Bayesian Reconstruction of the Density of States

Michael Habeck*

Max-Planck-Institute for Biological Cybernetics, Spemannstraße 38
and Max-Planck-Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany

(Received 31 January 2007; published 14 May 2007)

A Bayesian framework is developed to reconstruct the density of states from multiple canonical simulations. The framework encompasses the histogram reweighting method of Ferrenberg and Swendsen. The new approach applies to nonparametric as well as parametric models and does not require simulation data to be discretized. It offers a means to assess the precision of the reconstructed density of states and of derived thermodynamic quantities.

DOI: 10.1103/PhysRevLett.98.200601 PACS numbers: 05.10.Ln, 02.50.Cw, 02.70.Rr, 05.20.−y

The density of states is a key quantity in statistical mechanics. Because the temperature dependence of thermodynamic quantities such as the free energy can be derived from it, the density of states completely characterizes a thermodynamic system. Ferrenberg and Swendsen have shown that the precision of Monte Carlo estimates of the density of states and of the free energy improves dramatically if multiple canonical simulations are combined into a single consistent estimate [1]. The histogram reweighting method has been generalized to potentials of mean force and is now routinely used in biomolecular simulation [2,3].

One drawback of histogram reweighting is that it has been developed primarily for discrete systems and therefore requires binning of the variable of interest, even if the underlying system is large or continuous. A second drawback is that histogram reweighting lacks a probabilistic basis and can therefore only provide an *ad hoc* estimate of the error of the reconstructed density of states (Ref. [4] details an error analysis within the original histogram reweighting framework). Furthermore, it is not clear how to include properties like smoothness in the formalism.

This Letter introduces a Bayesian framework to infer the density of states. The new framework is completely general and applies to both discrete and continuous systems. In the following, simulations of a system with configuration x and energy $E(x)$ at inverse temperatures β are considered. Configurations are distributed according to the canonical ensemble $p(x|\beta) = \exp\{-\beta E(x)\}/Z(\beta)$ with partition function $Z(\beta)$. The density of states, $g(E)$, is defined as

$$g(E) = \int dx \delta[E - E(x)] \tag{1}$$

and quantifies the degree of energy degeneracy. Knowledge of g is desirable because it allows one to calculate the partition function by Laplace transformation, $Z(\beta) = \int dE g(E) \exp(-\beta E)$, and to derive the temperature dependence of important thermodynamic quantities.

The simulation data comprise M energy samples from canonical simulations at inverse temperatures $\beta_1, \ldots, \beta_M$ with sample sizes $N_1, \ldots, N_M$, respectively. The complete data set is $D = \{(\beta_i, E_{i1}, \ldots, E_{iN_i}), i = 1, \ldots, M\}$. A Bayesian approach to reconstruct the density of states starts from a probabilistic model for the simulation data. Assuming that the configurations have been simulated correctly, the likelihood of generating a single energy E at β is $p(E|\beta) = g(E)e^{-\beta E}/Z(\beta)$. Assuming further that the members of one energy sample are statistically independent, the probability of the whole data set is $p(D|g) = \prod_{i,j} p(E_{ij}|\beta_i)$. If constant terms are neglected, the likelihood of the data, viewed as a function of the density of states, becomes

$$L(g) = \prod_{i=1}^{M} \prod_{j=1}^{N_i} g(E_{ij}) \Big/ \prod_{i=1}^{M} [Z(\beta_i)]^{N_i}. \tag{2}$$

After pooling all data in one energy distribution $H(E) = \sum_{ij} \delta(E - E_{ij})$, the likelihood function can be written as

$$L(g) = \exp\left\{ \int dE H(E) \ln g(E) - \sum_i N_i \ln Z(\beta_i) \right\}. \tag{3}$$

$L(g)$ involves the cross entropy between the empirical energy distribution and the density of states and a weighted sum over the log-partition functions. Without this second term, reconstruction of g would be a standard density estimation problem for which H is the best estimate. The second term effectively reweights this estimate: $Z(\beta)$ measures the overlap between $e^{-\beta E}$ and $g(E)$ and becomes minimal for g accumulating probability mass in high-energy regions $E \gg 1/\min\{\beta_i\}$. The maximum likelihood estimate fulfills $\delta L(g)/\delta g(E) = 0$ leading to

$$g(E) = \frac{H(E)}{\sum_i N_i e^{-\beta_i E}/Z(\beta_i)}. \tag{4}$$

For discrete systems, H counts how often the different energy levels were observed, in which case Eq. (4) is identical to the histogram reweighting scheme of Ferrenberg and Swendsen. For continuous systems, H is a sum of delta peaks centered at the data, and therefore the estimated density of states the weighted superposition of

these peaks. This estimate suffers from overfitting because it only places nonzero probability at exactly the observed energies.

Overfitting will always be an issue for continuous systems because then one tries to estimate a density function, i.e., an infinite dimensional object, from finite data. But even for finite discrete systems, overfitting can arise if the data are of poor quality or sparse. A way to alleviate overfitting is to control the complexity of g. In a Bayesian framework, such a regularization is implemented via a prior probability $\pi(g)$ encoding general knowledge about g. Bayes' theorem [5] yields the posterior probability

$$p(g) \propto L(g)\pi(g) \tag{5}$$

defined on the space of all admissible densities of states. In the nonparametric setting, one does not assume a specific functional form but tries to determine g entirely from the data. Alternatively, one could model g parametrically using a family of functions involving parameters θ. Then the posterior distribution (5) is a probability density over the θ parameter space. The prior probability can be chosen freely, which may appear as a source of bias and prejudice. However, the prior probability should be understood as a representation of objective information on g including, for example, properties such as smoothness, convexity, or asymptotic behavior. Different prior assumptions can be compared quantitatively using Bayesian model comparison techniques [5].

Let us illustrate nonparametric estimation of g for an Ising model on a square lattice with L^2 spins and $K = L^2 - 1$ energy levels E_k, subject to periodic boundary conditions [6]. For this finite system, a nonparametric approach describing g as a K dimensional probability vector does not seem problematic. The Dirichlet prior, $\pi(g) \propto \prod_k g_k^{n_k - 1}$, references g to some initial guess n of the density of states via their cross entropy. Such a prior guess could be obtained, for example, from a previous simulation or from an approximate analytical result; the least informative choice would be constant n. Maximization of the posterior distribution (5) results in the modified histogram equations:

$$g_k = \frac{H_k + n_k - 1}{\sum_{i=1}^{M} N_i \exp\{-\beta_i E_k\}/Z(\beta_i)}. \tag{6}$$

The Dirichlet prior augments the histogram H_k with "pseudo counts" n_k. For the uninformative prior $n_k = 1$, these equations are identical to those of Ferrenberg and Swendsen.

The density of states for a system of size $L = 8$ was reconstructed from 11 independent simulations at inverse temperatures $\beta_i = -1 + i/5$, $i = 0, \ldots, 10$. At each β, 10^5 configurations were generated by flipping randomly selected spins; spin flips are accepted according to the Metropolis criterion; out of all 10^5 configurations, only a subset with autocorrelation less than 0.1 was kept. This resulted in 5464 energy values in total; N_i ranges from 100 to 2358. Figure 1 shows the true density of states g_{true} [6], the maximum posterior estimate g_{bayes} based on an uninformative prior $n_k = 1$, and the naive estimate obtained by unweighted averaging: $g_{\text{naive}}(E) \propto \sum_i H_i(E) e^{\beta_i E}$, where $H_i(E)$ is the empirical energy distribution of the ith heat bath. The Bayesian estimate is much closer to the exact density of states than the naive estimate. Moreover, the reconstructed energy distributions $g_{\text{bayes}}(E)e^{-\beta_i E}$ match the empirical distributions $H_i(E)$ very well.

When using an uninformative Dirichlet prior, maximization of the posterior probability [Eq. (5)] and histogram reweighting lead to the same results. A truly Bayesian analysis, however, does not rely on a single point estimate, but also takes account of the uncertainty in g. This uncertainty will lead to imprecisions in predicted quantities and is reflected by the posterior distribution. To explore how precisely g is defined by the data, one typically generates statistical samples using Markov chain Monte Carlo methods [7,8]. Figure 2 shows a posterior sample generated with the Hybrid Monte Carlo (HMC) algorithm [9]. The sample scatters around the exact density of states; the variance is mainly determined by the number of counts in each bin. It is also possible to derive an approximate analytical expression for the uncertainty by Laplace approximation. The inverse Hessian of $-\ln p(g)$ evaluated at the maximum posterior estimate quantifies the (co)variances of the g_k. The estimate is most reliable for $\beta \approx 0$, which concurs

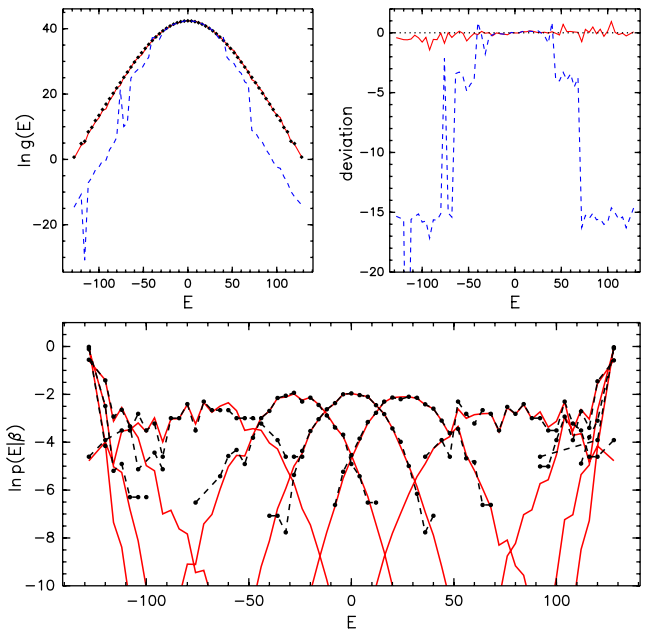

FIG. 1 (color online). Upper left: Comparison of the true density of states (dots), its Bayesian estimate obtained by posterior maximization (solid line), and the naive estimate (dashed line). Upper right: Deviation, $\ln[g_{\text{est}}(E)/g_{\text{true}}(E)]$, between the logarithms of the estimated and the exact density of states (solid line: g_{bayes}, dashed line: g_{naive}). Lower panel: Comparison of empirical energy distributions $H_i(E)$ (dots) with reconstructions based on g_{bayes} (solid line).

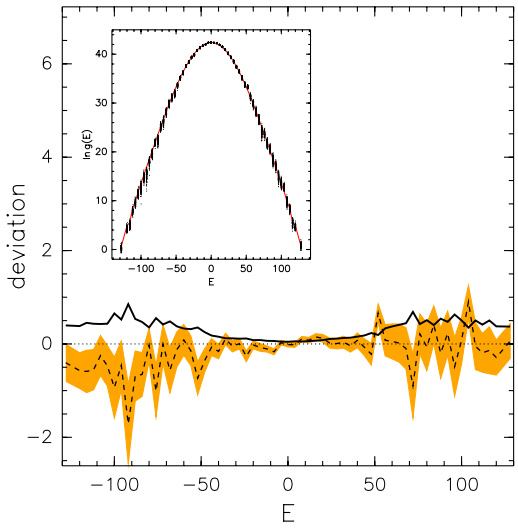

FIG. 2 (color online). Uncertainty of the density of states. The dashed line and the shaded range indicate the mean deviation of the Monte Carlo sample from the true density of states and its standard deviation, respectively. The solid line indicates the uncertainty of the density estimate calculated from the Monte Carlo sample. Inset: sampled g_k shown as dots; the solid line is the exact density of states.

with the fact that correct simulation is easiest in this temperature range. By averaging over the sample, predictive distributions of quantities such as specific heats can be estimated (Fig. 3). As one would expect, the error is largest in the critical region. But still the true curve is within 1 standard deviation.

Even for the finite Ising model, overfitting can occur. The left panel of Fig. 4 illustrates the effect of missing simulation data. The same data as in Figs. 1–3 were analyzed, except that in the energy histogram, one bin was set to zero. If one applies a Dirichlet prior or likewise histogram reweighting, the density of states drops to zero at the corrupt bin and exhibits an infinite uncertainty. This contradicts our intuition that g should be to some extent smooth and results from the fact that the Dirichlet prior does not capture correlations between neighboring bins. The "roughness penality" prior $\pi(g) \propto \exp\{-\int dE g(E)[\partial_E \ln g(E)]^2\}$ encodes such a notion of smoothness [10]. It penalizes densities with large variations in the first derivative, i.e., rough densities, but also other forms of regularization could be applied here. The right panel of Fig. 4 illustrates how the roughness penality prior helps to bridge the gap having zero counts and increases precision. This demonstrates that a suitable prior can exploit the information content of the data more efficiently by taking into account background information such as smoothness.

A Bayesian analysis is particularly helpful in situations where the nonparametric model runs into problems. With

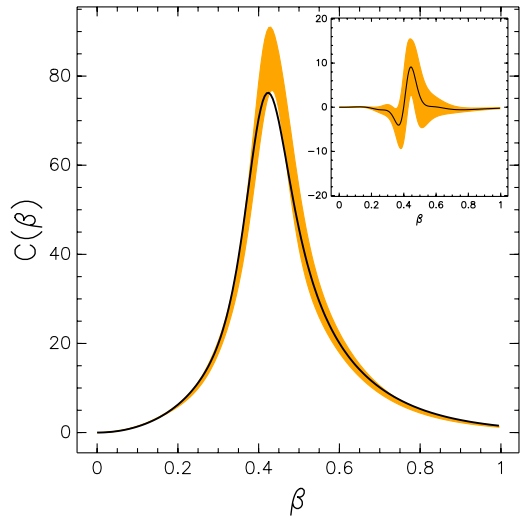

FIG. 3 (color online). Predictive distributions for the specific heat $C(\beta) = \beta^2 \partial_\beta^2 \ln Z(\beta)$. Ranges indicate 1 standard deviation around the mean curve computed from the Monte Carlo sample. The exact specific heat is shown as solid line. Inset: Deviation of the reconstructed specific heat from the exact curve with mean deviation shown as black line.

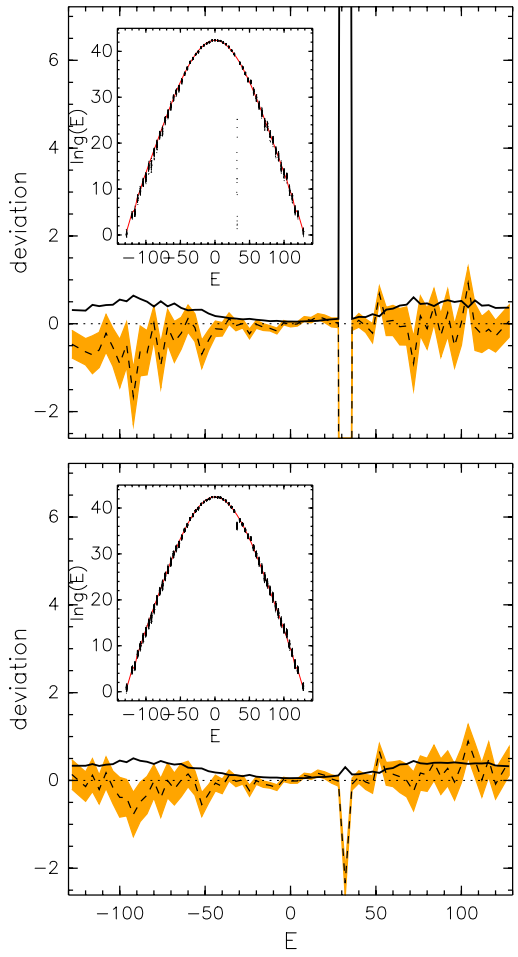

FIG. 4 (color online). Effect of a smoothing prior on corrupt data. Shown is the deviation g_k samples from the exact density of states as in Fig. 2. Top: results for the uninformative Dirichlet prior; bottom: results obtained for the smoothing prior.

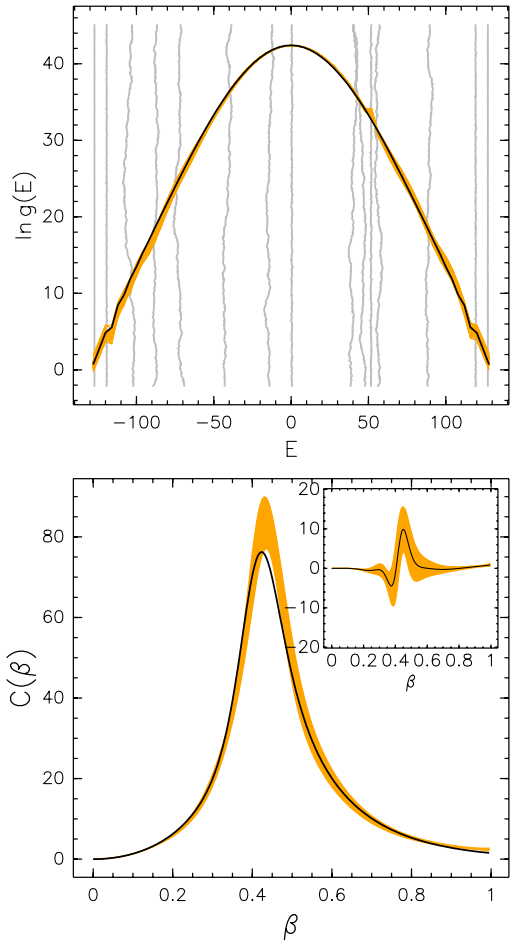

FIG. 5 (color online). Results for a finite mixture model using 15 Gaussian components. Top: exact density of states (solid curve) and the reconstructed density of states (shaded range indicating 1 standard deviation around the mean). The vertical traces indicate the energy levels ϵ_c sampled in the Monte Carlo run. Bottom: specific heat with local error (inset).

increasing system size, the number of parameters describing the density of states grows larger and larger, which results in an unfavorable ratio between the number of parameters and the number of data and increases the risk of overfitting. But also practical considerations may play a role if the size of the density of state becomes unwieldy. For large or continuous systems, it is therefore common practice to discretize the data and to apply the histogram equations to the binned energies. However, such an artificial discretization can lead to systematic errors [11]. Another possibility would be to approximate H with some continuous function such as a kernel density estimate. In either case, one must choose parameters such as the bin size of the histogram or the width of the kernel function beforehand. A Bayesian analysis, on the other hand, directly works with the raw data and does not require the data to be represented by histograms or other density estimates. All modeling assumptions only concern the density of states—one does not need to hypothesize about the functional form of the distribution of energies which is not a truly physical quantity but depends on the simulation algorithm.

An alternative to a nonparametric description of the density of states is provided by a parametric model such as a finite mixture of Gaussian distributions [12]:

$$g(E) = \sum_{c=1}^{C} \pi_c G(E; \epsilon_c, \sigma_c). \tag{7}$$

Here, C Gaussian components $G(E; \epsilon_c, \sigma_c)$ centered at energy levels ϵ_c with variances σ_c^2 are used to model the density of states; the weights π_c live on a simplex: $\sum_c \pi_c = 1$, $\pi_c \in [0, 1]$. Now, $3C - 1$ parameters $\theta = \{\pi_c, \epsilon_c, \sigma_c\}$ must be estimated, and the posterior probability becomes a density over the space of all θ. Note that the likelihood, Eq. (2), does not assume energies to be binned. We can therefore estimate the density of states without discretizing the energy samples.

Because of their reduced complexity, parametric continuous models usually require a smaller number of parameters when compared to their nonparametric counterparts. And even if the underlying system is discrete, a continuous model may be the better choice. Figure 5 shows results for the Ising data modeled with $C = 15$ Gaussian components. Although the number of parameters (44) is significantly smaller than for the histogram (62), the reconstructed continuous model captures the most salient features of the density of states equally well. The estimate of the density of states and of the specific heat is equally good as, if not better than, inferences based on a histogram.

*Electronic address: michael.habeck@tuebingen.mpg.de

[1] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[2] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[3] B. Roux, Comput. Phys. Commun. **91**, 275 (1995).
[4] A. M. Ferrenberg, D. P. Landau, and R. H. Swendsen, Phys. Rev. E **51**, 5092 (1995).
[5] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge UK, 2003).
[6] P. D. Beale, Phys. Rev. Lett. **76**, 78 (1996).
[7] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation* (Springer Verlag, Inc., New York, 2002).
[8] N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[9] S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).
[10] I. J. Good and R. A. Gaskin, Biometrika **58**, 255 (1971).
[11] M. N. Kobrak, J. Comput. Chem. **24**, 1437 (2003).
[12] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions* (Wiley, New York, 1985).