ell} = \frac{\ell^{2\alpha}}{GL^{2\alpha}}(\Lambda_{mn} + \delta_{mn}e^{-\phi_{\ell n}}). \qquad \text{(C1)}$$

The corresponding Laplace approximation to the path integral is therefore given by

$$Z_\ell \approx e^{-S_\ell[\phi_\ell]}\left\{\left(\frac{\ell^{2\alpha}}{2\pi G L^{2\alpha}}\right)^G \det[\Lambda + \eta e^{-\phi_\ell}]\right\}^{-1/2}. \qquad \text{(C2)}$$

Note that the operator Λ is unitless and has well-defined eigenvalues in the $G \to \infty$ limit. Also note that η is unitless. For these reasons, η will emerge as a natural perturbation parameter in the $\ell \to \infty$ limit.

Evaluating the partition function Z^0_ℓ requires more care because the regularized form of the action S^0_ℓ, given in Eq. (A1), has to be used in order for the equations we derive to make sense. We find that

$$Z^0_\ell = \left\{\left(\frac{\ell^{2\alpha}}{2\pi G L^{2\alpha}}\right)^G \det\left[\Lambda + \frac{\eta\epsilon}{N}\right]\right\}^{-1/2} \qquad \text{(C3)}$$

$$= \left\{\left(\frac{\ell^{2\alpha}}{2\pi G L^{2\alpha}}\right)^G N^{-\alpha}\eta^\alpha \epsilon^\alpha \det_{\text{row}}[\Lambda]\right\}^{-1/2}. \qquad \text{(C4)}$$

As in the main text, the subscript "row" on the determinant denotes restriction to the row space of Λ. Note: in moving from Eq. (C3) to Eq. (C4), we have used degenerate perturbation theory to calculate the determinant in the $\epsilon \to 0$ limit.

Putting these values for Z_ℓ and Z^0_ℓ back into Eq. (A8), we get

$$p(\text{data}|\ell) = \epsilon^{\frac{\alpha-1}{2}} \frac{\sqrt{2\pi}N^{N-\frac{\alpha}{2}}}{L^N \Gamma(N)} e^{-S_\ell[\phi_\ell]} \sqrt{\frac{\eta^\alpha \det_{\text{row}}[\Lambda]}{\det[\Lambda + \eta e^{-\phi_\ell}]}}. \qquad \text{(C5)}$$

Although both Z_ℓ and Z^0_ℓ 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 ϵ; specifically, it is proportional to $\epsilon^{\frac{\alpha-1}{2}}$. This is the basis for the claim in the main text that the evidence vanishes for $\alpha > 1$.

In the large ℓ limit, $\eta \to 0$, and so

$$\det[\Lambda + \eta e^{-\phi_\ell}] \to \eta^\alpha \det_{\text{ker}}[e^{-\phi_\infty}] \det_{\text{row}}[\Lambda], \qquad \text{(C6)}$$

where "ker" denotes restriction to the kernel of Λ. As a result,

$$p(\text{data}|\infty) = \epsilon^{\frac{\alpha-1}{2}} \frac{\sqrt{2\pi}N^{N-\frac{\alpha}{2}}}{L^N \Gamma(N)} \frac{e^{-S_\infty[\phi_\infty]}}{\sqrt{\det_{\text{ker}}[e^{-\phi_\infty}]}}. \qquad \text{(C7)}$$

Taking the ratio of these expressions for $p(\text{data}|\ell)$ and $p(\text{data}|\infty)$, we recover the formula for the evidence ratio E in Eq. (27). Note that E, unlike the evidence itself, does not depend on the regularization parameter ϵ.

APPENDIX D: DERIVATION OF THE K COEFFICIENT

The goal of this section is to clarify the large length scale ($\ell \to \infty$) behavior of

$$\ln E = S_\infty[\phi_\infty] - S_\ell[\phi_\ell]$$
$$+ \frac{1}{2}\ln\left\{\frac{\det_{\text{ker}}[e^{-\phi_\infty}]\det_{\text{row}}[\Lambda]}{\eta^{-\alpha}\det[\Lambda + \eta e^{-\phi_\ell}]}\right\}. \qquad \text{(D1)}$$

To do this we expand $\ln E$ as a power series in η about $\eta = 0$. We will find that $\ln E = K\eta + O(\eta^2)$ where K is a nontrivial coefficient, given by Eq. (29), that can be either positive or negative depending on the data.

We carry out this expansion in three steps:

(1) Expand ϕ_ℓ to first order in η.
(2) Expand $S_\ell[\phi_\ell]$ to first order in η.
(3) Expand $\ln \det[\Lambda + \eta e^{-\phi_\ell}]$ to first order in η.

In what follows we will use the bracket notation of Appendix B. The eigenvalues λ_i and corresponding eigenfunctions $\psi_i(x)$ of Λ are taken to satisfy

$$\langle \psi_i | \psi_j \rangle = \delta_{ij} \text{ for all } i,j, \qquad \text{(D2)}$$

$$\lambda_i = 0 \text{ for } i \leqslant \alpha, \qquad \text{(D3)}$$

and

$$\langle \psi_i | e^{-\phi_\infty} | \psi_j \rangle = \delta_{ij}\zeta_j \text{ for } i,j \leqslant \alpha \qquad \text{(D4)}$$

for some positive numbers ζ_j. We will also make use of the following indexed quantities:

$$v_i = L\langle \psi_i | Q_\infty - R \rangle = \int dx\, (Q_\infty - R)\psi_i, \qquad \text{(D5)}$$

$$z_{ij} = L\langle \psi_i | Q_\infty | \psi_j \rangle = \int dx\, Q_\infty \psi_i \psi_j, \qquad \text{(D6)}$$

$$z_{ijk} = L\langle \psi_i | Q_\infty \psi_j | \psi_k \rangle = \int dx\, Q_\infty \psi_i \psi_j \psi_k. \qquad \text{(D7)}$$

1. Expansion of ϕ_ℓ to first order in η

We begin by representing ϕ_ℓ as a power series in η. Abusing our subscript notation somewhat, we write

$$\phi_\ell = \phi_\infty + \eta\phi_1 + \eta^2\phi_2 + \cdots. \quad (D8)$$

Plugging this expansion into the equation of motion,

$$0 = \Lambda\phi_\ell + \eta(LR - e^{-\phi_\ell}), \quad (D9)$$

then collecting terms of equal order in η, we get

$$0 = \Lambda\phi_\infty + \eta(\Lambda\phi_1 + LR - e^{-\phi_\infty}) \\ + \eta^2(\Lambda\phi_2 + e^{-\phi_\infty}\phi_1) + \cdots. \quad (D10)$$

This expansion must vanish at each order in η. At lowest order in η we recover $\Lambda\phi_\infty = 0$, which just reflects the restriction of ϕ_∞ to the kernel of Λ. At first order in η,

$$0 = \Lambda\phi_1 + LR - e^{-\phi_\infty}, \quad (D11)$$

which we will now proceed to investigate.

To compute ϕ_1, we first write it in terms of the eigenfunctions of Λ, i.e.,

$$\phi_1(x) = \sum_i A_i\psi_i(x), \quad (D12)$$

for appropriate coefficients A_i. Taking the inner product of Eq. (D11) with $\langle\psi_i|$, we get

$$0 = \lambda_i A_i + L\langle\psi_i|R - Q_\infty\rangle \quad (D13)$$
$$= \lambda_i A_i - v_i. \quad (D14)$$

Since $\lambda_i > 0$ for $i > \alpha$, we find that

$$A_i = \frac{v_i}{\lambda_i}, \quad i > \alpha. \quad (D15)$$

As yet we have no information about the values A_i for $i \leqslant \alpha$. For this we need to consider the second-order term in Eq. (D10). Starting from

$$0 = \Lambda\phi_2 + e^{-\phi_\infty}\phi_1 \quad (D16)$$

and dotting each side with $\langle\psi_j|$ for $j \leqslant \alpha$, we find that

$$0 = \langle\psi_j|\Lambda|\phi_2\rangle + \langle\phi_i|e^{-\phi_\infty}|\phi_1\rangle \quad (D17)$$
$$= \sum_i A_i\langle\psi_j|e^{-\phi_\infty}|\psi_i\rangle \quad (D18)$$
$$= A_j\zeta_j + \sum_{i>\alpha} A_i z_{ij}. \quad (D19)$$

Applying Eq. (D15), we thus see that

$$A_j = -\sum_{i>\alpha}\frac{v_i z_{ij}}{\lambda_i\zeta_j}, \quad j \leqslant \alpha. \quad (D20)$$

This completes our computation of the A_i coefficients, from which we obtain

$$\phi_\ell = \phi_\infty + \eta\left[\sum_{i>\alpha}\frac{v_i}{\lambda_i}\psi_i - \sum_{\substack{i>\alpha \\ j\leqslant\alpha}}\frac{v_i z_{ij}}{\lambda_i\zeta_j}\psi_j\right] + O(\eta^2). \quad (D21)$$

2. Expansion of $S_\ell[\phi_\ell]$ to first order in η

The action $S_\ell[\phi]$ can be expressed as

$$S_\ell[\phi] = N\left\{\frac{\eta^{-1}}{2}\langle\phi|\Lambda|\phi\rangle + L\langle\phi|R\rangle + \int\frac{dx}{L}e^{-\phi}\right\}. \quad (D22)$$

Using this expression together with the expansion in Eq. (D8), we find that the value of the action S_ℓ at its minimum ϕ_ℓ is

$$S_\ell[\phi_\ell] = S_\infty[\phi_\infty] + N\eta\left\{\frac{1}{2}\langle\phi_1|\Lambda|\phi_1\rangle + L\langle\phi_1|R - Q_\infty\rangle\right\} \\ + O(\eta^2). \quad (D23)$$

The first inner product term in Eq. (D23) is

$$\frac{1}{2}\langle\phi_1|\Lambda|\phi_1\rangle = \frac{1}{2}\sum_{i>\alpha}\frac{v_i^2}{\lambda_i}, \quad (D24)$$

while the second is

$$L\langle\phi_1|R - Q_\infty\rangle = -\sum_{i>\alpha}\frac{v_i^2}{\lambda_i}. \quad (D25)$$

This gives the rather simple result

$$S_\ell[\phi_\ell] - S_\infty[\phi_\infty] = -\eta\sum_{i>\alpha}\frac{Nv_i^2}{2\lambda_i} + O(\eta^2). \quad (D26)$$

3. Expansion of $\ln\det[\Lambda + \eta e^{-\phi_\ell}]$ to first order in η

We first outline how we will go about computing $\ln\det\Gamma$ where $\Gamma = \Lambda + \eta e^{-\phi_\ell}$. Calculating this quantity requires calculating the eigenvalues of Γ. We will use γ_i and ρ_i to denote the eigenvalues and corresponding orthonormal eigenfunctions of Γ, i.e.,

$$\Gamma\rho_i = \gamma_i\rho_i \quad (D27)$$

and

$$\langle\rho_i|\rho_j\rangle = \delta_{ij}. \quad (D28)$$

Our primary task is to compute each eigenvalue γ_i as a power series in η:

$$\gamma_i = \lambda_i + \eta\gamma_i^1 + \eta^2\gamma_i^2 + \cdots. \quad (D29)$$

This task, as we will see, also requires computing the eigenfunctions ρ_i as a power series in η:

$$\rho_i = \psi_i + \eta\rho_i^1 + \eta^2\rho_i^2 + \cdots. \quad (D30)$$

As usual, it will help to expand ρ_i^1 in the eigenfunctions of Λ. We write

$$\rho_i^1(x) = \sum_j B_j^i\psi_j(x) \quad (D31)$$

and will soon proceed to compute the coefficients B_j^i.

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

$$\ln\det\Gamma = \ln\prod_i\gamma_i \quad (D32)$$

$$= \ln \left\{ \eta^\alpha \left(\prod_{j \leqslant \alpha} \gamma_j^1 \right) \left(\prod_{i > \alpha} \lambda_i \right) \right\}$$

$$+ \eta \left\{ \sum_{i > \alpha} \frac{\gamma_i^1}{\lambda_i} + \sum_{i \leqslant \alpha} \frac{\gamma_i^2}{\gamma_i^1} \right\} + O(\eta^2). \tag{D33}$$

So while the larger eigenvalues of Γ need only be computed to first order in η, the α lowest eigenvalues of Γ must actually be computed to second order in η. Performing this second-order calculation will require that we (partially) compute the eigenfunctions ρ_i to first order in η, i.e., compute (some of) the coefficients B_j^i in Eq. (D31).

Plugging Eq. (D29) and Eq. (D30) into Eq. (D27) and collecting terms by order in η, we get

$$0 = \Gamma \rho_i - \gamma_i \rho_i \tag{D34}$$

$$= (\Lambda + \eta e^{-\phi_\infty} + \eta^2 e^{-\phi_\infty} \phi_1)(\psi_i + \eta \rho_i^1 + \eta^2 \rho_i^2)$$
$$\quad - (\lambda_i + \eta \gamma_i^1 + \eta^2 \gamma_i^2)(\psi_i + \eta \rho_i^1 + \eta^2 \rho_i^2) + O(\eta^3) \tag{D35}$$

$$= (\Lambda \psi_i - \lambda_i \psi_i) + \eta (\Lambda \rho_i^1 + e^{-\phi_\infty} \psi_i - \lambda_i \rho_i^1 - \gamma_i^1 \psi_i)$$
$$\quad + \eta^2 (\Lambda \rho_i^2 + e^{-\phi_\infty} \rho_i^1 - e^{-\phi_\infty} \phi_1 \psi_i - \lambda_i \rho_i^2 - \gamma_i^1 \rho_i^1$$
$$\quad - \gamma_i^2 \psi_i) + O(\eta^3). \tag{D36}$$

The coefficient of each term in this expansion must vanish. From the zeroth-order term of Eq. (D36) we recover the eigenvalue equation $\Lambda \psi_i = \lambda_i \psi_i$, which we already knew. From the first-order term we get

$$0 = \Lambda \rho_i^1 + e^{-\phi_\infty} \psi_i - \lambda_i \rho_i^1 - \gamma_i^1 \psi_i. \tag{D37}$$

Dotting this with $\langle \psi_k |$ and using Eq. (D31) then gives

$$0 = \langle \psi_k | \Lambda | \rho_i^1 \rangle + \langle \psi_k | e^{-\phi_\infty} | \psi_i \rangle - \lambda_i \langle \psi_k | \rho_i^1 \rangle - \gamma_i^1 \langle \psi_k | \psi_i \rangle \tag{D38}$$

$$= (\lambda_k - \lambda_i) B_k^i + z_{ik} - \gamma_i^1 \delta_{ik}. \tag{D39}$$

If we set $k = i$, we recover the standard first-order correction to the eigenvalues, namely,

$$\gamma_i^1 = z_{ii} \text{ for all } i. \tag{D40}$$

In particular,

$$\gamma_j^1 = \zeta_j \text{ for } j \leqslant \alpha. \tag{D41}$$

We also see by inspection of Eq. (D39) that

$$B_k^i = -\frac{z_{ik}}{\lambda_k} \text{ for } i \leqslant \alpha, k > \alpha. \tag{D42}$$

Moreover, from the normalization requirement of Eq. (D28),

$$1 = \langle \rho_i | \rho_i \rangle \tag{D43}$$

$$= \langle \psi_i | \psi_i \rangle + 2\eta \langle \psi_i | \rho_i \rangle + O(\eta^2) \tag{D44}$$

$$= 1 + 2\eta B_i^i + O(\eta^2), \tag{D45}$$

from which we conclude that

$$B_i^i = 0 \text{ for all } i. \tag{D46}$$

Turning to the second-order term in Eq. (D36), we now consider the requirement

$$0 = \Lambda \rho_i^2 + e^{-\phi_\infty} \rho_i^1 - e^{-\phi_\infty} \phi_1 \psi_i - \lambda_i \rho_i^2 - \gamma_i^1 \rho_i^1 - \gamma_i^2 \psi_i. \tag{D47}$$

Choosing $i \leqslant \alpha$, dotting with $\langle \psi_i |$, and using the fact that $\lambda_i = 0$, we find that

$$0 = \langle \psi_i | e^{-\phi_\infty} | \rho_i^1 \rangle - \langle \psi_i | e^{-\phi_\infty} \phi_1 | \psi_i \rangle - \gamma_i^1 \langle \psi_i | \rho_i^1 \rangle - \gamma_i^2 \tag{D48}$$

$$= \sum_j B_j^i z_{ij} - \sum_j A_j z_{iij} - \gamma_i^1 B_i^i - \gamma_i^2. \tag{D49}$$

Now consider the first term of Eq. (D49). Because $B_i^i = 0$ and $z_{ij} = \zeta_i \delta_{ij}$ for $i, j \leqslant \alpha$, no $j \leqslant \alpha$ terms contribute to this sum. The third term also vanishes due to $B_i^i = 0$. So for $i \leqslant \alpha$,

$$\gamma_i^2 = \sum_{j > \alpha} B_j^i z_{ij} - \sum_{j > \alpha} A_j z_{iij} - \sum_{j \leqslant \alpha} A_j z_{iij} \tag{D50}$$

$$= -\sum_{j > \alpha} \frac{z_{ij}^2}{\lambda_j} - \sum_{j > \alpha} \frac{v_j z_{iij}}{\lambda_j} + \sum_{\substack{j \leqslant \alpha \\ k > \alpha}} \frac{v_k z_{jk} z_{iij}}{\lambda_k \zeta_j}. \tag{D51}$$

Having computed γ_i^1 for all i and γ_i^2 for $i \leqslant \alpha$, we can now find $\ln \det \Gamma$. Plugging in our results for γ_i^1 and γ_i^2 and using

$$\prod_{j \leqslant \alpha} \zeta_j = \det_{\ker}[e^{-\phi_\infty}], \quad \prod_{i > \alpha} \lambda_i = \det_{\text{row}}[\Lambda], \tag{D52}$$

we get what we set out to find:

$$\ln \det \Gamma = \ln \left\{ \eta^\alpha \det_{\ker}[e^{-\phi_\infty}] \det_{\text{row}}[\Lambda] \right\}$$

$$+ \eta \left\{ \sum_{i > \alpha} \frac{z_{ii}}{\lambda_i} - \sum_{\substack{j > \alpha \\ i \leqslant \alpha}} \frac{z_{ij}^2 + v_j z_{iij}}{\lambda_j \zeta_i} + \sum_{\substack{k > \alpha \\ i,j \leqslant \alpha}} \frac{v_k z_{jk} z_{iij}}{\lambda_k \zeta_i \zeta_j} \right\}$$

$$+ O(\eta^2). \tag{D53}$$

4. Putting it all together

Putting together our results from Eq. (D26) and Eq. (D53), we find that to first order in η,

$$\ln E = S_\infty[\phi_\infty] - S_\ell[\phi_\ell]$$

$$\quad - \frac{1}{2} \ln \left\{ \frac{\det \Gamma}{\eta^\alpha \det_{\ker}[e^{-\phi_\infty}] \det_{\text{row}}[\Lambda]} \right\} \tag{D54}$$

$$= \eta \sum_{i > \alpha} \frac{N v_i^2}{2\lambda_i}$$

$$\quad - \frac{\eta}{2} \left\{ \sum_{i > \alpha} \frac{z_{ii}}{\lambda_i} - \sum_{\substack{j > \alpha \\ i \leqslant \alpha}} \frac{z_{ij}^2 + v_j z_{iij}}{\lambda_j \zeta_i} + \sum_{\substack{k > \alpha \\ i,j \leqslant \alpha}} \frac{v_k z_{jk} z_{iij}}{\lambda_k \zeta_i \zeta_j} \right\} \tag{D55}$$

$$= K \eta, \tag{D56}$$

where

$$K = \sum_{i > \alpha} \frac{N v_i^2 - z_{ii}}{2\lambda_i} + \sum_{\substack{j > \alpha \\ i \leqslant \alpha}} \frac{z_{ij}^2 + v_j z_{iij}}{2\lambda_j \zeta_i} - \sum_{\substack{k > \alpha \\ i,j \leqslant \alpha}} \frac{v_k z_{jk} z_{iij}}{2\lambda_k \zeta_i \zeta_j}.$$
(D57)

The formula for K in Eq. (29) is obtained by renaming the indices i,j,k in the second and third terms.

APPENDIX E: PREDICTOR-CORRECTOR ALGORITHM

1. Computing the MaxEnt density

We saw in the main text that adopting the prior defined by the action in Eq. (3) renders ϕ_∞ a polynomial of order $\alpha - 1$, i.e.,

$$\phi_\infty(x) = \sum_{i=0}^{\alpha-1} a_i x^i$$
(E1)

for some vector of coefficients $\boldsymbol{a} = (a_0, a_1, \ldots, a_{\alpha-1})$. The problem of computing the MaxEnt density Q_∞ therefore reduces to finding the vector $\boldsymbol{a}$ that minimizes the posterior action

$$S_\infty(\boldsymbol{a}) = N \int \frac{dx}{L} \left\{ LR \sum_{i=0}^{\alpha-1} a_i x^i + \exp\left[-\sum_{i=0}^{\alpha-1} a_i x^i \right] \right\}.$$
(E2)

Following Ormoneit and White [15], we solve this optimization problem using the Newton-Raphson algorithm with backtracking. Starting at $\boldsymbol{a}^0 = \boldsymbol{0}$, we iterate

$$\boldsymbol{a}^n \to \boldsymbol{a}^{n+1} = \boldsymbol{a}^n + \gamma_n \delta \boldsymbol{a}^n,$$
(E3)

where the vector $\delta \boldsymbol{a}^n$ is the solution to

$$\sum_{j=0}^{\alpha-1} \left. \frac{\partial^2 S}{\partial a_i \partial a_j} \right|_{\boldsymbol{a}^n} \delta a_j^n = - \left. \frac{\partial S}{\partial a_i} \right|_{\boldsymbol{a}^n}$$
(E4)

and γ_n is some number in the interval $(0,1]$. From Eq. (E2),

$$\frac{\partial S}{\partial a_i} = N\mu_i - N \int \frac{dx}{L} x^i \exp\left[-\sum_{k=1}^{\alpha-1} a_k x^k \right],$$
(E5)

where $\mu_i = \int dx\, R\, x^i$ is the ith moment of the data, and

$$\frac{\partial^2 S}{\partial a_i \partial a_j} = N \int \frac{dx}{L} x^{i+j} \exp\left[-\sum_{k=1}^{\alpha-1} a_k x^k \right].$$
(E6)

The Hessian matrix H, with elements $H_{ij} = \frac{\partial^2 S}{\partial a_i \partial a_j}$, is positive definite at all $\boldsymbol{a}$. This is readily seen from the fact that for any vector $\boldsymbol{w}$,

$$\boldsymbol{w}^\top H \boldsymbol{w} = N \int \frac{dx}{L} \left(\sum_i x^i w_i \right)^2 e^{-\sum_k a_k x^k} > 0.$$
(E7)

Equation (E4) will therefore always yield a unique solution for $\delta \boldsymbol{a}^n$.

The scalar γ_n is chosen so that the change in the action in each iteration is commensurate with the linear approximation. Specifically, γ_n is first set to 1. Then, if

$$S_\infty(\boldsymbol{a}^n + \gamma_n \delta \boldsymbol{a}^n) - S_\infty(\boldsymbol{a}^n) < \frac{\gamma_n}{2} \sum_{i=0}^{\alpha-1} \left. \frac{\partial S}{\partial a_i} \right|_{\boldsymbol{a}^n} \delta a_i^n$$
(E8)

is not satisfied, γ_n is reduced by factors of $\frac{1}{2}$ until Eq. (E8) holds. This "dampening" of the Newton-Raphson method is sufficient to guarantee convergence [15,30]. The algorithm is terminated when the magnitude of the change in the action, $|S_\infty(\boldsymbol{a}^{n+1}) - S_\infty(\boldsymbol{a}^n)|$, falls below a specified tolerance.

2. Predictor step

The predictor step computes a change $\ell \to \ell'$ in the length scale, as well as an approximation to the corresponding change $\phi_\ell \to \phi_{\ell'}$ in the MAP field. Specifically, the predictor step returns a scalar δt and a function $\rho(x)$ such that

$$t' = t + \delta t$$
(E9)

and

$$\phi_{\ell'}(x) \approx \phi_\ell^{(0)}(x) = \phi_\ell(x) + \rho(x)\delta t,$$
(E10)

where $t = \ln \eta$ is a numerically convenient reparametrization of ℓ. To determine the function ρ, we examine the equation of motion, Eq. (D9), at ℓ':

$$0 = \Lambda \phi_{\ell'} + \eta'(LR - e^{-\phi_{\ell'}})$$
(E11)

$$= \Lambda(\phi_\ell + \rho\delta t) + \eta e^{\delta t}(LR - e^{-(\phi_\ell + \rho\delta t)})$$
(E12)

$$= \Lambda \phi_\ell + \eta(LR - e^{-\phi_\ell}) + \delta t \{ [\Lambda + \eta e^{-\phi_\ell}]\rho$$
$$+ \eta(LR - e^{-\phi_\ell}) \} + O(\delta t^2).$$
(E13)

The $O(1)$ term vanishes due to $\phi^{(n)}$ satisfying the equation of motion at ℓ. The $O(\delta t)$ term must therefore vanish as well. We thus obtain a linear equation,

$$[\Lambda + \eta e^{-\phi_\ell}]\rho = \eta(e^{-\phi_\ell} - LR),$$
(E14)

which can be numerically solved for ρ. The scalar δt is then chosen to satisfy the distance criterion,

$$\epsilon^2 = D_{\text{geo}}^2[Q_\ell, Q_{\ell'}]$$
(E15)

$$\approx \int dx \frac{(Q_\ell - Q_{\ell'})^2}{Q_\ell}$$
(E16)

$$\approx (\delta t)^2 \int dx\, Q_\ell\, \rho^2.$$
(E17)

We therefore set

$$\delta t = \pm \frac{\epsilon}{\sqrt{\int dx\, Q_\ell\, \rho^2}},$$
(E18)

with the sign of δt chosen according to the direction we wish to traverse the MAP curve.

3. Corrector step

The purpose of the corrector step is to accurately solve the equation of motion, Eq. (D9), at fixed length scale ℓ. This step is initially used to compute Q_{ℓ_0} at the starting length scale ℓ_0. It is then employed to home 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 step), the iteration

$$\phi^{(n)} \rightarrow \phi^{(n+1)} = \phi^{(n)} + \gamma_n \delta\phi^{(n)} \qquad (E19)$$

is performed. The function $\delta\phi^{(n)}$ and scalar γ_n are chosen so that this iteration converges to the true field ϕ_ℓ. To derive the field perturbation $\delta\phi^{(n)}$, we provisionally set $\gamma_n = 1$ and plug the above formula for $\phi^{(n+1)}$ into the equation of motion, Eq. (D9). Keeping only terms of order $\delta\phi^{(n)}$ or less, we see that the field perturbation $\delta\phi^{(n)}$ is the solution to the linear equation,

$$\left[\Lambda + \eta e^{-\phi^{(n)}}\right]\delta\phi^{(n)} = \eta(e^{-\phi} - LR) - \Lambda\phi^{(n)}, \qquad (E20)$$

which we solve numerically. As before, γ_n is then chosen so that the action decreases by an amount commensurate with the linear approximation, i.e.,

$$S_\ell[\phi^{(n+1)}] - S_\ell[\phi^{(n)}] < \frac{\gamma_n}{2} \int dx \left.\frac{\delta S_\ell}{\delta\phi(x)}\right|_{\phi^{(n)}} \delta\phi^{(n)}. \qquad (E21)$$

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

APPENDIX F: MAXIMUM PENALIZED LIKELIHOOD AND BAYESIAN FIELD THEORY

In statistics there is a class of nonparametric techniques for estimating smooth functions called "maximum penalized likelihood" estimation [3,18,19,29]. The central idea behind these estimation methods is to maximize the likelihood function modified by a heuristic roughness penalty. In this context, Silverman proposed using $-S_\ell[\phi]$, defined in Eq. (6), as the penalized likelihood function for probability density estimation [18]. 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 Ref. [18] in a number of critical ways. First, the connection with MaxEnt estimation was not recognized in Ref. [18], nor was the fact that the MAP density Q_ℓ matches the same moments as Q_∞ even at finite values of ℓ. Moreover, periodic boundary conditions on Q_ℓ were assumed in much of the analysis described in Ref. [18], and the contradiction between these boundary conditions and the results for $\ell = \infty$ was not discussed.

Perhaps most importantly, the shortcomings of the maximum penalized likelihood approach highlight the importance of adopting an explicit Bayesian interpretation. Although the penalized likelihood context of Ref. [18] and later work (see Ref. [3,19]) was sufficient to motivate the formula for $S_\ell[\phi]$, it provided no motivation for computing the evidence $p(\text{data}|\ell)$. Without the Bayesian notion of evidence, it is unclear how to determine the optimal smoothness length scale ℓ^* without resorting to empirical methods, such as cross-validation. By contrast, the Bayesian interpretation introduced by Bialek *et al.* [6] and built upon here transparently motivates the computation of $p(\text{data}|\ell)$, thereby providing an explicit criterion for choosing ℓ^*. In particular, this Bayesian interpretation is essential for our derivation of the K coefficient in Eq. (29).

[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York, 1986).

[2] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (Wiley, New York, 1992).

[3] P. P. B. Eggermont and V. N. LaRiccia, *Maximum Penalized Likelihood Estimation, vol. 1: Density Estimation* (Springer, New York, 2001).

[4] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[5] L. R. Mead and N. Papanicolaou, Maximum entropy in the problem of moments, J. Math. Phys. **25**, 2404 (1984).

[6] W. Bialek, C. G. Callan, and S. P. Strong, Field Theories for Learning Probability Distributions, Phys. Rev. Lett. **77**, 4693 (1996).

[7] I. Nemenman and W. Bialek, Occam factors and model independent Bayesian learning of continuous distributions, Phys. Rev. E **65**, 026137 (2002).

[8] J. B. Kinney, Estimation of probability densities using scale-free field theories, Phys. Rev. E **90**, 011301(R) (2014).

[9] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis, Phys. Rev. D **80**, 105005 (2009).

[10] J. C. Lemm, *Bayesian Field Theory* (Johns Hopkins University Press, Baltimore, 2003).

[11] T. E. Holy, Analysis of Data from Continuous Probability Distributions, Phys. Rev. Lett. **79**, 3545 (1997).

[12] V. Periwal, Reparametrization Invariant Statistical Inference and Gravity, Phys. Rev. Lett. **78**, 4671 (1997).

[13] T. Aida, Field Theoretical Analysis of On-Line Learning of Probability Distributions, Phys. Rev. Lett. **83**, 3554 (1999).

[14] D. M. Schmidt, Continuous probability distributions from finite data, Phys. Rev. E **61**, 1052 (2000).

[15] D. Ormoneit and H. White, An efficient algorithm to compute maximum entropy densities, Economet. Rev. **18**, 127 (1999).

[16] I. J. Good, Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, Ann. Math. Stat. **34**, 911 (1963).

[17] Available at https://github.com/jbkinney/14_maxent.

[18] B. W. Silverman, On the estimation of a probability density function by the maximum penalized likelihood method, Ann. Stat. **10**, 795 (1982).

[19] C. Gu, *Smoothing Spline ANOVA Models*, 2nd ed. (Springer, New York, 2013).

[20] Throughout this paper, integrals over x are restricted to the interval of length L.

[21] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003), Chap. 28.

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

[23] See Appendix D for a detailed derivation of Eq. (29).

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, New York, 2006).

[25] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction* (Springer, New York, 1990).

[26] J. Skilling, Probability and geometry, AIP Conf. Proc. **954**, 39 (2007).

[27] The one trial reported in Fig. 3(g) for which $\ell^* = \infty$ even though $K > 0$ is due to difficulties with the numerics in the $\ell \to \infty$ limit.

[28] See Appendix F for a discussion of earlier related work on the maximum penalized likelihood formulation of the density estimation problem, as well as a discussion of how this formulation relates to the Bayesian field theory approach.

[29] I. J. Good and R. A. Gaskins, Nonparametric roughness penalties for probability densities, Biometrika **58**, 255 (1971).

[30] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).