Density Estimation on Small Data Sets

Wei-Chia Chen, Ammar Tareen, and Justin B. Kinney*
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

(Received 12 April 2018; published 19 October 2018)

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 nonperturbative treatment, are found to play a major role in reducing distribution uncertainty. A software implementation of this method is provided.

DOI: 10.1103/PhysRevLett.121.160605

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 data sets, where standard large-data set 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.

Building on Ref. [2], previous work has 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 Refs. [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 length scale. 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 the Laplace approximation advocated in prior work can be catastrophic. This is because non-Gaussian features of the DEFT posterior are critical for suppressing "wisps"—large positive fluctuations 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]. These results are not specific to DEFT but rather reflect the fundamentally nonperturbative nature of the density estimation problem.

Happily, we find that importance resampling [7] can rapidly and effectively correct for the Laplace approximation. The resulting DEFT algorithm, which we have made available in robust and easy-to-use software, thus appears to satisfy all of the above requirements for an ideal density estimation method in one dimension. Tests of DEFT on simulated data show favorable performance relative to KDE and DPMM. We also illustrate the utility of DEFT on real data from the Large Hadron Collider [11] and from the World Health Organization (WHO) [12].

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 as an ensemble of plausible distributions with which to quantify the uncertainty in this estimate.

DEFT reparametrizes each candidate distribution Q in terms of a field ϕ via

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

After adopting a Bayesian prior that constrains the α-order x derivative of ϕ (denoted by $\partial^\alpha\phi$ in what follows), and accounting for the likelihood of the data given ϕ, one obtains a posterior distribution on ϕ. We represent this posterior as $p(Q|\text{data}, \ell) \propto \exp(-S_\ell[\phi])$ where

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

is the "posterior action" described in Ref. [9]. In Eq. (2), ℓ is a smoothness length scale that has yet to be determined, and $R(x) = (1/N)\sum_{i=1}^{N}\delta(x - x_i)$ is a histogram (of bin width zero) that summarizes the data. See Supplemental Material Section 1 (SM.1) [13] for details. The behavior of Q under this action $S_\ell[\phi]$ is the primary focus of the present Letter.

$S_\ell[\phi]$ is minimized at the maximum *a posteriori* (MAP) field ϕ_ℓ. The MAP field ϕ_ℓ is unique even in the absence of boundary conditions; see SM.2 [13] for details. Although ϕ_ℓ cannot be solved analytically, it is readily computed as the solution to a convex optimization problem 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. The optimal length scale ℓ^* is identified by maximizing the Bayesian evidence, $p(\text{data}|\ell)$; see SM.3 [13] for details. $Q^* = Q_{\ell^*}$ is then used as our best density estimate. Figures 1(a)–1(c) illustrate 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 approximation, i.e., approximating $p(Q|\text{data}, \ell)$ with a Gaussian distribution that has the same mean and Hessian. The corresponding action, $S_\ell^{\text{Lap}}[\phi]$, is thus quadratic in $\delta\phi = \phi - \phi_\ell$. This Laplace approximation has the advantage that posterior samples Q can be rapidly and independently generated [8].

Figure 1(d) shows multiple 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, others exhibit wisps that have substantial probability mass far removed from the data.

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 ϕ, we computed a weight

$$w_\ell[\phi] = \exp\left(S_\ell^{\text{Lap}}[\phi] - S_\ell[\phi]\right). \quad (3)$$

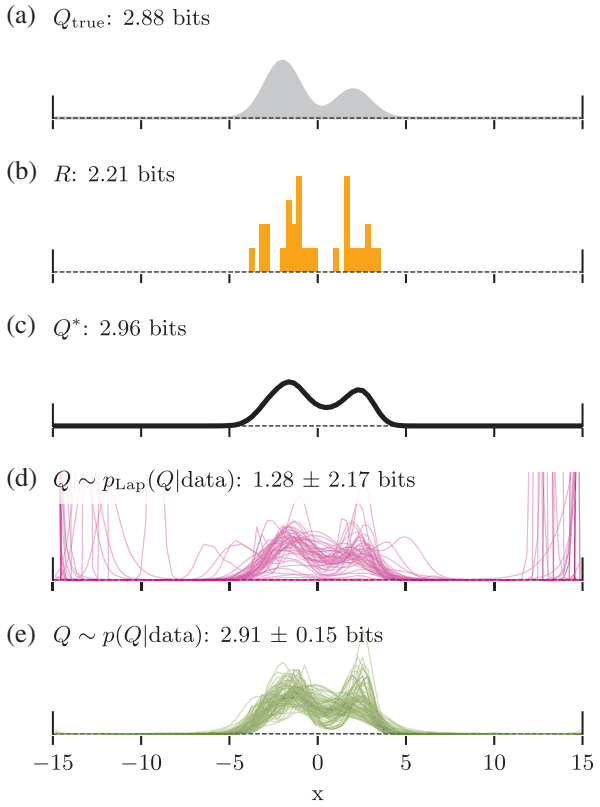

FIG. 1. Density estimation using field theory. (a) A Gaussian mixture distribution $Q_{\text{true}} = \frac{2}{3}\mathcal{N}(-2, 1) + \frac{1}{3}\mathcal{N}(2, 1)$ within the x interval $(-15, 15)$. (b) A histogram R of $N = 30$ data points sampled from Q_{true} and discretized to $G = 100$ grid points. (c) The corresponding estimate Q^* computed by DEFT using $\alpha = 3$ and the same grid as in (b). (d) One hundred 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) One hundred distributions generated using importance resampling of the Laplace ensemble. The differential entropies of the illustrated distributions are provided.

We then resampled the Laplace ensemble with replacement, selecting each ϕ (and thus Q) with a probability proportional to $w_\ell[\phi]$. A mixture of such resampled ensembles across length scales ℓ was then used to generate an ensemble reflecting $p(Q|\text{data})$; see SM.4 [13] for details. Figure 1(e) shows 100 distributions Q from this resampled posterior. Wisps no longer appear.

Eliminating wisps is especially important when estimating values for summary statistics, such as distribution entropy. In entropy estimation, 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 $\hat{H} \pm \delta\hat{H}$, where $\hat{H} = \langle H\rangle$ and $\delta\hat{H} = \sqrt{\langle H^2\rangle - \langle H\rangle^2}$, 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 $\hat{H}$ is far less accurate than the point estimates $H[R]$ or $H[Q^*]$, and $\delta\hat{H}$ is enormous [Fig. 1(d)]. Importance resampling fixes both problems: The resulting $\hat{H}$ is closer to H_{true} than either point estimate, and $\delta\hat{H}$ is remarkably small [Fig. 1(e)].

We now turn to the problem of understanding how wisps arise. To this end, we consider the variation in the action upon $\phi_\ell \to \phi_\ell + \delta\phi$. One finds that

$$\delta S_\ell[\phi_\ell + \delta\phi] = \int \frac{dx}{L}\frac{\ell^{2\alpha}}{2}(\partial^\alpha \delta\phi)^2 + \int \frac{dx}{L} V(\delta\phi), \tag{4}$$

where

$$V(\delta\phi) = NLQ_\ell[e^{-\delta\phi} - 1 + \delta\phi]. \tag{5}$$

The first (kinetic) term on the right-hand side of Eq. (4) imposes a smoothness constraint on $\delta\phi$, while the second (potential) term keeps $\delta\phi$ confined to a potential well consistent with the data. See SM.5 [13] for details. Note that V is convex, non-negative, and vanishes when $\delta\phi = 0$. By analogy to equipartition, we define n_{eff}, the effective number of degrees of freedom constrained by the data, as twice the value of the second term in Eq. (4) averaged over the posterior ensemble. Typical fluctuations $\delta\phi$ will therefore exhibit $V(\delta\phi) \sim n_{\text{eff}}/2$.

We now separately consider the "data-rich" regime of the x domain, which we define by $Q_\ell(x) \gg n_{\text{eff}}/2NL$, and the "data-poor" regime corresponding to $Q_\ell(x) \ll n_{\text{eff}}/2NL$. In the data-rich regime, fluctuations are small enough that V adheres well to its Laplace approximation, $V \approx NLQ_\ell\delta\phi^2/2$. Under this nearly symmetric potential, both positive fluctuations $\delta\phi^+$ and negative fluctuations $\delta\phi^-$ are constrained by

$$|\delta\phi^\pm| \sim \delta\phi_{\text{rich}} = \sqrt{\frac{n_{\text{eff}}}{NLQ_\ell}}. \tag{6}$$

By contrast, V is highly asymmetric in the data-poor regime and produces highly asymmetric fluctuations. Positive fluctuations satisfy $\delta\phi^+ \sim n_{\text{eff}}/2NLQ_\ell$, whereas negative fluctuations obey

$$-\delta\phi^- \sim \delta\phi^-_{\text{poor}} = \log\frac{n_{\text{eff}}}{2NLQ_\ell}. \tag{7}$$

See SM.5 [13] for more information.

The key point is that adopting $S_\ell^{\text{Lap}}[\phi]$ in place of $S_\ell[\phi]$ is equivalent to assuming the Laplace approximation for V throughout the entire x domain. Because $\delta\phi_{\text{rich}} \gg \delta\phi^-_{\text{poor}}$ in data-poor regions, the Laplace approximation greatly overestimates the size of downward fluctuations in ϕ. This results in the large upward fluctuations in Q that we identify as wisps. We note that wisps are especially prominent at the x-interval boundaries in Fig. 1 for two reasons: (i) Q_ℓ is especially small here, making these regions very data poor, and (ii) the kinetic term in Eq. (4), which is all that suppresses wisps in data-poor regions, is less effective at constraining $\delta\phi$ because data are present on only one side.

Feynman diagrams provide a general means of correcting for inaccuracies in Laplace approximations [14] and have been advocated in the context of some Bayesian field theory regression problems [4,5]. For density estimation, however, Feynman diagrams are ineffective if any region of the x interval is data poor. This is due to the action $S_\ell[\phi]$ being strongly coupled. For example, in the Bayesian evidence computations used to determine ℓ^*, DEFT estimates the action $Z_\ell = \int \mathcal{D}\phi e^{-S_\ell[\phi]}$ using the Laplace approximation $Z_\ell^{\text{Lap}} = \int \mathcal{D}\phi e^{-S_\ell^{\text{Lap}}[\phi]}$. See SM.3 [13] for details. At first, one might think it possible to correct for potential inaccuracies in this approximation using a series of vacuum diagrams (see SM.6 [13]), i.e.,

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \quad + \quad + \quad + \cdots. \tag{8}$$

However, as described in SM.8 [13], the number of diagrams needed to obtain accurate results is prohibitive when data-poor regions of the x interval are present. Fortunately, one can instead compute nonperturbative corrections to this log ratio using the importance resampling weights in Eq. (3) via

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \log\langle w_\ell\rangle_{\text{Lap}|\ell}. \tag{9}$$

See SM.7 [13] for details.

These results reflect a fundamental yet underappreciated aspect of density estimation: Unless data are observed throughout the x domain, the uncertainties in estimated probability densities require a nonperturbative treatment. Specifically, nonperturbative methods such as the Laplace approximation or Feynman diagrams can only be expected to work if $Q_{\text{true}}(x) \gtrsim 1/NL$ everywhere within the x domain. Very often, however, density estimation is applied to data like that in Fig. 1, which are localized far away from one or both x-interval boundaries. We argue that the analysis of such data will quite generally require a nonperturbative treatment.

To benchmark the performance of DEFT, we quantified its ability to estimate probability densities of known functional form. Specifically, we simulated data sets of varying size N from a variety of Q_{true} distributions, then asked two questions. First, how accurately does Q^* estimate Q_{true}?

Second, how typical is Q_{true} among the distributions in the Bayesian posterior? In both contexts, DEFT was compared to KDE and DPMM. See SM.9 [13] for details on how KDE and DPMM were implemented. Figure 2 shows the results of these performance tests for two different choices of Q_{true}. Figure S3 in Supplemental Material [13] provides analogous results for other Q_{true} distributions.

To answer the first question, we compared the Kullback-Leibler divergence $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ achieved by each estimator on each data set. Note that smaller values for these divergences indicate better method accuracy. As illustrated in Fig. 2(b), DEFT 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. Also unsurprising is that DEFT performs notably better with $\alpha = 2$, 3, and 4 than with $\alpha = 1$, since $\alpha = 1$ yields nonsmooth Q^* distributions with cusps at each data point [8,15].

To answer the second question, we computed where $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ falls within the distribution of divergences $D_{\text{KL}}(Q\|Q^*)$ observed for $Q \sim p(Q|\text{data})$. This location is naturally quantified by a p value corresponding to the null hypothesis that $Q_{\text{true}} \sim p(Q|\text{data})$. If Q_{true} is typical of plausible Qs, these p values should be uniformly distributed between 0 and 1. Alternatively, p values clustered close to 0 indicate that the posterior ensemble $p(Q|\text{data})$ overestimates how much Q_{true} diverges from Q^*, whereas p values clustered close to 1 indicate that $p(Q|\text{data})$ underestimates this uncertainty. Figure 2(c) shows our results for the two choices of Q_{true} in Fig. 2(a); the results for other choices of Q_{true} are shown in Fig. S3 [13]. In general, the p values for DEFT (with $\alpha = 2$, 3, and 4) were distributed with remarkable uniformity. DEFT with $\alpha = 1$ tended to overestimate uncertainties, whereas KDE and DPMM tended to underestimate uncertainties.

Finally, we illustrate the capabilities of DEFT using data reported in the initial observation of the Higgs boson [11] (see Fig. S4 [13] for an analysis of data from the WHO). Figure 3(a), which is a reconstruction of Fig. 4 of

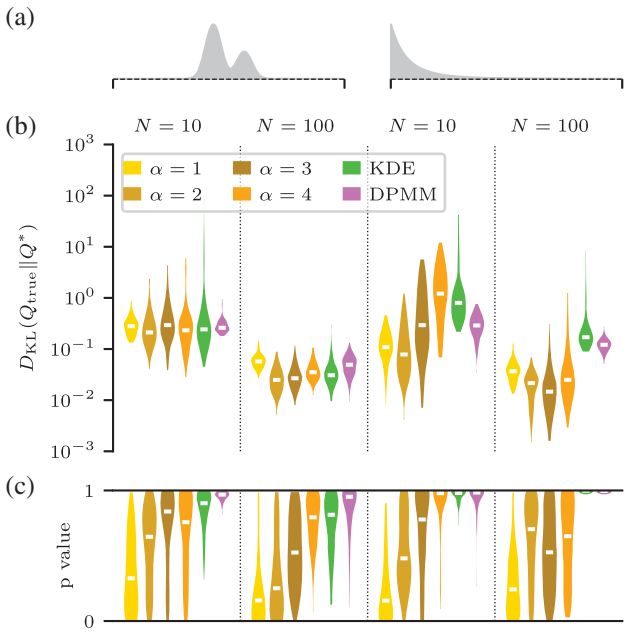

FIG. 2. Performance of DEFT. (a) DEFT, KDE, and DPMM were used to analyze data from two different Q_{true} distributions: the Gaussian mixture from Fig. 1(a) (left) and a Pareto distribution, $Q_{\text{true}}(x) = 3x^{-4}$, confined to the x interval (1,4) (right). (b) One hundred data sets of size $N = 10$ and one hundred data sets of size $N = 100$ were generated for each Q_{true}. For each data set, 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 Kullback-Leibler divergences $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$. (c) p values quantifying, for each simulated data set, the location of $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ within the distribution of $D_{\text{KL}}(Q\|Q^*)$ values observed for $Q \sim p(Q|\text{data})$.

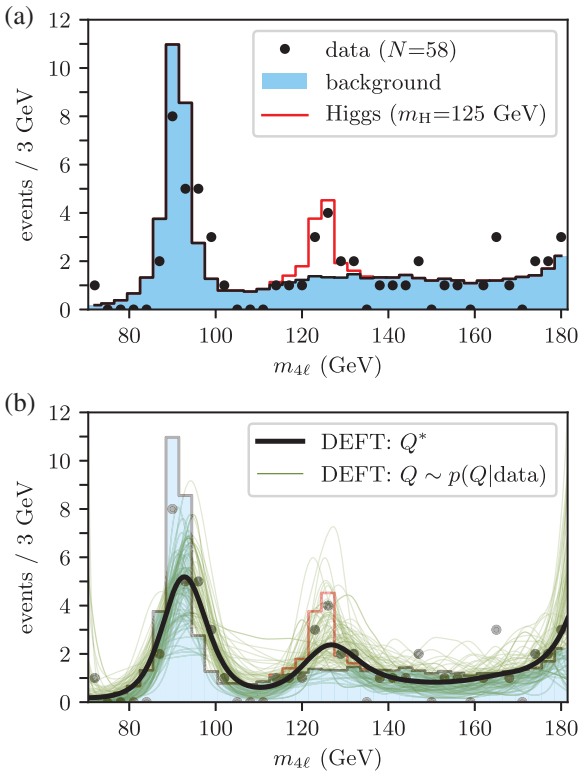

FIG. 3. DEFT applied to Higgs boson data. (a) A reconstruction of Fig. 4 from Ref. [11]. Dots (black) indicate the invariant masses of four-lepton decay events histogrammed across $G = 37$ bins of width 3 GeV each. Also shown are the number of events expected, based on Standard Model simulations, from either background decay processes (blue) or from the decay of a Higgs boson with mass of 125 GeV (red). (b) The optimal density estimate Q^* (black), along with 100 posterior samples $Q \sim p(Q|\text{data})$ (olive) computed by DEFT using the histogram data in panel (a).

Ref. [11], shows a histogram of the invariant masses of $N = 58$ four-lepton events observed by the CMS Collaboration at the Large Hadron Collider. Such events are generated by decays of the Higgs boson via $H \to ZZ \to 4\ell$, but they also arise from a variety of background decay processes. One of the challenges faced by the CMS Collaboration was determining whether these data exhibit a localized excess of events representing a possible Higgs resonance. Figure 3(b) shows DEFT applied to these data using default parameters. Despite Higgs decays representing only ∼10% of the observed events, DEFT detects a prominent local maximum near the Higgs resonance at $m_H = 125$ GeV. The confidence in this maximum can be quantified by sampling $Q \sim p(Q|\text{data})$: 81% of sampled Qs have exactly one local maximum between 110 GeV and 140 GeV (7% have no local maxima and 12% have multiple local maxima), and these maxima occurred at 127.1 ± 3.7 GeV.

Here we have shown that DEFT can effectively address density estimation needs on small data sets in one dimension. DEFT provides point estimates comparable to KDE and DPMM, but it does not suffer from the multiple drawbacks of these other methods. In particular, the only key parameter that the user must specify is a small positive integer α that defines the qualitative meaning of smoothness and which governs how DEFT relates to maximum entropy estimation (see Ref. [9]). In our experience, however, using $\alpha = 3$ seems to work well nearly all of the time. Other parameters, such as the number of grid points G, reflect computational practicalities. These parameters can be chosen automatically and have little effect on the results as long as reasonable values are used.

DEFT thus addresses a major outstanding need, not just in statistical learning theory but also in 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 it will include additional field-theory-based statistical methods. This implementation is sufficiently fast for routine use: The computations for Fig. 1 take about 0.25 seconds on a standard laptop computer (see SM.10 [13] for a discussion on computational complexity). SUFTWARE has minimal dependencies, is compatible with both PYTHON 2 and PYTHON 3, and is readily installed using the PIP package manager. SUFTWARE homepage [16] for installation and usage instructions.

We thank Kush Coshic for preliminary contributions to this project, as well as Serena Bradde, David McCandlish, and two anonymous referees for helpful feedback. This work was supported by a CSHL/Northwell Health Alliance grant to J. B. K. and by NIH Cancer Center Support Grant No. 5P30CA045508.

*Corresponding author.
jkinney@cshl.edu

[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
[2] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[3] J. C. Lemm, *Bayesian Field Theory* (Johns Hopkins, Baltimore, 2003).
[4] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D **80**, 105005 (2009).
[5] T. Enßlin, arXiv:1301.2556v1.
[6] P. Müller, F. A. Quintana, A. Jara, and T. Hanson, *Bayesian Nonparametric Data Analysis* (Springer, New York, 2015).
[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, and A. Vehtari, *Bayesian Data Analysis*, 3rd ed. (CRC Press, New York, 2013), Vol. 109.
[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] CMS Collaboration, Phys. Lett. B **716**, 30 (2012).
[12] World Health Organization Collaboration, *World Health Statistics 2017: Monitoring Health for the SDGs, Sustainable Development Goals* (World Health Organization, Geneva, 2017).
[13] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.121.160605 for a derivation of the DEFT algorithm as well as for other information.
[14] J. Zinn-Justin, *Path Integrals in Quantum Mechanics* (Oxford University, New York, 2010).
[15] I. Nemenman and W. Bialek, Phys. Rev. E **65**, 026137 (2002).
[16] SUFTWARE homepage http://suftware.readthedocs.io

Density estimation on small data sets: Supplemental material

Wei-Chia Chen, Ammar Tareen, and Justin B. Kinney

10 October 2018

Contents

List of Figures

SM.1 The posterior action $S_\ell[\phi]$

A derivation for Eq. 2 has already been reported in Ref. [1]. The derivation presented here, however, is more straightforward. The action in Eq. 2 is given by

$$S_\ell[\phi] = S_\ell^0[\phi] + S_{\text{data}}[\phi], \tag{S1}$$

where $S_\ell^0[\phi]$ is the "prior action", corresponding to a Bayesian prior $p(Q|\ell) \propto \exp(-S_\ell^0[\phi])$, while $S_{\text{data}}[\phi]$, the "likelihood action", is related to likelihood via $p(\text{data}|Q) \propto \exp(-S_{\text{data}}[\phi])$. DEFT uses a prior action of the form

$$S_\ell^0[\phi] = \int \frac{dx}{L} \frac{\ell^{2\alpha}}{2} (\partial^\alpha \phi)^2. \tag{S2}$$

The parameter α reflects a fundamental choice in how one defines "smoothness", and ℓ is a length scale below which fluctuations in ϕ are strongly damped. The derivation of $S_{\rm data}[\phi]$ is as follows. Suppose we are given N data points drawn from a probability distribution $Q_{\rm true}(x)$ that is confined to the interval $[x_{\rm min}, x_{\rm max}]$. Label these data in order of increasing value as $x_1, x_2, \ldots, x_N$. Next, imagine these data as being produced by a stochastic process in time, with x being the time variable and $r(x)$ being the instantaneous emission rate. The likelihood of the data is then given by

$$\begin{aligned}
(dx)^N p({\rm data}|r) &= \left[e^{-\int_{x_{\rm min}}^{x_1} dx\, r(x)}\right] \cdot [dx\, r(x_1)] \cdot \left[e^{-\int_{x_1}^{x_2} dx\, r(x)}\right] \cdot [dx\, r(x_2)] \;\cdots\; [dx\, r(x_N)] \cdot \left[e^{-\int_{x_N}^{x_{\rm max}} dx\, r(x)}\right] \\
&= (dx)^N \;\exp\left\{-\int_{x_{\rm min}}^{x_{\rm max}} dx\, r(x)\right\} \prod_{i=1}^{N} r(x_i) \\
&= (dx)^N \;\exp\left\{-\int dx\, r(x) + \sum_{i=1}^{N} \log\, r(x_i)\right\} \\
&= (dx)^N \;\exp\left\{-\int dx\, [r(x) - NR(x) \log\, r(x)]\right\}, \qquad \text{(S3)}
\end{aligned}$$

where $\int dx$ indicates integration over the entire x domain and $R(x) = N^{-1} \sum_{i=1}^{N} \delta(x - x_i)$ is the raw data density referred to in the main text. Next, we parametrize the emission rate $r(x)$ using the field $\phi(x)$ via

$$r(x) = \frac{N}{L} e^{-\phi(x)}. \qquad \text{(S4)}$$

The probability density corresponding to this rate is

$$Q(x) = \frac{r(x)}{\int dx'\, r(x')} = \frac{e^{-\phi(x)}}{\int dx'\, e^{-\phi(x')}}, \qquad \text{(S5)}$$

and so our definition of ϕ here is consistent with the definition of ϕ in the main text. We therefore see that the likelihood density in Eq. S3 is given by $p({\rm data}|\phi) \propto \exp(-S_{\rm data}[\phi])$ where the corresponding action (after dropping the constant term $N \log(L/N)$) is,

$$S_{\rm data}[\phi] = \int \frac{dx}{L} \left[NLR(x)\phi(x) + Ne^{-\phi(x)}\right]. \qquad \text{(S6)}$$

Plugging Eq. S2 and Eq. S6 into Eq. S1 gives Eq. 2 of the main text. Note the origin of the two terms in the integrand in Eq. S6: The term linear in ϕ comes from the exact locations of the N data points, whereas the nonlinear term (which leads to such interesting behavior) comes from regions of the x domain in which no data is observed.

We briefly discuss a subtle issue with the above derivation. The probability distribution $Q(x)$ is invariant under additive shifts in the underlying field, i.e., $\phi(x) \to \phi(x) + c$ for any constant c. By contrast, the likelihood action $S_{\rm data}[\phi]$ is not invariant under such transformations. This difference is due to Eq. S4 which, by specifying how the emission rate $r(x)$ relates to $\phi(x)$, introduces an additional assumption about how ϕ should be constrained by data. But although this additional assumption alters $p(\phi|{\rm data})$, it does not alter $p(Q|{\rm data})$. The more involved derivation of $S_\ell[\phi]$ provided in Ref. [1] demonstrates this fact explicitly.

SM.2 The MAP field ϕ_ℓ

To solve for ϕ_ℓ, the maximum *a posteriori* (MAP) field at length scale ℓ, we set $\delta S_\ell / \delta\phi = 0$. The resulting equation of motion is

$$\ell^{2\alpha} \Delta^\alpha \phi_\ell + NLR - Ne^{-\phi_\ell} = 0. \qquad \text{(S7)}$$

The operator Δ^α that appears here is the "bilateral Laplacian", which is described in Ref. [1]. Briefly, Δ^α is defined by the requirement that

$$\int dx\; \varphi \Delta^\alpha \phi = \int dx\; (\partial^\alpha \varphi)(\partial^\alpha \phi), \qquad \text{(S8)}$$

for any two fields φ and ϕ. This bilateral Laplacian is identical to the standard α-order Laplacian $(-1)^\alpha \partial^{2\alpha}$ in the interior of the x interval, but differs at the boundaries. Specifically, the standard α-order Laplacian requires the additional specification of 2α boundary conditions in order to be self-adjoint. By contrast, the bilateral Laplacian is self-adjoint without the specification of any boundary conditions. The equation of motion, Eq. S7, thus has a unique solution without the need to assume any boundary conditions on ϕ. See Ref. [1] for more information.

By integrating Eq. S7 we find that $\int dx\, e^{-\phi_\ell(x)} = L$, due to $\int dx\, R(x) = 1$ and $\int dx \Delta^\alpha \phi_\ell = \int dx (\partial^\alpha 1)(\partial^\alpha \phi_\ell) = 0$. The MAP density Q_ℓ thus has a simple form:

$$Q_\ell(x) = \frac{e^{-\phi_\ell(x)}}{L}. \tag{S9}$$

Similarly, multiplying Eq. S7 on the left by x^k for $k = 1, \ldots, \alpha - 1$ and integrating reveals that

$$\left\langle x^k \right\rangle_{Q_\ell} = \left\langle x^k \right\rangle_R, \tag{S10}$$

i.e., the first $\alpha - 1$ moments of Q_ℓ exactly match those of the data.

As described in Ref. [1], DEFT computes the MAP field ϕ_ℓ for a set of length scales $\{\ell_0, \ell_1, \ell_2, \ldots, \ell_K\}$, ranging from $\ell_0 = 0$ to $\ell_K = \infty$. These length scales are chosen so that neighboring MAP densities, Q_{ℓ_k} and $Q_{\ell_{k+1}}$, are approximately equally spaced along this "MAP curve", as quantified by the geodesic distance $D_{\text{geo}}(Q_{\ell_k}, Q_{\ell_{k+1}})$. We note that Q_0 is in fact the data histogram R, while Q_∞ is in fact the maximum entropy distribution consistent with the moment constraints in Eq. S10. See Ref. [1] for details.

SM.3 The evidence $p(\text{data}|\ell)$

The DEFT algorithm computes the MAP field at length scales spanning $\ell = 0$ to $\ell = \infty$. The optimal length scale ℓ^* is then computed by maximizing the Bayesian evidence $p(\text{data}|\ell)$. The key quantity needed for this procedure is the "evidence ratio", which is given by

$$E(\ell) = \frac{p(\text{data}|\ell)}{p(\text{data}|\infty)}. \tag{S11}$$

It can be shown that $E(\ell) = (Z_\ell / Z_\ell^0)/(Z_\infty / Z_\infty^0)$, where

$$Z_\ell = \int \mathcal{D}\phi\; e^{-S_\ell[\phi]} \quad \text{and} \quad Z_\ell^0 = \int \mathcal{D}\phi\; e^{-S_\ell^0[\phi]} \tag{S12}$$

respectively denote the posterior partition function and the prior partition function. The prior partition function Z_ℓ^0 can be computed analytically, although it has a divergence that must be regularized. By contrast, the posterior partition function Z_ℓ can only be analytically computed in the Laplace approximation. We therefore instead use the quantity

$$Z_\ell^{\text{Lap}} = \int \mathcal{D}\phi\; e^{-S_\ell^{\text{Lap}}[\phi]}, \tag{S13}$$

where $S_\ell^{\text{Lap}}[\phi]$ is the Laplace approximation of $S_\ell[\phi]$. The resulting evidence ratio in this approximation is found to be

$$E(\ell) = e^{S_\infty[\phi_\infty] - S_\ell[\phi_\ell]} \sqrt{\frac{\det_{\text{ker}}[e^{-\phi_\infty}] \det_{\text{row}}[L^{2\alpha}\Delta^\alpha]}{\eta^{-\alpha}\, \det[L^{2\alpha}\Delta^\alpha + \eta e^{-\phi_\ell}]}}, \tag{S14}$$

where $\eta = N(L/\ell)^{2\alpha}$, and "ker" and "row" respectively denote the kernel and row space of the bilateral Laplacian Δ^α. See Ref. [1] for details.

It should be emphasized that, although the Laplace approximation can be grossly innacurate when sampling $Q \sim p(Q|\text{data})$, it does not strongly affect the evidence ratio $E(\ell)$. This is because $\log E(\ell)$ typically varies over many orders of magnitude, whereas $\log(Z_\ell / Z_\ell^{\text{Lap}})$ varies with ℓ far less dramatically. This is demonstrated in Fig. S2 below. Nevertheless, the SUFTware implementation of DEFT includes an option to correct for this approximation using importance sampling, as described in Eq. 9 the main text.

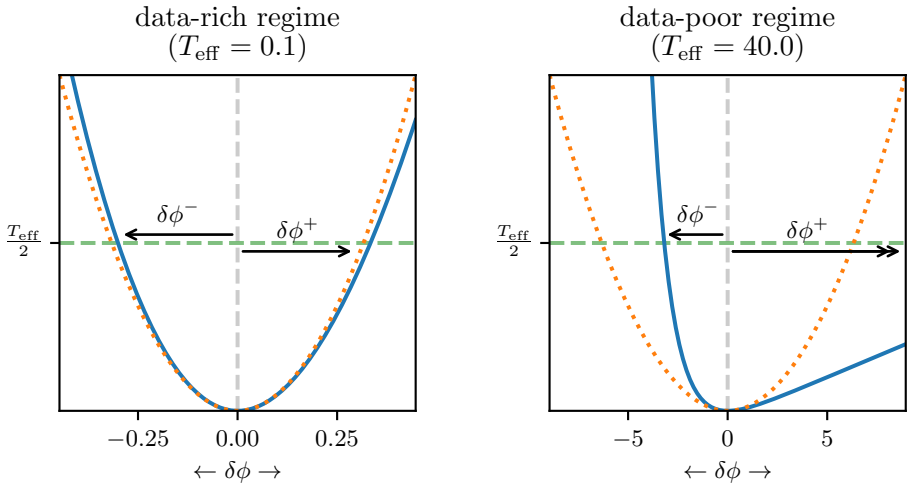

Figure S1: **Fluctuations in data-rich vs. data-poor regimes**. Solid blue lines indicate $f(\delta\phi)$ from Eq. S18. Dotted orange lines indicate the Laplace approximation $f_{\text{Lap}}(\delta\phi) = \delta\phi^2/2$. Dashed green lines indicate the value of half the effective temperature ($T_{\text{eff}}/2$) from Eq. S20. In the data-rich regime, $T_{\text{eff}}/2 \ll 1$ (left panel), resulting in nearly symmetric $\delta\phi^\pm$ fluctuations. In the data-poor regime, $T_{\text{eff}}/2 \gg 1$ (right panel), resulting in highly asymmetric fluctuations; in particular, the magnitude of $\delta\phi^-$ due to f is substantially less than would result from f_{Lap}. Note also that the very large positive fluctuations $\delta\phi^+$ in the data-poor regime have little noticeable effect on Q_ℓ, since they just push Q_ℓ closer to zero.

SM.4 Sampling the posterior $p(\phi, \ell|\text{data})$

The posterior probability $p(\phi, \ell|\text{data})$ can be decomposed as

$$p(\phi, \ell|\text{data}) = p(\phi|\ell, \text{data})\ p(\ell|\text{data}). \tag{S15}$$

This forms the basis for our posterior sampling procedure. First, we sample plausible ℓs from $p(\ell|\text{data})$. Note that $p(\ell|\text{data}) \propto p(\text{data}|\ell)\ p(\ell)$ by Bayes's Theorem. Assuming $p(\ell)$ is uniform over the length of the MAP curve as quantified by geodesic distance (see Ref. [1]), $p(\ell|\text{data})$ becomes proportional to the evidence ratio $E(\ell)$. We thus sample values of ℓ from the set $\{\ell_0, \ell_1, \ldots, \ell_K\}$ used to trace the MAP curve, each ℓ_k being selected with probability proportional to $E(\ell_k)$. For each of these ℓ values, we then sample plausible ϕs from $p(\phi|\ell, \text{data})$. Here we employ importance sampling. Specifically, we can rewrite the distribution $p(\phi|\ell, \text{data})$ as follows

$$p(\phi|\ell, \text{data}) = \frac{e^{-S_\ell[\phi]}}{Z_\ell} = \frac{e^{-S_\ell^{\text{Lap}}[\phi]}}{Z_\ell^{\text{Lap}}} \frac{w_\ell[\phi]}{\langle w_\ell \rangle_{\text{Lap}|\ell}} \propto p_{\text{Lap}}(\phi|\ell, \text{data}) w_\ell[\phi], \tag{S16}$$

where we have made use of Eq. S36 (derived below). Therefore, we first sample ϕs from the Laplace-approximated distribution $p_{\text{Lap}}(\phi|\text{data}, \ell)$, then correct for the non-Gaussian nature of the original distribution by resampling these ϕs using the importance weights $w_\ell[\phi]$.

SM.5 Origin of wisps

To derive Eqs. 4 and 5, it suffices to note that

$$S_\ell[\phi_\ell + \delta\phi] \quad = \quad S_\ell[\phi_\ell] + O(\delta\phi^2), \tag{S17}$$

because the equation of motion, Eq. S7, causes all first-order terms in $\delta\phi$ to cancel. Next, we express $V(\delta\phi) = NLQ_\ell f(\delta\phi)$ where

$$f(\delta\phi) = e^{-\delta\phi} - 1 + \delta\phi \tag{S18}$$

is just $e^{-\delta\phi}$ with the 0th and 1st order terms subtracted out. This function is plotted in Fig. S1. The key to deriving the magnitude of fluctuations $\delta\phi$ in different regimes is the relationship $\langle V(\delta\phi)\rangle = \frac{n_{\text{eff}}}{2}$, where n_{eff} is the effective number of degrees of freedom. We reexpress this as

$$\langle f(\delta\phi)\rangle = \frac{T_{\text{eff}}}{2}, \tag{S19}$$

where

$$T_{\text{eff}} = \frac{n_{\text{eff}}}{NLQ_\ell} \tag{S20}$$

is an effective temperature.

In the data-rich regime, $T_{\text{eff}}/2 \ll 1$. Therefore, $f(\delta\phi) \ll 1$ for typical fluctuations $\delta\phi$. As illustrated in Fig. S1 (left panel), the Laplace approximation works well in this regime. Setting

$$f(\delta\phi) \approx f_{\text{Lap}}(\delta\phi) = \frac{\delta\phi^2}{2} \sim \frac{T_{\text{eff}}}{2} \tag{S21}$$

and solving for $\delta\phi$ gives Eq. 6.

In the data-poor regime, $T_{\text{eff}}/2 \gg 1$. As illustrated in Fig. S1 (right panel), f is highly asymmetric in this regime and so the positive and negative fluctuations, $\delta\phi^+$ and $\delta\phi^-$, need to be treated separately. Specifically,

$$f(\delta\phi^+) \approx \delta\phi^+ \sim \frac{T_{\text{eff}}}{2}, \quad \text{whereas} \quad f(\delta\phi^-) \approx e^{-\delta\phi^-} \sim \frac{T_{\text{eff}}}{2}. \tag{S22}$$

Solving the latter condition for $\delta\phi^-$ gives Eq. 7. Note in Fig. S1 (right panel) how the Laplace approximation greatly overestimates the magnitude of negative fluctuations $\delta\phi^-$ in the data-poor regime.

SM.6 Computing $\log(Z_\ell/Z_\ell^{\text{Lap}})$ using Feynman diagrams

Here we show how Feynman diagrams can be used to compute $\log(Z_\ell/Z_\ell^{\text{Lap}})$, thereby obtaining corrections to the Laplace approximation. Our exposition closely follows that sketched by Zinn-Justin [2]. However, because Feynman diagrams are rarely used in the context of statistical inference, we have chosen to make these calculations explicit.

Upon discretization of the x interval using G grid points, the action in Eq. 2 becomes

$$S_\ell[\phi] = \frac{\ell^{2\alpha}}{2G}\sum_{ij}\Delta_{ij}^\alpha\phi_i\phi_j + \frac{NL}{G}\sum_i R_i\phi_i + \frac{N}{G}\sum_i e^{-\phi_i}, \tag{S23}$$

where $i, j = 1, 2, \ldots, G$. In what follows we represent fluctuations in ϕ about the MAP field (here denoted ϕ^ℓ) using the rescaled fluctuation $x = \sqrt{N}(\phi - \phi^\ell)$. The action can then be expanded in the following way:

$$S_\ell[\phi] = S_\ell^{\text{Lap}}[\phi] + \frac{1}{3!}\sum_{ijk}\frac{B_{ijk}}{\sqrt{N}}x_ix_jx_k + \frac{1}{4!}\sum_{ijkl}\frac{C_{ijkl}}{N}x_ix_jx_kx_l + \cdots, \tag{S24}$$

where the Laplace action is

$$S_\ell^{\text{Lap}}[\phi] = S_\ell[\phi^\ell] + \frac{1}{2}\sum_{ij}A_{ij}x_ix_j, \tag{S25}$$

and

$$A_{ij} = \frac{1}{N}\left.\frac{\partial^2 S_\ell}{\partial\phi_i\partial\phi_j}\right|_{\phi^\ell} = \frac{\ell^{2\alpha}}{NG}\Delta_{ij}^\alpha + \frac{1}{G}e^{-\phi_i^\ell}\delta_{ij}, \tag{S26}$$

$$B_{ijk} = \frac{1}{N}\left.\frac{\partial^3 S_\ell}{\partial\phi_i\partial\phi_j\partial\phi_k}\right|_{\phi^\ell} = -\frac{1}{G}e^{-\phi_i^\ell}\delta_{ijk}, \tag{S27}$$

$$C_{ijkl} = \frac{1}{N}\left.\frac{\partial^4 S_\ell}{\partial\phi_i\partial\phi_j\partial\phi_k\partial\phi_l}\right|_{\phi^\ell} = \frac{1}{G}e^{-\phi_i^\ell}\delta_{ijkl}. \tag{S28}$$

The quantity $\log(Z_\ell/Z_\ell^{\rm Lap})$ is conveniently given by the sum of connected vacuum diagrams. At $O(N^{-1})$, the relevant diagrams contain only 3rd-order and 4th-order vertices. From the expansion in Eq. S24 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

$$\log \frac{Z_\ell}{Z_\ell^{\rm Lap}} = \quad + \quad + \quad + O(N^{-2}), \tag{S29}$$

where the contribution from each diagram is

$$= \frac{1}{8}\sum_{ijkl}\left(-\frac{C_{ijkl}}{N}\right)P_{ij}P_{kl} = -\sum_i \frac{e^{-\phi_i^\ell}}{8NG}\left(P_{ii}\right)^2, \tag{S30}$$

$$= \frac{1}{8}\sum_{ijk}\sum_{lmn}\left(-\frac{B_{ijk}}{\sqrt{N}}\right)\left(-\frac{B_{lmn}}{\sqrt{N}}\right)P_{ij}P_{kl}P_{mn} = \sum_i\sum_l \frac{e^{-\phi_i^\ell-\phi_l^\ell}}{8NG^2}P_{ii}P_{il}P_{ll}, \tag{S31}$$

$$= \frac{1}{12}\sum_{ijk}\sum_{lmn}\left(-\frac{B_{ijk}}{\sqrt{N}}\right)\left(-\frac{B_{lmn}}{\sqrt{N}}\right)P_{il}P_{jm}P_{kn} = \sum_i\sum_l \frac{e^{-\phi_i^\ell-\phi_l^\ell}}{12NG^2}\left(P_{il}\right)^3. \tag{S32}$$

SM.7 Computing $\log(Z_\ell/Z_\ell^{\rm Lap})$ using importance sampling

Alternatively, the correction $\log(Z_\ell/Z_\ell^{\rm Lap})$ can be computed using importance sampling involving the weights w_ℓ in Eq. 3 of the main text. To see how, we express the partition function Z_ℓ as an average over the Laplace ensemble:

$$Z_\ell = \int \mathcal{D}\phi\; e^{-S_\ell[\phi]} \tag{S33}$$

$$= Z_\ell^{\rm Lap}\int \mathcal{D}\phi\; \frac{e^{-S_\ell^{\rm Lap}[\phi]}}{Z_\ell^{\rm Lap}}\; e^{S_\ell^{\rm Lap}[\phi]-S_\ell[\phi]} \tag{S34}$$

$$= Z_\ell^{\rm Lap}\int \mathcal{D}\phi\; p_{\rm Lap}(\phi|{\rm data},\ell)\; w_\ell[\phi] \tag{S35}$$

$$= Z_\ell^{\rm Lap}\;\langle w_\ell\rangle_{{\rm Lap}|\ell}, \tag{S36}$$

where $\langle\cdot\rangle_{{\rm Lap}|\ell}$ denotes the mean taken with respect to the Laplace posterior $p_{\rm Lap}(\phi|{\rm data},\ell)$, and w_ℓ denotes the importance sampling weights in Eq. 3. The quantity $\log(Z_\ell/Z_\ell^{\rm Lap})$ can thus be computed using Eq. 9 of the main text.

SM.8 Feynman diagrams vs. importance sampling

Perhaps disappointingly, Feynman diagrams generally do not work well in situations where wisps appear. This is because the posterior action in such cases is strongly coupled. To see this, consider an expansion of the potential V in Eq. 5 of the main text to m'th order in $\delta\phi$:

$$V_m(\delta\phi) = NLQ_\ell\sum_{n=2}^{m}\frac{(-\delta\phi)^n}{n!}. \tag{S37}$$

To produce accurate results, the potential V_m must include enough terms to sufficiently approximate V when evaluated at $\delta\phi = -\delta\phi^-_{\rm poor} = -\phi^* + \log(N/n_{\rm eff})$. This would require $m_{\rm min} = \phi^* - \log(N/n_{\rm eff})$ terms at the very least, since not until here do the (all positive) 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 $m_{\rm min}$'th term in Eq. S37 must contain an $m_{\rm min}$'th order vertex. But $m_{\rm min}$ can be quite large: For ϕ^* in Fig. 1, one

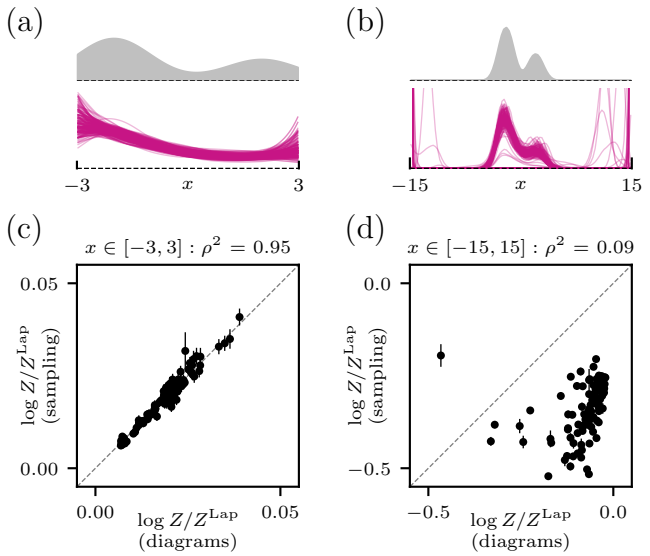

Figure S2: **Wisps appear only when the posterior action is strongly coupled**. The accuracy of Feynman diagrams was assessed using data drawn from the Q_{true} density in Fig. 1 confined to the intervals $[-3, 3]$ (a,c) or $[-15, 15]$ (b,d). (a,b) Q_{true} (gray) is shown along with 100 distributions Q (magenta) sampled at fixed $\ell = \ell^*$ from the Laplace-approximated posterior inferred from a data set 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 data sets, generated as above, using either Feynman diagrams (Eq. 8) or importance weights (Eq. 9). These two quantities agree well in (c) but poorly in (d). Squared Pearson correlations, ρ^2, are shown in the titles of (c,d).

finds $m_{\min} > 100$ near the boundaries of the x interval. Evaluating Feynman diagrams up to such high order is not feasible.

This expectation is confirmed in Fig. S2, which compares the two ways of computing $\log(Z_\ell / Z_\ell^{\text{Lap}})$ for two different choices of Q_{true}. The Feynman diagram approximation works well when Q_{true} fills the entire x interval, indicating that the action $S_\ell[\phi]$ is nearly quadratic and the corrections to the Laplace approximation are small. However, when Q_{true} vanishes in large regions of the x domain, the Feynman diagram approximation is very bad. In this case, the action $S_\ell[\phi]$ is strongly coupled and a fundamentally nonperturbative approach is required to compute the corrections.

Although the nonquadratic nature of the posterior action can lead to a partition function Z_ℓ differing from its Laplace-approximated value Z_ℓ^{Lap} by a large amount, we find that the Laplace approximation generally works well nevertheless for identifying the optimal length scale. This is because $\log Z_\ell^{\text{Lap}}$ typically varies by multiple orders of magnitude across different values of ℓ, thereby swamping potential inaccuracies in the $Z_\ell \approx Z_\ell^{\text{Lap}}$ assumption.

SM.9 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. S3.

SM.9.1 Kernel density estimation (KDE)

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

$$Q^*(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{w} K\left(\frac{x - x_i}{w}\right), \tag{S38}$$

where $K(z)$ is the kernel function and w is the "bandwidth". We used a Gaussian kernel,

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \tag{S39}$$

and chose the bandwidth w using cross-validation. Specifically, we considered 100 candidate bandwidths geometrically distributed between $w_{\min}$ (the minimum spacing between data points) and $w_{\max}$ (10 times the span of the data). We then chose the bandwidth w that maximized the jackknifed log likelihood

$$\mathcal{L}(w) = \sum_{i=1}^{N} \log Q^*_{-i}(x_i), \tag{S40}$$

where the subscript on Q^*_{-i} indicates the density Q^* computed as in Eq. S38 but using a data set missing the datum x_i.

KDE does not provide an explicit posterior on Q. Therefore, to compute p-values for Fig. 2 and Fig. S3, we approximated posterior samples $Q \sim p(Q|\text{data})$ by applying KDE to bootstrap-resampled data sets.

SM.9.2 Dirichlet process mixture modeling (DPMM)

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 kernel of pre-specified functional form.

In the computations for Fig. 2 and Fig. S3, we adopted the finite DPMM described in Refs. [3, 4]. Densities were assumed to be of the form

$$Q(x) = \sum_{h=1}^{H} w_h K_{m_h}(x), \tag{S41}$$

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:

$$p(\mu, \sigma^2) = \mathcal{N}(\mu|\hat{\mu}, \hat{\kappa}\sigma^2)\ \Gamma^{-1}(\sigma^2|\hat{\alpha}, \hat{\beta}), \tag{S42}$$

where $\hat{\kappa} = 1$, $\hat{\alpha} = 1$, $\hat{\beta} = \hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2, \quad \text{and} \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{S43}$$

The number of clusters was fixed at $H = 10$. For each data set, we used Gibbs sampling to obtain an ensemble of plausible densities representing $p(Q|\text{data})$. The optimal estimate Q^* was then defined as the mean density in this ensemble. Following Ref. [4], 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 selected according to the prior distribution $p(\mu, \sigma^2)$ in Eq. S42. 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
$$p(h|x_i) = \frac{w_h K_{m_h}(x_i)}{\sum_{h'=1}^{H} w_{h'} K_{m_{h'}}(x_i)}. \tag{S44}$$
2. The mean and variance of each cluster were updated using
$$m_h \sim \mathcal{N}(\mu_h|\hat{\mu}_h, \hat{\kappa}_h\sigma_h^2)\ \Gamma^{-1}(\sigma_h^2|\hat{\alpha}_h, \hat{\beta}_h), \tag{S45}$$

where

$$\hat{\mu}_h = \hat{\kappa}_h \left(\frac{\hat{\mu}}{\hat{\kappa}} + n_h \langle x \rangle_h \right), \tag{S46}$$

$$\hat{\kappa}_h = \frac{\hat{\kappa}}{1 + n_h \hat{\kappa}}, \tag{S47}$$

$$\hat{\alpha}_h = \hat{\alpha} + \frac{n_h}{2}, \tag{S48}$$

$$\hat{\beta}_h = \hat{\beta} + \frac{1}{2} \left(\sum_{i \epsilon h} (x_i - \langle x \rangle_h)^2 + \frac{n_h}{1 + n_h \hat{\kappa}} (\langle x \rangle_h - \hat{\mu})^2 \right). \tag{S49}$$

Here, $\langle x \rangle_h$ represents the mean value of data points belonging to cluster h, and n_h is the number of data points in this cluster.

3. The cluster weights were updated by sampling

$$w_1, \ldots, w_H \sim \text{Dirichlet}(1 + n_1, \ldots, 1 + n_H). \tag{S50}$$

SM.10 Computational complexity

An explicit expression for the algorithmic complexity of DEFT is not very helpful for understanding runtime performance. This is because DEFT involves multiple steps computed in series, the runtimes of which are governed by different parameters. In practice, we have found DEFT to be primarily limited by the number of grid points G. This is because a computation of the evidence ratio $E(\ell)$, as well as posterior sampling, requires a spectral decomposition of the $G \times G$ Hessian matrix at each length scale ℓ along the MAP curve. We note, however, that DEFT computations with $G = 100$ are generally quite fast (i.e., ~ 0.25 seconds on a standard laptop computer). Although DEFT requires histogramming the data, which is $\mathcal{O}(N)$, this is rarely the bottleneck in practice. In fact, we have found that the speed of DEFT often *increases* with N, since this leads to a shorter MAP curve, thus requiring fewer discrete length scales ℓ to be examined.

In our computations for Fig. 2 and Fig. S3, DEFT was often faster than our KDE and DPMM implementations. The use of jackknife cross-validation greatly slows down KDE in a manner that increases linearly with N. DPMM, on the other hand, is greatly slowed by its reliance on Gibbs sampling, which is necessitated by the nonconvexity of the parameter posterior. In fact, Gibbs sampling is needed not just to generate a posterior sample, but also to estimate Q^* (via a posterior mean). We note that the accuracy of KDE and DPMM is also very sensitive to the choice of kernel, especially when data is clustered near the x interval boundaries.

References

[1] J. B. Kinney, "Unification of field theory and maximum entropy methods for learning probability densities.," *Phys Rev E*, vol. 92, p. 032107, Sept. 2015.

[2] J. Zinn-Justin, *Path Integrals in Quantum Mechanics*. Oxford, 2010.

[3] P. Müller, F. A. Quintana, A. Jara, and T. Hanson, *Bayesian Nonparametric Data Analysis*. Springer, 2015.

[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, and A. Vehtari, *Bayesian Data Analysis*, vol. 109. CRC Press, 3rd ed., 2013.

[5] World Health Organization, *World health statistics 2017: Monitoring health for the SDGs, Sustainable Development Goals*. Geneva: World Health Organization, 2017.

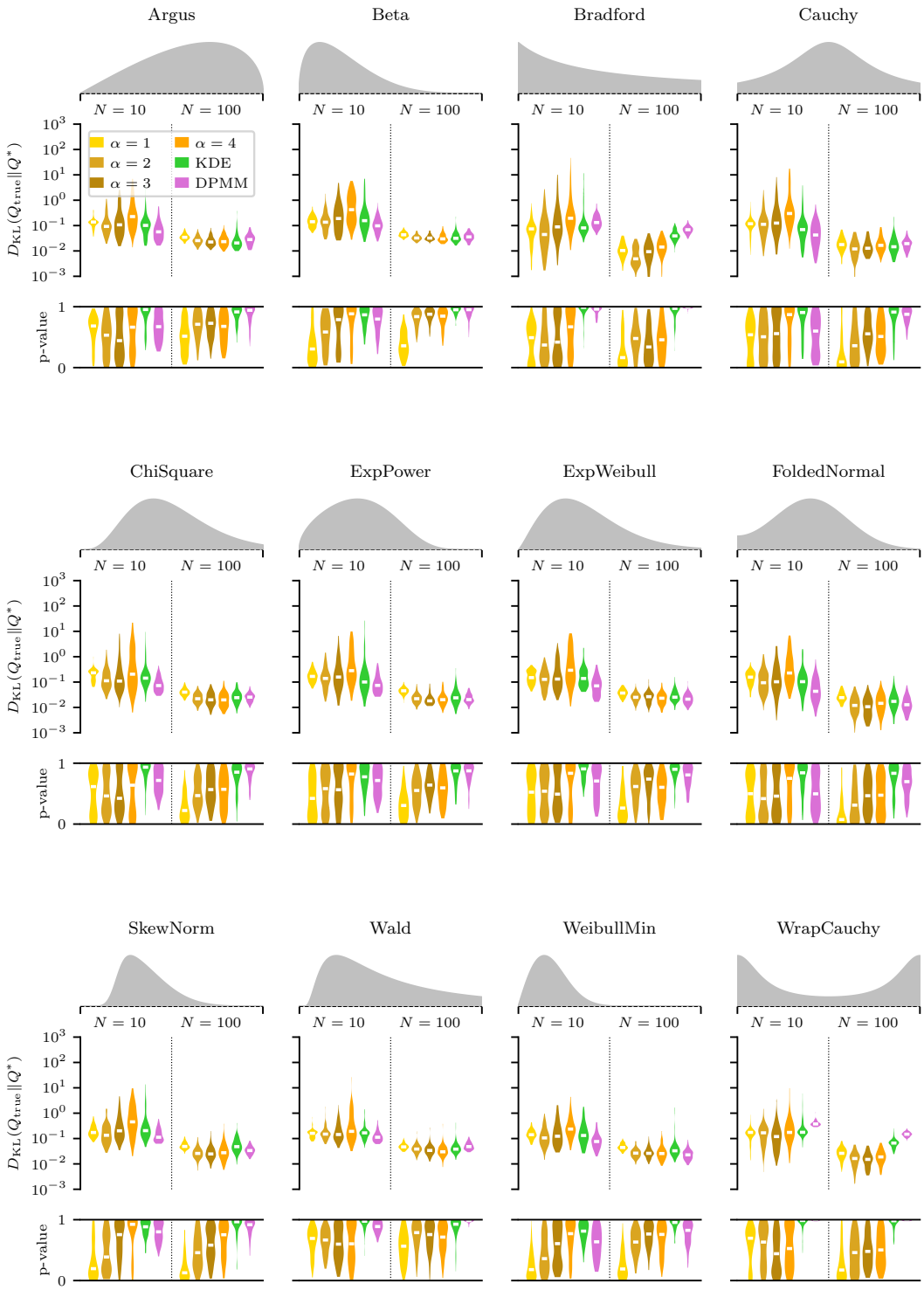

Figure S3: **Extension of Fig. 2 to other distributions**. The same analysis as in Fig. 2 was performed for twelve additional Q_{true} distributions, which were selected from the built-in distributions in the `scipy.stats` Python library.

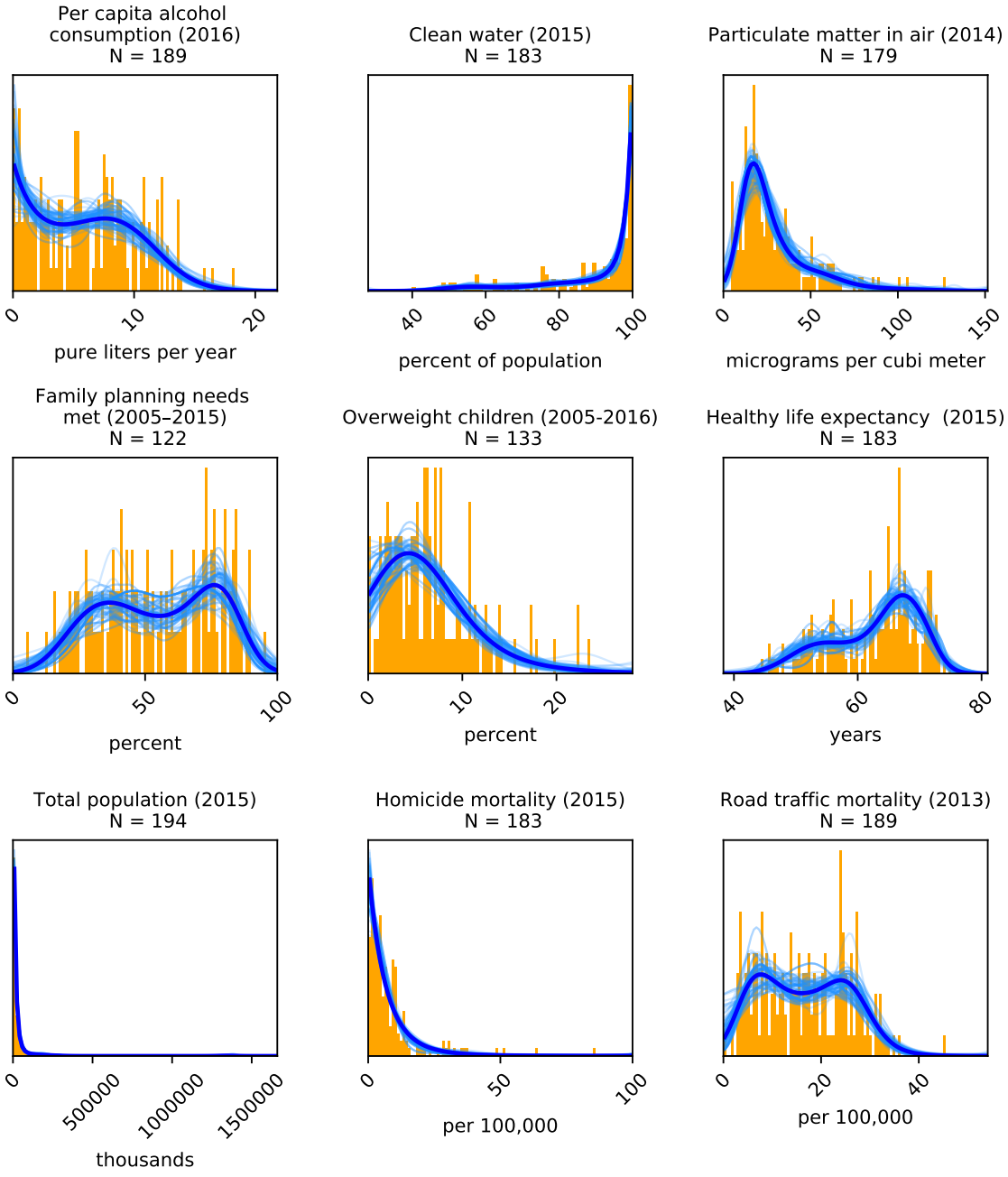

Figure S4: **Demonstration of DEFT on data from the World Health Organization (WHO)**. Densities were estimated for 9 different global health indicators reported by the WHO in [5]. Each datum corresponds to a different country; N varies between panels because of missing data in [5]. Orange shows a histogram of each global health indicator computed using $G = 100$ grid points. The best DEFT estimate Q^* is shown in dark blue, while 100 posterior-sampled densities $Q \sim p(Q|\text{data})$ are shown in light blue. As in Fig. 3, default DEFT parameters were used for all of these data sets.