Density estimation on small datasets

Wei-Chia Chen, Ammar Tareen, Justin B. Kinney[*]

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

How might a smooth probability distribution be estimated, with accurately quantified uncertainty, from a limited amount of sampled data? Here we describe a field-theoretic approach that addresses this problem remarkably well in one dimension, providing an exact nonparametric Bayesian posterior without relying on tunable parameters or large-data approximations. Strong non-Gaussian constraints, which require a non-perturbative treatment, are found to play a major role in reducing distribution uncertainty. A software implementation of this method is provided.

The need to estimate smooth probability distributions from a limited number of samples is ubiquitous in data analysis [1]. This "density estimation" problem also presents a fundamental conceptual challenge in statistical learning, important aspects of which remain unresolved. These outstanding problems are especially acute in the context of small datasets, where standard large-dataset approximations do not apply. Here we investigate the potential for Bayesian field theory, an area of statistical learning based on field-theoretic methods in physics [2–5], to estimate probability densities in this small data regime.

Density estimation requires answering two distinct questions. First, what is the *best* estimate for the underlying probability distribution? Second, what do other *plausible* distributions look like? Ideally, one would like to answer these questions by first considering all possible distributions (regardless of mathematical form), then identifying those that fit the data while satisfying a transparent notion of smoothness. Such an approach should not require one to manually identify values for critical parameters, specify boundary conditions, or make invalid mathematical approximations in the small data regime. However, the most common density estimation approaches, including kernel density estimation (KDE) [1] and Dirichlet process mixture modeling (DPMM)[6, 7], do not satisfy these requirements.

Previous work described a Bayesian field theory approach, called Density Estimation using Field Theory (DEFT) [8, 9], for addressing the density estimation problem in low dimensions. DEFT satisfies all of the above criteria except for the last one: in [8, 9], an appeal to the large data regime was used to justify a Laplace approximation (i.e., a saddle-point approximation) of the Bayesian posterior. This approximation facilitated the sampling of an ensemble of plausible densities, as well as the identification of an optimal smoothness lengthscale. Independent but closely related work [10] has also relied heavily on this approximation.

Here we investigate the performance of DEFT in the small data regime and find that Laplace approximation advocated in prior work can be catastrophic. This is because non-Gaussian aspects of the DEFT posterior are critical for suppressing "wisps" – large positive fluctua-

tions that otherwise occur in posterior-sampled densities. We further find that these non-Gaussian effects cannot be addressed perturbatively using Feynman diagrams, as has been suggested in other Bayesian field theory contexts [4, 5].

Happily, importance resampling [7] is found to rapidly and effectively correct for the Laplace approximation. The resulting DEFT algorithm thus appears to satisfy all of the above requirements for an ideal density estimation method in one dimension, and simulation tests show favorable performance relative to KDE and DPMM. A software implementation of this approach is provided.

We first recap the DEFT approach to density estimation [8, 9]. Consider N data points $\{x_i\}_{i=1}^N$ drawn from a smooth one-dimensional probability distribution $Q_{\text{true}}(x)$ that is confined to an x-interval of length L. From these data we wish to obtain a best estimate Q^* of Q_{true}, as well an ensemble of plausible distributions with which to quantify the uncertainty in this estimate. DEFT first reparametrizes each candidate distribution Q in terms of a field ϕ via $Q(x) = e^{-\phi(x)}/\int dx' e^{-\phi(x')}$. A Bayesian prior $p(Q|\ell) \propto \exp(-S_\ell^0[\phi])$ is then adopted, where the prior action

$$S_\ell^0[\phi] = \int \frac{dx}{L} \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2 \qquad (1)$$

is used to quantify the smoothness of ϕ. S_ℓ^0 involves two parameters: α, the order of x-derivative we wish to constrain, and ℓ, the smoothness lengthscale. α is chosen by the user, whereas ℓ (which we hold fixed for the moment) is ultimately selected based on the data. The resulting Bayesian posterior is given by $p(Q|\text{data}, \ell) \propto \exp(-S_\ell[\phi])$, which is defied by the posterior action

$$S_\ell[\phi] = S_\ell^0[\phi] + N \int dx \left[R\phi + \frac{e^{-\phi}}{L} \right], \qquad (2)$$

where $R(x) = \frac{1}{N} \sum_{i=1}^N \delta(x - x_i)$ is a (bin width zero) histogram summarizing the data. $S_\ell[\phi]$ is minimized at the maximum a posteriori (MAP) field ϕ_ℓ, which satisfies the equation of motion $0 = \ell^{2\alpha} \Delta^\alpha \phi_\ell + NLR - Ne^{-\phi_\ell}$ where Δ^α is the α-order bilateral Laplacian (described in [9]). The MAP field ϕ_ℓ is unique, even in the absence of boundary conditions, and is related to its cor-
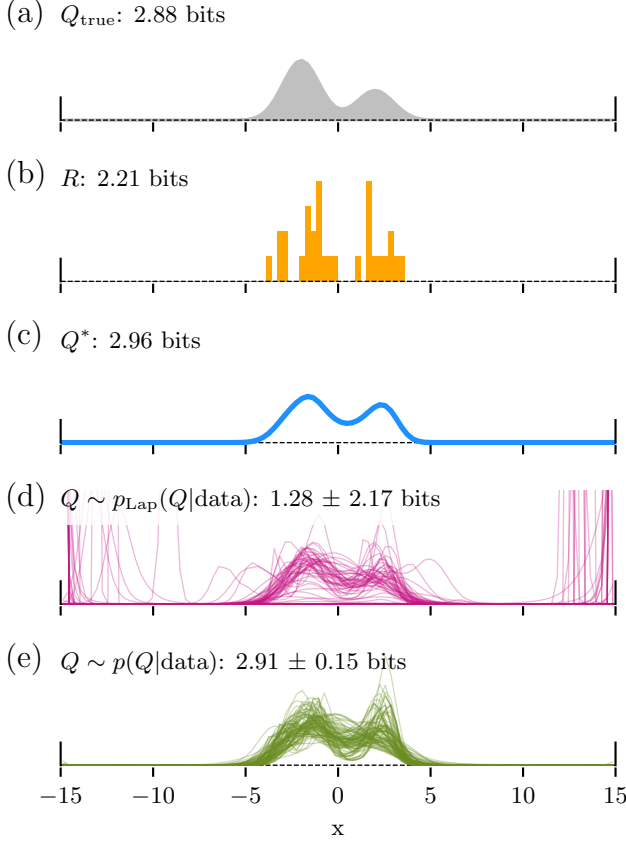
FIG. 1. (Color) **Density estimation using field theory**. (a) A Gaussian mixture distribution $Q_{\text{true}} = \frac{2}{3}\mathcal{N}(-2,1) + \frac{1}{3}\mathcal{N}(2,1)$ within the x-interval $(-15,15)$. (b) A histogram R of $N = 30$ data points sampled from Q_{true} and discretized to $G = 100$ grid points. (c) The corresponding estimate Q^* computed by DEFT using $\alpha = 3$ and the same grid as in (b). (d) 100 distributions sampled from the Laplace-approximated posterior $p_{\text{Lap}}(Q|\text{data})$, which accounts for uncertainty in ℓ as well as in Q. (e) 100 distributions generated using importance resampling of the Laplace ensemble. The differential entropies of the illustrated distributions are provided.

responding distribution via $Q_\ell(x) = e^{-\phi_\ell(x)}/L$. It cannot be found analytically but is readily computed after discretization of the x-domain at G equally-spaced grid points. In this discrete representation, R becomes a histogram with bin width $h = L/G$. As long as $h \ll \ell$, the choice of G will not greatly affect ϕ_ℓ. Finally, the optimal length scale ℓ^* is identified by maximizing the Bayesian evidence, $p(\text{data}|\ell)$. $Q^* = Q_{\ell^*}$ is then used as our best density estimate. Fig. 1(a-c) illustrates this procedure on simulated data.

To characterize the uncertainty in the DEFT estimate Q^*, we sample the Bayesian posterior $p(Q|\text{data}) = \int d\ell\, p(\ell|\text{data}) p(Q|\text{data},\ell)$. Each sample is generated by first drawing ℓ from $p(\ell|\text{data})$, then drawing Q from $p(Q|\text{data},\ell)$. Previous work [8] has suggested that this sampling task be performed using the Laplace approx-

imation, i.e., approximating $p(Q|\text{data},\ell)$ with a Gaussian that has the same mean and Hessian. Specifically, this approximation is given by $p_{\text{Lap}}(Q|\text{data},\ell) \propto \exp(-S_\ell^{\text{Lap}}[\phi])$ where

$$S_\ell^{\text{Lap}}[\phi] = S_\ell[\phi_\ell] + \frac{1}{2}\int dx\, \delta\phi\, \Lambda_\ell\, \delta\phi, \qquad (3)$$

$\delta\phi = \phi - \phi_\ell$, and $\Lambda_\ell = \frac{\ell^{2\alpha}}{L}\Delta^\alpha + NQ_\ell$. This Laplace approximation has the advantage that, after discretization of x and an eigendecomposition of the operator Λ_ℓ, posterior samples Q can be rapidly and independently generated (see [8]).

Fig. 1d shows Qs sampled from the Laplace posterior $p_{\text{Lap}}(Q|\text{data}) = \int d\ell\, p(\ell|\text{data}) p_{\text{Lap}}(Q|\text{data},\ell)$. Clearly something is very wrong. Although many of these Qs appear reasonable, many others exhibit wisps that have substantial probability mass far removed from the data. Note that these wisps primarily occur at the boundary of the x-interval, although some have local maxima in the bulk.

We hypothesized that wisps are an artifact of the Laplace approximation. To correct for potential inaccuracies of this approximation, we adopted an importance resampling approach [7]. For each sampled Q we computed a weight

$$w_\ell[Q] = \exp\left(S_\ell^{\text{Lap}}[\phi] - S_\ell[\phi]\right). \qquad (4)$$

We then resampled the Laplace ensemble with replacement, selecting each Q with a probability proportional to $w_\ell[Q]$. A mixture of such resampled ensembles across lengthscales ℓ was then used to generate an ensemble reflecting $p(Q|\text{data})$. Fig. 1e shows 100 distributions Q from this resampled posterior. Wisps no longer appear.

Eliminating wisps is especially important when estimating distribution entropy. Here the goal is to discern a value for the quantity $H_{\text{true}} = H[Q_{\text{true}}]$ where $H[Q] = -\int dx\, Q(x)\log_2 Q(x)$. Using the DEFT posterior ensemble, we can estimate H_{true} as $\widehat{H} \pm \widehat{\delta H}$, where $\widehat{H} = \langle H \rangle$ and $\widehat{\delta H} = \sqrt{\langle H^2 \rangle - \langle H \rangle^2}$, with $\langle \cdot \rangle$ denoting a posterior average. Previous work expressed hope that the ensemble provided by the Laplace approximation might serve this purpose [8]. But in this case we see that $\widehat{H}$ is far less accurate than the point estimates $H[R]$ or $H[Q^*]$, and $\widehat{\delta H}$ is enormous (Fig. 1d). Importance resampling fixes both problems: the resulting $\widehat{H}$ is closer to H_{true} than either point estimate, and $\widehat{\delta H}$ is remarkably small (Fig. 1e).

To further characterize the performance of DEFT, we simulated datasets of varying size N from a variety of Q_{true} distributions. We then asked two questions. First, how accurately does Q^* estimate Q_{true}? Second, how typical is Q_{true} among the distributions Q that are judged to be plausible? In both contexts, DEFT was compared to KDE and DPMM [11]. As in
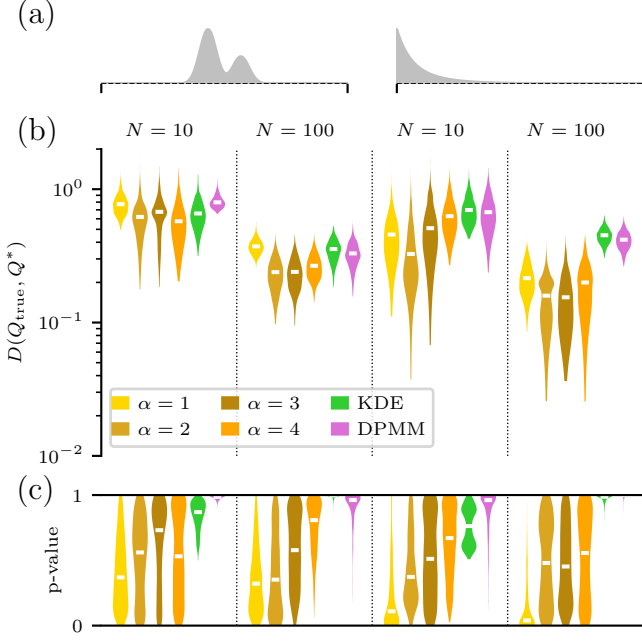
(a)

(b)

FIG. 2. (Color) **Performance of DEFT**. (a) DEFT, KDE, and DPMM were used to analyze data from two different Q_{true} distributions: the Gaussian mixture from Fig. 1a (left) or a Pareto distribution, $Q_{\text{true}}(x) = 5x^{-4}$, confined to the x-interval $(1,4)$ (right). (b) 100 datasets size $N = 10$ or $N = 100$ were generated for each Q_{true}. For each dataset, Q^* was computed by DEFT (using $G = 100$ and $\alpha = 1, 2, 3,$ or 4), by KDE, or by DPMM. Violin plots (with median indicated) show the resulting distances $D(Q^*, Q_{\text{true}})$. (c) Corresponding p-values relative to $D(Q, Q_{\text{true}})$ for plausible distributions Q.

[8] we quantified the difference between any two distributions Q and Q' using the geodesic distance $D(Q, Q') = 2\cos^{-1}\int dx \sqrt{Q(x)Q'(x)}$, which provides a well-behaved alternative to the Kullback-Leibler divergence; see [12**?**]. Fig. 2 shows the results of these performance tests for two different choices of Q_{true} (Fig. 2a). Fig. S1 in Supplemental Information provides analogous results for other Q_{true} distributions.

To answer the first question, we compared the distances $D(Q^*, Q_{\text{true}})$ obtained by each estimator on simulated datasets. Smaller values for these distances indicate better method accuracy. As illustrated in Fig. 2b, DEFT usually performed comparably to KDE and DPMM at $N = 10$, and somewhat better at $N = 100$. DEFT appears to have a particular advantage over both KDE and DPMM on Q_{true} distributions that bump up against one or both x-interval boundaries. One unsurprising exception is when Q_{true} is a mixture of Gaussians, in which case DPMM outperforms DEFT. Also unsurprising is that DEFT performs notably better with $\alpha = 2, 3,$ and 4 than with $\alpha = 1$, since $\alpha = 1$ yields non-smooth Q^* distributions with cusps at each data point [8, 13].

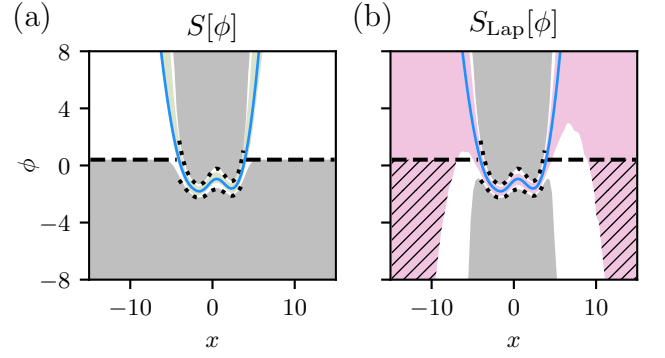To answer the second question, we computed the p-

FIG. 3. (Color) **Origin of wisps.** Panel (a) shows the field ϕ^* (blue) and its 68% confidence interval (olive) estimated using the exact action $S[\phi]$ based on the data in Fig. 1b. Gray indicates regions where $V > C$ with $C = 20$. $\delta\phi_{\text{rich}}^+$ and $\delta\phi_{\text{rich}}^-$ are indicated by dotted curves and $\delta\phi_{\text{poor}}^-$ by a dashed line. $\delta\phi_{\text{poor}}^+$ is not shown. Panel (b) shows the same quantities as in (a) but for the Laplace action $S^{\text{Lap}}[\phi]$. Specifically, gray indicates regions where $V_{\text{Lap}} > C$ and the 68% confidence interval of ϕ is computed in the Laplace approximation (magenta). Wisps arise from fields ϕ which, being constrained by V_{Lap} instead of V, fluctuate below $\delta\phi_{\text{poor}}^-$ (into the hatched region) and thus cause large upward fluctuations in Q.

value of $D(Q^*, Q_{\text{true}})$ relative to the $D(Q^*, Q)$ values observed for plausible distributions Q. If the plausible Qs accurately reflect distribution uncertainty, these p-values should be uniformly distributed between 0 and 1. Alternatively, p-values clustered close to 0 would indicate overestimated uncertainty, while values clustered close to 1 would indicate underestimated uncertainty. Fig. 2c shows our results. In general, the p-values for DEFT (with $\alpha = 2, 3,$ and 4) were distributed with remarkable uniformity. DEFT with $\alpha = 1$ tended to over-estimate uncertainties, whereas KDE and DPMM tended to underestimate uncertainties. Again, an exception was observed when Q_{true} was a Gaussian mixture; in this case DPMM gave nearly uniform p-values.

We now turn to the problem of understanding the origin of wisps. First we divide the action into three parts [14]: $S[\phi] = S[\phi^*] + S^0[\delta\phi] + \int \frac{dx}{L} V(\phi^*(x), \delta\phi(x))$, where

$$V(\phi^*, \delta\phi) = N e^{-\phi^*} f(\delta\phi) \qquad (5)$$

and $f(\delta\phi) = e^{-\delta\phi} - 1 + \delta\phi$. Here V acts as a potential landscape that constrains fluctuations in $\delta\phi$. The Laplace action S^{Lap} is recovered by replacing V with its quadratic approximation,

$$V_{\text{Lap}}(\phi^*, \delta\phi) = \frac{N}{2} e^{-\phi^*} \delta\phi^2. \qquad (6)$$

We now derive approximate expressions for the level curves of V, i.e., fluctuations $\delta\phi$ (expressed as functions of ϕ^*) for which V maintains a fixed value C. To facilitate this discussion we let $\delta\phi^+$ and $\delta\phi^-$ respectively
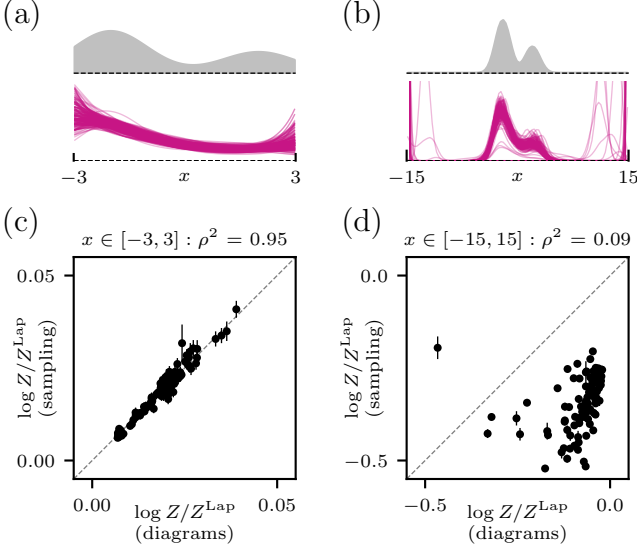
FIG. 4. (Color) **Wisps occur in a non-perturbative regime**. The accuracy of Feynman diagrams was assessed using density estimates of Q_{true} confined to the intervals $[-3, 3]$ (a,c) or $[-15, 15]$ (b,d). Q_{true} (gray) is shown along with 100 distributions Q (magenta) sampled at fixed $\ell = \ell^*$ from the Laplace-approximated posterior resulting from a dataset of size $N = 100$. Wisps are observed in (b) but not in (a). (c,d) Values for $\log(Z_\ell / Z_\ell^{\text{Lap}})$ computed for 100 different datasets using either importance resampling (Eq. 12) or Feynman diagrams (Eq. 11). These two quantities agree well in (c) but poorly in (d). Squared Pearson correlations are shown in the titles of (c,d).

denote positive and negative fluctuations. In the data rich regime of the x domain, defined by $Q^*(x) \gg C/NL$, we find $1 \gg f(\delta\phi^\pm) \approx \frac{1}{2}(\delta\phi^\pm)^2$. V_{Lap} is a good approximation of V here, and the corresponding level curves are

$$\delta\phi_{\text{rich}}^\pm \approx \pm\sqrt{\frac{2C}{N}}e^{\phi^*/2}. \tag{7}$$

In the data poor regime, defined by $Q^* \ll C/NL$, we find that $1 \ll f(\delta\phi^\pm)$. This implies that $\delta\phi^+ \gg 1 \Rightarrow f(\delta\phi^+) \approx \delta\phi^+$ and that $\delta\phi^- \ll -1 \Rightarrow f(\delta\phi^-) \approx e^{-\delta\phi^-}$. Using these approximations in Eq. 5 one obtains

$$\delta\phi_{\text{poor}}^+ \approx \frac{C}{N}e^{\phi^*}, \quad \delta\phi_{\text{poor}}^- \approx -\phi^* + \log\frac{N}{C}. \tag{8}$$

Unsurprisingly, the Laplace approximation works well in the data rich regime. Here, $\delta\phi^\pm \sim \pm e^{\phi^*/2}$ regardless of the specific value of C. In the data poor regime, however, the constraints on $\delta\phi$ are highly asymmetric: $\delta\phi^+ \sim e^{\phi^*}$ while $\delta\phi^- \sim -\phi^*$. Thus, the Laplace approximation exponentially overestimates the size of allowable downward fluctuations in ϕ, which correspond to large upward fluctuations in Q. This is how wisps arise (see Fig. 3).

Feynman diagrams provide a systematic approach for computing non-Gaussian corrections to the Laplace approximation. Might such diagrams provide a route for analytically countering the effect of wisps? The answer appears to be 'no'. Consider the potential V in Eq. 5 expanded to order m in $\delta\phi$:

$$V_m(\phi^*, \delta\phi) = Ne^{-\phi^*}\sum_{n=2}^{m}\frac{(-\delta\phi)^n}{n!}. \tag{9}$$

If the potential V_m is to suppress wisps, it must include enough terms to sufficiently approximate V when evaluated at $\delta\phi^- = -\phi^* + \log(N/C)$. This would require $m_{\text{min}} = \phi^* - \log(N/C)$ terms at the very least, since not until here do the terms in this power series begin to decrease. Thus, the number of terms that would be needed cannot be fixed *a priori*, but rather must increase with ϕ^*. This presents a major problem for Feynman-diagram-based expansions. Any diagram influenced by the the m_{min}'th term in Eq. 9 must contain an m_{min}'th order vertex. But m_{min} can be large; indeed, in Fig. 3 one finds $m_{\text{min}} > 100$ near the boundaries of the x-interval. Evaluating Feynman diagrams up to such high order is not feasible.

To empirically test the utility of low-order Feynman diagram expansions, we compared two different ways of computing the quantity $\log(Z_\ell / Z_\ell^{\text{Lap}})$, where

$$Z_\ell = \int \mathcal{D}\phi\, e^{-S_\ell[\phi]} \quad \text{and} \quad Z_\ell^{\text{Lap}} = \int \mathcal{D}\phi\, e^{-S_\ell^{\text{Lap}}[\phi]}, \tag{10}$$

are the partition functions for the exact and Laplace-approximated DEFT actions. At order N^{-1}, this log ratio can be estimated using three vacuum diagrams [15]:

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} \approx \mathord{\vcenter{\hbox{◯◯}}} + \mathord{\vcenter{\hbox{●—◯}}} + \mathord{\vcenter{\hbox{⬭}}}. \tag{11}$$

See Supplemental Information for the specific formulas used to evaluate these diagrams. Alternatively, one can compute this log ratio using the importance resampling weights in Eq. 4 via

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \log\langle w\rangle_{\text{Lap}|\ell}, \tag{12}$$

where the Laplace ensemble is generated at fixed ℓ.

Fig. 4 compares these two ways of computing $\log(Z_\ell/Z_\ell^{\text{Lap}})$ for two different choices of Q_{true}. The Feynman diagram approximation in Eq. 11 works well well when Q_{true} fills the entire x-interval, but in this case there are wisps (Fig. 4a,c). Alternatively, prominent wisps are observed when Q_{true} vanishes in large regions of the sampling domain, but in this case Eq. 11 proves to be a very bad approximation (Fig. 4b,d). These observations support our argument that suppressing wisps requires a fundamentally non-perturbative approach.

Finally, does the Laplace approximation adversely affect the identification of ℓ^*? Were this the case, Eq. 12 could be used to correct for the approximation $Z_\ell \approx Z_\ell^{\mathrm{Lap}}$ and thus remedy this problem. But this approximation seems to work well in practice when compared to the many-orders-of-magnitude-variation in Z_ℓ^{Lap} that are typically observed for different values of ℓ [8, 9]. Thus, correcting for the Laplace approximation appears to be non-essential when selecting a smoothness lengthscale.

Here we have shown that DEFT can effectively address density estimation needs on small datasets in one dimension. DEFT provides point estimates comparable to KDE and DPMM, but does not suffer from the multiple drawbacks of these other methods. In particular, the only fundamental parameter that the user must specify is a small positive integer α that defines the qualitative meaning of smoothness (via Eq. 1). As described in [9], α also controls the relationship of DEFT-based inference to standard maximum entropy methods [16]. But in our experience, using $\alpha = 3$ seems to work fine nearly all of the time. Other parameters, such as the number of grid points G, reflect computational practicalities. These can often be chosen automatically and have minuscule effects on the results as long as reasonable values are used.

DEFT also provides an ensemble of plausible distributions without requiring any large-data approximations. We have shown here that this capability requires a non-perturbative treatment of Bayesian posteriors that are highly non-Gaussian. This finding contrasts with prior literature, which has relied on the the Laplace approximation [8, 10] or on Feynman diagram corrections to this approximation [4, 5] for characterizing posterior ensembles.

DEFT therefore addresses a major outstanding need, not just in statistical learning theory, but also in the computational methods available for day-to-day data analysis. To this end we have developed an open source Python package called SUFTware. SUFTware allows users to apply DEFT in one dimension to their own data, and in the future will include additional field-theory-based statistical methods. This implementation is sufficiently fast for routine use; the computations for Fig. 1 takes about 0.25 seconds on a standard laptop computer. SUFTware has minimal dependencies, is compatible with both Python 2 and Python 3, and is readily installed using the `pip` package manager. See `http://suftware.readthedocs.io` for installation and usage instructions.

[*] Email correspondence to jkinney@cshl.edu
[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, 1986).
[2] W. Bialek, C. Callan, and S. Strong, Phys Rev Lett **77**, 4693 (1996).
[3] J. C. Lemm, *Bayesian Field Theory* (Johns Hopkins, 2003).
[4] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D **80**, 105005 (2009).
[5] T. Ensslin, arXiv [astro-ph.IM] (2013), 1301.2556v1.
[6] P. Müller, F. A. Quintana, A. Jara, and T. Hanson, *Bayesian Nonparametric Data Analysis* (Springer, 2015).
[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, and A. Vehtari, *Bayesian Data Analysis*, 3rd ed., Vol. 109 (CRC Press, 2013).
[8] J. B. Kinney, Phys Rev E **90**, 011301(R) (2014).
[9] J. B. Kinney, Phys Rev E **92**, 032107 (2015).
[10] J. Riihimäki and A. Vehtari, Bayesian Anal **9**, 425 (2014).
[11] For KDE, Q^* was defined as the estimate obtained from the full dataset, while estimates from bootstrap-resampled datasets were used as plausible densities Q. For DPMM, Gibbs sampling of the Bayesian posterior was used to generate plausible densities Q, and Q^* was defined as the mean of these densities. See Supplemental Information for details.
[12] J. Skilling, AIP Conf. Proc. **954**, 39 (2007).
[13] I. Nemenman and W. Bialek, Phys. Rev. E **65**, 026137 (2002).
[14] To simplify our discussion, we keep ℓ implicit and fix it to the optimal value ℓ^*.
[15] J. Zinn-Justin, *Path Integrals in Quantum Mechanics* (Oxford, 2010).
[16] L. R. Mead and N. Papanicolaou, J. Math. Phys. (1984).

SUPPLEMENTAL INFORMATION

Other density estimation methods

Here we describe the Kernel density estimation (KDE) and Dirichlet process mixture modeling (DPMM) algorithms used for the computations shown in Fig. 2 and Fig. S1.

Kernel Density Estimation

KDE is arguably the most common approach to density estimation in one dimension. Given data $\{x_i\}_{i=1}^N$, the KDE density estimate is given by

$$Q^*(x) = \frac{1}{N}\sum_{i=1}^N \frac{1}{w}K\left(\frac{x-x_i}{w}\right) \tag{13}$$

where $K(z)$ is the kernel function and w is the "bandwidth". In our computations we used the KDE estimator `scipy.stats.gaussian_kde` (SciPy v1.0.0) with default settings. This yielded KDE estimates computed using a Gaussian kernel,

$$K(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}, \tag{14}$$

and a bandwidth w chosen using Scott's rule:

$$w = 1.059AN^{-1/5}, \tag{15}$$

where N is the number of data points, $A = \min(\hat{\sigma}, \mathrm{IQR}/1.349)$, IQR is the interquartile range of the data,

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^N (x_i - \hat{\mu})^2, \quad \text{and} \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^N x_i. \tag{16}$$

KDE is a plug-in estimator, not a Bayesian estimator, and does not naturally specify a posterior that can be sampled. To quantify the uncertainty in KDE estimates, we therefore generated an ensemble of plausible KDE estimates Q by applying the same KDE algorithm to bootstrap-resampled versions of the dataset.

Dirichlet process mixture modeling

DPMM is arguably the most popular nonparametric Bayesian method for estimating probability densities. DPMMs have a hierarchical structure, in the sense that each data point is assumed to be drawn from one of a number of "clusters," with each cluster having a probability density defined by a fixed kernel function.

In our computations, we adopted the finite DPMM described in [6, 7]. Densities were assumed to be of the form

$$Q(x) = \sum_{h=1}^H w_h K_{m_h}(x), \tag{17}$$

where H is the number of clusters, w_h is the probability of cluster h, and m_h is the set of parameters defining the density of cluster h. $K_m(z)$ was assumed to be a Gaussian density specified by $m = (\mu, \sigma^2)$, i.e., a mean and a variance. A normal-inverse-gamma distribution was used as the prior on m:

$$p(\mu, \sigma^2) = \mathcal{N}(\mu|\hat{\mu}, \hat{\kappa}\sigma^2)\,\Gamma^{-1}(\sigma^2|\hat{\alpha}, \hat{\beta}), \tag{18}$$

where $\hat{\kappa} = 1$, $\hat{\alpha} = 1$, and $\hat{\beta} = \hat{\sigma}^2$. Here $\hat{\mu}$ and $\hat{\sigma}$ were defined as in Eq. 16. For the number of clusters we used $H = 10$. For each dataset, we used Gibbs sampling to sample an ensemble of plausible densities. The optimal estimate was then defined as the mean density in this ensemble. Following [7], our Gibbs sampling algorithm worked as follows. For each cluster $h = 1, 2, \ldots, H$, we chose an initial weight $w_h = 1/H$ and a set of kernel parameters $m_h \sim p(\mu, \sigma^2)$. The sampler was then run by iterating the following steps:

1. Data were redistributed across clusters. Specifically, each data point x_i was allocated to cluster h with probability

$$p(h|x_i) = \frac{w_h K_{m_h}(x_i)}{\sum_{h=1}^{H} w_h K_{m_h}(x_i)}. \tag{19}$$

2. The mean and variance of each cluster were updated using

$$m_h \sim \mathcal{N}(\mu_h|\hat{\mu}_h, \hat{\kappa}_h \sigma_h^2) \ \Gamma^{-1}(\sigma_h^2|\hat{\alpha}_h, \hat{\beta}_h), \tag{20}$$

where

$$\hat{\mu}_h = \hat{\kappa}_h \left(\frac{\hat{\mu}}{\hat{\kappa}} + n_h \langle x_h \rangle \right), \tag{21}$$

$$\hat{\kappa}_h = \frac{\hat{\kappa}}{1 + n_h \hat{\kappa}}, \tag{22}$$

$$\hat{\alpha}_h = \hat{\alpha} + \frac{n_h}{2}, \tag{23}$$

$$\hat{\beta}_h = \hat{\beta} + \frac{1}{2} \left(\sum_{i \in h} (x_i - \langle x_h \rangle)^2 + \frac{n_h}{1 + n_h \hat{\kappa}} (\langle x_h \rangle - \hat{\mu})^2 \right). \tag{24}$$

Here, x_h represents the set of data points belonging to cluster h and $n_h = |x_h|$.

3. The cluster weights were updated by sampling

$$w_1, \ldots, w_H \sim \text{Dirichlet}(1 + n_1, \ldots, 1 + n_H). \tag{25}$$

Feynman Diagrams

The action in Eq. 2 is given in the discrete space representation by

$$S_\ell[\phi] = \frac{\ell^{2\alpha}}{2G} \sum_{ij} \Delta_{ij}^\alpha \phi_i \phi_j + \frac{NL}{G} \sum_i R_i \phi_i + \frac{N}{G} \sum_i e^{-\phi_i}, \tag{26}$$

where N is the number of data points, G is the number of grid points, $i, j \in \{1, 2, \ldots, G\}$, L is the size of the bounding box, and ℓ is the smoothness length scale. We represent the departure of ϕ from the MAP field ϕ^ℓ using the rescaled fluctuation $x = \sqrt{N}(\phi - \phi^\ell)$. The action can then be expanded in the following way:

$$S_\ell[\phi] = S_\ell^{\text{Lap}}[\phi] + \frac{1}{3!} \sum_{ijk} \frac{B_{ijk}}{\sqrt{N}} x_i x_j x_k + \frac{1}{4!} \sum_{ijkl} \frac{C_{ijkl}}{N} x_i x_j x_k x_l + \cdots, \tag{27}$$

where the Laplace action is

$$S_\ell^{\text{Lap}}[\phi] = S_\ell[\phi^\ell] + \frac{1}{2} \sum_{ij} A_{ij} x_i x_j, \tag{28}$$

and

$$A_{ij} = \frac{1}{N} \frac{\partial^2 S_\ell}{\partial \phi_i \partial \phi_j} \bigg|_{\phi^\ell} = \frac{\ell^{2\alpha}}{NG} \Delta_{ij}^\alpha + \frac{1}{G} e^{-\phi_i^\ell} \delta_{ij}, \tag{29}$$

$$B_{ijk} = \frac{1}{N} \frac{\partial^3 S_\ell}{\partial \phi_i \partial \phi_j \partial \phi_k} \bigg|_{\phi^\ell} = -\frac{1}{G} e^{-\phi_i^\ell} \delta_{ijk}, \tag{30}$$

$$C_{ijkl} = \frac{1}{N} \frac{\partial^4 S_\ell}{\partial \phi_i \partial \phi_j \partial \phi_k \partial \phi_l} \bigg|_{\phi^\ell} = \frac{1}{G} e^{-\phi_i^\ell} \delta_{ijkl}. \tag{31}$$

Our goal is to compute $\log(Z_\ell/Z_\ell^{\text{Lap}})$ where, in the discrete space representation,

$$Z_\ell = \int d^G \phi \, e^{-S_\ell[\phi]} \quad \text{and} \quad Z_\ell^{\text{Lap}} = \int d^G \phi \, e^{-S_\ell^{\text{Lap}}[\phi]}. \tag{32}$$

The quantity $\log(Z_\ell/Z_\ell^{\text{Lap}})$ is conveniently given by the sum of connected vacuum diagrams [15]. At $O(N^{-1})$, the relevant diagrams contain only 3rd-order and 4th-order vertices. From the expansion we see that the values corresponding to these vertices are given by $-B_{ijk}/\sqrt{N}$ and $-C_{ijkl}/N$, respectively. We also need the propagator matrix P, which is given by the inverse of the Hessian A, i.e., $P_{ij} = (A^{-1})_{ij}$. We thus obtain

$$\log \frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \text{⧖} + \text{⊶⊷} + \text{⬭} + O(N^{-2}),$$

(33)

where the contribution from each diagram is

$$\text{⧖} = \frac{1}{8} \sum_{ijkl} \left(-\frac{C_{ijkl}}{N}\right) P_{ij} P_{kl} = -\sum_i \frac{e^{-\phi_i^\ell}}{8NG} (P_{ii})^2,$$

(34)

$$\text{⊶⊷} = \frac{1}{8} \sum_{ijk} \sum_{lmn} \left(-\frac{B_{ijk}}{\sqrt{N}}\right) \left(-\frac{B_{lmn}}{\sqrt{N}}\right) P_{ij} P_{kl} P_{mn} = \sum_i \sum_l \frac{e^{-\phi_i^\ell - \phi_l^\ell}}{8NG^2} P_{ii} P_{il} P_{ll},$$

(35)

$$\text{⬭} = \frac{1}{12} \sum_{ijk} \sum_{lmn} \left(-\frac{B_{ijk}}{\sqrt{N}}\right) \left(-\frac{B_{lmn}}{\sqrt{N}}\right) P_{il} P_{jm} P_{kn} = \sum_i \sum_l \frac{e^{-\phi_i^\ell - \phi_l^\ell}}{12NG^2} (P_{il})^3.$$

(36)

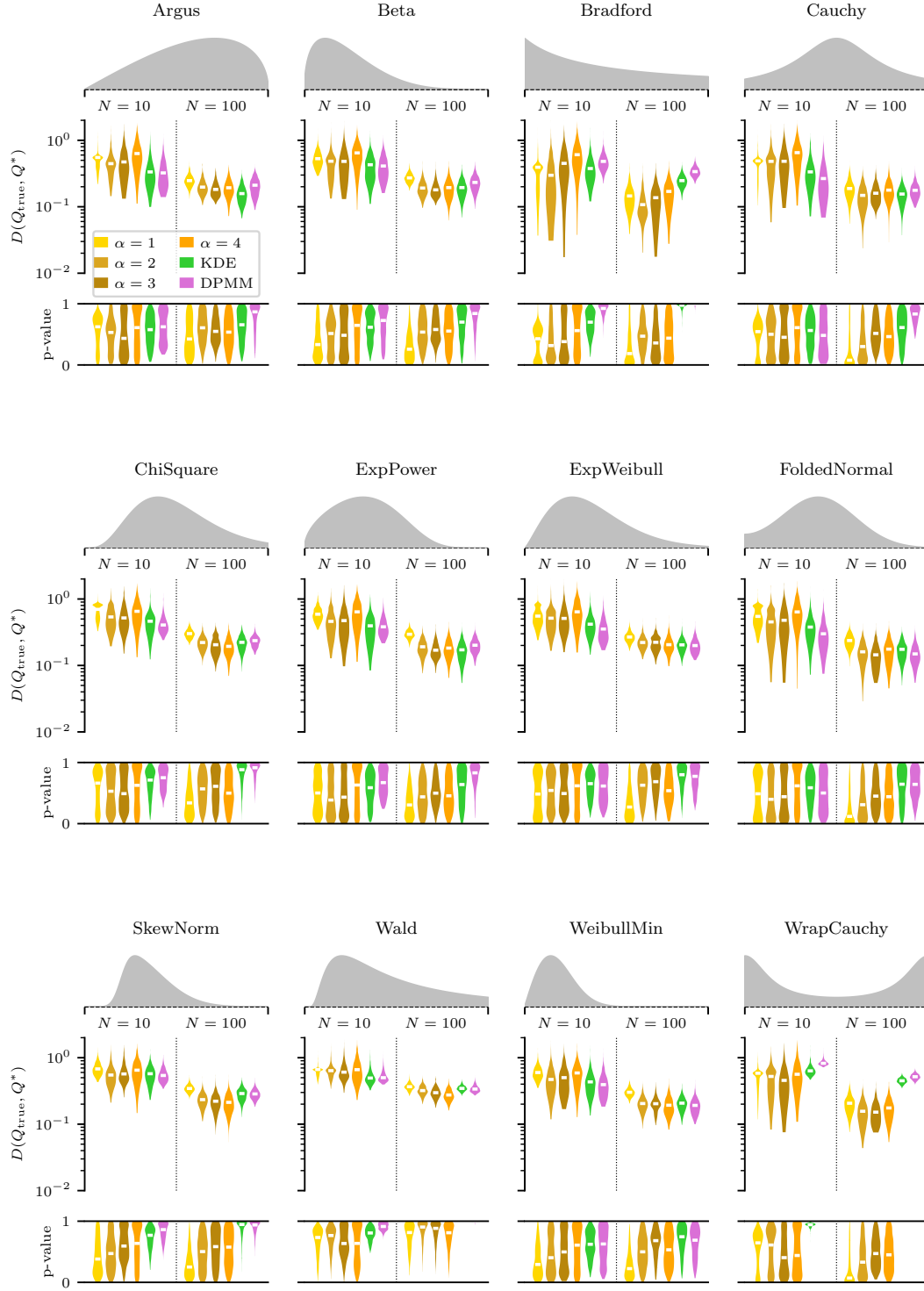These are the formulas used for the computations shown in Fig. 4c,d.

FIG. S1. (Color) **Comparison of density estimation methods across distributions**. The same analysis as in Fig. 2 was performed for twelve additional Q_{true} distributions, which were selected from among the built-in distributions in the `scipy.stats` library.