Bayesian update method for adaptive weighted sampling

Sanghyun Park, Daniel L. Ensign, and Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305, USA
(Received 2 September 2006; revised manuscript received 8 November 2006; published 27 December 2006)

Exploring conformational spaces is still a challenging task for simulations of complex systems. One way to enhance such a task is weighted sampling, e.g., by assigning high weights to regions that are rarely sampled. It is, however, difficult to estimate adequate weights beforehand, and therefore adaptive methods are desired. Here we present a method for adaptive weighted sampling based on Bayesian inference. Within the framework of Bayesian inference, we develop an update scheme in which the information from previous data is stored in a prior distribution which is then updated to a posterior distribution according to new data. The method proposed here is particularly well suited for distributed computing, in which one must deal with rapid influxes of large amounts of data.

DOI: 10.1103/PhysRevE.74.066703 PACS number(s): 05.10.−a

I. INTRODUCTION

In computer simulations of high-dimensional systems, such as liquids or macromolecules, it is often extremely difficult to thoroughly sample an entire configurational space of interest. One way of enhancing such sampling is to perform a set of simulations in which each simulation samples, by means of weighting, only a subspace; the entire space of interest is then covered by the union of the subspaces sampled in each simulation. To estimate relevant quantities from such simulations, one needs to combine data from simulations that used different weights, which is by no means a trivial task.

Ferrenberg and Swendsen [1] derived a method for optimally combining differently weighted histograms by minimizing an error estimate. This method is now commonly known as the weighted histogram analysis method (WHAM). The WHAM has become a ubiquitous tool for Monte Carlo and molecular dynamics simulations. One area where the WHAM has proven to be particularly useful is the analysis of umbrella sampling simulations, where weighted sampling is performed along a coordinate (or a set of coordinates) of interest [2,3]. The WHAM has also been used for combining simulation data obtained at different temperatures [4].

One crucial ingredient for a successful weighted-sampling simulation is to choose an efficient weighting scheme. Since it is difficult to guess an efficient weighting scheme before seeing any data, the need for an adaptive method is evident. The idea of adaptation is to continuously update the weighting scheme as new data are obtained; as more and more data are gathered, the weighting scheme becomes more efficient and estimates of parameters converge to the true values. Bartels and Karplus (BK) suggested an adaptive WHAM in such context [5]. Their method, however, requires analyzing all the data gathered up to the point where an update is performed, which could become quite demanding as the amount of data grows. During the revision of the manuscript, we were informed of other previous works concerning adaptive determination of weights [6–9]. Notably, the work of Smith and Bruce [6] adopted a Bayesian approach.

Here we present an adaptive Bayesian WHAM (ABWHAM). By following the framework of Bayesian inference, this method is able to perform a fast update using only the new data gathered since the last update. Another advantage is that, as in any Bayesian methods, it yields distributions of parameters from which error estimates can be obtained in a consistent manner. Below we describe the method, illustrate it using simple systems, and discuss its advantage over previous methods and its potential use in distributed computing.

II. THE METHOD

A. Motivation for Bayesian inference

Consider a system that can be in K different states, and let θ_i be the probability for the ith state $(\theta_1+\cdots+\theta_K=1)$. We want to estimate the parameters θ_i by means of weighted sampling. Typically, probabilities are related to free energies; in umbrella sampling, for example, $\theta_i \propto \exp(-F_i/k_BT)$ where F_i is the potential of mean force at the ith bin along the reaction coordinate, and for simulated tempering [10,11], $\theta_i \propto \exp(-F_i/k_BT_i)$ where F_i is the free energy at the ith temperature T_i.

We seek an adaptive weighted sampling scheme as outlined in Fig. 1. Based on the estimates $\theta_i^{(n-1)}$ from the previous iteration step, new weights $w_i^{(n)}$ are determined in a way that leads to efficient sampling of states. How exactly $w_i^{(n)}$ is determined from $\theta_i^{(n-1)}$ may differ case by case, but one natural choice is to set $w_i^{(n)}=1/\theta_i^{(n-1)}$ in order to ensure uniform sampling of the states. Weighted sampling is then performed with the new weights, and a histogram $h_i^{(n)}$ (the number of observations of each state) is recorded. From the new

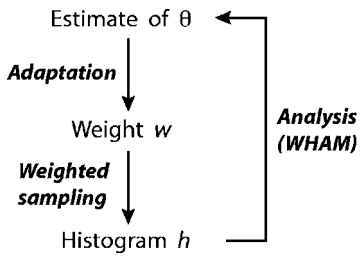

FIG. 1. The scheme for adaptive weighted sampling.

histogram $h_i^{(n)}$, new estimates $\theta_i^{(n)}$ are obtained, which completes a cycle. It is this last step where the WHAM is needed. The cycle of adaptation, sampling, and analysis is iterated starting with an initial guess $\theta_i^{(0)}$.

One way of performing the WHAM in the context of adaptive weighting is to shove all the available data into the WHAM, i.e.,

$$(w^{(1)},h^{(1)}),\ldots,(w^{(n)},h^{(n)}) \rightarrow \theta^{(n)} \tag{1}$$

which is the approach taken by BK [5]. Symbols with suppressed subscripts collectively denote entire sets of variables, e.g., $w^{(1)} \equiv (w_1^{(1)},\ldots,w_K^{(1)})$, $\theta^{(n)} \equiv (\theta_1^{(n)},\ldots,\theta_K^{(n)})$, etc. In this approach, accordingly, the cost of the WHAM calculation at each iteration step increases linearly with n, the iteration number. Although the computational bottleneck is usually in the generation of data (e.g., molecular dynamics simulations) rather than the analysis, this feature is certainly not attractive, especially for distributed computing where one has to deal with rapid influxes of large amounts of data. We will discuss further about distributed computing in Sec. IV.

Therefore, we attempt to develop a method in which only new data are needed for the update of estimates. Namely, we seek a way to determine a new estimate $\theta^{(n)}$ from the knowledge of the new histogram $h^{(n)}$, the new weight $w^{(n)}$, and the previous estimate $\theta^{(n-1)}$:

$$(\theta^{(n-1)},w^{(n)},h^{(n)}) \rightarrow \theta^{(n)}. \tag{2}$$

At this point, it is already evident that point estimates of θ will not suffice, because they do not signify uncertainties of estimates. With given $(\theta^{(n-1)},w^{(n)},h^{(n)})$, the new estimate $\theta^{(n)}$ should not be very different from $\theta^{(n-1)}$ if $\theta^{(n-1)}$ was already obtained from a large amount of data (i.e., after many iterations of weighted sampling), whereas it can be very different from $\theta^{(n-1)}$ if only a small amount of data were used for the estimation of $\theta^{(n-1)}$. With point estimates, $\theta^{(n)}$ will be the same in both cases because point estimates do not contain the information of how much data were used to obtain them. We need distributions, not just point estimates, which naturally leads to Bayesian inference.

B. General framework based on Bayesian inference

In Bayesian probability theory, a probability means a state of knowledge or a degree of belief [12]. In an abstract form, Bayesian inference operates as

$$P(p|d) = \frac{P(d|p)P(p)}{\sum_{p'} P(d|p')P(p')}, \tag{3}$$

where p represents a set of parameters that we want to estimate and d represents data. This equation is known as Bayes' rule; $P(d|p)$ is called the likelihood, $P(p)$ the prior probability (or simply "prior"), and $P(p|d)$ the posterior probability (or simply "posterior"). Namely, Bayesian inference updates the estimate of parameters, from the prior to the posterior, based on the data. The denominator in Eq. (3) can be considered a normalization constant, and Bayes' rule can be written as

$$P(p|d) \propto P(d|p)P(p) \tag{4}$$

under the implication that the proportionality constant is to be determined by $\Sigma_p P(p|d)=1$.

In the case of the adaptive WHAM, we want to estimate the parameters $\theta \equiv (\theta_1,\ldots,\theta_K)$ for K states. Let $f^{(n)}(\theta)$ denote the probability of θ after n iterations of weighted sampling:

$$f^{(n)}(\theta) \equiv P(\theta|w^{(n)},h^{(n)},\ldots,w^{(1)},h^{(1)}). \tag{5}$$

As a function of θ, $f^{(n)}(\theta)$ is the distribution that represents the estimate of θ after n iterations [in cases where states are are continuous, $f^{(n)}(\theta)$ is essentially a distribution of a distribution]. Our goal is to find a way to perform the update

$$(f^{(n-1)}(\theta),w^{(n)},h^{(n)}) \rightarrow f^{(n)}(\theta). \tag{6}$$

This is analogous to the update in Eq. (2), but we are now updating distributions of θ instead of θ itself. In the context of Bayesian inference, $f^{(n-1)}(\theta)$ and $f^{(n)}(\theta)$ can be considered the prior and the posterior, respectively, for the update at the nth iteration step. Therefore, using Bayes' rule [Eq. (4)] we find

$$f^{(n)}(\theta) \propto P(h^{(n)}|\theta,w^{(n)})f^{(n-1)}(\theta). \tag{7}$$

A more rigorous derivation is given in Appendix 1.

The likelihood $P(h|\theta,w)$, which plays a central role in the Bayesian update [Eq. (7)], is the probability of obtaining the histogram h from a weighted sampling given the state probability θ and the weight w. Let us define weighted state probabilities

$$\phi_i \equiv \frac{w_i\theta_i}{\sum_j w_j\theta_j}. \tag{8}$$

A weighted sampling of θ is then identical to an unweighted sampling of ϕ. Therefore, assuming that the histogram is collected from statistically independent measurements, the likelihood is given as a multinomial distribution

$$P(h|\theta,w) = \frac{H!}{h_1!\cdots h_K!}\phi_1^{h_1}\cdots\phi_K^{h_K} \tag{9}$$

or in terms of the original state probability θ,

$$P(h|\theta,w) = \frac{H!}{h_1!\cdots h_K!}\frac{(w_1\theta_1)^{h_1}\cdots(w_K\theta_K)^{h_K}}{(w_1\theta_1+\cdots+w_K\theta_K)^H}, \tag{10}$$

where $H \equiv h_1+\cdots+h_K$. The Bayesian update equation (7) along with the likelihood [Eqs. (9), (10)] constitutes the framework of the ABWHAM.

C. Bayesian inference using conjugate priors — weighted Dirichlet distributions

In the previous section, we laid out the general framework of the ABWHAM. In order to make it practical, however, we need to represent the distribution $f(\theta)$ in a parametric form. For that purpose, we employ the notion of the conjugate prior. A conjugate prior is a family of prior distributions, or a member therein, that share the same parametric form and

whose corresponding posterior distributions also belong to that same family. In the context of the Bayesian update [Eq. (7)], if we have a conjugate prior $f^{(n-1)}(\theta)$, the corresponding posterior $f^{(n)}(\theta)$ will have the same parametric form as $f^{(n-1)}(\theta)$ and we can formulate the Bayesian update as an update of a set of parameters, instead of an update of a function.

The likelihood for our Bayesian update is a multinomial distribution when it is written in terms of ϕ [Eq. (9)]. It is well known that Dirichlet distributions

$$f_D(\phi;\alpha) \equiv \frac{1}{B(\alpha)} \phi_1^{\alpha_1-1} \cdots \phi_K^{\alpha_K-1} \delta\left(\sum_j \phi_j - 1\right),$$

$$B(\alpha) \equiv \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(A)}, \quad A \equiv \alpha_1 + \cdots + \alpha_K \tag{11}$$

form a conjugate prior family for a sampling process whose likelihood is a multinomial distribution [13]. The parameters $\alpha \equiv (\alpha_1, \ldots, \alpha_K)$ uniquely specify a K-state Dirichlet distribution. The normalizing constant $B(\alpha)$ is known as the beta function. Some properties of the Dirichlet distribution, which we will be using here, are derived in Appendix 2. By transforming ϕ back to the original state probability θ, we obtain the corresponding distribution of θ,

$$f_{\mathrm{WD}}(\theta;\alpha,w) = \left|\frac{\partial\phi}{\partial\theta}\right| \frac{1}{B(\alpha)} \frac{(w_1\theta_1)^{\alpha_1-1}\cdots(w_K\theta_K)^{\alpha_K-1}}{(w_1\theta_1+\cdots+w_K\theta_K)^{A-K}} \times \delta\left(\sum_j \theta_j - 1\right) \tag{12}$$

which we call a "weighted" Dirichlet distribution. Fortunately, the explicit form of the Jacobian $|\partial\phi/\partial\theta|$ is not needed in developing our method.

It is straightforward to verify that weighted Dirichlet distributions are indeed conjugate priors for our Bayesian update. Suppose that $f^{(n-1)}(\theta)$ is a weighted Dirichlet distribution with respect to the weights $w^{(n)}$: $f^{(n-1)}(\theta) = f_{\mathrm{WD}}(\theta;\alpha^{(n-1)},w^{(n)})$. Then, from Eqs. (7), (10), and (12), the posterior is given as $f^{(n)}(\theta) = f_{\mathrm{WD}}(\theta;\alpha^{(n)},w^{(n)})$, where α is updated as

$$\alpha_i^{(n)} = \alpha_i^{(n-1)} + h_i^{(n)}. \tag{13}$$

D. Mapping between weighted Dirichlet distributions — the relative entropy

The difficulty, however, is that the likelihood [Eq. (10)] depends on the weights which change from one iteration step to another. Thus, even if $f^{(n-1)}(\theta)$ is a conjugate prior for the nth iteration step, the resulting posterior $f^{(n)}(\theta)$ will *not* be a conjugate prior for the $(n+1)$-th iteration step unless the weights are identical between the nth and $(n+1)$-th iteration steps. Accordingly, the Bayesian update chain cannot proceed beyond one cycle. This is in fact the trickiest problem in developing a Bayesian update scheme for adaptive weighted sampling.

Our solution to this problem is to devise a mapping between weighted Dirichlet distributions

$$f_{\mathrm{WD}}(\theta;\alpha,w) \to f_{\mathrm{WD}}(\theta;\alpha',w'), \tag{14}$$

where α' is the unknown. Given a weighted Dirichlet distribution $f_{\mathrm{WD}}(\theta;\alpha,w)$ and a new weight w', we want to find a weighted Dirichlet distribution $f_{\mathrm{WD}}(\theta;\alpha',w')$, with respect to the new weight, that is *closest* to the original distribution $f_{\mathrm{WD}}(\theta;\alpha,w)$. In other words, we want to map a Dirichlet distribution in $\phi_i \equiv w_i\theta_i/\Sigma_j w_j\theta_j$ to a Dirichlet distribution in $\phi_i' \equiv w_i'\theta_i/\Sigma_j w_j'\theta_j$ [14].

In order to formulate such a mapping, we need to quantify how close one distribution is to another. For this purpose, we choose the relative entropy (also known as the Kullback-Leibler divergence) which is a commonly used metric for distances between distributions [15]. For the two distributions at hand, $f(\theta) \equiv f_{\mathrm{WD}}(\theta;\alpha,w)$ and $f'(\theta) \equiv f_{\mathrm{WD}}(\theta;\alpha',w')$, the relative entropy is defined as

$$D(f\|f') \equiv \int d\theta f(\theta) \ln \frac{f(\theta)}{f'(\theta)}. \tag{15}$$

Notice that the relative entropy is not symmetric: $D(f\|f') \neq D(f'\|f)$. We choose $D(f\|f')$ because it immensely simplifies the algebra. The mapping [Eq. (14)] inevitably introduces an approximation, as it approximates a Dirichlet distribution in ϕ with a Dirichlet distribution in ϕ'. The choice of $D(f\|f')$ over $D(f'\|f)$, or over the symmetric form $\frac{1}{2}[D(f\|f')+D(f'\|f)]$, is essentially a choice of one approximation over another for the sake of the simplicity of the algebra.

By minimizing the relative entropy with respect to α', we find

$$\langle \ln \phi_i' \rangle_{f'} = \langle \ln \phi_i' \rangle_f, \tag{16}$$

where $\langle\cdots\rangle_f$ denotes an average with respect to the distribution f, or equivalently,

$$\psi(\alpha_i') - \psi(A') = \int d\phi f_D(\phi;\alpha) \ln \frac{w_i' w_i^{-1} \phi_i}{\sum_j w_j' w_j^{-1} \phi_j}, \tag{17}$$

where $\psi(x) \equiv \frac{d}{dx}\ln\Gamma(x)$ is the digamma function and $A' \equiv \alpha_1' + \cdots + \alpha_K'$. A detailed derivation is given in Appendix 3.

Equation (17) provides K equations that determine the K unknowns $\alpha_1', \ldots, \alpha_K'$ and thereby determine the mapping between weighted Dirichlet distributions. Solving Eq. (17) is, however, quite demanding. One has to evaluate the integral on the right hand side and then solve the equation involving the digamma function. Neither step can be done analytically; therefore, we turn to a heuristic approach.

E. Mapping between weighted Dirichlet distributions — a heuristic approach

We observe that a Dirichlet distribution $f_D(\phi;\alpha)$ is uniquely determined by the mean values $\bar{\phi}_i = \alpha_i/A$ (see Appendix 2) and A. (Due to the constraint $\Sigma_j \bar{\phi}_j = 1$, the mean values provide only $K-1$ independent conditions.) The mean values signify where the center of the distribution is located, and the value of A signifies how broad the distribution is. We

call A the "confidence" since a bigger A indicates a narrower distribution, hence more confidence in estimates.

We also observe that as more and more data are gathered from adaptive weighted sampling, the distribution $f(\theta)$ that represents the estimate of θ will become narrower. Thus, we would prefer heuristic schemes that are accurate for narrow distributions, although possibly inaccurate for broad distributions, to the ones that behave in the opposite way.

Based on these observations, we construct a heuristic scheme for the mapping of Eq. (14) as follows. We start with the zero-dispersion limit. At this limit, the means, $\bar{\phi} \equiv \langle \phi \rangle_f$ and $\bar{\phi}' \equiv \langle \phi' \rangle_{f'}$, obey the same relationship that ϕ and ϕ' obey. Therefore, recalling the relationship between ϕ and ϕ', we find

$$\bar{\phi}_i' = \frac{u_i \bar{\phi}_i}{\sum_j u_j \bar{\phi}_j}, \quad u_i \equiv w_i' w_i^{-1} \tag{18}$$

which provides $K-1$ independent conditions. In fact, it is crucial that the means are not sufficient to determine a Dirichlet distribution. Recall that the motivation for using distributions instead of point estimates was to encode uncertainty in estimates. If the means alone determined a Dirichlet distribution, it would contradict the whole purpose of using distributions.

To obtain the last condition, i.e., to determine the confidence A', we must go beyond the zero-dispersion limit. Here, we again employ the principle of minimizing the relative entropy. Now that A' is the only unknown left, we minimize the relative entropy [Eq. (15)] with respect to A', and obtain

$$\sum_i \bar{\phi}_i' \langle \ln \phi_i' \rangle_{f'} = \sum_i \bar{\phi}_i' \langle \ln \phi_i' \rangle_f. \tag{19}$$

A derivation is given in Appendix 3. The relationship between Eq. (19) and Eq. (16) is apparent: Eq. (19) (a single equation) is the average of Eq. (16) (K equations) with respect to $\bar{\phi}'$. Just as solving Eq. (16) is demanding, so is solving Eq. (19). Therefore, we seek an approximate solution by expanding Eq. (19) around the zero-dispersion limit. Keeping only the leading order, we find (see Appendix 4)

$$A' = \frac{K-1}{2(D_1 + D_2)} - 1,$$

$$D_1 \equiv \sum_i \frac{\bar{\phi}_i'(1-\bar{\phi}_i)}{2(A+1)\bar{\phi}_i},$$

$$D_2 \equiv \frac{\sum_{i \neq j} u_i u_j \bar{\phi}_i \bar{\phi}_j - \sum_i u_i^2 \bar{\phi}_i (1-\bar{\phi}_i)}{2(A+1)\sum_{i,j} u_i u_j \bar{\phi}_i \bar{\phi}_j}. \tag{20}$$

With A' at hand, we can completely determine α' by

$$\alpha_i' = A' \bar{\phi}_i'. \tag{21}$$

Equations (18), (20), and (21) constitute our heuristic scheme for the mapping between weighted Dirichlet distributions.

F. The algorithm

In some cases, the numerical precision of the variables that store state probabilities and weights could be an issue. This is because probabilities may be exponentially different among states (in statistical mechanical systems, probabilities are governed by the Boltzmann factor). Therefore, we suggest working with logarithms

$$\sigma_i \equiv \ln \theta_i + C_1, \quad g_i \equiv \ln w_i + C_2. \tag{22}$$

C_1 and C_2 are arbitrary constants. The presence of C_2 is easy to understand; w_i is only defined up to a multiplicative constant. The presence of C_1 means that we are interested in $\ln \theta_i$ only up to an additive constant, which is indeed the case in many applications where free energies or potentials of mean force are the main quantities of interest. If desired, of course, C_1 can be determined by the normalization $\Sigma_j \theta_j = 1$. Here, however, we simply choose these arbitrary constants such that the averages of σ and g over the states equal zero:

$$\frac{1}{K}\sum_j \sigma_j = 0, \quad \frac{1}{K}\sum_j g_j = 0. \tag{23}$$

We now summarize the algorithm of the ABWHAM.

(0) Initial setup: (i) Choose an initial guess σ^{guess}. Without any prior information about the system, one might simply take $\sigma_i^{\text{guess}} = 0$. (ii) Set the initial weight as $g_i = -\sigma_i^{\text{guess}}$. (iii) Set $\alpha_i = 1$. This amounts to choosing an initial distribution $f(\theta) = f_{\text{WD}}(\theta; \alpha, w)$ as the uniform distribution under the weighting $w_i \propto \exp(g_i)$. In general, unless the initial guess was obtained from a significant amount of information, a broad distribution (small α values) must be chosen in order to ensure that the resulting estimates will be dominated by the data rather than the initial guess. (iv) Set the initial point estimate as $\bar{\sigma}_i = \sigma_i^{\text{guess}}$. (v) Set $\sigma_i^{\text{ref}} = \bar{\sigma}_i$.

(1) Adapt the weights

$$g_i^{\text{new}} = -\bar{\sigma}_i + \frac{1}{K}\sum_j \bar{\sigma}_j. \tag{24}$$

(2) Perform weighted sampling with $w_i^{\text{new}} \propto \exp(g_i^{\text{new}})$ and obtain a histogram h. Depending on the problem at hand, this step is done through various simulation techniques such as umbrella sampling, simulated tempering, etc.

(3) Map between weighted Dirichlet distributions: $(\alpha, g) \to (\alpha', g^{\text{new}})$. Determine α' by the heuristic scheme [Eqs. (18), (20), and (21)]

$$\alpha_i' = A' \bar{\phi}_i', \tag{25}$$

where

$$\bar{\phi}_i' = \frac{u_i \bar{\phi}_i}{\sum_j u_j \bar{\phi}_j},$$

$$\bar{\phi}_i = \alpha_i / A, \quad A = \sum_j \alpha_j, \tag{26}$$

$$u_i = \exp(g_i^{\text{new}} - g_i)$$

and

$$A' = \frac{K-1}{2(D_1+D_2)} - 1,$$

$$D_1 = \sum_i \frac{\bar{\phi}_i'(1-\bar{\phi}_i)}{2(A+1)\bar{\phi}_i}, \tag{27}$$

$$D_2 = \frac{\sum_{i\neq j} u_i u_j \bar{\phi}_i \bar{\phi}_j - \sum_i u_i^2 \bar{\phi}_i(1-\bar{\phi}_i)}{2(A+1)\sum_{i,j} u_i u_j \bar{\phi}_i \bar{\phi}_j}.$$

(4) Perform the Bayesian update: $(\alpha', h) \to \alpha$. Determine α by

$$\alpha_i = \alpha_i' + h_i. \tag{28}$$

(5) Calculate new point estimates

$$\bar{\sigma}_i = \ln \alpha_i - g_i^{\text{new}} - \frac{1}{K}\sum_j (\ln \alpha_j - g_j^{\text{new}}). \tag{29}$$

Notice that although we deal with distributions throughout our method, we need point estimates here in order to decide on the weights; we choose the mean for this purpose. Also notice that, in line with our heuristic mapping scheme, we simply relate means of different variables at the zero-dispersion limit.

(6) If $\max_i |\bar{\sigma}_i - \sigma_i^{\text{ref}}| > \Omega$, then refresh: set $\alpha_i = 1$, $g_i = -\bar{\sigma}_i$, and $\sigma_i^{\text{ref}} = \bar{\sigma}_i$. Otherwise, set $g_i = g_i^{\text{new}}$. Ω is a predetermined constant.

(7) Go to step (1). Repeat a given number of times or until desired precision is achieved for relevant quantities, e.g., θ or σ.

In step (6), we introduced a refresh procedure, which deserves explanation. Our method uses a mapping [step (3)] between weighted Dirichlet distributions, which is an approximation. If σ^{guess} is close to the true value σ^{true}, the weights will not change very much throughout the iteration procedure and the mapping will be a good approximation. On the other hand, if σ^{guess} is far from σ^{true}, the weights will change a lot and the approximation will break down. Notice that this is due to the approximate nature of the mapping [Eq. (14)] itself; even if we directly solve Eq. (16) instead of using the heuristic scheme, the same issue will persist. In this step, we check whether $\bar{\sigma}$ has significantly deviated from σ^{ref} and, if it has, return to a uniform distribution as if starting anew with $\bar{\sigma}$ as an initial guess. There is no strict rule on how to choose Ω, but we have found that values between 1 and ln 10 yield reasonable performance. Smaller values of Ω lead to fast and noisy convergence while larger values lead to slow and smooth convergence. In the illustration section, we show results for $\Omega = 1$.

III. ILLUSTRATIONS

Using two simple models, we illustrate our method and compare it to the BK method. These two simple models represent two drastically different situations. In the first model, the state probabilities are more or less within the same order of magnitude, which represents an easy case. The second model represents a much more challenging case where state probabilities are different from each other by many orders of magnitude. We also show an application of our method to a simulated tempering simulation of Ala_{10} (decamer of alanine).

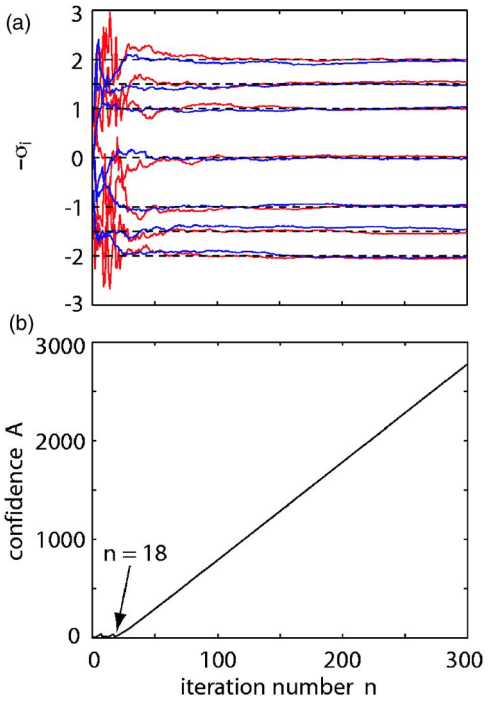

FIG. 2. (Color online) Convergence of point estimates (model 1). (a) Point estimates $-\bar{\sigma}_i^{(n)}$ obtained with the ABWHAM (red) are compared to those obtained with the BK method (blue). The true values $-\sigma_i^{\text{true}}$ are shown as dashed lines. (b) The confidence $A^{(n)}$. The last refresh step was executed at $n=18$.

A. Model 1

Let us consider a seven-state model whose states have the following free energies (measured in k_BT) associated with them:

$$-\sigma^{\text{true}} = \left(-2, -\frac{3}{2}, -1, 0, 1, \frac{3}{2}, 2\right). \tag{30}$$

Recall that the state probabilities θ_i are related to σ_i by Eq. (22). In this model, the ratio between the largest and the smallest probabilities is only $\theta_1/\theta_7 = \exp(4) \approx 55$. Thus, one can expect to get decent estimates for free energies even without any weighting. Nevertheless, we examine how adaptive weighting works in this case.

Starting with the initial guess $\sigma_i^{\text{guess}} = 0$, we performed adaptive weighted sampling following the algorithm outlined above. At each weighted sampling step, ten statistically independent samples were drawn according to the probabilities $\exp(\sigma_i^{\text{true}} + g_i^{(n)}) / \Sigma_j \exp(\sigma_j^{\text{true}} + g_j^{(n)})$. The cycle of adaptation, sampling, and analysis was iterated 300 times. Therefore, the total data amount to 3000 samples. The results are summarized in Figs. 2 and 3.

Figure 2(a) shows how the point estimates $-\bar{\sigma}_i^{(n)}$ converge to the true values as the iteration number n increases. The

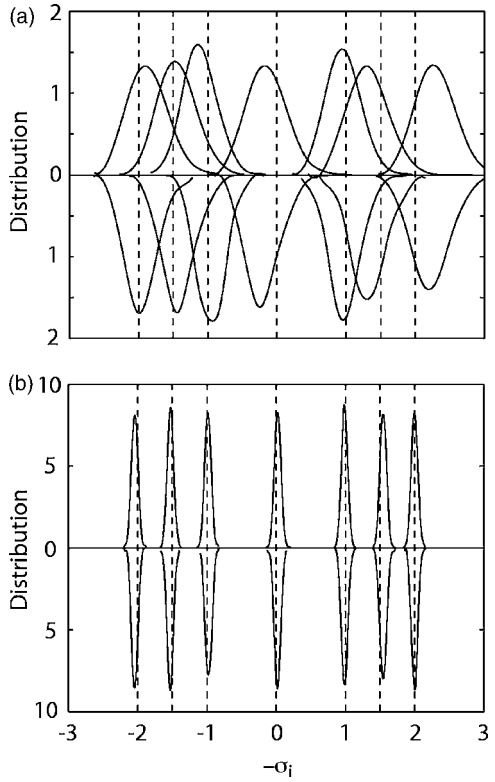

FIG. 3. Convergence of posterior distributions (model 1). The distributions of $-\sigma_i$ sampled from $f^{(n)}(\theta)=f_{\mathrm{WD}}(\theta;\alpha^{(n)},w^{(n)})$ obtained with the ABWHAM (upright figures) are compared to those obtained with the Monte Carlo posterior sampling (upside-down figures). Smooth curves were produced by the cubic spline interpolation of histograms of 10 bins. The true values $-\sigma_i^{true}$ are shown as dashed lines. (a) $n=28$. (b) $n=300$.

ABWHAM shows virtually the same convergence as the BK method. Figure 2(b) shows the confidence A as a function of n. Whenever the refresh step is executed, the confidence A drops down to $K=7$, the number of states. We can comprehend the entire iteration process by dividing it into two parts where the last refresh step was executed ($n=18$, in this example). The iterations before the last refresh step improve the initial guess, and those after refine the posterior distributions. In this example, σ^{guess} is already so close to σ^{true} that the refresh step is essentially immaterial; various choices of the threshold Ω, as well as the ABWHAM without the refresh step, lead to similar convergence. Only the results for $\Omega=1$ are shown here.

One advantage of Bayesian inference is that it yields distributions, not just point estimates, which contain information about uncertainties of estimates. Especially, since we have Dirichlet distributions of ϕ (i.e., weighted Dirichlet distributions of θ), distributions of any quantities that can be derived from θ or ϕ can be obtained by sampling from Dirichlet distributions, which is a straightforward process [16]. Figure 3 shows distributions of $-\sigma_i$ sampled from $f^{(n)}(\theta)=f_{\mathrm{WD}}(\theta;\alpha^{(n)},w^{(n)})$ at $n=28$ [Fig. 3(a)] and at $n=300$ [Fig. 3(b)]. The convergence of the estimates is evident; the free energies are not yet fully resolved at $n=28$, but they are at $n=300$.

Since our distributions are obtained through approximations (the mapping between weighted Dirichlet distributions), one may ask how accurately they represent the actual uncertainty. To address this question, we sampled state probabilities using a Monte Carlo technique as described in Appendix 5. This Monte Carlo sampling, although very expensive, can possibly be considered the most faithful Bayesian method. As can be seen in Fig. 3, there is no significant difference between the distributions obtained with the ABWHAM and those obtained with the Monte Carlo technique. Although it uses approximations, the ABWHAM still produces more or less accurate distributions.

B. Model 2

The second model for illustration is another seven-state model with free energies

$$-\sigma^{\mathrm{true}}=(-50,-48,-40,10,30,48,50). \tag{31}$$

This model presents a much more challenging case than the first one. The overall range of free energy that we must cover is as large as $100\,k_BT$ [the ratio between the largest and the smallest state probabilities is $\theta_1/\theta_7=\exp(100)\approx 3\times 10^{43}$], while states 1 and 2 (and states 6 and 7) are only separated by $2\,k_BT$. This is, however, not an uncommon situation. The resolution of k_BT is often needed in free energy calculations, and the range of $100\,k_BT$ is not too unrealistic either.

Again starting with the initial guess $\sigma_i^{\mathrm{guess}}=0$, we performed adaptive weighted sampling in the same way as with model 1 collecting ten samples per iteration over 300 iterations (3000 samples total). The results are summarized in Figs. 4 and 5.

The point estimates for the free energies are plotted in Fig. 4. The BK method and the ABWHAM with $\Omega=1$ again yield virtually the same convergence. Unlike in model 1, σ^{guess} is now so far from σ^{true} that the convergence of the ABWHAM is sensitive to the choice of the threshold Ω; with $\Omega=\ln(10)$ the convergence is slowed down by a factor of 1.5 (result not shown). Overall, the refresh step is critical when initial guesses are far off, because without it the convergence would be too slow.

Figure 5 shows distributions of $-\sigma_i$ obtained at $n=89$ and $n=300$. At $n=89$, a couple of iterations after the final refresh step, only 20 samples actually contribute to producing the distributions; all the previous data are used only for improving the initial guess. The distributions at $n=89$, accordingly, are rather broad [broader than those in Fig. 3(a)]. That they look sharp in Fig. 5(a) is due to the wide range of the free energies in this model. Distributions get refined through more iterations, and at $n=300$ we see that they are sharply peaked at the true values [Fig. 5(b)]. Again, we do not see any significant difference between the distributions obtained with the ABWHAM and those obtained with the Monte Carlo posterior sampling.

C. Simulated tempering of Ala$_{10}$

For a more realistic illustration, we have applied the ABWHAM to a simulated-tempering [10,11] molecular

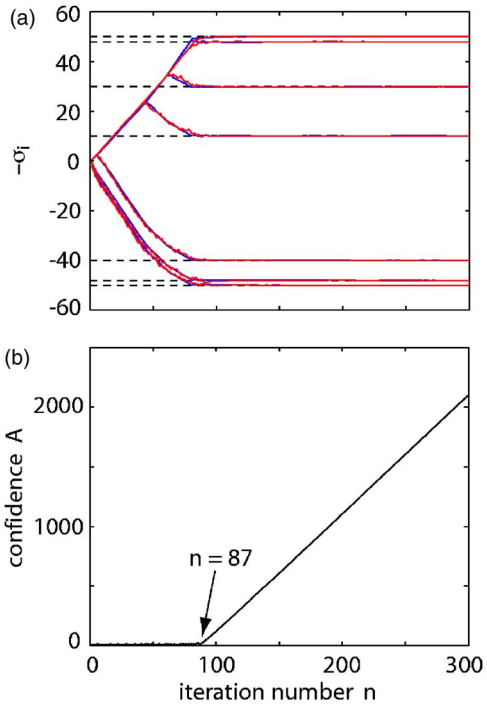

FIG. 4. (Color online) Convergence of point estimates (model 2). (a) Point estimates $-\bar{\sigma}_i^{(n)}$ obtained with the ABWHAM (red) are compared to those obtained with the BK method (blue). The true values $-\sigma_i^{true}$ are shown as dashed lines. (b) The confidence $A^{(n)}$. The last refresh step was executed at $n=87$.

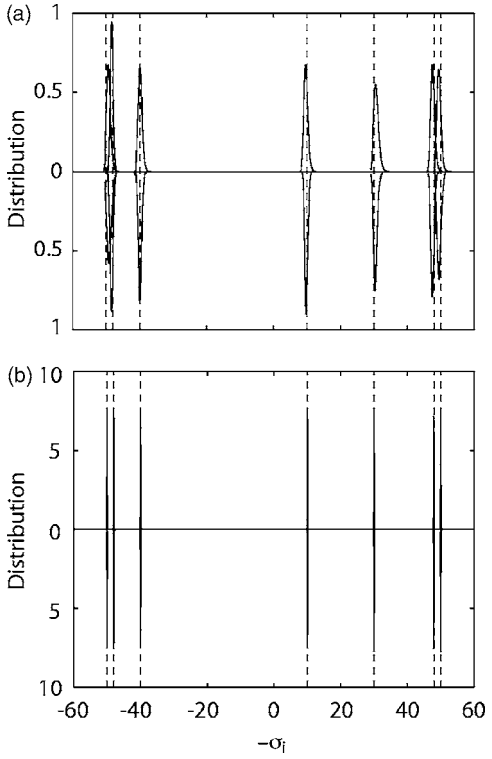

FIG. 5. Convergence of posterior distributions (model 2). The distributions of $-\sigma_i$ sampled from $f^{(n)}(\theta)=f_{\mathrm{WD}}(\theta;\alpha^{(n)},w^{(n)})$ obtained with the ABWHAM (upright figures) are compared to those obtained with the Monte Carlo posterior sampling (upside-down figures). Smooth curves were produced by the cubic spline interpolation of histograms of ten bins. The true values $-\sigma_i^{true}$ are shown as dashed lines. (a) $n=89$. (b) $n=300$.

dynamics simulation of Ala$_{10}$ peptide (capped with acetyl and *N*-methyl groups) in vacuum (Fig. 6). We chose seven temperatures 300, 325, 352, 381, 413, 448, and 485 K. Transitions between neighboring temperatures were attempted every 1 ps. Acceptance of a transition was decided based on the Metropolis criterion using the potential energy; when a transition was accepted, velocities were rescaled according to the temperature change [17]. Between transitions, the system was kept at a constant temperature using the Nose-Hoover thermostat [18,19]. Molecular dynamics simulations were performed with a modified version of GROMACS [20] in which we implemented the simulated tempering algorithm. Using the ABWHAM, we updated the weights for the temperatures every 100 ps (100 transition attempts), and iterated the adaptation-sampling-analysis cycle for 100 times (10 ns of total simulation time). For comparison, we performed another simulation where we used the BK method instead of the ABWHAM.

In the above two simple models, by design there was no time correlation in data, but that is not the case in this example. A histogram of 100 correlated counts, obtained at each iteration, does not contain the same amount of information as a histogram of 100 uncorrelated (i.e., statistically independent) counts. Using histograms of correlated counts without considering the correlation can lead to underestimation of errors, although point estimates are usually not affected significantly. A commonly adopted way of accounting for correlation is to reduce each count by a factor $\gamma=1+2\tau$ (τ is the correlation time) before processing histograms [3,4,21]. In general, determination of γ is nontrivial because the correlation time may depend on the weights that change over the course of adaptive weighting. In this example, we simply used a single value $\gamma=10$ (i.e., ten counts are considered effectively equivalent to one statically independent count) which seemed reasonable. We note that the choice of γ should not affect the comparison between the BK method and the ABWHAM as the same γ was used for both.

The results are shown in Figs. 7 and 8. Starting at the initial guess $\sigma_i^{\mathrm{guess}}=0$, the point estimates $\bar{\sigma}$ converge after about 30 iterations. Just as in the above two simple models, we see no significant difference in convergence between the BK method and the ABWHAM (Fig. 7), and no significant

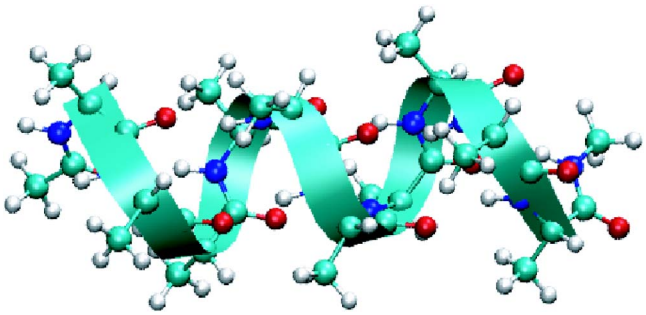

FIG. 6. (Color online) Ala$_{10}$ peptide in an α-helical form. Made with VMD [25].

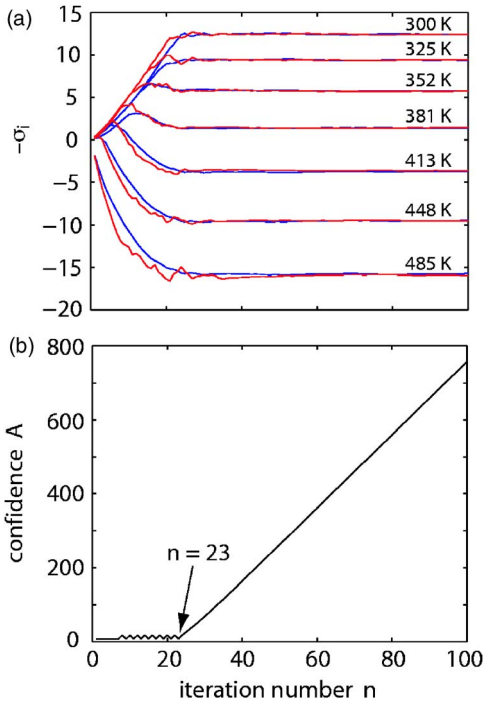

FIG. 7. (Color online) Convergence of point estimates (simulated tempering of Ala$_{10}$). (a) Point estimates $-\bar{\sigma}_i^{(n)}$ obtained with the ABWHAM (red) are compared to those obtained with the BK method (blue). (b) The confidence $A^{(n)}$. The last refresh step was executed at $n=23$.

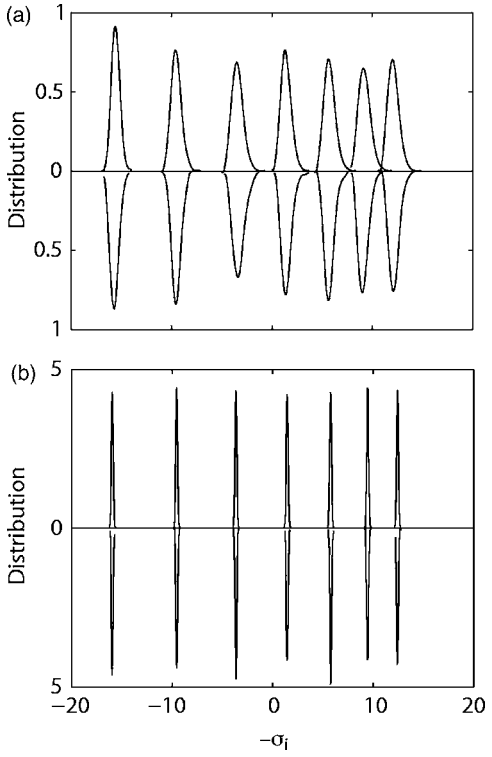

FIG. 8. Convergence of posterior distributions (simulated tempering of Ala$_{10}$). The distributions of $-\sigma_i$ sampled from $f^{(n)}(\theta)=f_{\mathrm{WD}}(\theta;\alpha^{(n)},w^{(n)})$ obtained with the ABWHAM (upright figures) are compared to those obtained with the Monte Carlo posterior sampling (upside-down figures). Smooth curves were produced by the cubic spline interpolation of histograms of 10 bins. (a) $n=25$. (b) $n=100$.

difference in posterior distributions between the ABWHAM and the Monte Carlo posterior sampling (Fig. 8). This is not surprising at all. The Ala$_{10}$ system is much more complicated than the above two simple models, but the complication occurs only in the sampling part of the adaptation-sampling-analysis cycle. The different methods that we are comparing, on the other hand, only affect the adaptation and analysis parts. It is therefore expected that the complexity of the Ala$_{10}$ system is largely irrelevant to this comparison.

IV. CONCLUDING REMARKS

One of the possible applications of adaptive weighting is simulated tempering [10,11], as we illustrated above with Ala$_{10}$. Recently, its parallel version, replica exchange [22,17] (also known as parallel tempering), has received a lot of attention. Replica exchange is commonly considered superior to simulated tempering, the reason being that replica exchange does not require determination of weights. However, with an adaptive weighting scheme (such as the BK method or the ABWHAM) determination of weights does not have to be tedious. The need for weights, perhaps, should not overshadow the advantages of simulated tempering. One obvious advantage of simulated tempering over replica exchange is that it is naturally suited for distributed computing because it does not require communications between CPUs.

Another possible application is the calculation of free energy along a parameter λ (possibly multidimensional), where states are defined by a discrete set of λ values, and simulations are done in a similar way as simulated tempering, with transitions between different λ values instead of different temperatures. In this case, the weights are by themselves crucial quantities because they are directly related to the free energy of interest. Therefore, adaptive weighting finds its natural place. Furthermore, compared to other free energy methods such as free energy perturbation, there may be advantages coming from making frequent transitions between λ values [10]. It will require further tests to assess the efficiency of free energy calculations using adaptive weighting.

The ABWHAM is based on a Bayesian update scheme that uses only the new data at each iteration step. The computational cost of each iteration, therefore, does not increase with the iteration number n; going from $n=10\,000$ to $10\,001$ requires the same amount of computation as going from $n=1$ to $n=2$. This is, along with the capability of producing consistent error estimates, the main advantage of the ABWHAM over the BK method. One may argue that the sampling part is usually the most costly part among the adaptation-sampling-analysis cycle and that reducing the computational cost for the adaptation and analysis parts will not help much. This is a legitimate objection, but the situation is different for large scale computing such as distributed computing. In distributed computing, the sampling part can be distributed over many CPUs on the network, but the adaptation and analysis parts must be done at a central CPU.

Whenever a new set of data comes in, the central CPU updates the weights and sends out new jobs with the updated weights. Therefore, the central CPU must be able to handle rapid influxes of data from all the CPUs on the network, and the ABWHAM should be very useful in this type of application.

In summary, we have developed a Bayesian update method for adaptive weighted sampling based on an approximation using weighted Dirichlet distributions. Two most important features of the method, in contrast to the previous BK method, are (i) that the computational cost of each iteration does not increase with the iteration number because the update scheme uses only the new data and (ii) that the method also yields error estimates based on Bayesian inference. Tests with simple systems indicate that, even though it is based on an approximation, the ABWHAM shows virtually the same convergence of point estimates as the BK method and yields reasonable error estimates. The ABWHAM seems to be an adequate alternative to the BK method, especially for distributed computing, and we hope that it will find many applications such as simulated tempering and free energy calculations.

ACKNOWLEDGMENTS

We have benefited tremendously from discussion with John D. Chodera, Paula Petrone, and Nina Singhal. This work was supported by a grant from NSF for Cyberinfrastructure.

APPENDIX

1. Derivation of the Bayesian update equation [Eq. (7)]

Let us start with the definition of $f^{(n)}(\theta)$:

$$f^{(n)}(\theta) \equiv P(\theta|w^{(n)},h^{(n)},\ldots,w^{(1)},h^{(1)}). \tag{A1}$$

Applying Bayes' rule to the $(\theta,h^{(n)})$ pair, we obtain

$$f^{(n)}(\theta) \propto P(h^{(n)}|\theta,w^{(n)},w^{(n-1)},h^{(n-1)},\ldots,w^{(1)},h^{(1)}) \times P(\theta|w^{(n)},w^{(n-1)},h^{(n-1)},\ldots,w^{(1)},h^{(1)}). \tag{A2}$$

This expression can be further simplified by eliminating redundant conditions. First, let us look at the first factor. When θ and $w^{(n)}$ are given, they completely determine the probability of $h^{(n)}$; all the previous data are redundant, and therefore can be eliminated:

$$P(h^{(n)}|\theta,w^{(n)},w^{(n-1)},h^{(n-1)},\ldots,w^{(1)},h^{(1)}) = P(h^{(n)}|\theta,w^{(n)}). \tag{A3}$$

For the second factor, $w^{(n)}$ is a redundant condition because, without $h^{(n)}$, $w^{(n)}$ itself has no implication on θ:

$$\begin{aligned} &P(\theta|w^{(n)},w^{(n-1)},h^{(n-1)},\ldots,w^{(1)},h^{(1)}) \\ &= P(\theta|w^{(n-1)},h^{(n-1)},\ldots,w^{(1)},h^{(1)}) \\ &= f^{(n-1)}(\theta) \end{aligned} \tag{A4}$$

where the last equality comes from Eq. (A1). Substituting Eqs. (A3) and (A4) into Eq. (A2), we obtain the Bayesian update equation

$$f^{(n)}(\theta) \propto P(h^{(n)}|\theta,w^{(n)})f^{(n-1)}(\theta). \tag{A5}$$

2. Some properties of the Dirichlet distribution

Here we derive some properties of the K-state Dirichlet distribution

$$f_D(\phi;\alpha) \equiv \frac{1}{B(\alpha)}\phi_1^{\alpha_1-1}\cdots\phi_K^{\alpha_K-1}\delta\left(\sum_j \phi_j - 1\right) \tag{A6}$$

that are used in this paper. All the following derivations are based on the normalization equation

$$B(\alpha) = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(A)} = \int d\phi\phi_1^{\alpha_1-1}\cdots\phi_K^{\alpha_K-1}\delta\left(\sum_j \phi_j - 1\right), \tag{A7}$$

where $A \equiv \alpha_1+\cdots+\alpha_K$. First, we calculate the moments:

$$\begin{aligned} \langle\phi_1^n\rangle &= \frac{1}{B(\alpha)}\int d\phi\phi_1^{\alpha_1+n-1}\phi_2^{\alpha_2-1}\cdots\phi_K^{\alpha_K-1}\delta\left(\sum_j \phi_j - 1\right) \\ &= \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\frac{\Gamma(\alpha_1+n)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}{\Gamma(A+n)} \\ &= \frac{\alpha_1(\alpha_1+1)\cdots(\alpha_1+n-1)}{A(A+1)\cdots(A+n-1)} \end{aligned} \tag{A8}$$

from which we obtain the mean

$$\langle\phi_1\rangle = \frac{\alpha_1}{A} \tag{A9}$$

and the variance

$$\mathrm{var}(\phi_1) = \langle\phi_1^2\rangle - \langle\phi_1\rangle^2 = \frac{\alpha_1(A-\alpha_1)}{A^2(A+1)} = \frac{\langle\phi_1\rangle(1-\langle\phi_1\rangle)}{A+1}. \tag{A10}$$

In a similar manner, we calculate the covariance

$$\begin{aligned} \langle\phi_1\phi_2\rangle &= \frac{1}{B(\alpha)}\int d\phi\phi_1^{\alpha_1}\phi_2^{\alpha_2}\phi_3^{\alpha_3-1}\cdots\phi_K^{\alpha_K-1}\delta\left(\sum_j \phi_j - 1\right) \\ &= \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\frac{\Gamma(\alpha_1+1)\Gamma(\alpha_2+1)\Gamma(\alpha_3)\cdots\Gamma(\alpha_K)}{\Gamma(A+2)} \\ &= \frac{\alpha_1\alpha_2}{A(A+1)} \end{aligned} \tag{A11}$$

$$\begin{aligned} \mathrm{cov}(\phi_1,\phi_2) &= \langle\phi_1\phi_2\rangle - \langle\phi_1\rangle\langle\phi_2\rangle \\ &= -\frac{\alpha_1\alpha_2}{A^2(A+1)} \\ &= -\frac{\langle\phi_1\rangle\langle\phi_2\rangle}{(A+1)}. \end{aligned} \tag{A12}$$

Finally, we calculate the mean log

$$\begin{aligned}\langle \ln \phi_1 \rangle &= \frac{1}{B(\alpha)} \int d\phi \ln \phi_1 \phi_1^{\alpha_1 - 1} \cdots \phi_K^{\alpha_K - 1} \delta\left(\sum_j \phi_j - 1\right) \\ &= \frac{1}{B(\alpha)} \frac{\partial}{\partial \alpha_1} \int d\phi \phi_1^{\alpha_1 - 1} \cdots \phi_K^{\alpha_K - 1} \delta\left(\sum_j \phi_j - 1\right) \\ &= \frac{1}{B(\alpha)} \frac{\partial}{\partial \alpha_1} B(\alpha) \\ &= \frac{\partial}{\partial \alpha_1} \ln B(\alpha) \\ &= \frac{\partial}{\partial \alpha_1} [\ln \Gamma(\alpha_1) + \cdots + \ln \Gamma(\alpha_K) \\ &\quad - \ln \Gamma(\alpha_1 + \cdots + \alpha_K)] \\ &= \psi(\alpha_1) - \psi(A), \end{aligned} \tag{A13}$$

where $\psi(x) \equiv \frac{d}{dx} \ln \Gamma(x)$ is the digamma function. Although we derived the above formulas for the first state, by symmetry similar formulas can be written for any other states.

3. Minimization of the relative entropy

In this paper we use the minimization of the relative entropy as a guiding principle for constructing a mapping between weighted Dirichlet distributions. Here we find the solutions for the minimization for two different cases.

The first case is where we are given a weighted Dirichlet distribution $f(\theta) \equiv f_{\mathrm{WD}}(\theta; \alpha, w)$ and a new weight w' and want to find a new weighted Dirichlet distribution $f'(\theta) \equiv f_{\mathrm{WD}}(\theta; \alpha', w')$ that minimizes the relative entropy $D(f \| f')$ shown in Eq. (15). Thus, we calculate the derivative of $D(f \| f')$ with respect to α':

$$\begin{aligned}\frac{\partial}{\partial \alpha_i'} D(f \| f') &= -\int d\theta f(\theta) \frac{\partial}{\partial \alpha_i'} \ln f'(\theta) \\ &= -\int d\theta f(\theta) \frac{\partial}{\partial \alpha_i'} [\ln \Gamma(\alpha_1' + \cdots + \alpha_K') \\ &\quad - \ln \Gamma(\alpha_i') + (\alpha_i' - 1) \ln \phi_i'], \end{aligned} \tag{A14}$$

where we have used Eqs. (11) and (12). Setting the derivative to zero, we obtain

$$\psi(\alpha_i') - \psi(A') = \int d\theta f(\theta) \ln \phi_i' \tag{A15}$$

which can be recast into Eq. (17) by changing the integration variable from θ to ϕ, or into Eq. (16) by using Eq. (A13).

The second case is where in addition to $f(\theta)$ and w', the means $\bar{\phi}_i' = \alpha_i'/A'$ are also given; A' is the only unknown. By replacing α_i' with $\bar{\phi}_i' A'$, $f'(\theta)$ can be written as

$$\begin{aligned} f'(\theta) = \left| \frac{\partial \phi'}{\partial \theta} \right| &\frac{\Gamma(A')}{\Gamma(\bar{\phi}_1' A') \cdots \Gamma(\bar{\phi}_K' A')} \phi_1'^{(\bar{\phi}_1' A' - 1)} \cdots \phi_K'^{(\bar{\phi}_K' A' - 1)} \\ &\times \delta\left(\sum_j \theta_j - 1\right), \end{aligned} \tag{A16}$$

where $\phi_i' = w_i' \theta_i / \Sigma_j w_j' \theta_j$. In a similar manner as in the first case above, we take the derivative of $D(f \| f')$ with respect to A':

$$\begin{aligned}\frac{\partial}{\partial A'} D(f \| f') &= \sum_j \bar{\phi}_j' \psi(\bar{\phi}_j' A') - \psi(A') - \sum_j \bar{\phi}_j' \langle \ln \phi_j' \rangle_f \\ &= \sum_j \bar{\phi}_j' [\psi(\bar{\phi}_j' A') - \psi(A')] - \sum_j \bar{\phi}_j' \langle \ln \phi_j' \rangle_f \end{aligned} \tag{A17}$$

and then set it to zero to obtain Eq. (19).

4. Expansion of Eq. (19) around the zero-dispersion limit

Near the zero-dispersion limit, equations can be simplified by exploiting the property that distributions are sharply peaked. Here we expand Eq. (19), namely,

$$\sum_i \bar{\phi}_i' \langle \ln \phi_i' \rangle_{f'} = \sum_i \bar{\phi}_i' \langle \ln \phi_i' \rangle_f \tag{A18}$$

around the zero-dispersion limit.

First, let us expand $\langle \ln \phi_i' \rangle_{f'}$ on the left hand side. Define $\Delta_i' \equiv \phi_i' - \bar{\phi}_i'$ to be the deviation from the mean. Then, the expansion around the zero-dispersion limit amounts to the expansion around $\Delta_i' = 0$. Keeping up to the leading nontrivial order, we find

$$\begin{aligned}\langle \ln \phi_i' \rangle_{f'} &= \langle \ln (\bar{\phi}_i' + \Delta_i') \rangle_{f'} \\ &\approx \left\langle \ln \bar{\phi}_i' + \frac{\Delta_i'}{\bar{\phi}_i'} - \frac{(\Delta_i')^2}{2(\bar{\phi}_i')^2} \right\rangle_{f'} \\ &= \ln \bar{\phi}_i' - \frac{\mathrm{var}_{f'}(\phi_i')}{2(\bar{\phi}_i')^2} \\ &= \ln \bar{\phi}_i' - \frac{1 - \bar{\phi}_i'}{2\bar{\phi}_i'(A' + 1)}. \end{aligned} \tag{A19}$$

In the last line, Eq. (A10) was used.

$\langle \ln \phi_i' \rangle_f$ on the right hand side is expanded in a similar, although more complicated, manner:

$$\begin{aligned}\langle \ln \phi_i' \rangle_f &= \left\langle \ln \frac{u_i \phi_i}{\sum_j u_j \phi_j} \right\rangle_f \\ &\approx \left\langle \ln \frac{u_i \bar{\phi}_i}{\sum_j u_j \bar{\phi}_j} + \frac{\Delta_i}{\bar{\phi}_i} - \frac{\sum_j u_j \Delta_j}{\sum_j u_j \bar{\phi}_j} - \frac{\Delta_i^2}{2\bar{\phi}_i^2} \right. \\ &\quad \left. + \frac{(\sum_j u_j \Delta_j)^2}{2(\sum_j u_j \bar{\phi}_j)^2} \right\rangle_f, \end{aligned} \tag{A20}$$

where $u_i \equiv w_i'/w_i$ and $\Delta_i \equiv \phi_i - \bar{\phi}_i$. All these terms can be calculated using the properties of the Dirichlet distribution described in Appendix 2. In particular, the last term is calculated using

$$\left\langle \left(\sum_j u_j \Delta_j \right)^2 \right\rangle_f = \left\langle \sum_{j,k} u_j u_k \Delta_j \Delta_k \right\rangle_f = \sum_j u_j^2 \mathrm{var}_f(\phi_j) + \sum_{j \neq k} u_j u_k \mathrm{cov}_f(\phi_j, \phi_k). \tag{A21}$$

Collecting all the terms, we find

$$\langle \ln \phi_i' \rangle_f \approx \ln \frac{u_i \bar{\phi}_i}{\sum_j u_j \bar{\phi}_j} - \frac{1 - \bar{\phi}_i}{2\bar{\phi}_i (A+1)} + \frac{\sum_j u_j^2 \bar{\phi}_j (1 - \bar{\phi}_j) - \sum_{j \neq k} u_j u_k \bar{\phi}_j \bar{\phi}_k}{2(A+1) \sum_{j,k} u_j u_k \bar{\phi}_j \bar{\phi}_k}. \tag{A22}$$

Substituting Eqs. (A19) and (A22) into Eq. (A18) and solving for A', we obtain Eq. (20).

5. Monte Carlo sampling of state probabilities

In Bayesian inference, Monte Carlo methods are often employed for sampling parameters from posterior distributions [13]. In the context of weighted sampling, we are interested in sampling the state probabilities θ_i from the posterior

$$f^{(n)}(\theta) = P(\theta | w^{(n)}, h^{(n)}, \ldots, w^{(1)}, h^{(1)}) \propto P(w^{(n)}, h^{(n)}, \ldots, w^{(1)}, h^{(1)} | \theta) P(\theta), \tag{A23}$$

where the likelihood is

$$P(w^{(n)}, h^{(n)}, \ldots, w^{(1)}, h^{(1)} | \theta) = \prod_{m=1}^{n} \frac{H^{(m)}!}{h_1^{(m)}! \cdots h_K^{(m)}!} \times \frac{(w_1^{(m)} \theta_1)^{h_1^{(m)}} \cdots (w_K^{(m)} \theta_K)^{h_K^{(m)}}}{(w_1^{(m)} \theta_1 + \cdots + w_K^{(m)} \theta_K)^{H^{(m)}}}. \tag{A24}$$

For the prior, let us assume $\sigma_i^{\mathrm{guess}} = 0$ (i.e., $w_i^{(0)} = 1$) for simplicity

$$f^{(0)}(\theta) = P(\theta) = f_{\mathrm{WD}}(\theta; \alpha^{(0)}, w^{(0)}) = f_D(\theta; \alpha^{(0)}) = \frac{1}{B(\alpha)} \theta_1^{\alpha_1^{(0)} - 1} \cdots \theta_K^{\alpha_K^{(0)} - 1} \delta\left(\sum_j \theta_j - 1 \right), \tag{A25}$$

where, typically, $\alpha_i^{(0)} = 1$. In case $w_i^{(0)}$ are not uniform, one can reformulate the entire problem in terms of $\phi_i = w_i^{(0)} \theta_i / \Sigma_j w_j^{(0)} \theta_j$, thereby turning the problem into one with uniform initial weights.

Recently, Gallicchio *et al.*[4] suggested a method of Monte Carlo posterior sampling in the context of weighted sampling. However, we have found that that method does not yield satisfactory acceptance ratios, especially when the posterior distribution is sharp due to a large amount of data. It is much more efficient, we have found, to make moves on the logarithmic scale of θ. Below we describe this method.

Our method uses the Metropolis-Hastings algorithm [23,24] which consists of two parts, proposing a move

$$\theta = (\theta_1, \ldots, \theta_K) \rightarrow \theta' = (\theta_1', \ldots, \theta_K') \tag{A26}$$

and accepting/rejecting the move based on

$$R = \frac{p(\theta \leftarrow \theta') q(\theta')}{p(\theta' \leftarrow \theta) q(\theta)}, \tag{A27}$$

where $p(\theta' \leftarrow \theta)$ is the proposal probability density and $q(\theta)$ is the distribution that we want to sample from, i.e., $q(\theta) = f^{(n)}(\theta)$. The move is accepted with the probability of $\min\{1, R\}$.

To generate a move, we randomly choose a state (without loss of generality, assume the first state was chosen), draw a random number ε from a distribution $g(\varepsilon)$, and compute θ' as

$$\theta_1' = \frac{e^{\varepsilon} \theta_1}{e^{\varepsilon} \theta_1 + 1 - \theta_1}, \qquad \theta_j' = \frac{\theta_j}{e^{\varepsilon} \theta_1 + 1 - \theta_1} \quad \text{for } j \neq 1. \tag{28}$$

Notice that this amounts to making a move of ε in $\ln \theta_1$ followed by normalization $\Sigma_j \theta_j = 1$. A simple choice for $g(\varepsilon)$ is the uniform distribution between $-\Delta$ and $+\Delta$; the value of Δ is to be determined by trial and error such that a good acceptance ratio is obtained.

Under this movement scheme, the proposal probabilities are not symmetric, $p(\theta' \leftarrow \theta) \neq p(\theta \leftarrow \theta')$, and do not get canceled out in Eq. (A27). Therefore, we need to calculate the ratio $p(\theta \leftarrow \theta') / p(\theta' \leftarrow \theta)$. We start by writing the proposal probability as

$$p(\theta' \leftarrow \theta) d\theta_1' \cdots d\theta_{K-1}' = g(\varepsilon) d\varepsilon \prod_{j=2}^{K-1} \left[\delta\left(\theta_j' - \frac{\theta_j}{e^{\varepsilon} \theta_1 + 1 - \theta_1} \right) d\theta_j' \right]. \tag{A29}$$

Notice that ε determines θ_1', which in turn determines $\theta_2', \ldots, \theta_K'$; hence the delta functions. Also notice that there is no delta function assigned to θ_K' because it is already determined by the normalization $\Sigma_j \theta_j' = 1$; effectively, we are dealing with moves on a $K-1$ dimensional space. From Eq. (A28), we obtain

$$\varepsilon = \ln \frac{\theta_1'(1 - \theta_1)}{\theta_1(1 - \theta_1')}, \qquad d\varepsilon = \frac{d\theta_1'}{\theta_1'(1 - \theta_1')} \tag{A30}$$

which is used to eliminate ε from Eq. (A29). After eliminating ε, we find

$$p(\theta' \leftarrow \theta) = g\left(\ln \frac{\theta_1'(1-\theta_1)}{\theta_1(1-\theta_1')}\right) \frac{1}{\theta_1'(1-\theta_1')} \times \prod_{j=2}^{K-1} \delta\left(\theta_j' - \frac{\theta_j(1-\theta_1')}{1-\theta_1}\right) = g\left(\ln \frac{\theta_1'(1-\theta_1)}{\theta_1(1-\theta_1')}\right) \frac{(1-\theta_1)^{K-2}}{\theta_1'(1-\theta_1')} \times \prod_{j=2}^{K-1} \delta[\theta_j'(1-\theta_1) - \theta_j(1-\theta_1')]. \quad \text{(A31)}$$

The ratio is thus given as

$$\frac{p(\theta \leftarrow \theta')}{p(\theta' \leftarrow \theta)} = \left[g\left(\ln \frac{\theta_1(1-\theta_1')}{\theta_1'(1-\theta_1)}\right) \Big/ g\left(\ln \frac{\theta_1'(1-\theta_1)}{\theta_1(1-\theta_1')}\right) \right] \times \frac{\theta_1'(1-\theta_1')^{K-1}}{\theta_1(1-\theta_1)^{K-1}}, \quad \text{(A32)}$$

where the delta functions have been cancelled out because they are even functions. If $g(\cdots)$ is also an even function, we get a simpler formula

$$\frac{p(\theta \leftarrow \theta')}{p(\theta' \leftarrow \theta)} = \frac{\theta_1'(1-\theta_1')^{K-1}}{\theta_1(1-\theta_1)^{K-1}} \quad \text{(A33)}$$

which, by generalization, becomes

$$\frac{p(\theta \leftarrow \theta')}{p(\theta' \leftarrow \theta)} = \frac{\theta_i'(1-\theta_i')^{K-1}}{\theta_i(1-\theta_i)^{K-1}} \quad \text{(A34)}$$

when state i is selected for a move. This completes our Monte Carlo sampling technique. In practice, in view of the issue of numerical precision, it is advantageous to calculate $L \equiv \ln R$ first, instead of R. If $L \geq 0$, the proposed move is accepted; if $L < 0$, it is accepted with the probability of e^L.

[1] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[2] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[3] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[4] E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy, J. Phys. Chem. B **109**, 6722 (2005).
[5] C. Bartels and M. Karplus, J. Comput. Chem. **18**, 1450 (1997).
[6] G. R. Smith and A. D. Bruce, J. Phys. A **28**, 6623 (1995).
[7] B. A. Berg, J. Stat. Phys. **82**, 323 (1996).
[8] J. S. Wang and R. H. Swendsen, J. Stat. Phys. **106**, 245 (2002).
[9] F. G. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[10] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992).
[11] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[12] E. T. Jaynes, *Probability Theory: The Logic of Science*, (Cambridge University Press, Cambridge, 2003).
[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. (CRC Press, Boca Raton, 2003).
[14] Notice that the mapping we seek is not just the change of variables from ϕ to ϕ'. This change of variables will map a Dirichlet distribution in ϕ to a *non*-Dirichlet distribution in ϕ'.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[16] L. Devroye, *Non-Uniform Random Variate Generation* (Springer-Verlag, Berlin, 1986).
[17] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[18] S. Nosé, Mol. Phys. **52**, 255 (1984).
[19] W. G Hoover, Phys. Rev. A **31**, 1695 (1985).
[20] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[21] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. (to be published).
[22] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[24] W. K. Hastings, Biometrika **57**, 97 (1970).
[25] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).