Using bijective maps to improve free-energy estimates

A. M. Hahn and H. Then
Institut für Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany
(Received 12 September 2008; published 13 January 2009)

We derive a fluctuation theorem for generalized work distributions, related to bijective mappings of the phase spaces of two physical systems, and use it to derive a two-sided constraint maximum likelihood estimator of their free-energy difference which uses samples from the equilibrium configurations of both systems. As an application, we evaluate the chemical potential of a dense Lennard-Jones fluid and study the construction and performance of suitable maps.

DOI: 10.1103/PhysRevE.79.011113 PACS number(s): 05.40.−a, 05.70.Ln

I. INTRODUCTION

Extracting free-energy differences from a suitable set of computer simulation data is an active field of research and of interest, e.g., for drug design [1] or nonperturbative quantum chromodynamics [2]. Concerning estimators for the free-energy difference, an extensive literature can be found. Probably the most elementary estimator is the traditional free-energy perturbation [3], which is briefly introduced in the following. Assume we have given two systems, arbitrarily labeled as system 0 and system 1, that are characterized by Hamiltonians $H_0(x)$ and $H_1(x)$, respectively, depending on the point x in phase space. Further, let $\rho_i(x)$ denote the thermal equilibrium phase space density of system i,

$$\rho_i(x) = \frac{e^{-\beta H_i(x)}}{Z_i}, \quad i = 0,1, \tag{1}$$

where $Z_i = \int e^{-\beta H_i(x)} dx$ denotes the partition function and $\beta = \frac{1}{kT}$ the inverse temperature. We are interested in the Helmholtz free-energy difference ΔF of the systems, defined as $\Delta F = -\frac{1}{\beta} \ln \frac{Z_1}{Z_0}$. Traditional free-energy perturbation [3] originates from the equality

$$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta[\Delta H(x) - \Delta F]}, \tag{2}$$

with $\Delta H(x) := H_1(x) - H_0(x)$. The latter quantity may be interpreted as the work performed during an infinitely fast switching process transforming system 0 to system 1, with initial configuration x [4]. A direct consequence of Eq. (2) is the perturbation identity

$$e^{-\beta \Delta F} = \int e^{-\beta \Delta H(x)} \rho_0(x) dx, \tag{3}$$

which is frequently used to obtain an estimate of ΔF in drawing a sample $\{x_1, \ldots, x_N\}$ from $\rho_0(x)$ (e.g., by Monte Carlo simulations) and evaluating the estimator

$$\widehat{\Delta F}_0^{\text{trad}} = -\frac{1}{\beta} \ln \overline{e^{-\beta \Delta H(x)}}. \tag{4}$$

The overbar denotes a sample average [i.e., $\overline{f(x)} = \frac{1}{N} \sum_{k=1}^{N} f(x_k)$ where f stands for an arbitrary function]. As can be seen by comparison with Eq. (2), the integrand appearing in Eq. (3) is proportional to ρ_1, and thus the main contributions to an accurate estimate of ΔF with Eq. (4) come from realizations x (drawn from ρ_0) that are typical for the density ρ_1. This means that the performance of such an estimate depends strongly on the degree of overlap of ρ_0 with ρ_1. If the overlap is small, the traditional free-energy perturbation is plagued with a slow convergence and a large bias. This can be overcome by using methods that bridge the gap between the densities ρ_0 and ρ_1, for instance the thermodynamic integration. Since thermodynamic integration samples a sequence of many equilibrium distributions, it soon becomes computationally expensive. Another method is umbrella sampling [5] which distorts the original distribution in order to sample regions that are important for the average. Because of the distortion, the latter method is in general restricted to answer only one given question, e.g., the value of the free-energy difference, but fails to give further answers. This is of particular concern, if in addition the values of some other thermodynamic variables are sought, for example, pressure or internal energy. There are dynamical methods [6] that make use of the Jarzynski work theorem [4]. They allow to base the estimator on work values of fast, finite time, nonequilibrium processes connecting system 0 with system 1. However, the dynamic simulation of the trajectories is typically very expensive.

Six years ago, the targeted free-energy perturbation method [7] was introduced; a promising method which is based on mapping equilibrium distributions close to each other in order to overcome the problem of insufficient overlap, without the need to draw from biased distributions. However, this method is rarely used in the literature [8,9]. An obstacle might be that there is no general description of how to construct a suitable map. A related idea [10] was applied in [11]. A recent improvement is the escorted free-energy simulation [12] which is a dynamical generalization of the targeted free-energy perturbation.

Any free-energy difference refers to two equilibrium ensembles. The above mentioned methods draw only from one of the two ensembles and propagate the system in direction of the other. Insofar, they are "one-sided" methods. However, it is of advantage to draw from both equilibrium distributions and combine the obtained "two-sided" information. Optimizing the elementary two-sided estimator for free-energy differences results in the acceptance ratio method [13–15]. The next step of improvement is to implement a two-sided targeted free-energy method that optimally employs the information of drawings from both equilibrium distributions. Our aim is to combine the advantages of the acceptance ratio

method with the advantages of the targeted free-energy perturbation.

The central result of this paper is a fluctuation theorem for the distributions of generalized work values that is derived and presented in Sec. III. From this fluctuation theorem, the desired optimal two-sided targeted free-energy estimator follows in Sec. IV. In Sec. V, appropriate measures are introduced which relate the overlap of $\tilde{\rho}_0$ with ρ_1 to the mean square errors of the one- and two-sided free-energy estimators. In Sec. VI, a convergence criterion for the two-sided estimator is proposed. From Sec. VII on, numerics plays an important part. In particular, Sec. VII A deals with explicit numerical applications. Based on the two-sided targeted free-energy estimator, in Sec. VII B, an estimator for the chemical potential of a high-density homogeneous fluid is established and applied to a dense Lennard-Jones fluid. Finally, the construction and performance of suitable maps is studied. In order to set some notation straight, we start by recalling the targeted free-energy perturbation method.

II. TARGETED FREE-ENERGY PERTURBATION

Let Γ_0 and Γ_1 denote the phase spaces of the systems 0 and 1, respectively. We require that Γ_i contains only those points x for which $\rho_i(x)$ is nonzero.

Mapping the phase space points of system 0, $x \to \phi(x)$, such that the mapped phase space $\tilde{\Gamma}_0 = \phi(\Gamma_0)$ coincides with the phase space Γ_1 and such that the mapped distribution $\tilde{\rho}_0$ overlaps better with the canonical distribution ρ_1 results in the *targeted free-energy perturbation* [7] where the samples are drawn *effectively* from $\tilde{\rho}_0$ instead.

Following the idea of Jarzynski [7], we introduce such a phase space map. If Γ_0 and Γ_1 are diffeomorph, there exists a bijective and differentiable map $\mathcal{M}$ from Γ_0 to Γ_1,

$$\mathcal{M}:\Gamma_0 \to \Gamma_1, \quad \mathcal{M}:x \to \phi(x), \tag{5}$$

where the absolute value of the Jacobian is

$$K(x) = \left| \left| \frac{\partial \phi}{\partial x} \right| \right|. \tag{6}$$

The inverse map reads

$$\mathcal{M}^{-1}:\Gamma_1 \to \Gamma_0, \quad \mathcal{M}^{-1}:y \to \phi^{-1}(y). \tag{7}$$

According to $\mathcal{M}$, the phase space density ρ_0 is mapped to the density $\tilde{\rho}_0$,

$$\tilde{\rho}_0(y) = \int_{\Gamma_0} \delta[y - \phi(x)]\rho_0(x)dx, \tag{8}$$

which can be written as

$$\tilde{\rho}_0(\phi(x)) = \frac{\rho_0(x)}{K(x)} \tag{9}$$

or

$$\int_{\phi(\Gamma)} \tilde{\rho}_0(y)dy = \int_{\Gamma} \rho_0(x)dx, \quad \forall \ \Gamma \subset \Gamma_0. \tag{10}$$

In analogy to Eq. (2), the targeted free-energy perturbation is based on the identity

$$\frac{\tilde{\rho}_0(\phi(x))}{\rho_1(\phi(x))} = e^{\beta[\widetilde{\Delta H}(x) - \Delta F]} \quad \forall \ x \in \Gamma_0, \tag{11}$$

which follows from the densities (1) and (9) with $\widetilde{\Delta H}$ being defined by

$$\widetilde{\Delta H}(x) := H_1(\phi(x)) - H_0(x) - \frac{1}{\beta} \ln K(x). \tag{12}$$

Multiplying Eq. (11) by $e^{-\beta\widetilde{\Delta H}(x)}\rho_1(\phi(x))K(x)$ and integrating over Γ_0 yields the targeted free-energy perturbation formula,

$$e^{-\beta\Delta F} = \int_{\Gamma_0} e^{-\beta\widetilde{\Delta H}(x)}\rho_0(x)dx. \tag{13}$$

An alternative derivation is given in [7]. The traditional free-energy perturbation formula (3) can be viewed as a special case of Eq. (13). The latter reduces to the former if $\mathcal{M}$ is chosen to be the identity map, $\phi(x)=x$. [This requires that $\Gamma_1=\Gamma_0$ holds.]

Now an obvious estimator for ΔF, given a sample $\{x_k\}$ drawn from $\rho_0(x)$, is

$$\widehat{\Delta F}_0 = -\frac{1}{\beta} \ln \overline{e^{-\beta\widetilde{\Delta H}(x)}}, \tag{14}$$

which we refer to as the targeted *forward* estimator for ΔF. The convergence problem of the traditional forward estimator, Eq. (4), in the case of insufficient overlap of ρ_0 with ρ_1 is overcome in the targeted approach by choosing a suitable map $\mathcal{M}$ for which the image $\tilde{\rho}_0$ of ρ_0 overlaps better with ρ_1. Indeed, suppose for the moment that the map is chosen to be ideal, namely such that $\tilde{\rho}_0(x)$ coincides with $\rho_1(x)$. Then, as a consequence of Eq. (11), the quantity $\widetilde{\Delta H}(x)$ is constant and equals ΔF, and the convergence of the targeted estimator (14) is immediate. Although the construction of such an ideal map is impossible in general, the goal of approaching an ideal map guides the design of suitably good maps.

To complement the *one-sided* targeted estimator, a second perturbation formula in the "reverse" direction is derived from Eq. (11),

$$e^{+\beta\Delta F} = \int_{\Gamma_1} e^{+\beta\widetilde{\Delta H}(\phi^{-1}(y))}\rho_1(y)dy, \tag{15}$$

leading to the definition of the targeted *reverse* estimator $\widehat{\Delta F}_1$ of ΔF,

$$\widehat{\Delta F}_1 = +\frac{1}{\beta} \ln \overline{e^{+\beta\widetilde{\Delta H}(\phi^{-1}(y))}}. \tag{16}$$

The index 1 indicates that the set $\{y_k\}$ is drawn from ρ_1. Using the identity map $\phi(x)=x$ in Eq. (16) gives the traditional reverse estimator, which is valid if $\Gamma_0=\Gamma_1$ holds.

It will prove to be beneficial to switch from phase space densities to one-dimensional densities which describe the value distributions of $\widetilde{\Delta H}(x)$ and $\widetilde{\Delta H}(\phi^{-1}(y))$, cf. Eqs. (13) and (15). This is done next and results in the fluctuation theorem for generalized work distributions.

III. FLUCTUATION THEOREM FOR GENERALIZED WORK DISTRIBUTIONS

We call $\widetilde{\Delta H}(x)$, $x \in \Gamma_0$, function of the *generalized work* in *forward* direction and $\widetilde{\Delta H}(\phi^{-1}(y))$, $y \in \Gamma_1$, function of the *generalized work* in *reverse* direction, having in mind that these quantities are the functions of the actual physical work for special choices of the map $\mathcal{M}$ [16].

The probability density $p(W|0;\mathcal{M})$ for the outcome of a specific value W of the generalized work in forward direction subject to the map $\mathcal{M}$ when sampled from ρ_0 is given by

$$p(W|0;\mathcal{M}) = \int_{\Gamma_0} \delta[W - \widetilde{\Delta H}(x)]\rho_0(x)dx. \tag{17}$$

Conversely, the probability density $p(W|1;\mathcal{M})$ for the observation of a specific value W of the generalized work in reverse direction when sampled from ρ_1 reads

$$p(W|1;\mathcal{M}) = \int_{\Gamma_1} \delta[W - \widetilde{\Delta H}(\phi^{-1}(y))]\rho_1(y)dy. \tag{18}$$

Relating the forward and reverse "work" probability densities to each other results in the fluctuation theorem

$$\frac{p(W|0;\mathcal{M})}{p(W|1;\mathcal{M})} = e^{\beta(W-\Delta F)}. \tag{19}$$

This identity provides the main basis for our further results. It is established by multiplying Eq. (11) with $\delta[W - \widetilde{\Delta H}(x)]\rho_1(\phi(x))$ and integrating with respect to $\phi(x)$. The left-hand side yields

$$\int_{\phi(\Gamma_0)} \delta[W - \widetilde{\Delta H}(x)]\tilde{\rho}_0(\phi(x))d\phi(x)$$
$$= \int_{\Gamma_0} \delta[W - \widetilde{\Delta H}(x)]\rho_0(x)dx = p(W|0;\mathcal{M}), \tag{20}$$

and the right-hand side gives

$$\int_{\phi(\Gamma_0)} e^{\beta[\widetilde{\Delta H}(x)-\Delta F]}\delta[W - \widetilde{\Delta H}(x)]\rho_1(\phi(x))d\phi(x)$$
$$= e^{\beta(W-\Delta F)} \int_{\Gamma_1} \delta[W - \widetilde{\Delta H}(\phi^{-1}(y))]\rho_1(y)dy$$
$$= e^{\beta(W-\Delta F)} p(W|1;\mathcal{M}). \tag{21}$$

It is worthwhile to emphasize that the fluctuation theorem (19) is an exact identity for *any* differentiable, bijective map $\mathcal{M}$ from Γ_0 to Γ_1. Especially, it covers known fluctuation theorems [17–20] related to the *physical* work applied to a system that is driven externally and evolves in time according to some deterministic equations of motion, e.g., those of Hamiltonian dynamics, Nosé-Hoover dynamics, or Gaussian isokinetic dynamics [16].

As an example, consider the time-reversible adiabatic evolution of a conservative system with Hamiltonian $H_\lambda(x)$, depending on an externally controlled parameter λ (e.g., the strength of an external field). Let $x(t)=\phi(x_0,t;\lambda(\cdot))$ with $x(0)=x_0$ be the flow of the Hamiltonian system which is a functional of the parameter $\lambda(t)$ that is varied from $\lambda(0)=0$ to $\lambda(\tau)=1$ according to some prescribed protocol that constitutes the forward process. The Hamiltonian flow can be used to define a map, $\mathcal{M}: x \to \phi(x) := \phi(x,\tau;\lambda(\cdot))$. Since the evolution is adiabatic and Hamiltonian, no heat is exchanged, $Q=0$, and the Jacobian is identical to one, $|\frac{\partial\phi}{\partial x}|=1$. Consequently, the generalized work in the forward direction reduces to the physical work applied to the system, $W^0 := \widetilde{\Delta H}(x)=H_1(\phi(x))-H_0(x)=W$. For each forward path $\{x(t),\lambda(t);0\leqslant t\leqslant\tau\}$ we have a reverse path $\{x^T(\tau-t),\lambda^T(\tau-t);0\leqslant t\leqslant\tau\}$, where the superscript T indicates that quantities that are odd under time reversal (such as momenta) have changed their sign. The generalized work in reverse direction reduces to the physical work done *by* the system

$$W^1 := \widetilde{\Delta H}(\phi^{-1}(y)) = H_1(y) - H_0(\phi^{-1}(y)) - \tfrac{1}{\beta}\ln K(\phi^{-1}(y))$$
$$= -[H_0(\phi^{-1}(y)^T) - H_1(y^T)] = -W.$$

Starting the forward process with an initial canonical distribution, $\rho_0(x)$, some probability distribution for the physical work in forward direction follows, $p(W|0;\mathcal{M}) =: p^F(W)$. Starting the reverse process with an initial canonical distribution, $\rho_1(y)$, some probability distribution for the physical work in reverse direction follows, $p(W|1;\mathcal{M}) =: p^R(-W)$. The distributions $p^F(W)$ and $p^R(-W)$ are related to each other by the identity (19) which coincides with the fluctuation theorem of Crooks [17].

From the fluctuation theorem (19) some important inequalities follow that are valid for any map $\mathcal{M}$. First of all we state that the targeted free-energy perturbation formulas (13) and (15) can be regarded as a simple consequence of the fluctuation theorem (19) and can be rewritten in terms of the generalized work distributions, $e^{-\beta\Delta F}=\langle e^{-\beta W}\rangle_0$ and $e^{+\beta\Delta F}=\langle e^{+\beta W}\rangle_1$, where the angular brackets with subscript i denote an ensemble average with respect to the density $p(W|i;\mathcal{M})$, $i=0,1$. The monotonicity and convexity of the exponential function appearing in the above averages allows the application of Jensen's inequality, $\langle e^{\mp\beta W}\rangle \geqslant e^{\mp\beta\langle W\rangle}$. From this follows the fundamental inequality

$$\langle W\rangle_1 \leqslant \Delta F \leqslant \langle W\rangle_0, \tag{22}$$

which shows that the values of the average work in forward and reverse direction constitute an upper and a lower bound on ΔF, respectively.

Concerning one-sided estimates of ΔF, the targeted forward and reverse estimators (14) and (16) can be written $\widehat{\Delta F}_0=-\frac{1}{\beta}\overline{e^{-\beta W^0}}$ and $\widehat{\Delta F}_1=\frac{1}{\beta}\overline{e^{\beta W^1}}$, where the overbar denotes a sample average according to a sample $\{W_k^0\}$ and $\{W_k^1\}$ of forward and reverse work values, respectively. Similarly to Eq. (22) one finds the inequalities $\widehat{\Delta F}_0 \leqslant \overline{W^0}$ and $\widehat{\Delta F}_1 \geqslant \overline{W^1}$.

Taking the ensemble averages $\langle \widehat{\Delta F}_i \rangle_i = \mp \frac{1}{\beta} \langle \ln \overline{e^{\mp \beta W^i}} \rangle_i$, $i=0,1$, of the one-sided estimates and applying Jensen's inequality to the averages of the logarithms, $\langle \ln \overline{e^{\mp \beta W^i}} \rangle_i \leqslant \ln \langle \overline{e^{\mp \beta W^i}} \rangle_i = \mp \beta \Delta F$, one obtains

$$\langle W \rangle_1 \leqslant \langle \widehat{\Delta F}_1 \rangle_1 \leqslant \Delta F \leqslant \langle \widehat{\Delta F}_0 \rangle_0 \leqslant \langle W \rangle_0. \tag{23}$$

In other words, the forward and reverse estimators are biased in opposite directions for any finite size N of the work samples, but their mean values form closer upper and lower bounds on ΔF than the values of the mean work do.

So far, we were concerned with one-sided estimates of ΔF only. However, the full power of the fluctuation theorem (19) will develop when dealing with a two-sided targeted free-energy estimator where a sample of forward *and* reverse work values is used simultaneously, since the fluctuation theorem relates the forward and reverse work probability densities to each other in dependence of the free-energy difference.

In the next section, we will not mention the target map $\mathcal{M}$ explicitly in order to simplify the notation. For instance, we will write $p(W|i)$, but mean $p(W|i;\mathcal{M})$ instead.

IV. TWO-SIDED TARGETED FREE-ENERGY ESTIMATOR

An important feature of the fluctuation theorem (19) is that it provides a way to answer the following question: Given a sample of n_0 work values $\{W_i^0\}=\{W_1^0,\ldots,W_{n_0}^0\}$ in the forward direction and a second sample of n_1 work values $\{W_j^1\}=\{W_1^1,\ldots,W_{n_1}^1\}$ in the reverse direction, what would be the best estimator of ΔF that utilizes the entire two samples?

If drawn from an ensemble that consists of forward and reverse work values, the elements are given by a pair of values (W,Y) of work and direction, where $Y=0$ indicates the forward and $Y=1$ the reverse direction. The probability density of the pairs (W,Y) is $p(W,Y)$. The probability density for the work is $p(W) \coloneqq p(W,0)+p(W,1)$, and that for the direction is $p_Y \coloneqq \int p(W,Y)dW$.

Bayes theorem,

$$p(W|Y)p_Y = p(Y|W)p(W), \tag{24}$$

implies the "balance" equation

$$p_1 \int p(0|W)p(W|1)dW = p_0 \int p(1|W)p(W|0)dW. \tag{25}$$

From the fluctuation theorem (19) and Bayes theorem (24) follows

$$\frac{p(0|W)}{p(1|W)} = e^{\beta(W-C)} \tag{26}$$

with

$$C = \Delta F + \frac{1}{\beta} \ln \frac{p_1}{p_0}. \tag{27}$$

Together with the normalization $p(0|W)+p(1|W)=1$, Eq. (26) determines the explicit form of the conditional direction probabilities [15],

$$p(Y|W) = \frac{e^{Y\beta(C-W)}}{1+e^{\beta(C-W)}}, \quad Y=0,1. \tag{28}$$

Replacing both, the ensemble averages by sample averages and the ratio $\frac{p_1}{p_0}$ by $\frac{n_1}{n_0}$, the balance equation, $p_1\langle p(0|W)\rangle_1 = p_0\langle p(1|W)\rangle_0$, results in the two-sided targeted free-energy estimator, $n_1 p(0|W^1) = n_0 p(1|W^0)$, which reads

$$\sum_{j=1}^{n_1} \frac{1}{1+e^{\beta\left(\widehat{\Delta F}_{01}+\frac{1}{\beta}\ln\frac{n_1}{n_0}-W_j^1\right)}} = \sum_{i=1}^{n_0} \frac{1}{1+e^{-\beta\left(\widehat{\Delta F}_{01}+\frac{1}{\beta}\ln\frac{n_1}{n_0}-W_i^0\right)}}. \tag{29}$$

It is worth it to emphasize that this estimator is *the* optimal two-sided estimator, a result that is shown with a constraint maximum likelihood approach in the Appendix. A derivation of this estimator is also given by Shirts *et al.* [15] in the framework of a maximum likelihood approach.

If samples of n_0 forward and n_1 reverse work values $\{W_i^0\}$ and $\{W_j^1\}$ are given, but no further information is present, it is the two-sided estimator (29) that yields the best estimate of the free-energy difference with respect to the mean square error. If needed, the samples $\{W_i^0\}$ and $\{W_j^1\}$ can be obtained indirectly by drawing samples $\{x_i\}$ and $\{y_j\}$ of ρ_0 and ρ_1 and setting $W_i^0 = \widehat{\Delta H}(x_i)$ and $W_j^1 = \widehat{\Delta H}(\phi^{-1}(y_j))$.

Opposed to the one-sided estimators (14) and (16), the two-sided targeted free-energy estimator (29) is an implicit equation that needs to be solved for $\widehat{\Delta F}_{01}$. Note however that the solution $\widehat{\Delta F}_{01}$ is unique.

Let us mention a subtlety concerning the choice of the ratio $\frac{p_1}{p_0}$. The mixed ensemble $\{(W,Y)\}$ is specified by the mixing ratio $\frac{p_1}{p_0}$, and by the conditional work probability densities $p(W|Y)$. With the mixed ensemble we are free to choose the mixing ratio. For instance, replacing the ensemble averages in the balance equation (25) by sample averages results in an estimator $p_1 p(0|W^1) = p_0 p(1|W^0)$ for ΔF that depends on the value of the mixing ratio. This raises the question of the optimal choice for $\frac{p_1}{p_0}$. As shown in the Appendix, it is optimal to choose the mixing ratio equal to the sample ratio, $\frac{p_1}{p_0} = \frac{n_1}{n_0}$. A result that may be clear intuitively, since then the mixed ensemble reflects the actual samples best.

Other free-energy estimators follow, if the explicit expressions (28) and the definition of the constant C, Eq. (27), are inserted in the balance equation (25). The latter can then be expressed as

$$e^{\beta\Delta F} = e^{\beta C} \frac{\int \frac{1}{1+e^{\beta(C-W)}} p(W|1)dW}{\int \frac{1}{1+e^{-\beta(C-W)}} p(W|0)dW}, \tag{30}$$

and results in the estimator

$$\widehat{\Delta F}_{\mathrm{B}}(C) = C + \frac{1}{\beta} \ln \frac{\frac{1}{n_1}\sum_{j=1}^{n_1} \frac{1}{1+e^{\beta(C-W_j^1)}}}{\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{1}{1+e^{-\beta(C-W_i^0)}}}. \tag{31}$$

The nontargeted version of this estimator, i.e., for $\mathcal{M}=id$, is due to Bennett [13] who used a variational principle in order to find the estimator for the free-energy difference that minimizes the mean square error.

Equation (30) is an identity for any value of C, since with the ratio $\frac{p_1}{p_0}$ the value of $C=\Delta F+\frac{1}{\beta}\ln\frac{p_1}{p_0}$ can be chosen arbitrarily. However, concerning the estimator (31), different values of C yield different estimates. Bennett's choice is

$$C_{\mathrm{B}} = \Delta F + \frac{1}{\beta} \ln \frac{n_1}{n_0}, \tag{32}$$

i.e., $\frac{p_1}{p_0}=\frac{n_1}{n_0}$, which results from minimizing the mean square error $\langle(\widehat{\Delta F}_{\mathrm{B}}-\Delta F)^2\rangle$, where the angular brackets denote an average over infinitely many repetitions of the estimation process (31) with n_0 and n_1 being fixed. According to the Appendix, Bennett's choice is also optimal for any target map $\mathcal{M}$.

With $C=C_{\mathrm{B}}$, Eq. (31) has to be solved in a self-consistent manner which is tantamount to solve the two-sided targeted estimator (29). In other words, $\widehat{\Delta F}_{\mathrm{B}}(C_{\mathrm{B}})$ is the unique root $\widehat{\Delta F}_{01}$ of Eq. (29).

V. OVERLAP MEASURES AND MEAN SQUARE ERRORS

In this section we introduce measures for the overlap of $\tilde{\rho}_0$ with ρ_1, or, equivalently, of $p(W|0;\mathcal{M})$ with $p(W|1;\mathcal{M})$ and relate them to the mean square error of one- and two-sided estimators.

The estimators (14), (16), and (29) are subject to both, bias and variance. Taking both errors into account results in the mean square error. Let us consider the mean square errors of the one-sided targeted estimators first. They read $X_0 := \langle(\widehat{\Delta F}_0-\Delta F)^2\rangle_0=\langle(\ln \overline{e^{-\beta(W^0-\Delta F)}})^2\rangle_0$ in forward direction, and analogously in backward direction. In the forward direction, it can be quantified by expanding the logarithm into a power series about the mean value of its argument, $\langle e^{-\beta(W^0-\Delta F)}\rangle_0=1$, and neglecting terms of higher order in $\frac{1}{N}$, which gives

$$\beta^2 X_0 \approx \frac{1}{N}\langle(e^{-\beta(W-\Delta F)}-1)^2\rangle_0. \tag{33}$$

Equation (33) is valid for a sufficiently large sample size N (large N limit) [21]. With the use of the fluctuation theorem (19), the variance appearing on the right-hand side of Eq. (33) can be written $\langle(e^{-\beta(W-\Delta F)}-1)^2\rangle_0=\langle e^{-\beta(W-\Delta F)}\rangle_1-1 \geqslant e^{-\beta(\langle W\rangle_1-\Delta F)}-1$. This yields the inequality

$$\beta^2 X_0 \geqslant \frac{1}{N}(e^{\beta(\Delta F-\langle W\rangle_1)}-1). \tag{34}$$

In the same manner as above the inequality

$$\beta^2 X_1 \geqslant \frac{1}{N}(e^{\beta(\langle W\rangle_0-\Delta F)}-1) \tag{35}$$

is obtained for the mean square error X_1 of the reverse estimator $\widehat{\Delta F}_1$.

The inequalities (34) and (35) specify the minimum sample size N that is required to obtain a forward and reverse estimate $\widehat{\Delta F}$, respectively, whose root mean square error $\sqrt{X}$ is not larger than kT. Namely, $N \geqslant e^{\beta(\Delta F-\langle W\rangle_1)}$ is required for a forward, and $N \geqslant e^{\beta(\langle W\rangle_0-\Delta F)}$ for a reverse estimate. Similar expressions are found in Ref. [22]. Since the required sample size N depends exponentially on the dissipation, it is good to choose a target map $\mathcal{M}$ which reduces the dissipation in the opposite direction.

The dissipation is related to the overlap of $\tilde{\rho}_0$ with ρ_1. The overlap of two probability densities $\pi_a(z)$ and $\pi_b(z)$ of a random variable z can be quantified with the Kullback-Leibler divergence

$$D(\pi_a\|\pi_b) := \int \pi_a(z)\ln\frac{\pi_a(z)}{\pi_b(z)}dz, \tag{36}$$

a positive semidefinite measure that yields zero if and only if π_a is identical to π_b. Applied to the densities ρ_1 and $\tilde{\rho}_0$, the Kullback-Leibler divergence turns out to be identical with the Kullback-Leibler divergence of $p(W|1;\mathcal{M})$ with $p(W|0;\mathcal{M})$ and results in the generalized dissipated work in reverse direction,

$$D(\rho_1\|\tilde{\rho}_0) = D(p(W|1;\mathcal{M})\|p(W|0;\mathcal{M})) = \beta(\Delta F - \langle W\rangle_1), \tag{37}$$

which is established with the use of Eqs. (11) and (18), and the fluctuation theorem (19). Similarly, we have

$$D(\tilde{\rho}_0\|\rho_1) = D(p(W|0;\mathcal{M})\|p(W|1;\mathcal{M})) = \beta(\langle W\rangle_0 - \Delta F). \tag{38}$$

For the one-sided targeted free-energy estimators this means that choosing a target map which reduces the dissipation in the opposite direction is the same as choosing a target map which enhances the overlap of $\tilde{\rho}_0$ with ρ_1.

Now, we proceed with the overlap measure and the mean square error of the two-sided free-energy estimator (29). In order to keep the notation simple, we assume that the samples of forward and reverse work values are of equal size, $n_0=n_1=\frac{N}{2}$, i.e., $n_0+n_1=N$. (A generalization to $n_0 \neq n_1$ is straightforward possible, but not given in this paper.)

Consider the overlap density $p_{\mathrm{ol}}(W|\mathcal{M})$,

$$p_{\mathrm{ol}}(W|\mathcal{M}) := \frac{1}{A_{\mathrm{ol}}} \frac{p(W|0;\mathcal{M})p(W|1;\mathcal{M})}{\frac{1}{2}[p(W|0;\mathcal{M})+p(W|1;\mathcal{M})]}, \tag{39}$$

where the normalization constant A_{ol} reads

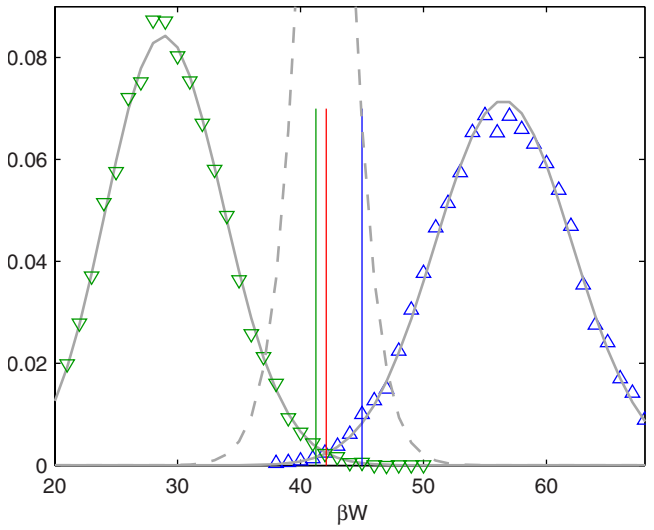

FIG. 1. (Color online) Targeted work probability distributions for the expansion of a cavity in an ideal gas and the associated overlap distribution. The up (down) triangles display the normalized histogram of a sample of forward (reverse) work values. The smooth solid curves are the exact analytic work distributions $p(W|0;\mathcal{M})$ (right) and $p(W|1;\mathcal{M})$ (left), and the dashed curve shows their overlap distribution $p_{\rm ol}(W|\mathcal{M})$. The straight vertical lines show the values of the targeted estimates of ΔF on the abscissa. From left to right: the reverse, the two-sided (which is indistinguishable from the exact analytic value), and the forward estimate.

$$A_{\rm ol} = \int \frac{p(W|0;\mathcal{M})p(W|1;\mathcal{M})}{\frac{1}{2}[p(W|0;\mathcal{M}) + p(W|1;\mathcal{M})]} dW$$

$$= \int \frac{\tilde{\rho}_0(y)\rho_1(y)}{\frac{1}{2}[\tilde{\rho}_0(y) + \rho_1(y)]} dy. \tag{40}$$

$A_{\rm ol}$ is a measure for the overlap area of the distributions and takes its maximum value 1 in case of coincidence. Using the fluctuation theorem (19), the two-sided overlap measure can be written

$$\frac{1}{2}A_{\rm ol} = \left\langle \frac{1}{1+e^{\beta(\Delta F - W)}} \right\rangle_1 = \left\langle \frac{1}{1+e^{-\beta(\Delta F - W)}} \right\rangle_0. \tag{41}$$

Comparing Eq. (41) with the two-sided targeted free-energy estimator (29), one sees that the two-sided targeted free-energy estimation method readily estimates the two-sided overlap measure. The accuracy of the estimate depends on how good the sampled work values reach into the main part of the overlap distribution $p_{\rm ol}(W|\mathcal{M})$. By construction, the overlap region is sampled far earlier than the further distant tail that lies in the peak of the other distribution, cf. Fig. 1. This is the reason why the two-sided estimator is superior if compared to the one-sided estimators.

In the large N limit the mean square error $X_{01}(N) = \langle(\widehat{\Delta F}_{01} - \Delta F)^2\rangle$ of the two-sided estimator can be expressed in terms of the overlap measure and reads

$$X_{01}(N) = \frac{4}{N}\left(\frac{1}{A_{\rm ol}} - 1\right), \tag{42}$$

cf. [13,15]. Note that if an estimated value $\hat{A}_{\rm ol}$ is plugged in, this formula is valid in the limit of large N only, but it is not clear *a priori* when this limit is reached. Therefore, we develop a simple convergence criterion for the two-sided estimate.

VI. CONVERGENCE

In this section, a measure for the convergence of the two-sided estimate is developed, again for the special case $n_0 = n_1 = \frac{N}{2}$. First, we define the estimate $\hat{A}_{\rm ol}$ of the overlap measure $A_{\rm ol}$ with

$$\frac{1}{2}\hat{A}_{\rm ol}(N) = \frac{1}{n_1}\sum_{j=1}^{n_1} \frac{1}{1+e^{\beta(\widehat{\Delta F}_{01} - W_j^1)}}, \tag{43}$$

which is equal to $\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{1}{1+e^{-\beta(\widehat{\Delta F}_{01}-W_i^0)}}$, as we understand the estimate $\widehat{\Delta F}_{01}$ to be obtained according to Eq. (29) with the same samples of forward and reverse work values. Since the accuracy of the estimated value $\hat{A}_{\rm ol}$ is unknown, we need an additional quantity to compare with.

Another expression for the overlap measure is

$$\frac{1}{2}A_{\rm ol} = \left\langle \left(\frac{1}{1+e^{\beta(\Delta F - W)}}\right)^2 \right\rangle_1 + \left\langle \left(\frac{1}{1+e^{-\beta(\Delta F - W)}}\right)^2 \right\rangle_0, \tag{44}$$

which can be verified with the fluctuation theorem (19). Based on Eq. (44), we define the overlap estimator of second order,

$$\frac{1}{2}\hat{A}_{\rm ol}^{(II)}(N) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{1}{1+e^{\beta(\widehat{\Delta F}_{01} - W_j^1)}}\right)^2 + \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{1}{1+e^{-\beta(\widehat{\Delta F}_{01} - W_i^0)}}\right)^2. \tag{45}$$

Because $\widehat{\Delta F}_{01}$ converges to ΔF, both $\hat{A}_{\rm ol}$ and $\hat{A}_{\rm ol}^{(II)}$ converge to $A_{\rm ol}$ in the limit $N \to \infty$. However, the second-order estimator $\hat{A}_{\rm ol}^{(II)}$ converges slower and is for small N typically much smaller than $\hat{A}_{\rm ol}$, since the main contributions to the averages appearing in Eq. (44) result from work values that lie somewhat further in the tails of the work distributions.

We use the relative difference

$$a(N) = \frac{\hat{A}_{\rm ol} - \hat{A}_{\rm ol}^{(II)}}{\hat{A}_{\rm ol}} \tag{46}$$

to quantify the convergence of the two-sided estimate $\widehat{\Delta F}_{01}$, where $\hat{A}_{\rm ol}$, $\hat{A}_{\rm ol}^{(II)}$, and $\widehat{\Delta F}_{01}$ are understood to be calculated with the same two samples of forward and reverse work values.

From Eqs. (45), (43), and (29) follows that $0 \leqslant \hat{A}_{\rm ol}^{(II)} \leqslant 2\hat{A}_{\rm ol}$ holds. Hence, the convergence measure $a(N)$ is bounded by

$$-1 \leqslant a(N) \leqslant 1 \tag{47}$$

for any N. A necessary convergence condition is $a(N) \to 0$. This means that only if $a(N)$ is close to zero, the two-sided overlap estimators can have converged. Typically, $a(N)$ being close to zero is also a sufficient convergence condition. Hence, if $a(N)$ is close to zero, the mean square error of $\widehat{\Delta F}_{01}$ is given by Eq. (42) with $A_{\mathrm{ol}} \approx \hat{A}_{\mathrm{ol}}$. As can be seen from Eq. (42), the mean square error and in turn the variance and the bias are reduced by both, by taking a larger sample size N and by choosing a map $\mathcal{M}$ that enhances the overlap of $\tilde{\rho}_0$ with ρ_1.

With the targeted free-energy estimators at hand, together with their mean square errors, we are now ready to compute free-energy differences numerically.

VII. NUMERICAL EXAMPLES

We investigate two numerical applications. One is the free-energy difference of a fluid subject to the expansion of a cavity which allows the comparison with published results [7]. The other is the chemical potential of a fluid in the high density regime.

Beneath an ideal gas, the fluid is chosen to be a Lennard-Jones fluid with pairwise interaction

$$V(r_{kl}) = 4\epsilon\left[\left(\frac{\sigma}{r_{kl}}\right)^{12} - \left(\frac{\sigma}{r_{kl}}\right)^{6}\right], \tag{48}$$

where r_{kl} is the distance between the kth and lth particle, $r_{kl} = |\mathbf{r}_k - \mathbf{r}_l|$. The parameters used are those of argon, $\sigma = 3.542$ Å, and $\epsilon/k = 93.3$ K [24].

In all applications, the samples from the densities ρ_0 and ρ_1 are simulated with the Metropolis algorithm [23]. In order to simulate macroscopic behavior with a small number N_p of particles, periodic boundary conditions and the minimum image convention [6] are used. Pairwise interactions are truncated at half of the box length $R_{\mathrm{box}} = L/2$, but are not shifted, and the appropriate cutoff corrections are applied [6].

A. Expansion of a cavity in a fluid

The expansion of a cavity in a fluid is given by the following setup: Consider a fluid of N_p point molecules with pairwise interaction $V(r_{kl})$ confined in a cubic box of side length $2R_{\mathrm{box}}$, but excluded from a sphere of radius $R \leqslant R_{\mathrm{box}}$, compare with Fig. 2. Both the box and the sphere are centered at the origin $\mathbf{r}=0$. A configurational microstate of the system is given by a set $x = (\mathbf{r}_1, \ldots, \mathbf{r}_{N_p})$ of particle positions $\mathbf{r}_k$. Growing the sphere from $R=R_0$ to $R=R_1$ decreases the volume accessible to the particles and the fluid is compressed. We are interested in the increase of free-energy ΔF subject to the compression of the fluid. Since the kinetic contribution to the free-energy is additive and independent of R, the difference ΔF depends only on the configurational part of the Hamiltonian. The latter reads

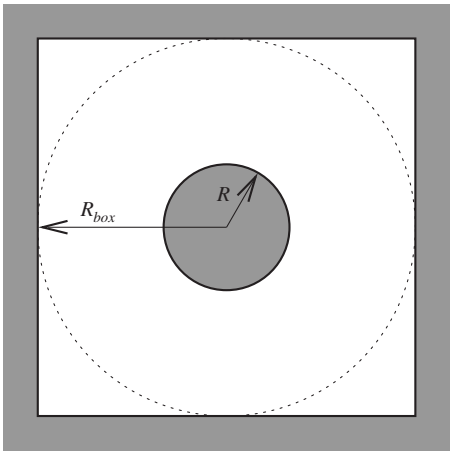

FIG. 2. The geometric setup.

$$H_i(x) = \begin{cases} \sum_{k<l} V(r_{kl}) & \text{if } x \in \Gamma_i, \\ \infty & \text{if } x \notin \Gamma_i, \end{cases} \tag{49}$$

with $i=0,1$. Γ_0 and Γ_1 denote the accessible parts of configuration space of the system 0 ($R=R_0$) and 1 ($R=R_1$), respectively. We assume that $R_0 < R_1$ holds which implies $\Gamma_1 \subset \Gamma_0$.

Drawing a sample $\{x_k\}$ from ρ_0 and applying the traditional forward estimator (4) results in the following: $e^{-\beta\Delta H(x_k)}$ takes the values one and zero depending on whether $x_k \in \Gamma_1$ or not, i.e., whether the region between the two spheres of radius R_0 and R_1 is found vacant of particles or not. A comparison with Eq. (3) reveals that $e^{-\beta\Delta F}$ is the probability for the spherical shell being observed devoid of particles [7]. Hence, the rate of convergence of $\overline{e^{-\beta\Delta H}}$ decreases with the latter probability and will in general be poor.

Conversely, drawing a sample y_k from ρ_1 and applying the traditional reverse estimator $\widehat{\Delta F}_1^{\mathrm{trad}} = \frac{1}{\beta}\overline{e^{\beta\Delta H(y)}}$ [Eq. (16) with $\phi(x)=x$] figures out to be invalid, because the term $e^{\beta\Delta H(y_k)}$ takes always the value one. In consequence, the traditional reverse estimator is inconsistent. The deeper reason for this is that $\Gamma_1 \subset \Gamma_0$ holds: Eq. (2) is valid only for $x \in \Gamma_1$. By the same reason, the traditional two-sided estimator is inconsistent, too.

The mentioned shortcomings are avoided with a well chosen target map. Consider mapping each particle separately according to

$$\phi(x) = \left(\psi(r_1)\frac{\mathbf{r}_1}{r_1}, \ldots, \psi(r_{N_p})\frac{\mathbf{r}_{N_p}}{r_{N_p}}\right), \tag{50}$$

where $r_k = |\mathbf{r}_k|$ is the distance of the kth particle with respect to the origin, and $\psi: (R_0, R_{\max}] \to (R_1, R_{\max}]$ is a bijective and piecewise smooth radial mapping function. In order not to map particles out of the confining box, it is required that $\psi(r)=r$ holds for $r > R_{\mathrm{box}}$. The Jacobian for the radial map (50) reads

$$\left|\frac{\partial\phi}{\partial x}\right| = \prod_{j=1}^{N_p} \frac{\psi(r_j)^2}{r_j^2}\frac{\partial\psi(r_j)}{\partial r_j}. \tag{51}$$

(This formula is immediately clear when changing to polar coordinates.) We use the map of Ref. [7] which is designed to uniformly compress the volume of the shell $R_0 < r \leq R_{\mathrm{box}}$ to the volume of the shell $R_1 < r \leq R_{\mathrm{box}}$. Thus, for $r \in (R_0, R_{\mathrm{box}}]$ the radial mapping function $\psi(r)$ is defined by

$$\psi(r)^3 - R_1^3 = c(r^3 - R_0^3), \tag{52}$$

with the compression factor $c=(R_{\mathrm{box}}^3-R_1^3)/(R_{\mathrm{box}}^3-R_0^3)$. According to Eq. (51), we have $\ln K(x)=\nu(x)\ln c$, where $\nu(x)$ is the number of particles in the shell $R_0 < r \leq R_{\mathrm{box}}$.

1. Ideal gas

As a first illustrative and exact solvable example we choose the fluid to be an ideal gas, $V(r_{kl})=0$. In this case the free-energy difference is solely determined by the ratio of the confined volume $V_i=8R_{\mathrm{box}}^3-\frac{4}{3}\pi R_i^3$, $i=0,1$, and is given by $\beta\Delta F=-N_p \ln(V_1/V_0)$. Using the radial map (52), the work in the forward direction as a function of x reads $\widetilde{\Delta H}(x)=-\frac{1}{\beta}\nu(x)\ln c$ and takes discrete values only, as $\nu(x)=n$ holds with $n \in \{0,1,\ldots,N_p\}$. Consequently, the probability $p(W_n|0;\mathcal{M})$ of observing the work $W_n=-\frac{n}{\beta}\ln c$ in forward direction is binomial,

$$p(W_n|0;\mathcal{M}) = \binom{N_p}{n} q_0^n (1-q_0)^{N-n}, \tag{53}$$

where $q_0=\frac{4}{3}\pi(R_{\mathrm{box}}^3-R_0^3)/V_0$ is the probability of any fixed particle to be found in the shell $R_0 < r \leq R_{\mathrm{box}}$. In analogy, the probability distribution $p(W_n|1;\mathcal{M})$ for observing the work $W=W_n$ in reverse direction is given by replacing the index 0 with 1 in Eq. (53). Finally, the work probability distributions (rather then the densities) obey the fluctuation theorem (19) for any $n=0,1,\ldots,N_p$,

$$\frac{p(W_n|0;\mathcal{M})}{p(W_n|1;\mathcal{M})} = \frac{1}{c^n}\left(\frac{V_1}{V_0}\right)^{N_p} = e^{\beta(W_n-\Delta F)}. \tag{54}$$

A simple numerical evaluation highlights the convergence properties. Choosing the parameters to be $2R_{\mathrm{box}}=22.28$ Å, $R_0=7$ Å, $R_1=10$ Å, and $N_p=125$ (β arbitrary), the free-energy difference takes the value $\beta\Delta F=42.1064$. Because $e^{-\beta\widetilde{\Delta H}(x)}$ can take only the numbers zero and one, the probability of observing a configuration x with nonvanishing contribution in the traditional forward estimator of ΔF is $e^{-\beta\Delta F} \approx 10^{-19}$. Hence, in practice it is impossible to use the traditional method successfully, since it would require at least $N_p \times 10^{19}$ Monte Carlo trial moves. However, the targeted approach already gives reasonable estimates with a sample size of just a few thousands. Figure 1 shows estimates of the targeted work probability distributions for samples of size $n_i=10^4$ ($i=0,1$) from ρ_0 and ρ_1 each. While the forward distribution $p(W|0;\mathcal{M})$ is obviously well sampled in the central region, the sampling size is too small in order to reach the small values of βW where the reverse distribution $p(W|1;\mathcal{M})$ is peaked. Exactly the latter values would be required for an *accurate* exponential average in the targeted forward estimator, Eq. (14). Therefore, the targeted forward estimate of ΔF is still inaccurate; it yields $\beta\widehat{\Delta F}_0=45.0\pm0.3$. The same is true for the targeted reverse estimate (16) which gives $\beta\widehat{\Delta F}_1=41.3\pm0.5$. The errors are calculated using root mean squares and propagation of uncertainty. A more accurate estimate follows from the targeted two-sided estimator (29) which yields $\beta\widehat{\Delta F}_{01}=42.1\pm0.1$ ($n_0=n_1=10^4$). This is clear, as for the two-sided estimate it is sufficient yet that the forward and reverse work values sample the region where the overlap distribution $p_{\mathrm{ol}}(W|\mathcal{M})$, Eq. (39), is peaked, which is obviously the case, cf. Fig. 1.

The ideal gas is an exactly solvable model. This raises the question of whether a "perfect" or an ideal map can be constructed. The answer is yes, however such an ideal map would not be in the set of radial maps as defined with Eq. (50). Instead, the ideal map would also depend on the angles and would have a more complicated structure. The reason for this is the geometry of the simulation box: An ideal map needs to compress the fluid of uniform density with $R=R_0$ to the fluid of uniform density with $R=R_1$. The radial mapping function $\psi(r)$, Eq. (52), can be viewed as a good approximation to the ideal map within the set of radial maps.

2. Lennard-Jones fluid

We now focus on particles with Lennard-Jones interaction (48). The parameters are chosen to coincide with those of Ref. [7], i.e., $2R_{\mathrm{box}}=22.28$ Å, $R_0=9.209$ Å, $R_1=9.386$ Å, $N_p=125$, and $T=300$ K. In Lennard-Jones units, the reduced densities $\rho_i^*=\sigma^3 N_p/V_i$ of the systems 0 ($R=R_0$) and 1 ($R=R_1$) are $\rho_0^*=0.713$ and $\rho_1^*=0.731$, respectively, and $T^*=1/(\beta\epsilon)=3.215$ holds for both. If we had an ideal gas, the probability of observing the space between the spheres of radius R_0 and R_1 to be vacant of particles would be $(V_1/V_0)^{N_p}=0.044$. Because of the strong repulsive part of interaction, this probability is much smaller in case of a dense Lennard-Jones fluid.

We generate samples of $\rho_0(x)$ and $\rho_1(x)$ with a Metropolis Monte Carlo simulation. Each run starts with 1000 equilibration sweeps, followed by the production run. In the production run the configurational microstate x is being sampled every fourth sweep only in order to reduce correlations between successive samples. The use of decorrelated data is of particular importance for the self-consistent two-sided estimate $\widehat{\Delta F}_{01}$ because it depends intrinsically on the ratio $\frac{n_1}{n_0}$ of the numbers of *uncorrelated* samples, cf. Eq. (29).

Figure 3 gives an overview of independent runs with different sample sizes N, where the one- and two-sided targeted estimators can be compared with each other and with the traditional forward estimator. Displayed is the estimated mean $\overline{\widehat{\Delta F}(N)}$ in dependence of the sample size N. The error bars reflect the estimated standard deviation

$$\overline{[\widehat{\Delta F}(N) - \overline{\widehat{\Delta F}(N)}]^2}^{1/2}.$$

Each mean and each standard deviation is estimated using $z(N)$ independent estimates $\widehat{\Delta F}(N)$. In ascending order of N,

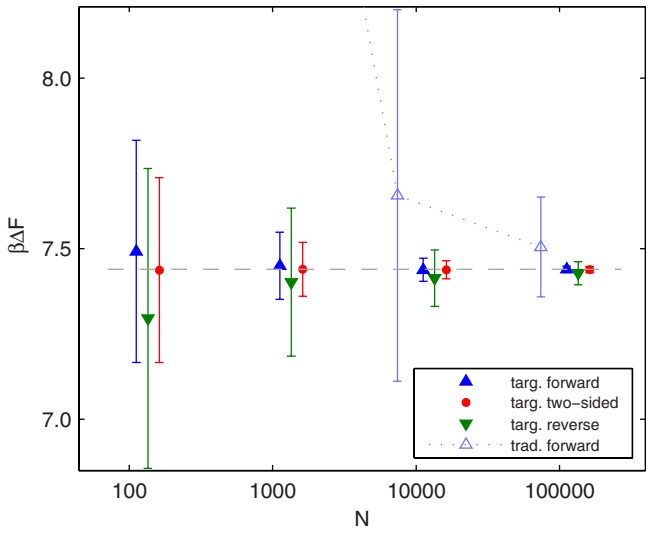

FIG. 3. (Color online) Free-energy estimates for the expansion of a cavity in a Lennard-Jones fluid. Shown are the average values of traditional and targeted estimates of ΔF in dependence of the sample size N, with an errorbar of one standard deviation. In order to distinguish the data points those corresponding to targeted estimates are shifted to the right and are spread, whereas those corresponding to traditional estimates are shifted to the left. For example, all four data points in the vicinity of $N=10\,000$ refer to $N=10\,000$. The dashed horizontal line represents a targeted two-sided estimate with $N=10^6$, see Table I.

$z(N)$ reads 440, 170, 40, 13. For the two-sided estimates, $n_0=n_1=\frac{N}{2}$ is used and Eq. (29) is solved.

Note that the theoretical mean of *traditional* forward estimates of ΔF is infinite *for any finite* N, because of the finite probability of observing a sequence of length N of solely vanishing contributions to the exponential average $\overline{e^{-\beta\Delta H}}$. Strictly spoken, the estimator $\widehat{\beta\Delta F_0^{\mathrm{trad}}}=-\ln \overline{e^{-\beta\Delta H}}$ is not well defined, because $\Gamma_1 \subset \Gamma_0$. Nevertheless, in Fig. 3 there are two finite *observed* mean values of traditional forward estimates displayed, what by no means is a contradiction. Infinite values are observed in the cases where $N<10^4$ holds. This is symbolized by the rising dotted line. The mentioned ill definiteness of the traditional estimator is removed by using the map (52). Figure 3 shows that all three targeted estimators are consistent even for small N in the sense that the error bars overlap. Whereas the targeted forward and reverse estimators show to be decreasingly biased with increasing N, the targeted two-sided estimator does not show any noticeable bias at all. This example demonstrates how worth it can be to take all three estimators, forward, reverse, and two-sided, into account. The one-sided estimators are biased in opposite directions and may serve as upper and lower bounds for ΔF, Eq. (23), whereas the two sided is typically placed in between the one sided.

We conclude this example with explicit estimates obtained from a single run with $N=10^6$, which are summarized in Table I. The errors are derived using block averages [25] and propagation of uncertainty.

TABLE I. Cavity in a Lennard-Jones fluid. Estimated free-energy differences $\widehat{\beta\Delta F}$ for the expansion of a cavity, using targeted and traditional estimators. $N=10^6$.

Method	$\widehat{\beta\Delta F}$
Traditional forward	7.500 ± 0.043
Targeted forward	7.439 ± 0.003
Targeted two sided	7.440 ± 0.002
Targeted reverse	7.420 ± 0.009

B. Chemical potential of a homogeneous fluid

Consider a fluid of N_p particles confined within a cubic box of volume $V_c=(2R_{\mathrm{box}})^3$ with pairwise interaction $V(r_{ij})$. The configurational Hamiltonian for the N_p-particle system at $x=(\mathbf{r}_1,\ldots,\mathbf{r}_{N_p})$ reads

$$H_{N_p}(x)=\sum_{i<j}^{N_p} V(r_{ij}). \tag{55}$$

The configurational density for the N_p system is given by

$$\rho_{N_p}(x)=e^{-\beta H_{N_p}(x)}/Z_{N_p}, \tag{56}$$

with the partition function $Z_{N_p}=\int e^{-\beta H_{N_p}(x)}dx$. Now consider one particle is added: the position of this new particle may be $\mathbf{r}_{N_p+1}$. The equilibrium density of the (N_p+1)-particle system reads

$$\rho_{N_p+1}(x)=e^{-\beta H_{N_p+1}(x,\mathbf{r}_{N_p+1})}/Z_{N_p+1}. \tag{57}$$

Taking the ratio of the densities (56) and (57) leads to Widom's particle insertion method [26] for estimating the excess chemical potential μ^{ex} of the N_p system, defined as the excess of the chemical potential μ to that of an ideal gas at the same temperature and density. For sufficiently large N_p, μ^{ex} can be approximated with

$$\mu^{\mathrm{ex}}=-\frac{1}{\beta}\ln\frac{Z_{N_p+1}}{Z_{N_p}V_c}. \tag{58}$$

Turning the tables, we use Eq. (58) to be the definition of the quantity μ^{ex}. The particle insertion method inserts at a random position an extra particle to the N_p system and measures the increase of energy that results from this particle. Since we consider a homogeneous fluid, we may as well fix the position of insertion arbitrarily, for instance at the origin, what is done in the following. We define system 1 through the configuration-space density $\rho_1(x)$ as follows:

$$\rho_1(x)=V_c\int \delta(\mathbf{r}_{N_p+1})\rho_{N_p+1}(x,\mathbf{r}_{N_p+1})d\mathbf{r}_{N_p+1}. \tag{59}$$

The factor V_c ensures normalization. Written in the usual form $\rho_1(x)=e^{-\beta H_1(x)}/Z_1$, we have

$$H_1(x)=H_{N_p}(x)+\sum_{k=1}^{N_p} V(r_k) \tag{60}$$

and $Z_1=Z_{N_p+1}/V_c$. System 1 can be understood as an equilibrium system of N_p interacting particles in the external potential $\Sigma_{k=1}^{N_p}V(r_k)$, due to one extra particle fixed at the origin $\mathbf{r}$

$=0$. Further, we identify system 0 with the N_p-particle system and rewrite

$$\rho_0(x) = \rho_{N_p}(x), \quad H_0(x) = H_{N_p}(x) \tag{61}$$

and $Z_0 = Z_{N_p}$. The ratio of ρ_0 and ρ_1 has the familiar form of Eq. (2), with ΔF being identical to μ^{ex},

$$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta[\Delta H(x) - \mu^{\mathrm{ex}}]}. \tag{62}$$

The energy difference $\Delta H(x) = H_1(x) - H_0(x)$ is the increase of energy due to an added particle at the origin $\mathbf{r}=0$,

$$\Delta H(x) = \sum_{k=1}^{N_p} V(r_k). \tag{63}$$

Assume a finite potential $V(r)$ for nonvanishing r (i.e., no hard-core potential), but with a strong repulsive part for $r \to 0$ (a so-called soft-core potential), e.g., a Lennard-Jones potential. In this case, the configuration spaces of system 0 and 1 coincide, i.e., $\Gamma_0 = \Gamma_1$. Thus a traditional estimate of μ^{ex} is in principle valid in both directions, forward and reverse. In the forward direction we have the equivalent to the particle insertion method [26], $\beta\widehat{\mu^{\mathrm{ex}}}_0^{\mathrm{trad}} = -\ln \overline{e^{-\beta\Delta H(x)}}$, but with the fixed position of insertion $\mathbf{r}=0$. Here x is drawn from ρ_0 and we will typically find a particle in a sphere of radius $\bar{r}$ centered at the origin. $\bar{r}$ can roughly be estimated by the mean next-neighbor distance $(V_c/N_p)^{1/3}$ of an ideal gas. The dominant contributions to the exponential average come from realizations x that resemble typical realizations of system 1 [22]. However, typical realizations x of system 1 do not contain any particle within a sphere of some radius r_{hc} centered at the origin, because of the extra particle fixed at the origin and the strong repulsive part of the interaction. r_{hc} may be regarded as a temperature-dependent effective hard-core radius of the interaction $\beta V(r)$. We conclude that the insertion method is accurate and fast convergent only if $r_{hc}^3 \ll \bar{r}^3$, i.e., for low densities. Concerning the reverse traditional estimator $\beta\widehat{\mu^{\mathrm{ex}}}_1^{\mathrm{trad}} = \ln \overline{e^{\beta\Delta H(y)}}$, where y is drawn from ρ_1, the same argumentation reveals the impossibility of obtaining an accurate estimate in this way. Effectively, the particles of system 1 cannot access the vicinity of the origin, no matter how large the sample size will be. In this sense, Γ_1 can be substituted with an effective $\Gamma_1^{\mathrm{eff}} \subset \Gamma_1 = \Gamma_0$, implying that the traditional reverse estimator tends to be inconsistent.

1. Constructing a map

Again, we use a map that changes each particle's distance to the origin separately, $\phi(x) = (\mathbf{R}_1, \ldots, \mathbf{R}_{N_p})$, with $\mathbf{R}_k = \psi(r_k)\frac{\mathbf{r}_k}{r_k}$. In searching a suitable radial mapping function $\psi(r)$, we are guided by the mean radial properties of the systems themselves. The radial probability density $g_0(r)$ of finding a particle in distance r from origin in system 0 is

$$g_0(r) = \frac{1}{N_p}\sum_{k=1}^{N_p} \int \delta(r_k - r)\rho_0(x)dx, \tag{64}$$

and that for system 1 is

$$g_1(r) = \frac{1}{N_p}\sum_{k=1}^{N_p} \int \delta(r_k - r)\rho_1(x)dx. \tag{65}$$

Due to the interaction with the extra particle fixed at the origin in system 1, $g_1(r)$ will in general be quite different from $g_0(r)$. The latter is related to a homogeneous fluid and is proportional to r^2 (for $r < R_{\mathrm{box}}$), whereas the former refers to an inhomogeneous one and is proportional to $r^2 e^{-\beta V(r)}$ in the limit $r \to 0$ [26]. For large r, however, the influence of the extra particle vanishes and $g_1(r) \to g_0(r)$. Evaluation of the definition (64) of g_0 yields

$$g_0(r) = \frac{r^2}{V_c} h_0(r), \tag{66}$$

where $h_0(r)$ accounts for the decay of volume in the corners of the confining box and is given by $h_0(r) = \int\int_{A(r)} \sin\theta d\phi d\theta$. The integration extends over the fraction of surface $A(r)$ of a sphere with radius r that lies inside the confining box. Note that $h_0(r) = 4\pi$ for $r < R_{\mathrm{box}}$. In contrast to g_0, g_1 depends on the interaction $V(r)$. After some transformations of the right-hand side of Eq. (65), g_1 can be written

$$g_1(r) = \frac{r^2 e^{-\beta V(r)}}{V_c} h_1(r). \tag{67}$$

The function $h_1(r)$ can be written (cf. [26])

$$h_1(r) = e^{2\beta\mu^{\mathrm{ex}}} h_0(r) \left\langle \exp\left(-\beta \sum_{k=1}^{N_p-1} [V(r_k) + V(|\mathbf{r}_k - \mathbf{r}_{N_p}|)] \right) \right\rangle_{(N_p-1)}, \tag{68}$$

where the angular brackets denote an average with a $N_p - 1$ particle density according to Eq. (56) and the vector $\mathbf{r}_{N_p}$ is arbitrarily fixed, but of magnitude r. Further, the approximation $V_c^2 Z_{N_p-1}/Z_{N_p+1} \approx e^{2\mu^{\mathrm{ex}}}$ is used.

We note that the ratio of $g_1(r)/g_0(r)$ equals the well-known radial distribution function of the N_p+1-particle fluid,

$$\Pi(r) = \frac{g_1(r)}{g_0(r)}. \tag{69}$$

Figure 4 shows estimates of g_0 and g_1 for a dense Lennard-Jones fluid with parameter values of argon [see below Eq. (48)], obtained from Monte Carlo simulations.

Now define a function $\psi^*(r)$ by requiring that it maps the mean radial behavior of system 0 to that of system 1. This is done by demanding

$$\int_0^{\psi^*(r)} g_1(t)dt = \int_0^r g_0(t)dt, \tag{70}$$

which yields

$$\frac{\partial\psi^*}{\partial r} = \frac{g_0(r)}{g_1(\psi^*(r))}. \tag{71}$$

In the limiting case of an ideal gas, $g_1 = g_0$ holds and the map becomes an identity, $\psi^*(r) = r$. Of practical interest are the

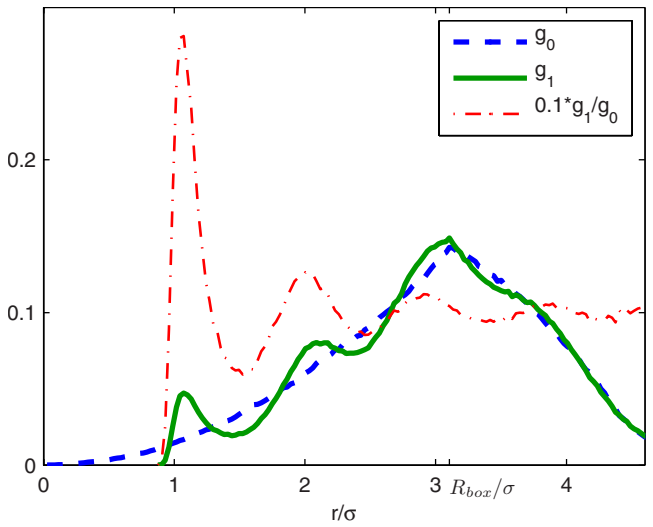

FIG. 4. (Color online) The radial densities $g_1(r)$ and $g_0(r)$ for a dense Lennard-Jones fluid ($\rho^*=0.9$ and $T^*=1.2$), estimated from simulated data. The ratio $g_1(r)/g_0(r)$ equals the radial distribution function.

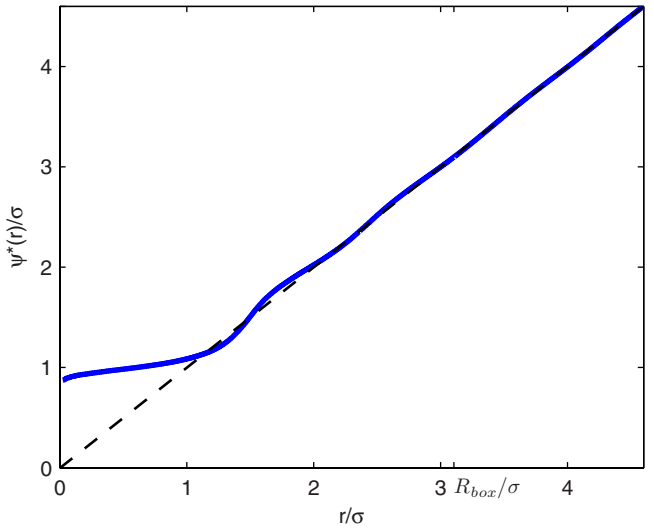

FIG. 5. (Color online) Simulated radial mapping function $\psi^*(r)$ for a dense Lennard-Jones fluid (solid). ψ^* maps the radial density $g_0(r)$ to $g_1(r)$, cf. Fig. 4. For the ideal gas, ψ^* is the identity map (dashed).

cases where g_1 is unknown and thus Eq. (70) cannot be used to derive $\psi^*(r)$. However, the function ψ^* can be estimated with Monte Carlo simulations without knowledge of g_1 and g_0 as follows.

Take a sufficiently large amount n of samples $x_j=(\mathbf{r}_{1j},\ldots,\mathbf{r}_{N_pj})$, $j=1,\ldots,n$, drawn from $\rho_0(x)$ together with the same number of samples $y_j=(\mathbf{R}_{1j},\ldots,\mathbf{R}_{N_pj})$ drawn from $\rho_1(y)$. Calculate the distances to the origin $r_{ij}=|\mathbf{r}_{ij}|$ and $R_{ij}=|\mathbf{R}_{ij}|$ and combine all r_{ij} to the set $(r_a,r_b,r_c\ldots)$, as well as all R_{ij} to the set $(R_a,R_b,R_c,\ldots)$. Provided in both sets the elements are ordered ascending, $r_a\leqslant r_b\leqslant r_c\leqslant\ldots$ and $R_a\leqslant R_b\leqslant R_c\leqslant\ldots$, ψ^* is simulated by constructing a one-to-one correspondence $r_a\to R_a$, $r_b\to R_b,\ldots$ and estimating $\psi^*(r_\alpha)$ to be R_α, $\alpha=a,b,c,\ldots$. In effect, we have drawn the r_α and R_α from the densities $g_0(r)$ and $g_1(r)$, respectively, and have established a one-to-one correspondence between the ordered samples. We refer to this scheme as the simulation of the map of g_0 to g_1.

The solid curve shown in Fig. 5 is the result of a simulation of the function ψ^* for a Lennard-Jones fluid (parameters of argon, $\rho^*=0.9$, $T^*=1.2$). The corresponding densities g_0 and g_1 are plotted in Fig. 4. Noticeable is the sudden "start" of ψ^* with a value of roughly σ. This is due to the strong repulsive part of the interaction that keeps particles in system 1 approximately a distance σ away from the origin. Therefore, the behavior of $\psi^*(r)$ for $r\to 0$ is not obtainable from finite-time simulations. However, the definition of ψ^* implies that for any soft-core potential $\psi^*(0)=0$ holds. To model ψ^* for small r, the limit $g_1(r)\overset{r\to 0}{\to} ar^2e^{-\beta V(r)}4\pi/V_c$ can be used, where a is a constant. Thus, Eq. (70) can be written

$$[\psi^{*-1}(r)]^3=3a\int_0^r r'^2e^{-\beta V(r')}dr' \tag{72}$$

in the limit $r\to 0$, with ψ^{*-1} being the inverse of ψ^*. The constant a is in general unknown, but here it can be chosen such that a continuous fit to the simulated part of ψ^{*-1} is obtained.

When the function ψ^* is used in the configuration space map ϕ according to Eq. (50), then, by definition of ψ^*, the radial density $\tilde{g}_0(r)$ of the mapped distribution $\tilde{\rho}_0(x)$, Eq. (8), is identical to the one of $\rho_1(x)$,

$$\begin{aligned}\tilde{g}_0(R) &:= \frac{1}{N_p}\sum_k\int\delta(|\mathbf{R}_k|-R)\tilde{\rho}_0(\phi)d\phi\\ &=\int\delta[\psi(r_1)-R]\rho_0(x)dx\\ &=\int\int\delta[\psi(r)-R]\delta(r_1-r)\rho_0(x)dxdr\\ &=\int\delta[\psi(r)-R]g_0(r)dr=g_1(R).\end{aligned} \tag{73}$$

Therefore we expect that the overlap of the mapped distribution $\tilde{\rho}_0$ with ρ_1 is larger than the overlap of the unmapped distribution ρ_0 with ρ_1. However, it must be noted that the use of ψ^* in the map ϕ is in general valid only in the limit of an infinite large system ($N,V_c\to\infty$; $N/V_c=$const), since we have not yet taken into account the requirement that particles may not be mapped out of the confining box. If R_{box} is chosen large enough, this might not be a serious problem, cf. Fig. 5.

2. Application of the radial map ψ^*

We now apply ψ^* and estimate the chemical potential of a dense Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$, parameters of argon) with $R_{\text{box}}=3.1056\sigma$ and $N_p=216$ particles. Configurations are drawn from ρ_0 and ρ_1 using a Metropolis algorithm with seven decorrelation sweeps between successive drawings. From every drawn configuration there results one value for the traditional work and one for the work related to

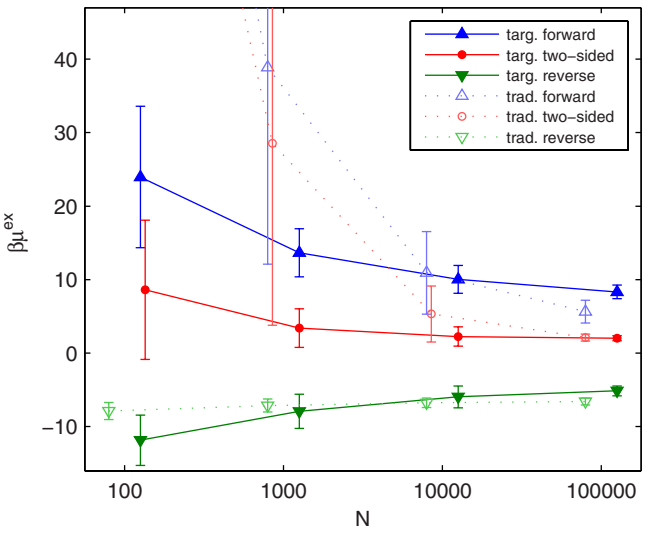

FIG. 6. (Color online) Targeted estimates of the excess chemical potential μ^{ex} of a dense Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) compared to traditional estimates.

the map. The usual cutoff corrections [6] are applied. To avoid mapping particles out of the confining box, we simulate the map on the interval $0 \leqslant r \leqslant R_{\mathrm{box}}$ subject to the condition $\psi^*(R_{\mathrm{box}})=R_{\mathrm{box}}$ and use $\psi^*(r)=r$ for $r>R_{\mathrm{box}}$. The derivatives of ψ^* and ψ^{*-1} are obtained numerically. For the calculation of the work values in the simulation, the functions $\psi^*(r)$ and $\psi^{*-1}(r)$ as well as their derivatives are discretized in steps Δr with $R_{\mathrm{box}}/\Delta r=11\times 10^4$.

A comparison of the behavior of the targeted and traditional forward, reverse, and two-sided estimators in dependence of the sample size N is given in Fig. 6 (for the two-sided estimators $n_0=n_1=\frac{N}{2}$ is used). Each data point represents the average value of $z(N)$ independent estimates $\widetilde{\mu^{\mathrm{ex}}}(N)$. The error bars display one standard deviation. $z(N)$ reads $z(N)=450,250,45,5$ for $N=100,1000,10\,000,100\,000$, respectively.

As can be seen from Fig. 6, the traditional one-sided estimators behave quite different. The reverse estimator converges extremely slow in comparison to the forward estimator. This can be understood by comparing the average work values $\overline{W^i}$ in forward ($i=0$) and reverse ($i=1$) direction, see Table II. Since the absolute value of $\beta\Delta F=\beta\mu^{\mathrm{ex}}$ is small, the traditional reverse estimator practically never converges, whereas for an accurate traditional forward estimate we need some 10^5 work values, cf. Eqs. (34) and (35). In contrast, the targeted one-sided estimators both show a similar convergence behavior if compared with each other. However, the convergence is slow.

The two-sided estimators converge much faster, in particular, the targeted two-sided estimator converges fastest,

TABLE II. Estimated values of the mean forward and reverse work, obtained from $N=10^5$ sampled work values each.

	$\beta\overline{W^0}$	$\beta\overline{W^1}$
Traditional	10^{20}	−9.8
Targeted	10^5	-10^6

see Fig. 6. The convergence of the latter was checked with the convergence measure $a(N)$, Eq. (46). For instance, the convergence measure $a(N)$ takes the values 0.08 and 0.01 for the traditional and the targeted estimator, respectively, if $n_0=n_1=\frac{N}{2}=10^5$ configurations per direction are sampled.

Investigating the histograms of the generalized work distributions in the traditional and the targeted case visualizes the effectiveness of the mapping. The histograms are similar to those displayed in Fig. 8 for $m=0$ (traditional) and $m=0.005$ (targeted).

A moderate gain in precision for the two-sided targeted estimator is found if compared to the precision of the two-sided traditional estimator which can be quantified with the overlap measure $\hat{A}_{\mathrm{ol}}$ (43). Namely, $\hat{A}_{\mathrm{ol}}=3.0\times 10^{-4}$ for the targeted case, and $\hat{A}_{\mathrm{ol}}=2.2\times 10^{-4}$ for the traditional case.

We also studied other radial mapping functions ψ. Some of them turned out to give much better results and are easier to deal with.

3. Other radial mapping functions

The radial mapping function ψ^* was obtained from simulations, because the distribution $g_1(r)$ is analytically unknown. However, we are free to use any radial mapping function $\psi(r)$ and can thus in turn fix the function g_1 appearing in Eq. (70). To do this, we introduce the normalized, positive definite function $g_1'(r)$,

$$g_1'(r)=\frac{r^2}{c_1}e^{-\beta[V(r)+Q(r)]},\quad r\in[0,R_{\mathrm{box}}]. \tag{74}$$

$Q(r)$ is an arbitrary finite function over $(0,R_{\mathrm{box}}]$ and $c_1=\int_0^{R_{\mathrm{box}}}r^2e^{-\beta[V(r)+Q(r)]}dr$ a normalization constant. Further, let $g_0'(r)$ be a normalized quadratic density,

$$g_0'(r)=\frac{r^2}{c_0},\quad r\in[0,R_{\mathrm{box}}], \tag{75}$$

with $c_0=R_{\mathrm{box}}^3/3$.

The general (monotonically increasing) radial mapping function $\psi(r)$ can be expressed in terms of the equation

$$\int_0^{\psi(r)}g_1'(t)dt=\int_0^r g_0'(t)dt \tag{76}$$

for $r\in[0,R_{\mathrm{box}}]$. For $r>R_{\mathrm{box}}$ it shall be understood that $\psi(r)=r$. Given the function $Q(r)$, ψ and ψ^{-1} are determined uniquely by Eq. (76). An advantage of defining ψ with Eq. (76) is that the derivative $\partial\psi/\partial r$ is given in terms of V and Q,

$$\frac{\partial\psi(r)}{\partial r}=\frac{r^2}{\psi(r)^2}e^{\beta\{V(\psi(r))+Q(\psi(r))-f\}}, \tag{77}$$

with $f=-\frac{1}{\beta}\ln\frac{c_1}{c_0}$. Using ψ in the configuration space map $\phi(x)$ according to Eq. (50) yields the work function

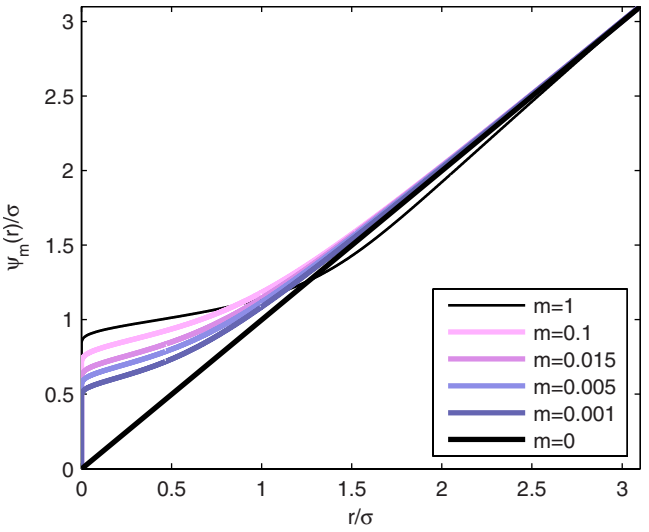

FIG. 7. (Color online) Members of the family of radial mapping functions ψ_m for the Lennard-Jones potential. For $m \to 0$, ψ_m converges to the identity map $\psi_0(r)=r$.

$$\widetilde{\Delta H}(x) = \sum_{i<j}^{N_p} \{V(|\mathbf{R}_i - \mathbf{R}_j|) - V(|\mathbf{r}_i - \mathbf{r}_j|)\} - \sum_{r_i \leqslant R_{\text{box}}} \{Q(\psi(r_i)) - f\}. \tag{78}$$

Here $\mathbf{R}_i$ is understood to be $\mathbf{R}_i = \psi(r_i)\frac{\mathbf{r}_i}{r_i}$, and the sum in the second line extends only over those particles for which $r \leqslant R_{\text{box}}$ holds. Note that the potential-energy contribution of the extra particle fixed at the origin is eliminated in the work function, due to the definition of ψ. However, in Eq. (78) we have already assumed $V(r)$ to be cut off at $r=R_{\text{box}}$, i.e., $V(r)=0$ for $r \geqslant R_{\text{box}}$. Otherwise we had to add $\Sigma_{r_i>R_{\text{box}}} V(\psi(r_i)) = \Sigma_{r_i>R_{\text{box}}} V(r_i)$ to the right-hand side of Eq. (78).

4. A family of maps

We now introduce a family $\{\psi_m\}$ of radial mapping functions, where each member ψ_m is defined by Eq. (76) with the choice

$$Q(r) = (m-1)V(r) \tag{79}$$

in the expression (74). This choice is motivated by the following: Consider a one particle system, $N_p=1$. In this case the optimal radial map can be computed analytically and results in $g_1'(r) = \frac{r^2}{c_1} e^{-\beta V(r)}$. We formally use this map for a system of $N_p \gg 1$ particles, but weaken the potential $V(r)$ by multiplying it with a small parameter m, i.e., $g_1'(r) = \frac{r^2}{c_1} e^{-\beta m V(r)}$, since the potential is screened by the N_p-1 other particles. This results in Eq. (79). Useful maps are obtained for $m \in [0,1]$. Since we have no *a priori* knowledge on the optimal value of m, we determine the best value of m numerically.

Figure 7 depicts some members of the family $\{\psi_m\}$ for Lennard-Jones interaction (with parameters of argon). Again, we apply these functions discretized (in steps Δr with $R_{\text{box}}/\Delta r = 11 \times 10^4$) in the calculation of the targeted forward and reverse work $\widetilde{\Delta H}(x)$ and $\widetilde{\Delta H}(\phi^{-1}(x))$. Any pair of forward and reverse targeted work distributions belonging to the same value of m obeys the fluctuation theorem (19). In particular they cross at $W=\mu^{\text{ex}}$ ($\Delta F=\mu^{\text{ex}}$ here), see Fig. 9. Nevertheless, the shape of these distributions is sensitive to the value of m. This is demonstrated in Fig. 8. There, normalized histograms of βW are shown. They result from 10^4 work values per m and per direction. We emphasize that all of the targeted forward (reverse) work values were obtained with *one* sample of $N=10^4$ configurations x from ρ_0 (ρ_1). Figure 9

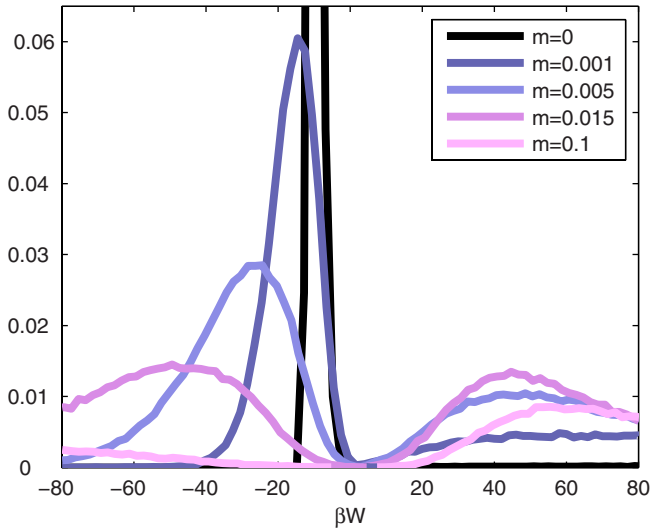

FIG. 8. (Color online) Forward (right) and reverse (left) work distributions of a Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) for different radial mapping functions ψ_m. ($m=0$ results in the traditional work distributions.)

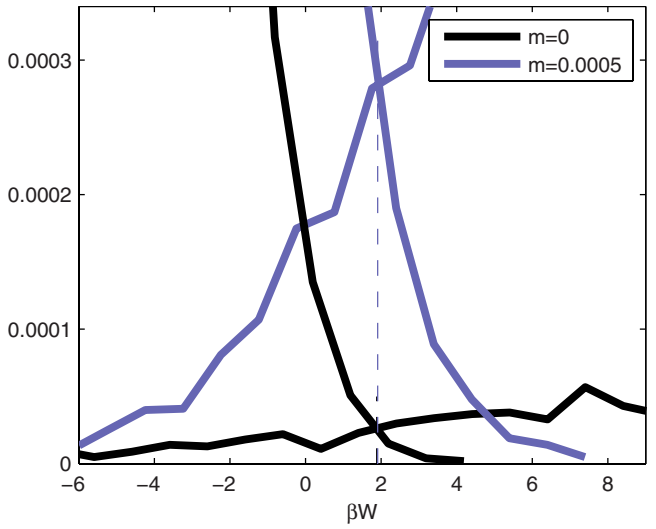

FIG. 9. (Color online) A detailed enlargement of forward (ascending lines) and reverse (descending lines) work distributions of the Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) for two different radial mapping functions ψ_m. Notice the enhancement of overlap for $m=0.0005$. The vertical dashed line displays the estimated two-sided targeted value, $\widehat{\beta\mu^{\text{ex}}}_{01}=1.91$.

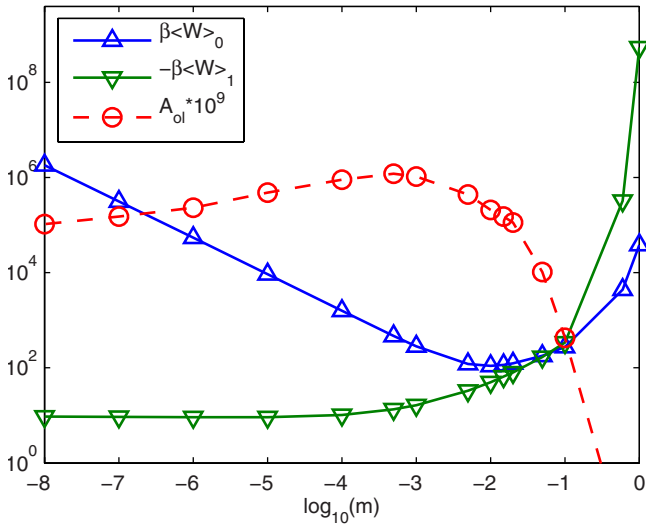

FIG. 10. (Color online) The average generalized work $\langle W\rangle_0$ and $\langle W\rangle_1$ in forward and reverse direction, respectively, and the two-sided overlap measure A_{ol} in dependence of the mapping parameter m. The forward dissipation is reduced up to 18 orders of magnitude if compared with the traditional dissipation, cf. Table II. Among the one-sided estimators the best is found for $m=0$ in forward direction. The optimal two-sided estimator results from using the m that maximizes A_{ol}.

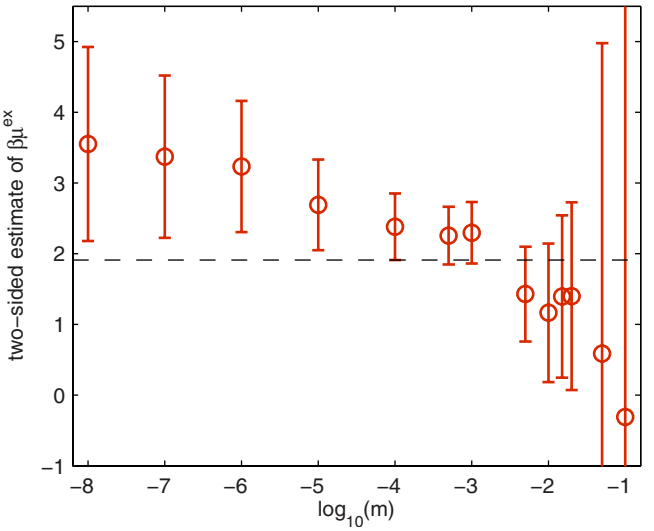

FIG. 11. (Color online) Two-sided estimates of μ^{ex} as function of the mapping parameter m out of $n_0=n_1=\frac{N}{2}=10^4$ work values per direction for each m. The value of the traditional estimate ($m=0$) is $\widehat{\mu^{\mathrm{ex}}}_{01}=4.0\pm2.0$. The error bars show the square root of the estimated mean square errors X_{01}. For comparison, the dashed line represents a two-sided estimate with $\frac{N}{2}=10^6$ and $m=0.0005$ (standard deviation 0.03).

is a detailed enlargement where a sample of $n_0=n_1=10^6$ forward (reverse) configurations is used.

Instructive is the comparison of the mean work $\langle W\rangle$ related to different values of m. In Fig. 10 estimated values of mean work are shown in dependence of m. From these values one sees that the dissipation is minimal for $m=0$ in the reverse direction. Therefore, the best one-sided targeted estimate of μ^{ex} among the family $\{\psi_m\}$ is obtained with $m=0$ in forward direction, i.e., with the traditional particle insertion. However, the same is not true for two-sided estimates. Using the same data as before and performing two-sided estimates with $n_0=n_1=\frac{N}{2}=10^4$ work values per direction, we obtain the displayed values $\widehat{\mu^{\mathrm{ex}}}_{01}$ of Fig. 11. In order to compare the performance of two-sided estimators for different maps, we estimate the overlap measures A_{ol}. The latter are shown in Fig. 10. The maximum value for A_{ol} is found with m being 0.0005. This indicates that $m\approx0.0005$ is the optimal choice for m. The estimates $\hat{A}_{\mathrm{ol}}$ are used to calculate the mean square errors X_{01} of the estimates $\widehat{\mu^{\mathrm{ex}}}_{01}$. The square roots of the X_{01} enter in Fig. 11 as error bars.

We are left to check the convergence properties of two-sided estimators. Figure 12 displays the convergence measure $a(N)$ for some parameter values m. Best convergence is found for $m=0.0005$ (not shown in Fig. 12, but very similar to $m=0.001$). The same value of the mapping parameter m was found to maximize the overlap A_{ol}.

Employing the optimal value 0.0005 for the mapping parameter and using $n_0=n_1=\frac{N}{2}=10^6$ forward and reverse samples, we have computed the chemical potential. The results are given in Table III. The listed error is the square root of the X_{01} according to Eq. (42) with $A_{\mathrm{ol}}=\hat{A}_{\mathrm{ol}}$. This is justified with the observed values of the convergence measure a which are listed in the table, too.

It should be mentioned that the optimal value of m found here is not universal, but depends on the density ρ^*. If another value is chosen for ρ^*, the optimal m can again be found from numerical simulations. Note that the maps used here can be applied to simulations where particles are inserted and deleted at random [26], too. One simply has to use the point of insertion (deletion) as temporary origin of the coordinate system and apply the map there. This might enhance the efficiency of the simulation.

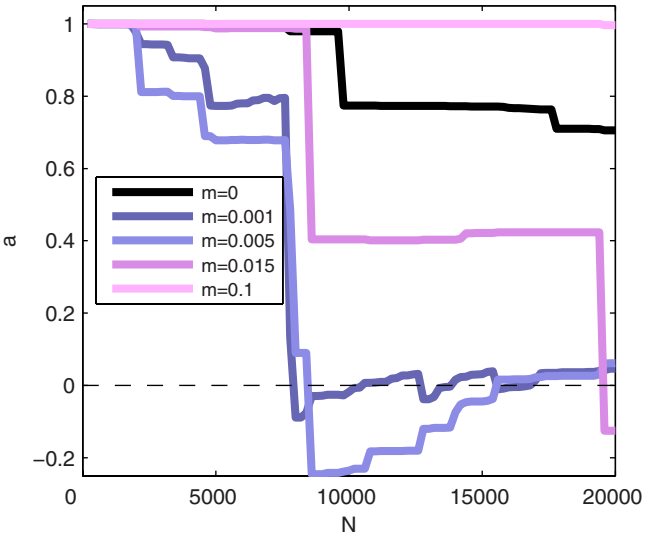

FIG. 12. (Color online) Convergence measure $a(N)$ of two-sided estimates for some parameter values m, depending on the sample size N. A faster decrease of a towards the value 0 indicates a faster convergence of the two-sided estimator.

TABLE III. Two-sided estimates $\widehat{\mu^{\mathrm{ex}}}_{01}$ of the excess chemical potential of a Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$). Also listed is the two-sided overlap measure A_{ol} and the convergence measure a. For the targeted estimate the radial mapping function ψ_m with $m=0.0005$ is used. The number of work values in each direction is 10^6 and the number of particles in the simulation is $N_p=216$.

	$\beta\widehat{\mu^{\mathrm{ex}}}_{01}$	$10^4\hat{A}_{\mathrm{ol}}$	a
Traditional	1.88 ± 0.08	2.4	0.05
Targeted	1.91 ± 0.03	19	–0.02

VIII. CONCLUSION

The central result of this paper, a fluctuation theorem for generalized work distributions, allowed us to establish an optimal targeted two-sided estimator of the free-energy difference ΔF. We have numerically tested this estimator and found it to be superior with respect to one-sided and nontargeted estimators. In addition we have demonstrated that this estimator can be applied successfully to estimate the chemical potential of a Lennard-Jones fluid in the high density regime.

In order to use the targeted two-sided estimator it is however crucial to use a suitable map. We have investigated the construction of maps and developed appropriate measures which enabled a quantitative comparison of the performance of different maps. Especially, a measure for the convergence of the two-sided estimate was designed. This points the way for better results when free-energy differences or chemical potentials need to be estimated numerically.

ACKNOWLEDGMENT

We thank Andreas Engel for helpful hints and discussions.

APPENDIX: CONSTRAINT MAXIMUM LIKELIHOOD DERIVATION OF THE TWO-SIDED ESTIMATOR

Deriving the optimal estimator of ΔF, given a collection of n_0 forward $\{W_i^0\}$ and n_1 reverse $\{W_j^1\}$ work values drawn from $p(W|0)$ and $p(W|1)$, respectively, leads to Bennett's acceptance ratio method [13] with the target map included.

In Sec. IV, the mixed ensemble is introduced, where the elements are given by pairs of values (W,Y) of work and direction, and which is specified by the probabilities of direction p_Y and the densities $p(W|Y)$. With the mixture ensemble, the mixing ratio $\frac{p_1}{p_0}$ can be chosen arbitrarily. Crucial about the mixture ensemble is that, according to the fluctuation theorem (19), the analytic form of the conditional probabilities $p(Y|W)$ can be derived explicitly, regardless of whether $p(W|Y)$ is known, see Sec. IV. This provides a natural way to construct a constraint maximum likelihood estimator [27–29] for ΔF.

Since it is only possible to draw from the ensembles $p(W|Y)$, but not from $p(Y|W)$, $Y=0,1$, the proper log-likelihood is

$$\ln \mathcal{L} = \sum_{i=1}^{n_0} \ln p(W_i^0|0) + \sum_{j=1}^{n_1} \ln p(W_j^1|1). \tag{A1}$$

A *direct* maximization of Eq. (A1) with respect to ΔF is impossible without knowledge of the analytic form of the probability densities $p(W|Y)$. However, according to Bayes theorem (24) the likelihood can be split into

$$\ln \mathcal{L} = \ln \mathcal{L}_{\mathrm{post}}(\Delta F) + \ln \mathcal{L}_{\mathrm{prior}} + \ln \mathcal{L}_{p_Y} \tag{A2}$$

with

$$\ln \mathcal{L}_{\mathrm{post}}(\Delta F) = \sum_{i=1}^{n_0} \ln p(0|W_i^0) + \sum_{j=1}^{n_1} \ln p(1|W_j^1), \tag{A3}$$

$$\ln \mathcal{L}_{\mathrm{prior}} = \sum_{k=1}^{n_0+n_1} \ln p(W_k), \tag{A4}$$

and

$$\ln \mathcal{L}_{p_Y} = n_0 \ln \frac{1}{p_0} + n_1 \ln \frac{1}{p_1}, \tag{A5}$$

where the sum in the prior likelihood (A4) runs over all n observed forward and reverse work values.

Since the definite form of $p(W)$ is unknown, we treat it in the manner of an unstructured prior distribution and maximize (A2) with respect to the constant ΔF *and* to the function $p(W)$ [29]. Thereby,

$$1 = \int p(W)dW \tag{A6}$$

and

$$p_1 = \int p(1|W)p(W)dW \tag{A7}$$

enter as constraints. Using Lagrange parameters λ and μ, the constrained likelihood reads

$$\ln \mathcal{L}^{\mathrm{c}} = \ln \mathcal{L} + \lambda\left(p_1 - \int p(1|W)p(W)dW\right) + \mu\left(1 - \int p(W)dW\right). \tag{A8}$$

The conditional direction probabilities $p(Y|W)$ are known explicitly in dependence of ΔF, Eq. (28), and their partial derivatives read $\frac{1}{\beta}\frac{\partial}{\partial \Delta F}\ln p(0|W) = -p(1|W)$ and $\frac{1}{\beta}\frac{\partial}{\partial \Delta F}\ln p(1|W) = p(0|W) = 1-p(1|W)$. This allows one to extremize the constraint likelihood (A8) with respect to ΔF,

$$0 = \frac{1}{\beta}\frac{\partial}{\partial \Delta F}\ln \mathcal{L}^{\mathrm{c}} = n_1 - \sum_{k=1}^{n_0+n_1} p(1|W_k) - \lambda \int [1-p(1|W)]p(1|W)p(W)dW. \tag{A9}$$

Extremizing the conditional likelihood (A8) with respect to the function $p(W)$ gives

$$0 = \frac{\delta}{\delta p(W)} \ln \mathcal{L}^{c} = \frac{1}{p(W)} \sum_{k=1}^{n} \delta(W - W_k) - \lambda p(1|W) - \mu, \tag{A10}$$

which can be solved in $p(W)$,

$$p(W) = \frac{\Sigma_k \delta(W - W_k)}{\lambda p(1|W) + \mu}, \tag{A11}$$

or written as

$$\lambda p(1|W) p(W) = -\mu p(W) + \sum_k \delta(W - W_k). \tag{A12}$$

If interested in the values of the Lagrange multipliers λ and μ, one multiplies Eq. (A10) with $p(W)$ and integrates. This yields

$$0 = n - \lambda p_1 - \mu. \tag{A13}$$

A second independent equation follows from inserting Eq. (A12) into Eq. (A9) which results in

$$0 = n_1 + \mu - \mu p_1 - n, \tag{A14}$$

and the Lagrange multipliers take the values

$$\mu = \frac{n_0}{p_0} \quad \text{and} \quad \lambda = \frac{n p_0 - n_0}{p_0 p_1}. \tag{A15}$$

With the distribution (A11) the constraints (A6) and (A7) read

$$1 = \sum_k \frac{1}{\lambda p(1|W_k) + \mu} \tag{A16}$$

and

$$p_1 = \sum_k \frac{p(1|W_k)}{\lambda p(1|W_k) + \mu} = \frac{p_1}{n_1} \sum_k p^{B}(1|W_k), \tag{A17}$$

where $p^{B}(1|W)$ denotes $p(1|W)$ with $C = \Delta F + \frac{1}{\beta} \ln \frac{n_1}{n_0}$. Whenever the constraint (A17) is fulfilled, the constraint (A16) and the variational equations (A9) and (A10) are automatically satisfied. In consequence, Eq. (A17) defines the constrained maximum likelihood estimate of ΔF. Note that the estimator (A17) is *independent* of the choice of $\frac{p_1}{p_0}$. Moreover, Eq. (A17) is equivalent to Eq. (29) regardless of the choice of $\frac{p_1}{p_0}$.

An alternative derivation of the estimator (A17) was presented by Shirts *et al.* [15]. There, the specific choice $\frac{p_1}{p_0} = \frac{n_1}{n_0}$ was *necessary*. With this choice, the Lagrange parameter λ is identical to zero. Hence, there is no need to take any constraint into consideration and the posterior likelihood (A3) results directly in the estimator of ΔF.

[1] M. R. Reddy and M. D. Erion, *Free Energy Calculations in Rational Drug Design* (Kluwer Academic, New York, 2001).
[2] T. Schäfer and E. V. Shuryak, Rev. Mod. Phys. **70**, 323 (1998).
[3] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[4] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[5] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[6] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, London, 2002).
[7] C. Jarzynski, Phys. Rev. E **65**, 046122 (2002).
[8] W. Lechner, H. Oberhofer, C. Dellago, and P. L. Geissler, J. Chem. Phys. **124**, 044113 (2006).
[9] H. Oberhofer, C. Dellago, and S. Boresch, Phys. Rev. E **75**, 061106 (2007).
[10] A. F. Voter, J. Chem. Phys. **82**, 1890 (1985).
[11] F. M. Ytreberg and D. M. Zuckerman, J. Phys. Chem. B **109**, 9096 (2005).
[12] S. Vaikuntanathan and C. Jarzynski, Phys. Rev. Lett. **100**, 190601 (2008).
[13] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[14] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).
[15] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[16] E. Schöll-Paschinger and C. Dellago, J. Chem. Phys. **125**, 054105 (2006).
[17] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[18] C. Jarzynski, J. Stat. Phys. **98**, 77 (2000).
[19] D. J. Evans, Mol. Phys. **101**, 1551 (2003).
[20] M. A. Cuendet, Phys. Rev. Lett. **96**, 120602 (2006).
[21] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **100**, 12564 (2003).
[22] C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).
[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[24] R. C. Reid, J. M. Prausnitz, and T. K. Sherwood, *The Properties of Gases and Liquids*, 3rd ed. (McGraw-Hill, New York, 1977), Appendixes A and C.
[25] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).
[26] B. Widom, J. Chem. Phys. **39**, 2808 (1963).
[27] J. Aitchison and S. D. Silvey, Ann. Math. Stat. **29**, 813 (1958).
[28] J. A. Anderson, Biometrika **59**, 19 (1972).
[29] R. L. Prentice and R. Pyke, Biometrika **66**, 403 (1979).