Title	Distribution-function approach to free energy computation.
Author(s)	Sakuraba, Shun; Matubayasi, Nobuyuki
Citation	The Journal of chemical physics (2011), 135(11)
Issue Date	2011-09
URL	http://hdl.handle.net/2433/147965
Right	© 2011 American Institute of Physics
Type	Journal Article
Textversion	publisher

Distribution-function approach to free energy computation

Shun Sakuraba and Nobuyuki Matubayasi

Citation: J. Chem. Phys. **135**, 114108 (2011); doi: 10.1063/1.3637036
View online: http://dx.doi.org/10.1063/1.3637036
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v135/i11
Published by the American Institute of Physics.

Distribution-function approach to free energy computation

Shun Sakuraba[1] and Nobuyuki Matubayasi[1,2,a)]
[1]*Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan*
[2]*Japan Science and Technology Agency (JST), CREST, Kawaguchi, Saitama 332-0012, Japan*

(Received 19 May 2011; accepted 22 August 2011; published online 19 September 2011)

Connections are explored between the free energy difference of two systems and the microscopic distribution functions of the energy difference. On the basis of a rigorous relationship between the energy distribution functions and the free energy, the scheme of error minimization is introduced to derive accurate and simple methods of free energy computation. A set of distribution-function approaches are then examined against model systems, and the newly derived methods exhibit state-of-art performance. It is shown that the notion of error minimization is powerful to improve the free energy calculation using distribution functions. © *2011 American Institute of Physics*. [doi:10.1063/1.3637036]

I. INTRODUCTION

Liquid, biomolecule, and lipid membrane have both order and randomness. To describe these, weakly ordered systems, a statistical-mechanical viewpoint is indispensable. A method of distribution function is highly useful in statistical-mechanical description. A few physical variables are chosen to characterize the state of the system of interest, and its probability distribution (with appropriate normalization) provides ensemble-averaged information often on the molecular and microscopic level.

A distribution function is even more useful when it constructs a macroscopic or thermodynamic observable. The distribution function can then serve as a bridge between molecular understanding and realistic observation. In chemical applications, a most important observable is the (standard) free energy change for a process of interest. A number of approximate theories are thus formulated which compute the free energy in terms of distribution function, for example, of distance, spatial coordinate, or energy.[1–13]

A fundamental expression connecting the free energy difference and distribution functions was provided by Shing and Gubbins.[14] In this expression, the distribution functions of the energy difference between two systems of interest constitute an exact functional for the free energy difference between the two systems. In practice, the histograms of the energy difference are constructed for the two systems and the overlapping region of the two histograms is used to estimate the free energy difference; the method is often called overlapping distribution method.

The distribution-function based approach to free energy computation was significantly strengthened by a recent work of Yokogawa.[15] On the basis of a hypernetted-chain-like formulation used in equilibrium theory of liquids, he put forward an exact, analytical functional for the free energy expressed in terms of distribution functions of energy. His work showed that the statistical mechanics of solutions can improve the numerical evaluation of free energy (and entropy).

In the present work, we investigate distribution-function methods for evaluating the free energy difference between two systems of interest. We examine the original overlapping distribution method and Yokogawa's method, and propose new and improved schemes formulated by minimizing the estimated error of free energy computation. The error due to finite sampling is analyzed with statistical methods, and is used to determine optimized functionals for exact computation of the free energy.

Apart from distribution-function based approaches, several methods were developed to obtain the free energy difference.[16–20] In the method of thermodynamic integration,[21] the "mean force" is calculated along reaction coordinate and is integrated to provide the free energy difference. The free energy perturbation method[22] uses the forward or reverse work over typically small changes of the system. The idea of introducing the statistical error analysis into free energy estimation appears in the weighted histogram analysis method (WHAM) (Refs. 23 and 24) and Bennett's acceptance ratio (BAR) (Ref. 25) method. The former improves the free energy estimation from observations in multiple systems, and the latter does so by combining both the forward and reverse works. The key notion in WHAM and BAR is the error minimization, and it is employed in the present work to optimize the free energy calculation based on distribution function.

The organization of the present paper is as follows. In Sec. II, we review Shing-Gubbins formula for the free energy difference and express a few free energy estimators as functionals of distribution function. We then derive new estimators based on error analysis. In Sec. III, we test the free energy schemes against model systems of harmonic oscillator and small biomolecules of ~40 atoms. We conclude in Sec. IV.

II. THEORY

The target quantity of the present developments is the free energy difference between two systems, called systems

a)Author to whom correspondence should be addressed. Electronic mail: nobuyuki@scl.kyoto-u.ac.jp.

0 and 1. The reduced potential function defined as[26] $U_\lambda(\boldsymbol{x}) = (k_B T_\lambda)^{-1} V_\lambda(\boldsymbol{x})$ is provided at $\lambda = 0$ and 1, where $V_\lambda(\boldsymbol{x})$ is the potential function defined over the configuration space $\boldsymbol{x}$, k_B is the Boltzmann constant, and T_λ is the temperature. The configuration space $\boldsymbol{x}$ is the same for both systems of $\lambda = 0$ and 1. The configurations in system λ are sampled in accordance with the Boltzmann distribution proportional to $\exp(-U_\lambda(\boldsymbol{x}))$. The ith sample or snapshot from the simulation of system λ is denoted as $\boldsymbol{x}_{\lambda i}$. The total number of samples in system λ is N_λ. To be exact, N_λ is the number of uncorrelated samples.[27,28] The dimensionless free energy difference Δf, which is the target quantity, is given by

$$\Delta f = -\log \frac{\int \mathrm{d}\boldsymbol{x} \exp\left(-U_1(\boldsymbol{x})\right)}{\int \mathrm{d}\boldsymbol{x} \exp\left(-U_0(\boldsymbol{x})\right)}. \tag{1}$$

We introduce the distribution function of reduced energy difference as[14,15]

$$\rho_\lambda(\varepsilon) = \frac{\int \mathrm{d}\boldsymbol{x} \delta\left(U_1(\boldsymbol{x}) - U_0(\boldsymbol{x}) - \varepsilon\right) \exp\left(-U_\lambda(\boldsymbol{x})\right)}{\int \mathrm{d}\boldsymbol{x} \exp\left(-U_\lambda(\boldsymbol{x})\right)}$$
$$\text{for } \lambda = 0 \text{ and } 1, \tag{2}$$

where $\delta(\cdot)$ is the Dirac delta function. This distribution function is the density of total energy difference ε between the two systems. Since the number of configurations from a simulation is finite, the true energy distribution $\rho_\lambda(\varepsilon)$ cannot be obtained in practice. Instead, the distribution can be obtained only from finite samples, and furthermore, the energy difference needs to be discretized into a set of finite bins.

The approximate probability distribution from finite samples and finite bins is introduced as

$$\widetilde{\rho_{\lambda i}} = \frac{1}{N_\lambda} \int_{e_{i-1}}^{e_i} \mathrm{d}\varepsilon \sum_{i=1}^{N_\lambda} \delta\left(U_1(\boldsymbol{x}_{\lambda i}) - U_0(\boldsymbol{x}_{\lambda i}) - \varepsilon\right), \tag{3}$$

where $\widetilde{\rho_{\lambda i}}$ is the discretized distribution function in system λ for the ith bin with $1 \le i \le B$, and B is the number of bins. The binning of the energy difference is conducted with $(B + 1)$ energy values of

$$e_0 < e_1 < \ldots < e_B, \tag{4}$$

and they are typically partitioned with equal intervals. With $N_\lambda \to \infty$, $\widetilde{\rho_{\lambda i}}$ converges to

$$\int_{e_{i-1}}^{e_i} \mathrm{d}\varepsilon \rho_\lambda(\varepsilon), \tag{5}$$

and thus $\widetilde{\rho_{\lambda i}}$ is a finite-sample approximation to the discretized distribution function. $\widetilde{\rho_{\lambda i}}$ is usually obtained from a simulation by

$$\widetilde{\rho_{\lambda i}} = \frac{n_{\lambda i}}{N_\lambda}, \tag{6}$$

where $n_{\lambda i}$ is the number of samples (snapshots) falling into the ith bin.

A. Linear thermodynamic integration

The method of linear thermodynamic integration (LTI) was introduced by Mezei and Beveridge[29] to obtain the free energy difference between two systems. LTI approximates the thermodynamic integration linearly through

$$\Delta f_{\mathrm{LTI}} = \frac{1}{2} \int \mathrm{d}\varepsilon\, \varepsilon \left(\rho_0(\varepsilon) + \rho_1(\varepsilon)\right). \tag{7}$$

In this approximation, the free energy difference does not converge to the exact value even in the limit of bringing both N_0 and N_1 to infinity. Thus, the LTI is a biased estimator. The difference between the exact free energy difference Δf and the LTI estimation Δf_{LTI} is given as[15,30]

$$\Delta f - \Delta f_{\mathrm{LTI}} = \frac{1}{2} \int \mathrm{d}\varepsilon \left(\rho_0(\varepsilon) + \rho_1(\varepsilon)\right) \log \frac{\rho_1(\varepsilon)}{\rho_0(\varepsilon)}. \tag{8}$$

This difference will increase when $\rho_1(\varepsilon)$ differs more from $\rho_0(\varepsilon)$.

B. Overlapping distribution method and Yokogawa's method

Shing and Gubbins proposed the overlapping distribution method.[14] This method estimates the free energy difference from the two distribution functions $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$ of the energy difference. The key equation is written as[14,25]

$$\Delta f = \log \rho_1(\varepsilon) - \log \rho_0(\varepsilon) + \varepsilon. \tag{9}$$

Equation (9) is valid for any ε when the numbers of samples N_0 and N_1 are infinite. A single, arbitrary ε does not provide an accurate Δf in practice, however, due to limited number of samples. A weighting function $w(\varepsilon)$ is then introduced to stabilize the calculation. The free energy can be calculated as a functional with arbitrary weighting function $w(\varepsilon)$ through

$$\Delta f = \int \mathrm{d}\varepsilon w(\varepsilon) \left(\log \rho_1(\varepsilon) - \log \rho_0(\varepsilon) + \varepsilon\right), \tag{10}$$

where $w(\varepsilon)$ satisfies the normalization condition of

$$\int \mathrm{d}\varepsilon\, w(\varepsilon) = 1. \tag{11}$$

Equation (10) involves the integration over the continuous ε. In numerical implementation, a discretized version needs to be employed. Using a set of bins given by Eq. (4), we rewrite Eq. (10) into

$$\Delta f = \Delta f^{\mathrm{binned}} + \Delta f^{\mathrm{rest}} \tag{12}$$

$$\equiv \sum_{i=1}^{B} w_i \left(\log \rho_{1i} - \log \rho_{0i} + \varepsilon_i\right) + \Delta f^{\mathrm{rest}} \tag{13}$$

$$\simeq \sum_{i=1}^{B} w_i \left(\log \widetilde{\rho_{1i}} - \log \widetilde{\rho_{0i}} + \varepsilon_i\right) + \Delta f^{\mathrm{rest}}, \tag{14}$$

where the discretized weight w_i is

$$w_i = \int_{e_{i-1}}^{e_i} \mathrm{d}\varepsilon w(\varepsilon). \tag{15}$$

ε_i is the representative energy of the ith bin and is usually set to

$$\varepsilon_i = \frac{1}{2}\left(e_{i-1} + e_i\right), \tag{16}$$

where e_i is introduced in Eq. (4) and defines the "edge" of bin. Equation (16) expresses the mid-point convention and will be justified in Appendix A. The discretized probability distribution $\rho_{\lambda i}$ within Eq. (13) is introduced through Eq. (2) as

$$\rho_{\lambda i} = \int_{e_{i-1}}^{e_i} \mathrm{d}\varepsilon \rho_\lambda(\varepsilon). \tag{17}$$

The free energy difference from the discretized set of distribution is defined as the first term of Eq. (13), and the term is written as $\Delta f^{\mathrm{binned}}$. The extra term Δf^{rest} represents the error from the discretization of the continuous energy difference ε, and will be discussed further in Sec. II E and Appendix A. We write Eqs. (13) and (14) distinctively to emphasize the finiteness of the number of samples N_λ. $\rho_{\lambda i}$ corresponds to the binned probability distribution with infinite number of samples, while $\widetilde{\rho_{\lambda i}}$ is its counterpart with finite samples. Some of $\widetilde{\rho_{\lambda i}}$ may vanish even when the corresponding $\rho_{\lambda i}$ is non-zero, thus the divergent $\log \widetilde{\rho_{\lambda i}}$ can be problematic in Eq. (14).

To avoid the problem of vanishing $\widetilde{\rho_{\lambda i}}$, the overlapping distribution (hereafter OD) method adopts a weighting function $w(\varepsilon)$ in the form of[14]

$$w_{\mathrm{OD}}(\varepsilon) = \begin{cases} (e_m - e_l)^{-1} & e_l \le \varepsilon \le e_m \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The minimum and maximum energies of e_l and e_m are determined to be within "good" overlapping region of $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$, in which both of $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$ are obtained with good statistics. The corresponding discretized weight is given by

$$w_{\mathrm{OD},i} = \begin{cases} (m-l)^{-1} & l+1 \le i \le m \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

when the binning is done with an equal interval.

Yokogawa proposed a variant of OD method[15,30] (hereafter Yokogawa's method), with a weighting function expressed as

$$w_{\mathrm{Yokogawa}}(\varepsilon) = \frac{1}{2}\left(\rho_0(\varepsilon) + \rho_1(\varepsilon)\right). \tag{20}$$

The corresponding discretized weight is

$$w_{\mathrm{Yokogawa},i} = \frac{1}{2}\left(\widetilde{\rho_{0i}} + \widetilde{\rho_{1i}}\right). \tag{21}$$

With this setting, however, the weight may be non-zero even in the bin with $\widetilde{\rho_{\lambda i}} = 0$. To avoid the problem of divergent $\log \widetilde{\rho_{\lambda i}}$, Yokogawa introduced a smoothed estimate of the histogram $\rho^\star_{\lambda i}$ as

$$\rho^\star_{\lambda i} \propto P_{\lambda,i}\widetilde{\rho_{\lambda i}} + P_{1-\lambda,i}\frac{\exp(-\varepsilon_i)\,\widetilde{\rho_{1-\lambda,i}}}{\sum_{j=1}^{B}\exp\left(-\varepsilon_j\right)\widetilde{\rho_{1-\lambda,j}}}, \tag{22}$$

$$P_{\lambda,i} = \frac{\widetilde{\rho_{\lambda i}}}{\widetilde{\rho_{0i}} + \widetilde{\rho_{1i}}}. \tag{23}$$

The normalization factor is determined from $\sum_{i=1}^{B}\rho^\star_{\lambda i} = 1$ and $\rho^\star_{\lambda i}$ is used in place of $\widetilde{\rho_{\lambda i}}$ in Eq. (14).

C. Error minimization

The weighting function $w(\varepsilon)$ is essential to compute the free energy difference accurately. In this subsection, we derive new forms of $w(\varepsilon)$ by minimizing the error of estimated free energy difference. The statistical error of $\Delta f^{\mathrm{binned}}$ is estimated as

$$\sigma^2(\Delta f^{\mathrm{binned}}) = \sum_{i=1}^{B} w_i^2\left(\sigma^2(\log\rho_{0i}) + \sigma^2(\log\rho_{1i})\right), \tag{24}$$

where $\sigma^2(\log\rho_{\lambda i})$ denotes the estimated variance of $\log\widetilde{\rho_{\lambda i}}$ of Eq. (14) and w_i satisfies the normalization constraint of

$$\sum_{i=1}^{B} w_i = 1. \tag{25}$$

The total error $\sigma^2(\Delta f^{\mathrm{binned}})$ can then be minimized by choosing w_i as

$$w_i = \frac{C}{\sigma^2(\log\rho_{0i}) + \sigma^2(\log\rho_{1i})}, \tag{26}$$

where C is the normalizing constant and is determined from Eq. (25). With w_i of Eq. (26), the total statistical error is

$$\sigma^2(\Delta f^{\mathrm{binned}}) = C = \left(\sum_{i=1}^{B}\frac{1}{\sigma^2(\log\rho_{0i}) + \sigma^2(\log\rho_{1i})}\right)^{-1}. \tag{27}$$

The concept of determining the weight according to the error minimization is used in well-known methodologies such as the WHAM.[23,24]

The remaining problems are (1) to estimate $\sigma^2(\log\rho_{\lambda i})$ and (2) to circumvent the divergence of $\log\widetilde{\rho_{\lambda i}}$. For this purpose, we resort to the Bayesian analysis of the probability to fall into each bin and derive the estimated mean $\overline{\log\rho_{\lambda i}}$ and variance $\sigma^2(\log\rho_{\lambda i})$ of $\log\widetilde{\rho_{\lambda i}}$ as

$$\overline{\log\rho_{\lambda i}} = -\sum_{k=n_{\lambda i}+1}^{N_\lambda+B-1}\frac{1}{k}, \tag{28}$$

$$\sigma^2(\log\rho_{\lambda i}) = \sum_{k=n_{\lambda i}+1}^{N_\lambda+B-1}\frac{1}{k^2}, \tag{29}$$

where $n_{\lambda i}$ is the number of samples falling into the ith bin and appeared with respect to Eq. (6). The derivation of Eqs. (28) and (29) is given in Appendix B. Note that $\overline{\log\rho_{\lambda i}}$ will converge to $\log\rho_{\lambda i}$ with $N_\lambda \to \infty$ and does not diverge even when $n_{\lambda i} = 0$. The free energy difference is then written as

$$\Delta f^{\mathrm{binned}}_{\mathrm{EROD}} = \sum_{i=1}^{B} w_{\mathrm{EROD},i}\left(\overline{\log\rho_{1i}} - \overline{\log\rho_{0i}} + \varepsilon_i\right), \tag{30}$$

where $w_{\mathrm{EROD},i}$ is given by Eq. (26) combined with Eq. (29). In the following, we call this method error-resilient overlapping distribution (EROD) method because the method focuses on heavily sampled region and is insensitive to outliers.

Another weighting strategy can be formulated from an error analysis of the number of samples in the bin. Let

$\sigma^2(n_{\lambda i})$ be the variance of the number of samples falling into the ith bin. When N_λ is fixed, $\sigma^2(\log \rho_{\lambda i})$ is estimated as $\sigma^2(n_{\lambda i})/n_{\lambda i}^2$, by virtue of Eq. (6) and the chain rule for the variance. Further, $\sigma^2(n_{\lambda i})$ is equal to $n_{\lambda i}$ by assuming the Poisson distribution, as adopted in the derivation of WHAM.[23,24] We then have

$$\sigma^2(\log \rho_{\lambda i}) \simeq \frac{1}{n_{\lambda i}}, \tag{31}$$

and derive an easy-to-use scheme with the weight of

$$w_{\mathrm{HMOD},i} = C_{\mathrm{HMOD}} \left(\frac{1}{n_{0i}} + \frac{1}{n_{1i}} \right)^{-1} = C_{\mathrm{HMOD}} \frac{n_{0i} n_{1i}}{n_{0i} + n_{1i}}$$
$$= C_{\mathrm{HMOD}} \frac{N_0 N_1 \widetilde{\rho_{0i}} \widetilde{\rho_{1i}}}{N_0 \widetilde{\rho_{0i}} + N_1 \widetilde{\rho_{1i}}}, \tag{32}$$

where C_{HMOD} is the normalizing constant expressed through Eq. (27) as

$$C_{\mathrm{HMOD}} = \left(\sum_{i=1}^{B} \frac{N_0 N_1 \widetilde{\rho_{0i}} \widetilde{\rho_{1i}}}{N_0 \widetilde{\rho_{0i}} + N_1 \widetilde{\rho_{1i}}} \right)^{-1}. \tag{33}$$

The continuous counterpart of Eqs. (32) and (33) is

$$w_{\mathrm{HMOD}}(\varepsilon) = C_{\mathrm{HMOD}} \frac{N_0 N_1 \rho_0(\varepsilon) \rho_1(\varepsilon)}{N_0 \rho_0(\varepsilon) + N_1 \rho_1(\varepsilon)}, \tag{34}$$

$$C_{\mathrm{HMOD}} = \left(\int \mathrm{d}\varepsilon \frac{N_0 N_1 \rho_0(\varepsilon) \rho_1(\varepsilon)}{N_0 \rho_0(\varepsilon) + N_1 \rho_1(\varepsilon)} \right)^{-1}. \tag{35}$$

We call this method harmonic mean overlapping distribution (HMOD) method. Its obvious advantage is the simplicity. Furthermore, the weighting factor $w_{\mathrm{HMOD},i}$ serves to emphasize the ε domain in which both $\widetilde{\rho_{0i}}$ and $\widetilde{\rho_{1i}}$ are non-zero and large. The bin with $\widetilde{\rho_{0i}} = 0$ or $\widetilde{\rho_{1i}} = 0$ is weighed as $w_{\mathrm{HMOD},i} = 0$ and is excluded from numerical calculation. Using Eq. (14), the free energy difference is calculated as

$$\Delta f_{\mathrm{HMOD}}^{\mathrm{binned}} = \sum_{i=1}^{B} w_{\mathrm{HMOD},i} \left(\log \widetilde{\rho_{1i}} - \log \widetilde{\rho_{0i}} + \varepsilon_i \right), \tag{36}$$

and the statistical error is estimated as Eq. (27) through Eq. (33) in discrete case and through Eq. (35) in continuous case.

The assumptions of the chain rule and the Poisson distribution are independent from the Bayesian assumption. The two schemes of EROD and HMOD were thus derived independently through error minimization with the different estimates of errors. Both are unbiased estimators for the free energy difference, and their rates of convergence determine which is superior or inferior.

D. Bennett's acceptance ratio method

BAR (Ref. 25) estimates the free energy difference on the basis of the variance minimization. It directly computes the free energy difference from the set of energy differences corresponding to snapshots of the system. The BAR is derived as the maximum likelihood solution of Eq. (9) as[31]

$$\Delta f_{\mathrm{BAR}} = C_{\mathrm{BAR}} - \log \frac{N_1}{N_0}, \tag{37}$$

$$C_{\mathrm{BAR}} = \Delta f_{\mathrm{BAR}} + \log \frac{N_1 \sum_{i=1}^{N_0} \mathcal{F}(U_1(\boldsymbol{x}_{0i}) - U_0(\boldsymbol{x}_{0i}) - C_{\mathrm{BAR}})}{N_0 \sum_{i=1}^{N_1} \mathcal{F}(U_0(\boldsymbol{x}_{1i}) - U_1(\boldsymbol{x}_{1i}) + C_{\mathrm{BAR}})}, \tag{38}$$

where $\mathcal{F}$ is the Fermi function defined as

$$\mathcal{F}(x) = \frac{1}{1 + \exp(x)}, \tag{39}$$

and the sum is taken over systems 0 and 1, respectively, for the numerator and denominator within the logarithm of Eq. (38). The free energy difference Δf_{BAR} is calculated self-consistently from Eqs. (37) and (38). In the limit of large N_0 and N_1, Eq. (38) reduces to an identity expressed as

$$C_{\mathrm{BAR}} = \Delta f_{\mathrm{BAR}} + \log \frac{\int \mathrm{d}\varepsilon \rho_0(\varepsilon) \mathcal{F}(\varepsilon - C_{\mathrm{BAR}})}{\int \mathrm{d}\varepsilon \rho_1(\varepsilon) \mathcal{F}(-\varepsilon + C_{\mathrm{BAR}})}, \tag{40}$$

which is valid for any C_{BAR} by virtue of Eq. (9); note that C_{BAR} of Eqs. (37) and (38) is the optimum at finite N_0 and N_1. Equation (40) has only a simple (linear) average over $\rho_\lambda(\varepsilon)$, and does not involve any non-linear operation on $\rho_\lambda(\varepsilon)$. Thus, the binning of ε is not necessary, and the binning error is absent in the form of Eq. (38). In fact, LTI has this property, too, as is evidenced in Eq. (7). As the BAR does not use bins, a straightforward implementation requires $O(N_0 + N_1)$ memory space, while distribution-function based methods require only $O(B)$ space, where B is the number of bins and is typically much smaller than N_0 and N_1.

Still, the statistical error exists in the estimation of BAR, which is provided in the literature.[25,31] It is given by

$$\sigma^2(\Delta f_{\mathrm{BAR}}) = \left(\int \mathrm{d}\varepsilon \frac{N_0 N_1 \rho_0(\varepsilon) \rho_1(\varepsilon)}{N_0 \rho_0(\varepsilon) + N_1 \rho_1(\varepsilon)} \right)^{-1} - \frac{N_0 + N_1}{N_0 N_1}. \tag{41}$$

If the error is expressed as the variance, no method based only on the set of energy difference shall provide a better estimation than BAR, without a specific model for $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$. The error of HMOD given by Eq. (35) is equal to the first term of Eq. (41). The second term $(N_0 + N_1)/N_0 N_1$ in Eq. (41) vanishes at $N_\lambda \to \infty$ and is usually negligible in practice. The errors from HMOD and BAR are thus expected to coincide with each other at large enough N_0 and N_1.

E. Systematic error due to binning

The systematic error due to the binning, represented as Δf^{rest} in Eq. (12), arises from the discretization of the continuous energy difference ε. Δf^{rest} depends on the number of bins B and on the definition of bin "edges" $e_0, e_1, \ldots, e_B$ given by Eq. (4). Generally speaking, the systematic error is larger if the probability distribution $\rho_0(\varepsilon)$ and/or $\rho_1(\varepsilon)$ is less flat within the bin. More precisely, the error from the ith bin is expressed as

$$\Delta f_i^{\mathrm{rest}} \simeq -\frac{1}{24} w_i (e_i - e_{i-1})^2$$

$$\times \left[\left(\frac{\partial \log \rho_1(\varepsilon)}{\partial \varepsilon} \right)^2_{\varepsilon=\varepsilon_i} - \left(\frac{\partial \log \rho_0(\varepsilon)}{\partial \varepsilon} \right)^2_{\varepsilon=\varepsilon_i} \right], \tag{42}$$

$$\Delta f^{\text{rest}} = \sum_{i=1}^{B} \Delta f_i^{\text{rest}}. \tag{43}$$

The derivation of Eq. (42) is described in Appendix A. Note that although the systematic error will converge to zero with the increase of the number of bins B, it does not vanish only with the increase of the number of samples N_λ. The systematic error given by Eq. (42) is enhanced especially when $\rho_\lambda(\varepsilon)$ is "rough" in the sense that the relative variation of $\rho_\lambda(\varepsilon)$ within the bin is large.

III. RESULTS AND DISCUSSION

In this section, we test the free energy methods against harmonic oscillator and biomolecules of ~40 atoms. The results will be all shown in dimensionless form in accordance with the developments in Sec. II.

A. One-dimensional harmonic oscillator

We examine the free energy difference of a pair of one-dimensional harmonic oscillators, represented by $U = (x \pm \alpha/2)^2$. The two harmonics have different minimum coordinates at $x = \pm\alpha/2$ and the free energy difference is exactly 0. The statistical ensemble was generated by the Box-Muller transform[32] over the number of samples $N_\lambda = 100\,000$ for each of the two harmonic systems. The free energy differences were computed for all the methods described in Sec. II. For each numerical method, we performed the simulation for 1000 times, to investigate the statistical errors. When the free energy difference is calculated to be $\Delta f(i)$ at the ith simulation, the statistical error is given by

$$\sqrt{\frac{1}{1000} \sum_{i=1}^{1000} \Delta f(i)^2}$$

since the exact value of the free energy difference is 0. The OD, EROD, and HMOD methods used $B = 100$ bins, equally partitioning the overlapping region of non-zero $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$. In the OD method, only the bins with more than 10 sample counts were used with Eq. (19). Yokogawa's method used 100 bins equally partitioning between the minimum and maximum energies of non-zero $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$.

The results are summarized in Fig. 1. In Fig. 1, LTI shows the best convergence. The reason of LTI's stability is that the bias term in Eq. (8) vanishes in the present case. The HMOD and BAR methods follow at the second in tie, presenting accurate and robust estimation of the free energy. The EROD method is also robust, although it has slightly larger error when $\alpha \geq 9$. The Yokogawa value increases steeply with α as the overlap becomes insufficient. The reason for this is considered to come from the treatment of low-counts region in the histogram. The OD estimation has a large error even at small α. This is due to the presence of low-counts bins. Outer bins in the overlapping region of $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$ have generally low sample counts, whereas even low-counts bins have the same weights as the others and amplify the error, as shown with Eqs. (24), (29), and (31).

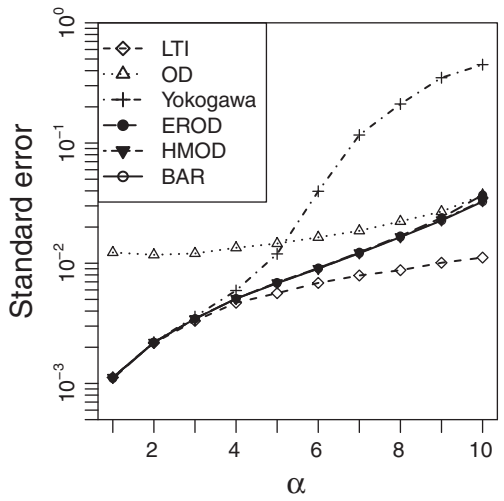

FIG. 1. The standard error of the free energy estimation as a function of the distance between the minima α of two harmonics. Note the log-scale of the ordinate. The lines are drawn for eye guide. The HMOD and BAR values are barely distinguishable from each other for all α. The EROD value is barely distinguishable from the HMOD and BAR values when $\alpha \leq 8$, and is slightly larger than them for $\alpha \geq 9$.

We consider the effect of the number of bins B. Figure 2 represents the dependence of the statistical error on B, for $U = (x \pm \alpha/2)^2$ with $\alpha = 5$. The abscissa of Fig. 2 is expressed in the form of N/B, which corresponds to the number of sample counts in each bin if the counts in the bins are constant. LTI is removed from comparison as it is irrespective to the binning, while BAR is left in the figure for comparison. With enough number of samples in each bin ($N/B \gtrsim 5 \times 10^2$), the errors of EROD and HMOD are smaller than those of OD and Yokogawa's methods, and are hardly distinguishable from that of BAR. Furthermore, HMOD remains almost as accurate as BAR for any number of bins.

With the distribution-function based methods derived from error minimization, therefore, the statistical error can be made competitive to the state-of-art method. The midpoint convention of ε_i, as specified in Eq. (16) and justified in Appendix A, reduces the systematic error due to binning, furthermore, which turns out to be inappreciable in practice. EROD's instability compared to HMOD's for small N/B is considered to be caused mainly from assigning non-zero weights to the bins with $n_{\lambda i} = 0$ in EROD, since the error is significant only when N/B is small.

B. Realistic chemical species

We examine the free energy differences of ATP (ATP^{4-}; adenosine 5′-triphosphate) and ADP (ADP^{3-}; adenosine 5′-diphosphate) ions between 300 K and a variety of temperatures T K with $T = 150 \cdot (300/150)^{i/10}$ ($i = 0, 1, \ldots, 9$). The all-atom, classical force field parameters by Meagher *et al.*[33] were used for ATP and ADP. The system was modeled as a

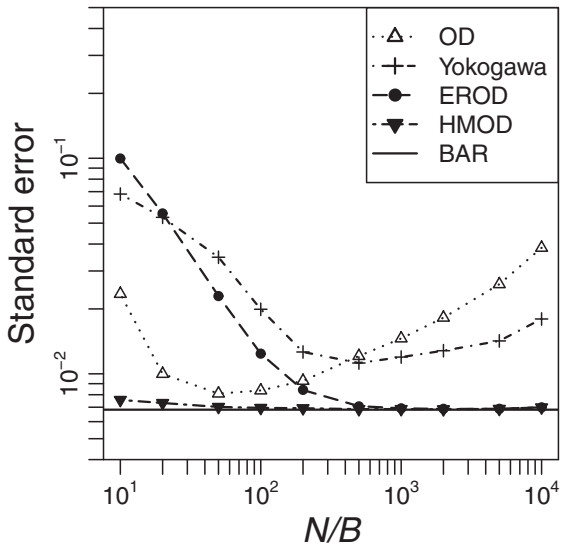

FIG. 2. The dependence of the standard error on the number of bins B for a pair of harmonics $U = (x \pm \alpha/2)^2$ with $\alpha = 5$. The abscissa is expressed in the form of N/B with the number of samples $N = N_\lambda = 100\,000$. N/B is equal to the number of sample counts in each bin if the counts are constant. LTI is removed from the comparison as it is unaffected by the number of bins. Since BAR is unaffected by the number of bins, its result is represented as a horizontal line without symbols. The lines for the other methods are drawn for eye guide. Note the log scales of both axes. The HMOD data overlaps the line for BAR when $N/B \geq 10^2$.

single particle *in vacuo*, and the simulation was performed with GROMACS 4.0 at double precision.[34] The Langevin dynamics was employed at constant temperature with a time step of 0.5 fs. The replica-exchange method[35] was adopted to generate the ensembles at a set of temperatures ranging from 140 to 800 K with 16 replicas at an exchange interval of 1 ps. After 20 ns of equilibration, 10 independent runs of 400 ns molecular dynamics simulation were performed to obtain the variance of the free energy difference. The potential energy was recorded with 1 ps interval and used for the free energy analysis. Thus, the number of samples is $N_0 = N_1 = 400\,000$. We used MBAR (Ref. 26) to calculate the accurate free energy difference. All the methods presented in Sec. II were examined except for LTI, which is a biased estimator. The error was calculated by

$$\sqrt{\frac{1}{10}\sum_{i=1}^{10}(\Delta f(i) - \Delta f_{\mathrm{MBAR}})^2}, \tag{44}$$

where $\Delta f(i)$ is the free energy estimation from the ith run by the examined method, and Δf_{MBAR} is the free energy difference estimated from MBAR. To obtain $\Delta f(i)$, the binning strategies were set similarly to those in Sec. III A. The error of Δf_{MBAR} is several times smaller than the others, and Δf_{MBAR} is used as the reference for assessing the performance of the methods in Sec. II. It should be noted for MBAR that a multiple set of intermediate systems are explicitly handled to connect the two systems of which the free energy difference is the target of computation. In the methods in Sec. II, on the other hand, inputs are needed only from the two systems of interest. MBAR cannot thus be treated on an equal footing, and is used only as the reference to evaluate the statistical error by Eq. (44).

The results for ATP and ADP are shown in Fig. 3. The number of bins B is $B = 100$ in Fig. 3, and the effect of the B value is described in Appendix C. The EROD, HMOD, and BAR methods exhibit excellent convergences for both ions. The scheme of error minimization works well in practice for distribution-function based method. The OD method shows moderate error, however, the error does not drop even when the temperature of system 0 increases and the difference between systems 0 and 1 diminishes. Such a behavior is also observed in Fig. 1; even low-counts bins have the equal weights in OD and amplify the error. The error in Yokogawa's method is the largest in Fig. 3 and is rather irregular. The cause of this is considered to be outliers in the sampled potential energy, which are used to estimate the smoothed histogram $\rho^\star$

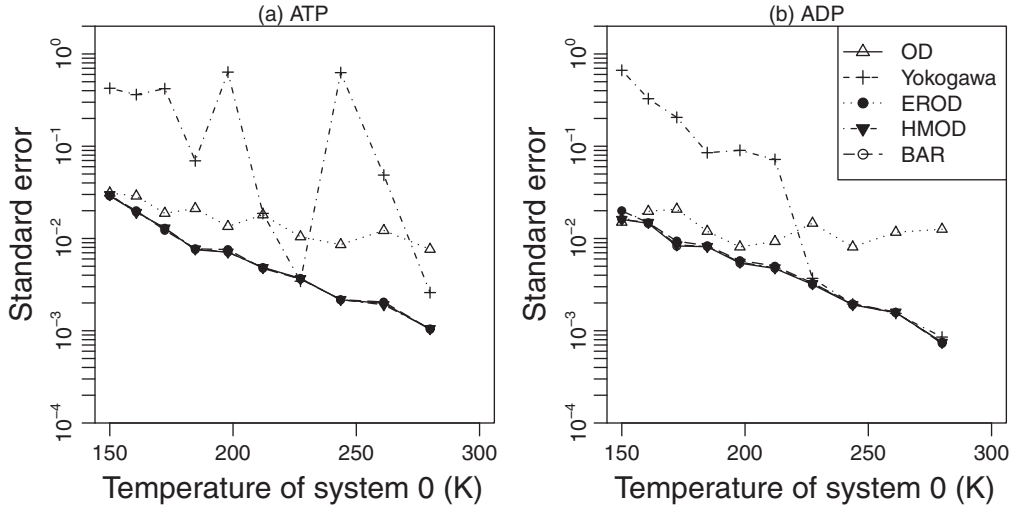

FIG. 3. The standard error of the free energy difference between system 0 (variable temperature, abscissa) and system 1 (temperature set to 300 K) of (a) ATP and (b) ADP. The data symbols are common in (a) and (b). Note the log-scale of the ordinate. The error was computed as the standard deviation from the numerically exact value, which was obtained with the MBAR method (see main text for details). The lines are drawn for eye guide. The EROD, HMOD, and BAR values are barely distinguishable among one another for both species.

in Eq. (22). In the extrapolation scheme given by Eq. (22), the presence of exponential coefficient may give rise to significant numerical variation in $\rho^\star$ and thus in Δf when the extrapolation is conducted over large change of ε. When the system of interest is ATP or ADP, the ε range is significantly larger than in the case of alanine dipeptide treated in Ref. 15. Yokogawa's method thus deteriorates when the domain of extrapolation is extended, and may improve the accuracy when combined with different extrapolation strategy.

In the developments of Sec. II, the number of samples N refers actually to the number of statistically uncorrelated samples. Configurations sampled from a molecular simulation are correlated, on the other hand, if the sampling interval is too short. A short interval reduces the effective number of independent samples, and may increase the statistical error when the total number of samples is fixed. To illustrate the effect of sampling interval, we examined the error estimated by Eq. (44) for the free energy difference between 196 and 300 K of the ATP system. The sampling interval was varied at 0.1, 0.2, 0.5, 1.0, and 2.0 ps with the number of samples fixed at 200 000 and the number of bins fixed at $B = 100$. The free energy errors from the EROD, HMOD, and BAR schemes were then 1×10^{-2}, 9×10^{-3}, 8×10^{-3}, 1×10^{-2}, and 8×10^{-3}, respectively, at the intervals examined. The error depends weakly on the sampling interval, and there is not much gain from a large interval in the present case (note that the computation time is determined by the sampling interval multiplied by the number of samples, which is fixed at 200 000 here).

The free energy schemes in Secs. II C and II D do not show optimal performances when the correlated samples are used without any correction of the effect of statistical interference and N_0 and N_1 in Eqs. (28), (29), (32), (37), and (38) refer to the numbers of snapshots read from simulations. An unbiased estimator such as EROD, HMOD, and BAR provides a correct value at the limit of large sample size, however, even when the sampling interval is too short and the correlation persists. The only, possible effect of the uncorrected use of N_0 and N_1 is a deteriorated rate of convergence.

IV. CONCLUSION

In the present work, we systematically analyzed several schemes to calculate the free energy difference on the basis of distribution function. The statistical error due to finite sampling was estimated, and improved schemes of free energy computation were derived through error minimization. We then tested the free energy methods against harmonic oscillator and biomolecules of ~40 atoms. The newly derived methods exhibit clear improvement over the previously established, overlapping distribution method and Yokogawa's method. Their performance is further seen to be comparable to Bennett's acceptance ratio method. Since the distribution function has clear, physical meanings, its utility in accurate and robust free energy calculation is desirable from both the analytical and computational viewpoints.

The notion of error minimization is a mature yet useful one for numerically determining physical quantities accurately. The present work applies this notion to distribution-function based approach to free energy computation. Within the context of error minimization, in fact, the "error" is set to the variance of estimated free energy difference. As far as the measure of error is defined as the variance, no more improvement beyond the variance-minimized scheme is possible within that definition of error measure. Further improvement and/or modification would then be possible by changing the measure of error, for example, by incorporating some information of higher-order moments to strongly suppress low-counts data.

ACKNOWLEDGMENTS

This work is supported by the Grants-in-Aid for Scientific Research (Grant Nos. 21300111 and 23651202) from the Japan Society for the Promotion of Science, and by the Grant-in-Aid for Scientific Research on Priority Areas (Grant No. 20038034), the Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 20118002), and the Next-Generation Supercomputing Project, Nanoscience Program from the Ministry of Education, Culture, Sports, Science, and Technology. Numerical calculations were conducted partly using computers at Center for Computational Sciences in University of Tsukuba, at Institute for Solid State Physics in the University of Tokyo, and at Research Center for Computational Science in Institute of Molecular Science.

APPENDIX A: DERIVATION OF EQ. (42)

The ith bin is defined as $e_{i-1} \le \varepsilon \le e_i$. The mid-point ε_i and the bin width W are then expressed, respectively, as

$$\varepsilon_i = \frac{e_{i-1} + e_i}{2}, \tag{A1}$$

$$W = e_i - e_{i-1}, \tag{A2}$$

and the ith bin is rewritten as

$$\varepsilon_i - W/2 \le \varepsilon \le \varepsilon_i + W/2. \tag{A3}$$

The systematic error term Δf_i^{rest} of the free energy difference contributed from the ith bin is given by taking Eqs. (10) and (15) into Eq. (13) as

$$\begin{aligned}\Delta f_i^{\text{rest}} &= \left(\int_{e_{i-1}}^{e_i} d\varepsilon\, w(\varepsilon)\,(\log \rho_1(\varepsilon) - \log \rho_0(\varepsilon) + \varepsilon)\right) \\ &\quad - w_i\,(\log \rho_{1i} - \log \rho_{0i} + \varepsilon_i) \\ &= w_i \Delta f - w_i \left[\log\left(\int_{\varepsilon_i - W/2}^{\varepsilon_i + W/2} \rho_1(\varepsilon)\, d\varepsilon\right)\right. \\ &\quad \left. - \log\left(\int_{\varepsilon_i - W/2}^{\varepsilon_i + W/2} \rho_0(\varepsilon)\, d\varepsilon\right) + \varepsilon_i\right].\end{aligned} \tag{A4}$$

We treat Eq. (A4) by approximating $\rho_\lambda(\varepsilon)$ in quadratic form as

$$\rho_\lambda(\varepsilon) \simeq \overline{\rho}_{\lambda i} + a_{\lambda i}(\varepsilon - \varepsilon_i) + \frac{b_{\lambda i}}{2}(\varepsilon - \varepsilon_i)^2, \tag{A5}$$

$$\overline{\rho}_{\lambda i} = \rho_\lambda(\varepsilon_i), \tag{A6}$$

$$a_{\lambda i} = \left(\frac{\partial \rho_\lambda (\varepsilon)}{\partial \varepsilon} \right)_{\varepsilon = \varepsilon_i}, \tag{A7}$$

$$b_{\lambda i} = \left(\frac{\partial^2 \rho_\lambda (\varepsilon)}{\partial \varepsilon^2} \right)_{\varepsilon = \varepsilon_i}, \tag{A8}$$

for $\varepsilon_i - W/2 \le \varepsilon \le \varepsilon_i + W/2$. Since Eq. (9) is valid for any ε, $\overline{\rho}_{\lambda i}$ and ε_i satisfy

$$\log \overline{\rho}_{1i} - \log \overline{\rho}_{0i} + \varepsilon_i = \Delta f, \tag{A9}$$

since $\overline{\rho}_{0i}$ and $\overline{\rho}_{1i}$ are the values at ε_i. With Eqs. (A5) and (A9), Eq. (A4) reduces to

$$\Delta f_i^{\text{rest}} \simeq w_i \left[-\log \left(1 + \frac{b_{1i} W^2}{24 \overline{\rho}_{1i}} \right) + \log \left(1 + \frac{b_{0i} W^2}{24 \overline{\rho}_{0i}} \right) \right]. \tag{A10}$$

The factor $a_{\lambda i}$ vanishes in Eq. (A10) and Δf_i^{rest} is of $O\left(W^2\right)$ since ε_i is set to the mid-point of the ith bin. If the representative value of the ith energy bin is not taken to be the mid-point value ε_i, Δf_i^{rest} would be of $O(W)$. Thus, the mid-point convention for ε_i adopted in the present work reduces the error due to the bin width.

When we further assume $b_{\lambda i} W^2 \ll \overline{\rho}_{\lambda i}$, the dominant term of Δf_i^{rest} is

$$\Delta f_i^{\text{rest}} \simeq -\frac{w_i W^2}{24} \left(\frac{b_{1i}}{\overline{\rho}_{1i}} - \frac{b_{0i}}{\overline{\rho}_{0i}} \right) \tag{A11}$$

$$= -\frac{w_i W^2}{24} \left[\left(\frac{\partial \log \rho_1 (\varepsilon)}{\partial \varepsilon} \right)^2_{\varepsilon = \varepsilon_i} - \left(\frac{\partial \log \rho_0 (\varepsilon)}{\partial \varepsilon} \right)^2_{\varepsilon = \varepsilon_i} \right]. \tag{A12}$$

From Eq. (A11) to Eq. (A12), we used the relations of

$$\frac{\partial^2 \log \rho_1 (\varepsilon)}{\partial \varepsilon^2} - \frac{\partial^2 \log \rho_0 (\varepsilon)}{\partial \varepsilon^2} = \frac{\partial^2 (\log \rho_1 (\varepsilon) - \log \rho_0 (\varepsilon))}{\partial \varepsilon^2} = 0, \tag{A13}$$

$$\frac{1}{\rho_\lambda (\varepsilon)} \frac{\partial^2 \rho_\lambda (\varepsilon)}{\partial \varepsilon^2} = \frac{\partial^2 \log \rho_\lambda (\varepsilon)}{\partial \varepsilon^2} + \left(\frac{\partial \log \rho_\lambda (\varepsilon)}{\partial \varepsilon} \right)^2. \tag{A14}$$

Equation (A13) holds by virtue of Eq. (9). Δf_i^{rest} treated in the present Appendix is a systematic error, not a statistical error. The total systematic error Δf^{rest} is thus given by the (linear) sum of Δf_i^{rest} over i, as shown in Eq. (43).

APPENDIX B: DERIVATION OF EQS. (28) AND (29)

In this Appendix, we derive Eqs. (28) and (29) on the basis of the statistical error of observed count in the sampling of the distribution functions. Using Bayesian statistics, we derive the *mean* and *variance* of $\log \widetilde{\rho_{\lambda i}}$ for the ith bin, namely $\overline{\log \rho_{\lambda i}}$ and $\sigma^2 (\log \rho_{\lambda i})$, respectively.

Let $\rho^\circ_{\lambda i}$ be the true probability of observing the energy difference ε to fall into the ith bin ($1 \le i \le B$). The true probability is an idealistic one and can be obtained only from infinite number of sampling. In practice, the probability is estimated from finite sampling expressed as a set of the count $n_{\lambda i}$ of the ε values falling into the ith bin. We denote the finite-sample estimate as $\rho_{\lambda i}$. $\rho_{\lambda i}$ in the present Appendix is different from $\rho_{\lambda i}$ of Eq. (17), which is obtained from infinite sampling. This expression is adopted here only for the sake of notational simplicity of the following equations. $\rho^\circ_{\lambda i}$ and $\rho_{\lambda i}$ satisfy the normalization conditions given by

$$\sum_{i=1}^{B} \rho^\circ_{\lambda i} = 1, \tag{B1}$$

$$\sum_{i=1}^{B} \rho_{\lambda i} = 1. \tag{B2}$$

For notational convenience, we define the vectors of $\boldsymbol{\rho}^\circ_\lambda = (\rho^\circ_{\lambda 1}, \rho^\circ_{\lambda 2}, \ldots, \rho^\circ_{\lambda B})$, $\boldsymbol{\rho}_\lambda = (\rho_{\lambda 1}, \rho_{\lambda 2}, \ldots, \rho_{\lambda B})$, and $\boldsymbol{n}_\lambda = (n_{\lambda 1}, n_{\lambda 2}, \ldots, n_{\lambda B})$. Since different simulations performed over finite period provide different $\boldsymbol{\rho}_\lambda$, $\boldsymbol{\rho}_\lambda$ is stochastic and its "probability distribution" is treated in the following.

Since a finite-length sampling inevitably contains statistical errors, the true probability $\boldsymbol{\rho}^\circ_\lambda$ is unable to obtain directly from simulation. We resort to Bayesian statistics and consider the probability of $\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda$ when the observed count in the ith bin is $n_{\lambda i}$. According to Bayes rule,

$$\begin{aligned} &\Pr\left(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda \mid \text{bin counts} = \boldsymbol{n}_\lambda\right) \\ &\quad \propto \Pr\left(\text{bin counts} = \boldsymbol{n}_\lambda \mid \boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda\right) \Pr\left(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda\right) \\ &\quad \propto \rho_{\lambda 1}^{n_{\lambda 1}} \rho_{\lambda 2}^{n_{\lambda 2}} \ldots \rho_{\lambda B}^{n_{\lambda B}} \Pr\left(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda\right), \end{aligned} \tag{B3}$$

where $\Pr(X)$ is the probability that X occurs and $\Pr(X \mid Y)$ is the conditional probability that X occurs when Y occurs. The second factor $\Pr(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda)$ is called prior distribution, since it corresponds to our prior knowledge on the system; if some $\rho^\circ_{\lambda i}$ value is more plausible than others, this prior is set so that $\Pr(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda)$ is high around the corresponding $\boldsymbol{\rho}^\circ_\lambda$. In the present work, we assume that $\Pr(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda)$ is constant over $0 \le \rho_{\lambda i} \le 1$ for each i, meaning that there are no biases in the value of $\rho_{\lambda i}$. The proportionality factor of Eq. (B3) can be determined from

$$\int \mathrm{d}\boldsymbol{\rho}_\lambda \Pr\left(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda \mid \text{bin counts} = \boldsymbol{n}_\lambda\right) = 1. \tag{B4}$$

Equation (B3) is then rewritten as

$$\Pr\left(\boldsymbol{\rho}^\circ_\lambda = \boldsymbol{\rho}_\lambda \mid \text{bin counts} = \boldsymbol{n}_\lambda\right) = \text{Mult}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{n}_\lambda\right), \tag{B5}$$

where $\mathrm{Mult}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{n}_\lambda\right)$ is the multinomial distribution expressed as

$$\mathrm{Mult}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{n}_\lambda\right) = \frac{N_\lambda!}{n_{\lambda 1}! n_{\lambda 2}! \ldots n_{\lambda B}!} \prod_{i=1}^{B} \rho_{\lambda i}^{n_{\lambda i}}, \tag{B6}$$

$$N_\lambda = \sum_{i=1}^{B} n_{\lambda i}. \tag{B7}$$

Using Eq. (B6), the mean value $\overline{\log \rho_{\lambda i}}$ and the variance $\sigma^2(\log \rho_{\lambda i})$ are estimated, respectively, as

$$\overline{\log \rho_{\lambda i}} = \int \mathrm{d}\boldsymbol{\rho}_\lambda \left(\log \rho_{\lambda i}\right) \mathrm{Mult}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{n}_\lambda\right), \tag{B8}$$

$$\sigma^2\left(\log \rho_{\lambda i}\right) = \int \mathrm{d}\boldsymbol{\rho}_\lambda \left(\log \rho_{\lambda i} - \overline{\log \rho_{\lambda i}}\right)^2 \mathrm{Mult}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{n}_\lambda\right). \tag{B9}$$

These integrations can be calculated as the moments of exponential family distribution,[36] yielding Eqs. (28) and (29).

In the Bayesian treatment, it is also a common practice to use the Dirichlet distribution $\Pr(\boldsymbol{\rho}_\lambda^\circ = \boldsymbol{\rho}_\lambda) = \mathrm{Dir}(\boldsymbol{\rho}_\lambda \mid \boldsymbol{\alpha}_\lambda)$ as the prior. This is known as the conjugate prior of the multinomial distribution[37] and is given by

$$\mathrm{Dir}\left(\boldsymbol{\rho}_\lambda \mid \boldsymbol{\alpha}_\lambda\right) = \frac{\Gamma\left(\sum_{i=1}^{B} \alpha_{\lambda i}\right)}{\prod_{j=1}^{B} \Gamma\left(\alpha_{\lambda j}\right)} \prod_{k=1}^{B} \rho_{\lambda k}^{\alpha_{\lambda k}-1}, \tag{B10}$$

where $\Gamma(\cdot)$ is the gamma function and $\boldsymbol{\alpha}_\lambda = (\alpha_{\lambda 1}, \alpha_{\lambda 2}, \ldots, \alpha_{\lambda B})$ is the set of parameters subject to the Bayesian update. Setting $\boldsymbol{\alpha}_\lambda = 1$ is equivalent to choosing the uniform prior (constant $\Pr(\boldsymbol{\rho}_\lambda^\circ = \boldsymbol{\rho}_\lambda)$), which is a noninformative prior for the multinomial distribution[37] and was used to derive Eqs. (28) and (29). With the Dirichlet prior, the equations corresponding to Eqs. (28) and (29) are derived as

$$\overline{\log \rho_{\lambda i}} = \psi\left(n_{\lambda i} + \alpha_{\lambda i}\right) - \psi\left(N_\lambda + \sum_{j=1}^{B} \alpha_{\lambda j}\right), \tag{B11}$$

$$\sigma^2\left(\log \rho_{\lambda i}\right) = \psi'\left(n_{\lambda i} + \alpha_{\lambda i}\right) - \psi'\left(N_\lambda + \sum_{j=1}^{B} \alpha_{\lambda j}\right), \tag{B12}$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions, respectively, expressed as

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x), \tag{B13}$$

$$\psi'(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \Gamma(x). \tag{B14}$$

Since the prior distribution cannot be determined *a priori*, we used the uniform prior in the present work for simplicity. In fact, we tested the Dirichlet prior with $\alpha_{\lambda i} = 1 + 1/B$, and found that the errors of the free energy difference do not deviate appreciably (within ± 1 %) from those from the uniform prior.

APPENDIX C: FREE ENERGY DIFFERENCES OF ATP AND ADP WITH VARIOUS NUMBERS OF BINS

The free energy differences of ATP and ADP ions between 300 K and a variety of temperatures were investigated in Sec. III B using the number of bins $B = 100$. In this Appendix, we show the B dependence of the ATP and ADP results. The system setup and parameters are identical to those in Sec. III B except for the number of bins. Figure 4 shows the computed results of the standard errors at $B = 500$. The error of HMOD is almost the same as the case of $B = 100$ for both ATP and ADP. The error of EROD is larger than in Fig. 3

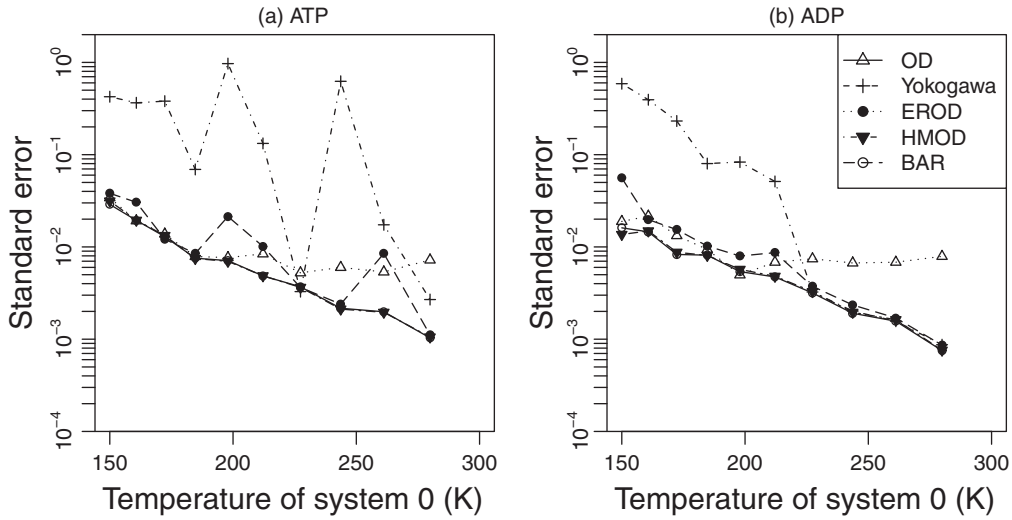

FIG. 4. The standard error of the free energy difference between system 0 (variable temperature, abscissa) and system 1 (temperature set to 300 K) of (a) ATP and (b) ADP. The number of bins is $B = 500$, and the other setups are identical to those in Fig. 3. The data symbols are common in (a) and (b). The lines are drawn for eye guide. The HMOD and BAR are barely distinguishable from each other for both species.

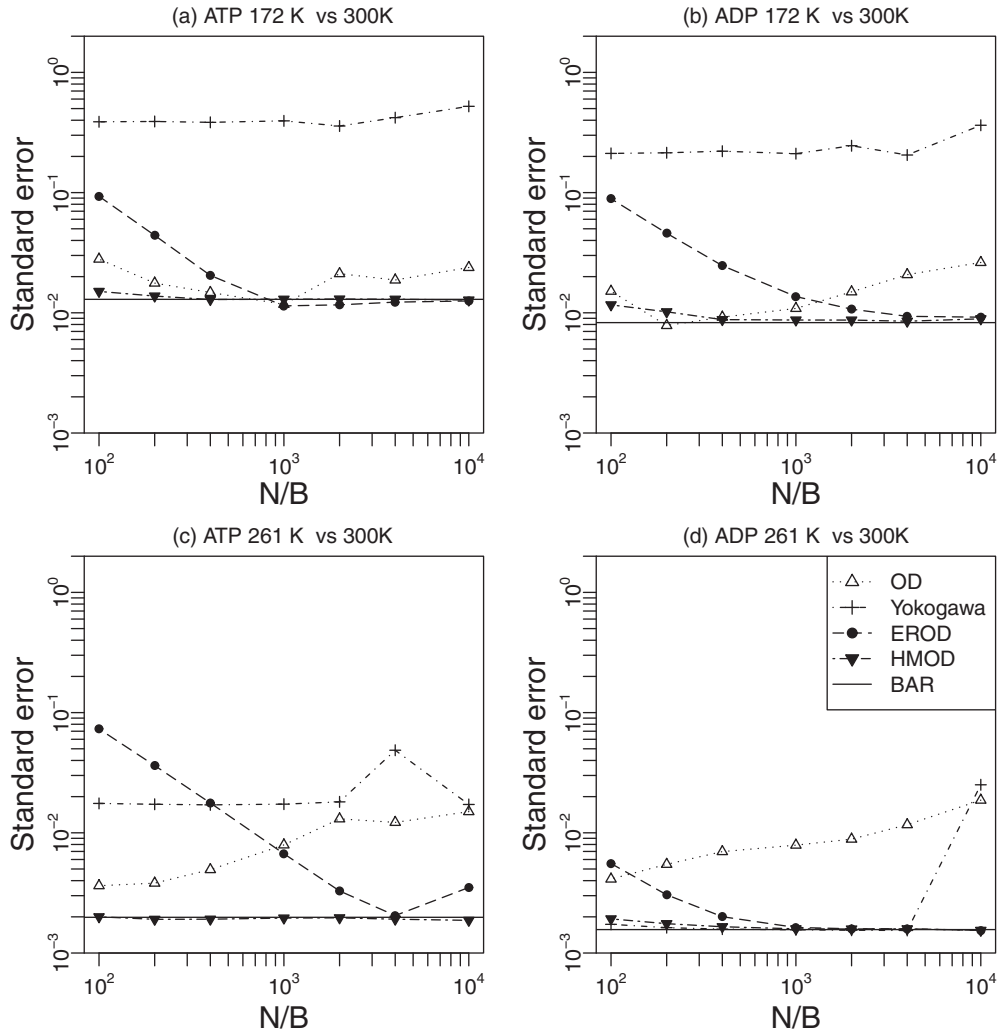

FIG. 5. The dependence of the standard error on the number of bins B for the free energy differences of ATP and ADP between system 1 at 300 K and system 0 at two temperatures of 172 and 261 K. The abscissa is expressed in the form of N/B with the number of samples $N = 400\,000$. The data symbols are common in (a) to (d). Since BAR is unaffected by the number of bins, its result is represented as a horizontal line without symbols. The lines for the other methods are drawn for eye guide. Note the log scale of both axes.

at all the temperatures, and the error of OD is slightly smaller than in Fig. 3. These trends coincide with the results in Fig. 2, where the OD error is susceptible to the number of bins and the EROD gives larger error when B is larger.

The dependence on the number of bins B is presented in Fig. 5 at two temperatures of system 0. The plot is shown as a function of N/B, where N is the number of snapshots in systems 0 and 1 and $N = 400\,000$. The overall trends are similar to those in Fig. 2. The HMOD shows a good agreement with BAR at both 172 K and 261 K, whereas the EROD's estimation becomes worse when N/B is small. The OD method shows good agreement with BAR over limited N/B at 172 K. The Yokogawa's results are irregular also in Fig. 5, and agree with BAR only in (d).

[1]J. D. Weeks, D. Chandler, and H. C. Andersen, J. Chem. Phys. **54**, 5237 (1971).
[2]F. Hirata and P. J. Rossky, Chem. Phys. Lett. **83**, 329 (1981).
[3]J. Perkyns and B. M. Pettitt, J. Chem. Phys. **97**, 7656 (1992); **100**, 8556 (1994) (erratum).
[4]M. Ikeguchi and J. Doi, J. Chem. Phys. **103**, 5011 (1995).
[5]D. Beglov and B. Roux, J. Phys. Chem. B **101**, 7821 (1997).
[6]A. Kovalenko and F. Hirata, J. Chem. Phys. **112**, 10391 (2000).
[7]N. Matubayasi and M. Nakahara, J. Chem. Phys. **113**, 6070 (2000).
[8]N. Matubayasi and M. Nakahara, J. Chem. Phys. **117**, 3605 (2002); **118**, 2446 (2003) (erratum).
[9]N. Matubayasi and M. Nakahara, J. Chem. Phys. **119**, 9686 (2003).
[10]G. N. Chuev, M. V. Fedorov, and J. Crain, Chem. Phys. Lett. **448**, 198 (2007).
[11]D. S. Palmer, A. I. Frolov, E. L. Ratkova, and M. V. Fedorov, J. Phys. Condens. Matter **22**, 492101 (2010).
[12]Y. Karino, M. V. Fedorov, and N. Matubayasi, Chem. Phys. Lett. **496**, 351 (2010).
[13]A. I. Frolov, E. L. Ratkova, D. S. Palmer, and M. V. Fedorov, J. Phys. Chem. B **115**, 6011 (2011).
[14]K. S. Shing and K. E. Gubbins, Mol. Phys. **46**, 1109 (1982).
[15]D. Yokogawa and T. Ikegami, J. Chem. Phys. **131**, 221101 (2009).
[16]D. Frenkel and B. Smit, *Understanding Molecular Simulation, 2nd ed.: From Algorithms to Applications*, Computational Science Series, Vol. 1 (Academic, New York, 2001).
[17]N. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001).
[18]N. Lu, J. K. Singh, and D. A. Kofke, J. Chem. Phys. **118**, 2977 (2003).
[19]M. R. Shirts and V. S. Pande, J. Chem. Phys. **122**, 144107 (2005).
[20]F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman, J. Chem. Phys. **125**, 184114 (2006).

[21]J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[22]R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[23]S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[24]M. Souaille and B. Roux, Comput. Phys. Commun. **135**, 40 (2001).
[25]C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[26]M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[27]H. Müller-Krumbhaar and K. Binder, J. Stat. Phys. **8**, 1 (1973).
[28]J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. **3**, 26 (2007).
[29]M. Mezei and D. Beveridge, Ann. N. Y. Acad. Sci. **482**, 1 (1986).
[30]Equation (11) in Ref. 15 has an additional term to Eq. (8) represented as $-\int d\varepsilon(\rho_1(\varepsilon) - \rho_0(\varepsilon))$. This term vanishes because the energy distribution function is normalized as $\int d\varepsilon \rho_\lambda(\varepsilon) = 1$. The term seems incorporated in Ref. 15 for the compatibility to the hypernetted-chain form. In the present work, we removed the term for simplicity.
[31]M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[32]G. E.P. Box and M. E. Muller, Ann. Math. Stat. **29**, 610 (1958).
[33]K. L. Meagher, L. T. Redman, and H. A. Carlson, J. Comput. Chem. **24**, 1016 (2003).
[34]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[35]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[36]L. D. Brown, *Fundamentals of Statistical Exponential Families* (Institute of Fundamentals Statistic, Hayward, CA, 1987).
[37]J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, New York, 1994).