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 better-fitting nonparametric density estimate when the maximum entropy assumption is rejected.

PACS numbers: 02.50.-r, 11.10.Lm, 89.70.Cf, 02.60.-x

Research in nearly all fields of science routinely calls for the estimation of smooth probability densities from finite sampled data [1, 2]. Even in one dimension, however, basic questions remain about how best to accomplish this task. For instance, statistical physics has inspired two disparate approaches to this problem: maximum entropy density estimation [3, 4] and Bayesian field theory [5–13]. Despite their common origin, the relationship between these approaches has remained unclear.

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 [4, 14]. 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. It is therefore useful to 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 [15].

In the Bayesian field theory approach, one first defines a prior on the space of continuous probability densities. This is done 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 ℓ when quantifying smoothness, the optimal value for ℓ can be learned from the data in a natural way [5–7].

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 far [5–7] have imposed 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 ℓ is infinite.

More generally, every Bayesian field theory prior reduces to a MaxEnt hypothesis in the infinite smoothness limit. This finding provides a natural way to test the validity of any MaxEnt hypothesis on one's data. If the optimal ℓ is infinite, the corresponding MaxEnt density estimate is recovered. If instead the optimal ℓ is finite, the MaxEnt hypothesis is rejected and a better-fitting nonparametric density estimate is obtained.

A predictor-corrector algorithm allows the optimal value for ℓ, as well as the corresponding density estimate, to be rapidly computed. Perturbation theory in the large ℓ limit further provides a plug-in formula capable of rejecting the MaxEnt hypothesis. Open source software for performing these calculations in one dimension available at `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_{true} from these data .

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

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

This parametrization ensures that Q is positive and normalized [16]. We then adopt a field theory prior on ϕ. Specifically, we consider priors of the form

$$p(\phi|\ell) = \frac{e^{-S^0_\ell[\phi]}}{Z^0_\ell} \quad \text{where} \quad S^0_\ell[\phi] = \int \frac{dx}{L} \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2. \tag{2}$$

Here, S^0_ℓ plays the role of an action in statistical field theory, $Z^0_\ell = \int \mathcal{D}\phi \; e^{-S^0_\ell[\phi]}$ is the corresponding partition

function, ℓ is a length scale below which fluctuations in ϕ are strongly damped, and α 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 ϕ of the form

$$p(\phi|\text{data},\ell) = \frac{e^{-S_\ell[\phi]}}{Z_\ell}, \tag{3}$$

where

$$S_\ell[\phi] = \int \frac{dx}{L}\left\{\frac{\ell^{2\alpha}}{2}(\partial^\alpha\phi)^2 + NLR\phi + Ne^{-\phi}\right\} \tag{4}$$

is a nonlinear action, Z_ℓ 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 ϕ_ℓ, minimizes the action $S_\ell[\phi]$. To obtain a differential equation for ϕ_ℓ, previous work [5–7] imposed periodic boundary conditions on ϕ and used integration by parts to derive

$$\ell^{2\alpha}(-1)^\alpha\partial^{2\alpha}\phi_\ell + NLR - Ne^{-\phi_\ell} = 0. \tag{5}$$

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 ϕ_ℓ that has a unique solution.

To do so, we first define a differential operator Δ^α by the requirement that

$$\phi\Delta^\alpha\psi \equiv (\partial^\alpha\phi)(\partial^\alpha\psi) \tag{6}$$

for any two fields ϕ and ψ. This operator is Hermitian [18] even without boundary conditions on ϕ and ψ. We refer to Δ^α as the "bilateral Laplacian of order α."

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

$$\ell^{2\alpha}\Delta^\alpha\phi_\ell + NLR - Ne^{-\phi_\ell} = 0. \tag{7}$$

Any solution of Eq. (7) also satisfies Eq. (5) in the interior of the x-interval. However, Eq. (7) also specifies the behavior of ϕ_ℓ 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 ϕ_ℓ.

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 ϕ_ℓ becomes a G-dimensional vector, and the derivative operator ∂_G becomes a $(G-1)\times G$ matrix having elements $(\partial_G)_{ij} = \frac{G}{L}(-\delta_{i,j} + \delta_{i+1,j})$. Similarly, the standard α-order Laplacian $(-1)^\alpha\partial^{2\alpha}$ is given by $(-1)^\alpha\partial_{G-2\alpha+1}\cdots\partial_{G-1}\partial_G$, a $(G-2\alpha)\times G$ matrix. Eq. (5) therefore provides only $G-2\alpha$ equations for the G unknown values of ϕ_ℓ at the grid points. 2α boundary conditions are thus needed to obtain a unique solution. By contrast, the α-order bilateral Laplacian is represented by the $G\times G$ matrix $\Delta^\alpha_G = (\partial^\alpha_G)^\top\partial^\alpha_G$, where $\partial^\alpha_G = \partial_{G-\alpha+1}\cdots\partial_{G-1}\partial_G$. Eq. (7) therefore provides G equations for the G unknown elements of ϕ_ℓ, and has a unique solution without additional constraints.

Connection to maximum entropy – The kernel of Δ^α is spanned by polynomials of order $\alpha-1$. Multiplying Eq. (7) on the left by such polynomials and integrating, we find that the MAP density Q_ℓ, corresponding to the field ϕ_ℓ, exactly matches the first α moments of the data:

$$\int dx\, Q_\ell\, x^k = \int dx\, R\, x^k, \quad k = 0, 1, \ldots, \alpha-1. \tag{8}$$

This moment-matching behavior also holds in the discrete representation, since the kernel of Δ^α_G is spanned by vectors whose elements are the values taken by these same polynomials at the chosen grid points.

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

$$Q_\infty(x) = \frac{1}{L}\exp\left(-\sum_{k=0}^{\alpha-1}a_k x^k\right) \tag{9}$$

where the values of the coefficients a_k are determined by Eq. (8). We therefore see that the MAP density is identical to the MaxEnt density that matches these same moments [4]. This equivalence holds exactly in the discrete representation as well.

Choosing the length scale – To determine the optimal value for ℓ, we compute the evidence $p(\text{data}|\ell) = \int\mathcal{D}\phi\, p(\text{data}|\phi)p(\phi|\ell)$ [5, 6, 20, 21]. This quantity is infinitesimal when $\alpha > 1$, due to $p(Q|\ell)$ being an improper prior. 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 sufficiently large N, we find that

$$E = e^{S_\infty[\phi_\infty]-S_\ell[\phi_\ell]}\sqrt{\frac{\det_{\text{ker}}[e^{-\phi_\infty}]\det_{\text{row}}[L^{2\alpha}\Delta^\alpha]}{\eta^{-\alpha}\det[L^{2\alpha}\Delta^\alpha + \eta e^{-\phi_\infty}]}} \tag{10}$$

where $\eta = N(L/\ell)^{2\alpha}$ [22].

The evidence ratio E, by construction, approaches unity in the large ℓ 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 in the vicinity of $\eta = 0$, we find that

$$\ln E = K\eta + O(\eta^2), \tag{11}$$

where the coefficient K is [23]

$$K = \sum_{i>\alpha}\frac{Nv_i^2 - z_{ii}}{2\lambda_i} + \sum_{\substack{i>\alpha\\ j\le\alpha}}\frac{z_{ij}^2 + v_i z_{ijj}}{2\lambda_i\zeta_j} - \sum_{\substack{i>\alpha\\ j,k\le\alpha}}\frac{v_i z_{ij} z_{jkk}}{2\lambda_i\zeta_j\zeta_k}. \tag{12}$$

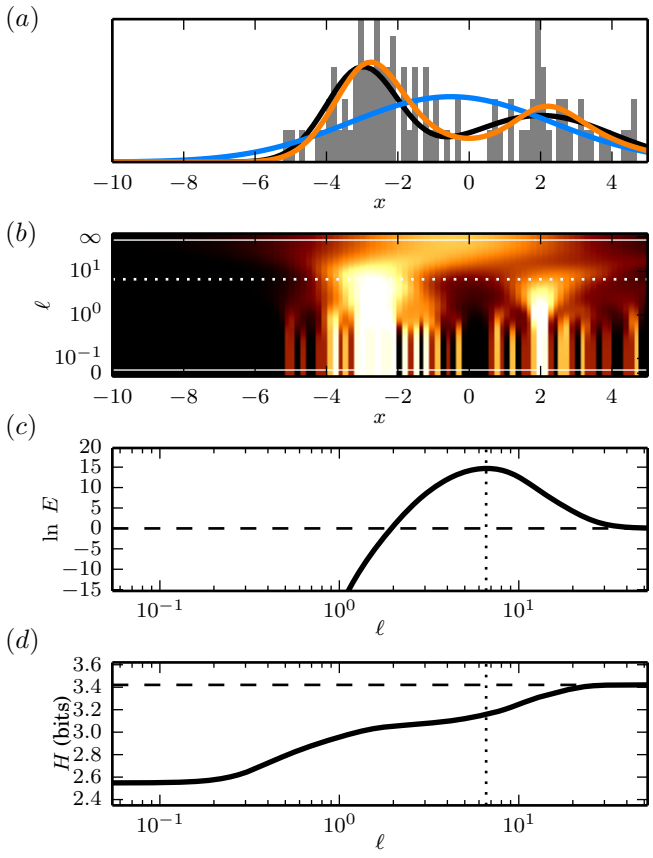

FIG. 1. Density estimation without boundary conditions in one dimension. (a) Shown are the histogram (gray) of N = 100 data points drawn from a simulated density $Q_{\rm true}$ (black) and binned at $G = 100$ grid points, the resulting field theory density estimate Q_{ℓ^*} (orange) computed using $\alpha = 3$, and the corresponding MaxEnt density Q_∞ (blue). (b) Heat map showing Q_ℓ at a range of ℓ values chosen by the homotopy algorithm, as well as at $\ell = \infty$ and $\ell = 0$. (c) The log evidence ratio E as a function of ℓ. (d) The differential entropy $H = -\int dx\, Q_\ell \ln Q_\ell$ [19] as a function of ℓ; the entropy at $\ell = \infty$ is indicated by the dashed line. Dotted lines in (b-d) indicate the value of ℓ^*.

Here, λ_i and $\psi_i(x)$ $(i = 1, 2, \ldots)$ respectively denote the eigenvalues and eigenfunctions of $L^{2\alpha}\Delta^\alpha$, and are indexed so that $\lambda_i = 0$ for $i \le \alpha$. The eigenfunctions are normalized so that $\int dx\, L^{-1}\psi_i\psi_j = \delta_{ij}$, and in the degenerate subspace $(i, j \le \alpha)$ 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 $z_{ijk} = \int dx\, Q_\infty \psi_i\psi_j\psi_k$.

Eq. (12) 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 larger evidence than $\ell = \infty$. In this case the MaxEnt estimate is seen to be non-optimal. On the other hand, if $K < 0$, then $\ell = \infty$ will be a local optimum that may or may not be a global optimum as well. Numerical computation of E over all length scales is needed to resolve this ambiguity.

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 [24] is used to compute the value of ϕ_ℓ over all length scales ℓ [25]. As in [7], this computation can be performed rapidly and deterministically due to the sparsity of Δ_G^α and the convexity of S_ℓ. 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 ℓ^*, 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 ℓ^* is finite, however, the estimated density Q_{ℓ^*} fits the data more closely and has lower entropy than Q_∞ (Figs. 1a,d). This improved fit and reduced entropy reflects the use of addition information in the data beyond the first α moments.

Bayesian field theory can therefore be used to test whether $Q_{\rm true}$ has a hypothesized functional form. For example, the value of ℓ^* that results from using $\alpha = 3$ provides a test of whether $Q_{\rm true}$ 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 α can be used to test other functional forms for $Q_{\rm true}$.

The "empirical Bayes" method used to select ℓ^* both here and in previous work [5, 6] 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 ℓ 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 $Q_{\rm true}$ 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 ℓ^*: 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 ℓ^* data sets; $K > 0$ and finite ℓ^* were found for all of the data sets analyzed in Fig. 2f .

Summary and discussion – The minimization of the action in Eq. (4), as well as the recovery of the MaxEnt density at $\ell = \infty$, was described early on by Silverman in the context of "penalized likelihood" [26, 27]. Later, Bialek et al. [5] had the critical insight that computing the optimum value of ℓ requires understanding not just the minimum of this action, but also the fluctuations about this minimum. Following standard practice in

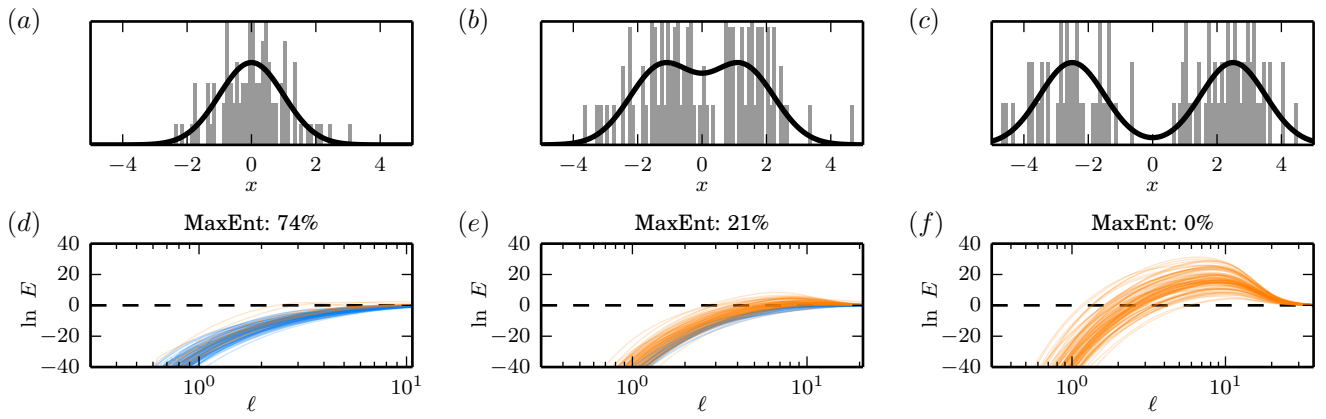

FIG. 2. The optimal estimated density may or may not have maximum entropy. Panels (a-c) show three difference choices for $Q_{\rm true}$ (black), along with a histogram (gray) of $N = 100$ data points binned at $G = 100$ grid points. $Q_{\rm true}$ was chosen to be the sum of two equally weighted normal distributions separated by a distance (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 $Q_{\rm true}$ 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 $\ell^* = \infty$.

field theory, periodic boundary conditions were imposed on candidate densities. This restriction made the standard Laplacian a Hermitian operator, facilitating further analysis. However, the MaxEnt density typically violates these boundary conditions. The ability of Bayesian field theory to subsume MaxEnt density estimation thus went unrecognized in [5] and in follow-up studies [6, 7].

Here we have seen that such boundary conditions are unnecessary. The bilateral Laplacian, defined in Eq. (6), 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.

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.

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 using Feynman diagrams, as in [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 reflects a lingering ambiguity in what one truly means by "density estimation." Different situations may ultimately call for the use of different priors, but understanding which situations call for which priors will require further investigation.

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, 7], 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.

* 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. (4).
[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] The subscripts "row" and "ker" respectively indicate restriction to the row space and kernel of Δ^α; see SM for a derivation of Eq. (10).
[23] See SM for a derivation of Eq. (12).
[24] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction* (Springer, 1990).
[25] This predictor-corrector algorithm is more robust and precise than the integration-based algorithm of [7]. See SM for algorithm details.
[26] B. W. Silverman, Ann. Stat. **10**, 795 (1982).
[27] P. P. B. Eggermont and V. N. LaRiccia, *Maximum Penalized Likelihood Estimation: Volume 1: Density Estimation* (Springer, 2001).
[28] I. J. Good and R. A. Gaskins, Biometrika **58**, 255 (1971).