Escorted free energy simulations

Suriyanarayanan Vaikuntanathan, and Christopher Jarzynski

View Online

Export Citation

ARTICLES YOU MAY BE INTERESTED IN

Replica exchange with nonequilibrium switches: Enhancing equilibrium sampling by increasing replica overlap
The Journal of Chemical Physics **136**, 194101 (2012); https://doi.org/10.1063/1.4712028

Statistically optimal analysis of samples from multiple equilibrium states
The Journal of Chemical Physics **129**, 124105 (2008); https://doi.org/10.1063/1.2978177

Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality
The Journal of Chemical Physics **119**, 3559 (2003); https://doi.org/10.1063/1.1590311

AIP Publishing

Escorted free energy simulations

Suriyanarayanan Vaikuntanathan[1,a)] and Christopher Jarzynski[1,2]
[1]*Chemical Physics Program, Institute for Physical Science and Technology, University of Maryland,
College Park, Maryland 20742, USA*
[2]*Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, USA*

We describe a strategy to improve the efficiency of free energy estimates by reducing dissipation in nonequilibrium Monte Carlo simulations. This strategy generalizes the targeted free energy perturbation approach [C. Jarzynski, Phys. Rev. E **65**, 046122 (2002)] to nonequilibrium switching simulations, and involves generating artificial, "escorted" trajectories by coupling the evolution of the system to updates in external work parameter. Our central results are: (1) a generalized fluctuation theorem for the escorted trajectories, and (2) estimators for the free energy difference ΔF in terms of these trajectories. We illustrate the method and its effectiveness on model systems. © *2011 American Institute of Physics*. [doi:10.1063/1.3544679]

I. INTRODUCTION

The computation of free energy differences is an essential component of computer studies of biological, chemical, and molecular processes, with applications to topics such as phase coexistence and phase equilibria, ligand binding events, and solvation of small molecules.[1,2] Given the importance of free energy calculations in computational thermodynamics, there is a need for robust, efficient, and accurate methods to estimate free energy differences.

In a standard formulation of the free energy estimation problem, we consider two equilibrium states of a system, corresponding to the same temperature T but different values of an external parameter, $\lambda = A, B$, and we are interested in the free energy difference between the two states, $\Delta F = F_B - F_A$. While many widely used free energy estimation methods, such as thermodynamic integration and free energy perturbation rely on equilibrium sampling, there has been considerable interest in methods for estimating ΔF that make use of nonequilibrium simulations.[1,2] In the most direct implementation of this approach, a number of independent simulations are performed in which the external parameter is varied at a finite rate from $\lambda = A$ to $\lambda = B$, with initial conditions sampled from the equilibrium state A. The free energy difference ΔF can then be estimated using the nonequilibrium work relation[3,4]

$$e^{-\beta \Delta F} = \langle e^{-\beta W} \rangle, \tag{1}$$

where W denotes the work performed on the system during a particular realization (i.e., simulation) of the process, angular brackets $\langle \ldots \rangle$ denote an average over the realizations of the process and $\beta = 1/T$. In principle, this approach allows one to compute ΔF from trajectories of arbitrarily short duration. However, the number of realizations required to obtain a reliable estimate of ΔF grows rapidly with the dissipation, $\langle W_{\text{diss}} \rangle \equiv \langle W \rangle - \Delta F$, that accompanies fast

switching simulations.[5–7] The dissipation is positive as a consequence of the second law of thermodynamics, and reflects the lag that builds up as the system pursues—but is unable to keep pace with—the equilibrium distribution corresponding to the continuously changing parameter λ.[8–11] This idea is illustrated schematically in Fig. 1.

In Ref. 12, we described a strategy to improve the efficiency of free energy estimates obtained with nonequilibrium molecular-dynamics simulations. This strategy involved adding nonphysical terms to the equations of motion, to reduce the lag and therefore the dissipation. As illustrated in Ref.12 using a simple model system, when these terms successfully "escorted" the system through a near-equilibrium sequence of states, the convergence of the free energy estimate improved dramatically. In the present paper we extend these results to simulations evolving according to Monte Carlo dynamics. We then show that the escorted trajectories satisfy a fluctuation theorem, and we discuss and illustrate the application of this result to the estimation of free energy differences.

In Sec. II we introduce escorted nonequilibrium switching simulations for systems evolving according to Monte Carlo dynamics. The approach we take here is motivated by previous work[12–15] and involves generating artificial, or "escorted," trajectories, Eq. (10), by modifying the dynamics with terms that directly couple the evolution of the system to changes in the external parameter. The central result of this section is an identity for ΔF in terms of these escorted trajectories, Eq. (19). In Sec. III we extend this result by showing that these trajectories satisfy a fluctuation relation analogous to Crooks's fluctuation relation.[16–18] This in turn allows us to combine our approach with Bennett's acceptance ratio method[19] which provides an optimal, asymptotically unbiased estimator, Eq. (38), for ΔF (Ref. 20). In Sec. IV, we show that while Eqs. (19) and (38) are identities for all escorted simulations, they are particularly effective as estimators of ΔF when the modified dynamics successfully reduce the lag described above. In particular, if these terms eliminate the lag entirely, then Eqs. (19) and (38) provide

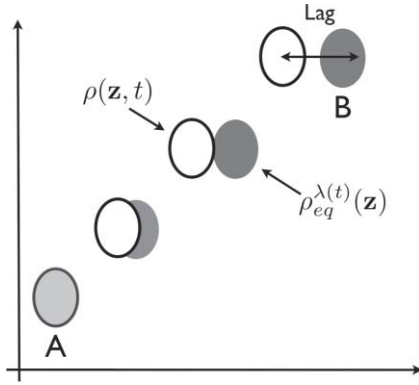a)Author to whom correspondence should be addressed. Electronic address: svaikunt@umd.edu.

FIG. 1. The axes schematically represent configuration space (**z**-space). The unshaded ovals denote the statistical state of the system, $\rho(\mathbf{z}, t)$, and the shaded ovals denote the equilibrium state, $\rho_{\text{eq}}^{\lambda(t)}(\mathbf{z})$, corresponding to the value of external parameter, $\lambda(t)$, at various instants of time. As the work parameter λ is switched from A to B, a lag builds up as the state of the system, $\rho(\mathbf{z}, t)$, pursues the equilibrium distribution corresponding to the changing work parameter, $\rho_{\text{eq}}^{\lambda(t)}(\mathbf{z})$.

perfect (zero variance) estimators: $W = \Delta F$ for every realization. Finally in Sec. V, we illustrate the effectiveness of our approach on two model systems.

II. ESCORTED NONEQUILIBRIUM SIMULATIONS

Consider a system whose energy is given by a classical Hamiltonian, $H_\lambda(\mathbf{z})$, where $\mathbf{z}$ denotes a microstate, that is a point in the D-dimensional configuration space of the system,[21] and λ is an external work parameter. At a temperature β^{-1}, the equilibrium state of this system is described by the distribution

$$\rho_{\text{eq}}^\lambda(\mathbf{z}) = \frac{e^{-\beta H_\lambda(\mathbf{z})}}{Z_\lambda}, \tag{2}$$

with the free energy $F_\lambda = -\beta^{-1} \ln Z_\lambda$. We wish to compute the free energy difference $\Delta F = F_B - F_A$ between two equilibrium states at the same temperature, β^{-1}, but different values of the work parameter, $\lambda = A, B$.

To estimate the value of ΔF, we assume we have at our disposal a discrete-time Monte Carlo algorithm, parameterized by the value of λ and defined by the transition probability $P_\lambda(\mathbf{z}|\mathbf{z}_0)$: if $\mathbf{z}_0$ represents the microstate of the system at one time step, then the next microstate $\mathbf{z}$ is sampled randomly from $P_\lambda(\mathbf{z}|\mathbf{z}_0)$. We assume this algorithm satisfies the conditions of *detailed balance*,

$$\frac{P_\lambda(\mathbf{z}|\mathbf{z}_0)}{P_\lambda(\mathbf{z}_0|\mathbf{z})} = \frac{e^{-\beta H_\lambda(\mathbf{z})}}{e^{-\beta H_\lambda(\mathbf{z}_0)}}, \tag{3}$$

and *ergodicity*.[22] Routinely used Monte Carlo schemes such as the Metropolis algorithm[1] satisfy these conditions. Equation (3) implies the somewhat weaker condition of *balance*,

$$\int d\mathbf{z}_0 \, P_\lambda(\mathbf{z}|\mathbf{z}_0) \, e^{-\beta H_\lambda(\mathbf{z}_0)} = e^{-\beta H_\lambda(\mathbf{z})}, \tag{4}$$

which we will use in the analysis below. With this Monte Carlo algorithm in place, we first describe a standard procedure for estimating ΔF using nonequilibrium simulations,

Eqs. (5)–(9) below, and then we introduce our modified version of this approach.

Imagine a process in which the system is initially prepared in equilibrium, at $\lambda = A$ and temperature β^{-1}, and then the system evolves under the Monte Carlo dynamics described above, as the value of λ is switched from A to B in N steps according to some predetermined protocol. This evolution generates a trajectory $\gamma = \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{N-1}\}$ that can be represented in more detail using the notation

$$[\mathbf{z}_0, \lambda_0] \Rightarrow [\mathbf{z}_0, \lambda_1] \rightarrow [\mathbf{z}_1, \lambda_1] \Rightarrow \cdots \rightarrow [\mathbf{z}_{N-1}, \lambda_{N-1}]$$
$$\Rightarrow [\mathbf{z}_{N-1}, \lambda_N]. \tag{5}$$

Here, the symbol $\Rightarrow$ denotes an update in the value of λ, with the microstate held fixed, while $\rightarrow$ denotes a Monte Carlo step at fixed λ, e.g., the microstate $\mathbf{z}_1$ is sampled from the distribution $P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0)$. Moreover,

$$\lambda_0 \equiv A, \qquad \lambda_N \equiv B, \tag{6}$$

and the initial point $\mathbf{z}_0$ is sampled from $\rho_{\text{eq}}^A(\mathbf{z}_0)$.

Because it is specified by the sequence of microstates $\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}$, the trajectory γ can be viewed as a point in a DN-dimensional trajectory space, with $d\gamma = d\mathbf{z}_0, \ldots, d\mathbf{z}_{N-1}$. For the process described in the previous paragraph, the probability density for generating this trajectory is

$$p[\gamma] = P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}_{N-2}) \cdots P_{\lambda_2}(\mathbf{z}_2|\mathbf{z}_1) \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}_0) \, \rho_{eq}^A(\mathbf{z}_0), \tag{7}$$

where the factors $P_{\lambda_i}(\mathbf{z}_i|\mathbf{z}_{i-1})$ in this equation (read from right to left) correspond to the symbols $\rightarrow$ in Eq. (5) (read from left to right). The work performed on the system during this process is the sum of energy changes due to updates in λ,[4,17,23,24]

$$W[\gamma] = \sum_{i=0}^{i=N-1} \delta W_i \equiv \sum_{i=0}^{i=N-1} [H_{\lambda_{i+1}}(\mathbf{z}_i) - H_{\lambda_i}(\mathbf{z}_i)]. \tag{8}$$

Using Eqs. (3), (7) and (8), we arrive at the nonequilibrium work relation for Monte Carlo dynamics[4,17]

$$\langle e^{-\beta W} \rangle \equiv \int d\gamma \, p[\gamma] e^{-\beta W[\gamma]} = e^{-\beta \Delta F}. \tag{9}$$

Thus we can estimate ΔF by repeatedly performing simulations to generate trajectories of the sort described by Eq. (5), computing the work associated with each trajectory, Eq. (8), and finally constructing the exponential average, Eq. (9). As mentioned in the Introduction, however, this average converges poorly when the process is highly dissipative.

To address the issue of poor convergence, let us now assume that for every integer $0 \leq i < N$, we have a deterministic function $M_i : \mathbf{z} \rightarrow \mathbf{z}'$ that takes any point $\mathbf{z}$ in configuration space and maps it to a point $\mathbf{z}'$. We assume that each of these functions is invertible (M_i^{-1} exists), but otherwise the functions are arbitrary. These M_i's then constitute a set of *bijective mappings*, which we use to modify the procedure for generating trajectories, as follows. When the value of the work parameter is switched from λ_i to λ_{i+1}, the configuration space coordinates are simultaneously subjected to the mapping M_i.

Equation (5) then becomes

$$[\mathbf{z}_0, \lambda_0] \overset{M_0}{\Rightarrow} [\mathbf{z}'_0, \lambda_1] \rightarrow [\mathbf{z}_1, \lambda_1] \overset{M_1}{\Rightarrow} \cdots \rightarrow [\mathbf{z}_{N-1}, \lambda_{N-1}]$$
$$\overset{M_{N-1}}{\Rightarrow} [\mathbf{z}'_{N-1}, \lambda_N], \tag{10}$$

where

$$\mathbf{z}'_i \equiv M_i(\mathbf{z}_i), \tag{11}$$

as indicated by the notation $\overset{M_i}{\Rightarrow}$. (As before, the symbol $\rightarrow$ denotes a Monte Carlo move at fixed λ.) The bijective maps effectively escort the system by directly coupling increments in λ to changes in the microstate. This is similar to the "metric scaling" approach introduced by Miller and Reinhardt,[15] in which each update in λ is accompanied by a linear scaling of coordinates; however, in the present paper we do not assume the M_i's are linear in $\mathbf{z}$.

In the escorted trajectory [Eq. (10)], the system visits a sequence of $2N$ points in configuration space: the N "primary" microstates $\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}$, alternating with the N "secondary" microstates $\mathbf{z}'_0, \ldots, \mathbf{z}'_{N-1}$. Since each $\mathbf{z}'_i$ is uniquely determined from $\mathbf{z}_i$ [Eq. (11)], the sequence of primary microstates $\gamma = \{\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}\}$ fully specifies the trajectory; that is, trajectory space remains DN-dimensional, with $d\gamma = d\mathbf{z}_0, \ldots, d\mathbf{z}_{N-1}$. The probability density for generating a trajectory γ is given by the following modification of Eq. (7):

$$p[\gamma] = P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2}) \cdots P_{\lambda_2}(\mathbf{z}_2|\mathbf{z}'_1) \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) \, \rho^A_{\text{eq}}(\mathbf{z}_0). \tag{12}$$

Taking a cue from Refs. 13 and 15, let us now define

$$W'[\gamma] = \sum_{i=0}^{N-1} \delta W'_i \equiv \sum_{i=0}^{N-1} \Big[H_{\lambda_{i+1}}(\mathbf{z}'_i) - H_{\lambda_i}(\mathbf{z}_i)$$
$$- \beta^{-1} \ln J_i(\mathbf{z}_i) \Big], \tag{13}$$

where $J_i(\mathbf{z}) = |\partial \mathbf{z}'/\partial \mathbf{z}|$ is the Jacobian associated with the map $M_i : \mathbf{z} \rightarrow \mathbf{z}'$. Averaging $\exp(-\beta W'[\gamma])$ over the ensemble of trajectories, we have

$$\langle e^{-\beta W'} \rangle = \int d\gamma \, p[\gamma] \, e^{-\beta W'[\gamma]}$$
$$= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \cdots \int d\mathbf{z}_0 \, e^{-\beta \sum_{i=0}^{N-1} \delta W'_i}$$
$$\times P_{\lambda_{N-1}}(\mathbf{z}_{N-1}|\mathbf{z}'_{N-2}) \ldots P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) \, e^{-\beta H_{\lambda_0}(\mathbf{z}_0)}. \tag{14}$$

To evaluate this expression, we first identify all factors in the integrand that do not depend on $\mathbf{z}_0$ or $\mathbf{z}'_0$, and we pull these outside the innermost integral, $\int d\mathbf{z}_0$, which gives us (for that integral):

$$\int d\mathbf{z}_0 \, e^{-\beta \delta W'_0} \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) \, e^{-\beta H_{\lambda_0}(\mathbf{z}_0)} \tag{15}$$

$$= \int d\mathbf{z}_0 \, J_0(\mathbf{z}_0) \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) \, e^{-\beta H_{\lambda_1}(\mathbf{z}'_0)} \tag{16}$$

$$= \int d\mathbf{z}'_0 \, P_{\lambda_1}(\mathbf{z}_1|\mathbf{z}'_0) \, e^{-\beta H_{\lambda_1}(\mathbf{z}'_0)} = e^{-\beta H_{\lambda_1}(\mathbf{z}_1)}. \tag{17}$$

We have used Eq. (13) to get to the second line, followed by a change in the variables of integration to get to the third line, $d\mathbf{z}_0 J_0(\mathbf{z}_0) \rightarrow d\mathbf{z}'_0$, and we have invoked Eq. (4) to arrive at the final result. This process can be repeated for the integrals $\int d\mathbf{z}_1$ to $\int d\mathbf{z}_{N-2}$, which brings us to

$$\langle e^{-\beta W'} \rangle = \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \, e^{-\beta \delta W'_{N-1}} \, e^{-\beta H_{\lambda_{N-1}}(\mathbf{z}_{N-1})}$$
$$= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}_{N-1} \, J_{N-1}(\mathbf{z}_{N-1}) \, e^{-\beta H_{\lambda_N}(\mathbf{z}'_{N-1})}$$
$$= \frac{1}{Z_{\lambda_0}} \int d\mathbf{z}'_{N-1} \, e^{-\beta H_{\lambda_N}(\mathbf{z}'_{N-1})} = \frac{Z_{\lambda_N}}{Z_{\lambda_0}}, \tag{18}$$

and therefore

$$\langle e^{-\beta W'} \rangle = e^{-\beta \Delta F}. \tag{19}$$

Equation (19) is an identity for ΔF in terms of escorted trajectories, generated as per Eq. (10). For the special case in which each mapping is the identity, $M_i = I$, we recover the usual scheme, Eq. (5), and then Eq. (19) reduces to the nonequilibrium work relation, Eq. (9). Following Miller and Reinhardt,[15] we will find it convenient to interpret W' as the work done during the switching process and simply denote it by W. As we will discuss in Sec. IV below, when the mappings $\{M_i\}$ are chosen so as to reduce the dynamic lag illustrated in Fig. 1, then the efficiency of the estimate of ΔF improves, often dramatically.

III. FLUCTUATION THEOREM

Let us now consider not only the switching process described by Eq. (10), which we will henceforth designate the *forward* process, but also its time-reversed analogue, the *reverse* process. In the reverse process, the system is prepared in equilibrium at $\lambda = B$ and temperature β^{-1}. The work parameter is then switched to $\lambda = A$ in N steps, following a sequence $\{\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_N\}$ that is the reversal of the protocol used during the forward process:

$$\tilde{\lambda}_i \equiv \lambda_{N-i}. \tag{20}$$

During the reverse process, changes in λ are coupled to the system's evolution through the inverse mapping functions, $\tilde{M}_i \equiv M_{N-1-i}^{-1}$, generating a trajectory

$$[\tilde{\mathbf{z}}'_{N-1}, \tilde{\lambda}_N] \overset{\tilde{M}_{N-1}}{\Longleftarrow} [\tilde{\mathbf{z}}_{N-1}, \tilde{\lambda}_{N-1}] \leftarrow \cdots \overset{\tilde{M}_1}{\Longleftarrow} [\tilde{\mathbf{z}}_1, \tilde{\lambda}_1]$$
$$\leftarrow [\tilde{\mathbf{z}}'_0, \tilde{\lambda}_1] \overset{\tilde{M}_0}{\Longleftarrow} [\tilde{\mathbf{z}}_0, \tilde{\lambda}_0], \tag{21}$$

where $\tilde{\mathbf{z}}'_i \equiv \tilde{M}_i(\tilde{\mathbf{z}}_i)$, and the initial state $\tilde{\mathbf{z}}_0$ is sampled from ρ^B_{eq}. The direction of the arrows indicates the progression of time. The probability density for obtaining a trajectory $\tilde{\gamma} = \{\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{N-1}\}$ is

$$p[\tilde{\gamma}] = P_{\tilde{\lambda}_{N-1}}(\tilde{\mathbf{z}}_{N-1}|\tilde{\mathbf{z}}'_{N-2}) \cdots P_{\tilde{\lambda}_2}(\tilde{\mathbf{z}}_2|\tilde{\mathbf{z}}'_1) \, P_{\tilde{\lambda}_1}(\tilde{\mathbf{z}}_1|\tilde{\mathbf{z}}'_0) \, \rho^B_{\text{eq}}(\tilde{\mathbf{z}}_0), \tag{22}$$

with $d\tilde{\gamma} = d\tilde{z}_0 \cdots d\tilde{z}_{N-1}$. Following Eq. (13), the work performed during this process is

$$W_R[\tilde{\gamma}] = \sum_{i=0}^{N-1} [H_{\tilde{\lambda}_{i+1}}(\tilde{z}_i') - H_{\tilde{\lambda}_i}(\tilde{z}_i) - \beta^{-1} \ln \tilde{J}_i(\tilde{z}_i)], \quad (23)$$

where $\tilde{J}_i(\tilde{z}) = |\partial \tilde{z}'/\partial \tilde{z}|$ is the Jacobian for the mapping $\tilde{M}_i$. Here and below we use the subscripts F and R to specify the forward and reverse processes, respectively.

We will now show that the work distributions corresponding to these two processes satisfy Crooks's fluctuation relation,[16–18] namely

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(W-\Delta F)}, \quad (24)$$

where

$$P_F(W) = \int d\gamma \, p_F[\gamma] \, \delta(W - W_F[\gamma]), \quad (25)$$

denotes the distribution of work values for the forward process, and $P_R(W)$ is similarly defined for the reverse process.

To establish this result, consider a *conjugate pair* of trajectories, γ and γ^*, related by time-reversal. Specifically, if $\gamma = \{z_0, \ldots, z_{N-1}\}_F$ is a trajectory generated during the forward process, that visits the sequence of microstates

$$z_0 \overset{M_0}{\Rightarrow} z_0' \to z_1 \overset{M_1}{\Rightarrow} z_1' \to \cdots \to z_{N-1} \overset{M_{N-1}}{\Rightarrow} z_{N-1}', \quad (26)$$

then its conjugate twin, $\gamma^* = \{z_{N-1}', \ldots, z_0'\}_R$, generated during the reverse process, visits the same microstates, in reverse order:

$$z_0 \overset{\tilde{M}_{N-1}}{\Leftarrow} z_0' \leftarrow z_1 \overset{\tilde{M}_{N-2}}{\Leftarrow} z_1' \leftarrow \cdots \leftarrow z_{N-1} \overset{\tilde{M}_0}{\Leftarrow} z_{N-1}', \quad (27)$$

that is, $\tilde{z}_i = z_{N-1-i}'$ and $\tilde{z}_i' = z_{N-1-i}$ [see Eq. (21)]. Note that the primary microstates of γ are the secondary microstates of γ^*, and *vice versa*, and the work function is odd under time-reversal:

$$W_F[\gamma] = -W_R[\gamma^*]. \quad (28)$$

We wish to evaluate the quantity

$$P_F(W) e^{-\beta(W-\Delta F)}$$
$$= \int d\gamma \, p_F[\gamma] \, e^{-\beta(W_F[\gamma]-\Delta F)} \delta(W - W_F[\gamma]), \quad (29)$$

with $p_F[\gamma]$ given by Eq. (12). To this end, we first decompose $W_F[\gamma]$ as follows:

$$W_F[\gamma] = \Delta E_F[\gamma] - Q_F[\gamma] - \beta^{-1} S_F[\gamma], \quad (30)$$

where

$$\Delta E_F[\gamma] \equiv H_{\lambda_N}(z_{N-1}') - H_{\lambda_0}(z_0), \quad (31a)$$

$$Q_F[\gamma] \equiv \sum_{i=1}^{N-1} \left[H_{\lambda_i}(z_i) - H_{\lambda_i}(z_{i-1}') \right], \quad (31b)$$

$$S_F[\gamma] \equiv \sum_{i=0}^{N-1} \ln J_{\lambda_i}(z_i) = \ln \prod_{i=0}^{N-1} \left| \frac{\partial z_i'}{\partial z_i} \right| = \ln \left| \frac{\partial \gamma^*}{\partial \gamma} \right|. \quad (31c)$$

Here $\Delta E_F[\gamma]$ is the total change in the energy of the system as it evolves along the trajectory γ, $Q_F[\gamma]$ can be interpreted as the heat transferred to the system from the reservoir,[15] and $S_F[\gamma]$ is an entropylike term, which arises because the mappings M_i need not preserve volume. The quantities defined in Eq. (31) satisfy the properties

$$P_{\lambda_{N-1}}(z_{N-1}|z_{N-2}') \cdots P_{\lambda_1}(z_1|z_0')$$
$$= P_{\lambda_{N-1}}(z_{N-2}'|z_{N-1}) \cdots P_{\lambda_1}(z_0'|z_1) e^{-\beta Q_F[\gamma]}, \quad (32a)$$

$$\rho_{eq}^{\lambda_0}(z_0) = \rho_{eq}^{\lambda_N}(z_{N-1}') e^{\beta(\Delta E_F[\gamma]-\Delta F)}, \quad (32b)$$

where we have used Eqs. (2) and (3). These properties then give us

$$p_F[\gamma] = P_{\lambda_{N-1}}(z_{N-1}|z_{N-2}') \cdots P_{\lambda_1}(z_1|z_0') \rho_{eq}^{\lambda_0}(z_0)$$
$$= P_{\lambda_{N-1}}(z_{N-2}'|z_{N-1}) \cdots P_{\lambda_1}(z_0'|z_1) e^{-\beta Q_F[\gamma]}$$
$$\times \rho_{eq}^{\lambda_N}(z_{N-1}') e^{\beta(\Delta E_F[\gamma]-\Delta F)}$$
$$= p_R[\gamma^*] e^{\beta(W_F[\gamma]-\Delta F)} e^{S_F[\gamma]}, \quad (33)$$

hence

$$p_F[\gamma] e^{-\beta(W_F[\gamma]-\Delta F)} = p_R[\gamma^*] \left| \frac{\partial \gamma^*}{\partial \gamma} \right|. \quad (34)$$

Substituting this result into the integrand on the right side of Eq. (29), then changing the variables of integration from $d\gamma$ to $d\gamma^*$, and invoking Eq. (28), we finally arrive at the result we set out to establish:

$$P_F(W) e^{-\beta(W-\Delta F)} = P_R(-W). \quad (35)$$

Equation (35) in turn implies that the average of any function $f(W)$ over work values generated in the forward process, can be related to an average over work values obtained in the reverse process:[16]

$$\frac{\langle f(W) \rangle_F}{\langle f(-W) e^{-\beta W} \rangle_R} = e^{-\beta \Delta F}. \quad (36)$$

In principle, this result can be used with any $f(W)$ to estimate ΔF. The problem of determining the optimal choice of $f(W)$ was solved by Bennett in the context of equilibrium sampling,[19] and this solution can be applied directly to the nonequilibrium setting.[16,20] Specifically, if we have n_F work values from the forward simulation, and n_R work values from the reverse simulation, then the optimal choice is

$$f(W) = \frac{1}{1 + \exp(\beta W + \beta K)}, \quad (37)$$

where $K = -\Delta F + \beta^{-1} \ln(n_F/n_R)$. The value of ΔF is then estimated by recursively solving the equation,

$$e^{-\beta \Delta F} = \frac{\langle 1/(1 + e^{\beta(W+K)}) \rangle_F}{\langle 1/(1 + e^{\beta(W-K)}) \rangle_R} e^{\beta K}, \quad (38)$$

as described in detail in Ref. 19. This procedure for estimating ΔF is known as Bennett's acceptance ratio method (BAR).

IV. COMPUTATIONAL EFFICIENCY AND FIGURES OF MERIT

While Eqs. (19) and (38) are valid for any set of invertible mapping functions, $\{M_i\}$, the efficiency of using escorted simulations to estimate ΔF depends strongly on the choice of these functions. Since the convergence of exponential averages such as Eq. (19) deteriorates rapidly with dissipation,[5–7] which in turn correlates with the lag illustrated in Fig. 1, it is reasonable to speculate that a choice of mappings that decreases the lag will improve the convergence of estimator [Eq. (19)].

To pursue this idea, let us first consider the extreme case of a set of mapping functions $\{M_i^*\}$ that entirely eliminates the lag. By this we mean the following: for an ensemble of trajectories generated using Eq. (10), with $\mathbf{z}_0$ sampled from $p_{\text{eq}}^A(\mathbf{z}_0)$, the subsequent microstates $\mathbf{z}_i$ are distributed according to $p_{\text{eq}}^{\lambda_i}(\mathbf{z}_i)$, for all $1 \leq i < N$. That is, the shaded and unshaded ovals coincide in Fig. 1. This occurs if under the bijective mapping $M_i^* : \mathbf{z} \to \mathbf{z}'$, the equilibrium distribution $\rho_{\text{eq}}^{\lambda_i}(\mathbf{z})$ transforms to the distribution $\rho_{\text{eq}}^{\lambda_{i+1}}(\mathbf{z}')$,[13] in other words

$$\rho_{\text{eq}}^{\lambda_{i+1}}(\mathbf{z}') = \frac{\rho_{\text{eq}}^{\lambda_i}(\mathbf{z})}{J_{\lambda_i}^*(\mathbf{z})}. \tag{39}$$

[Under a bijective map $M : \mathbf{x} \to \mathbf{y}$, a distribution $f(\mathbf{x})$ is transformed to the distribution $\eta(\mathbf{y}) = f(\mathbf{x})/J(\mathbf{x})$, where $J(\mathbf{x}) = |\partial\mathbf{y}/\partial\mathbf{x}|$.] When all the $M_{\lambda_i}^*$'s satisfy this condition, we will say that the set of mappings is *perfect*. Using $\rho_{\text{eq}}^\lambda = e^{\beta(F_\lambda - H_\lambda)}$, and taking the logarithm of both sides of Eq. (39), we obtain (for a perfect set of mappings)

$$\delta W_i \equiv H_{\lambda_{i+1}}(\mathbf{z}') - H_{\lambda_i}(\mathbf{z}) - \beta^{-1}\ln J_{\lambda_i}^*(\mathbf{z}) = F_{\lambda_{i+1}} - F_{\lambda_i}, \tag{40}$$

hence $W[\gamma] = \Delta F$ *for every trajectory* γ [Eq. (13)]. Thus for a perfect set of mappings we have $P_F(W) = \delta(W - \Delta F)$, and Eq. (19) provides a zero-variance estimate of the free energy difference. It is straightforward to show that if the M_i^*'s form a set of perfect mappings for the forward process, then the $\tilde{M}_i^*$'s form a set of perfect mappings for the reverse process, and $P_R(W) = \delta(W + \Delta F)$.

The considerations of the previous paragraph support the idea that reducing lag improves convergence. While we generally cannot expect to be able to construct a *perfect* set of mapping functions (this is likely to be far more difficult than the original problem of estimating ΔF![12]), in many cases it might be possible to use either intuition or prior information about a system to construct a set of M_i's that reduce the lag substantially. In such cases the dissipation accompanying the escorted simulations is less than that for the unescorted simulations, leading to improved convergence of the free energy estimate.

As an example of a strategy that can be used to construct good mappings, consider a system of identical, mutually interacting particles, in an external potential $U_\lambda(\mathbf{r})$:

$$H_\lambda(\mathbf{z}) = \sum_k U_\lambda(\mathbf{r}_k) + \sum_{k<l} V(\mathbf{r}_k, \mathbf{r}_l). \tag{41}$$

The probability distribution of a single, tagged particle is then given by the single-particle density

$$\rho_\lambda^{(1)}(\mathbf{r}) = \frac{1}{Z_\lambda}\int d\mathbf{z}\,\delta[\mathbf{r}_k(\mathbf{z}) - \mathbf{r}]\,e^{-\beta H_\lambda(\mathbf{z})}, \tag{42}$$

where $\mathbf{r}_k(\mathbf{z})$ specifies the coordinates of the tagged particle as a function of the microstate $\mathbf{z}$. Now consider a reference system of noninteracting particles, described by a Hamiltonian

$$\bar{H}_\lambda(\mathbf{z}) = \sum_k \bar{U}_\lambda(\mathbf{r}_k), \tag{43}$$

with a similarly defined single-particle density $\bar{\rho}_\lambda^{(1)}(\mathbf{r})$; and imagine that $\bar{U}_\lambda$ is chosen so that these single-particle densities are identical or nearly identical: $\rho_\lambda^{(1)}(\mathbf{r}) \approx \bar{\rho}_\lambda^{(1)}(\mathbf{r})$. In this case a set of mappings $\{M_i\}$ that are perfect or near-perfect for the reference system ($\bar{H}_\lambda$), might be quite effective in reducing lag in the original system (H_λ). We will illustrate this mean-field-like approach in Sec. V B, and we note that a similar strategy was explored by Hahn and Then in the context of targeted free energy perturbation.[14]

It will be useful to develop a figure of merit, allowing us to compare the efficiency of our method for different sets of mappings. One approach would be simply to compare the error bars associated with the statistical fluctuations in the respective free energy estimates. Unfortunately, estimates of ΔF obtained from convex nonlinear averages such as the one obtained from Eq. (19), are systematically biased for any finite number of realizations.[7,25] This bias can be large, and as a result the statistical error bars by themselves might not be sufficiently reliable to quantify the efficiency of the mapping. In the following paragraphs we discuss alternative figures of merit.

We begin by noting that when the unidirectional estimator, Eq. (19), is used in conjunction with simulations of the forward process, then the number of realizations (N_s) required to obtain a reliable estimate of ΔF is roughly given by[5,6]

$$N_s \sim e^{\beta(\langle W\rangle_R + \Delta F)}, \tag{44}$$

where $\langle W\rangle_R + \Delta F$ is the dissipation accompanying the reverse process. While this provides some intuition for the convergence of Eq. (19), its usefulness as a figure of merit is somewhat limited as it requires simulations of both the forward and the reverse processes, and in that case we are better off using a bidirectional estimator such as Eq. (38).

When we do have simulations of both processes, then an easily computed figure of merit is the hysteresis, $\langle W_{\text{diss}}\rangle_F + \langle W_{\text{diss}}\rangle_R = \langle W\rangle_F + \langle W\rangle_R$. The value of this quantity is zero if the mappings are perfect, otherwise it is positive. It is interesting to note that the hysteresis can be related to an information-theoretic measure of overlap between the forward and reverse work distributions $P_F(W)$ and $P_R(-W)$:[26]

$$D[P_F||P_R] + D[P_R||P_F] = \beta(\langle W\rangle_F + \langle W\rangle_R). \tag{45}$$

Here $D[p||q] \equiv \int p\ln(p/q) \geq 0$ denotes the relative entropy between the distributions p and q, and the symmetrized quantity $D[p||q] + D[q||p]$ (also known as the *Jeffreys divergence*[27]) provides a measure of the difference, or more precisely the lack of overlap, between the distributions.

The right side of Eq. (45) can be estimated from a modest sample of forward and reverse simulations. If a set of mappings reduces the hysteresis, $\langle W \rangle_F + \langle W \rangle_R$, then this indicates increased overlap between the work distributions, and therefore improved convergence.[6]

When $n_F = n_R = N_s \gg 1$, the mean square error of the Bennett estimator is[14,19,20,28]

$$\langle (F_{\text{BAR}}^{\text{est}} - \Delta F)^2 \rangle = \frac{2}{\beta^2 N_s} \left(\frac{1}{2C} - 1 \right). \tag{46}$$

Here $F_{\text{BAR}}^{\text{est}}$ denotes the estimate of ΔF obtained from Eq. (38), and

$$C \equiv \int dW \frac{P_F(W) P_R(-W)}{P_F(W) + P_R(-W)} = \left\langle \frac{1}{1 + \exp[\beta(W - \Delta F)]} \right\rangle_F$$
$$= \left\langle \frac{1}{1 + \exp[\beta(W + \Delta F)]} \right\rangle_R \tag{47}$$

[this result can be generalized to the case $n_F \neq n_R$ (Ref. 14).] As discussed by Bennett[19] and Hahn and Then,[14,28] the value of C measures the overlap between $P_F(W)$ and $P_R(-W)$, and provides a rough figure of merit for the Bennett estimator. When lag is eliminated and the two distributions coincide, then C attains its maximum value, $C = 1/2$, whereas when there is poor overlap, $C \approx 0$. Thus we expect that the higher the value of the overlap function C, the smaller the number of realizations N_s required to estimate ΔF from Eq. (38) with a prescribed accuracy. Indeed, Eq. (46) suggests a lower bound on the number of realizations needed to achieve a mean square error less than β^{-2}: $N_s > 1/C$. Note that since C is an ensemble average [Eq. (47)], it can readily be estimated from available simulation data.

In the Appendix, we derive an upper bound on the number of realizations needed to obtain a reliable estimate of ΔF using Bennett's method, N_s [Eq. (A4)]. Combining these bounds gives us

$$\frac{1}{C} < N_s < \frac{1}{C^2}. \tag{48}$$

While Eq. (48) cannot be used to obtain a good estimate for N_s (Ref. 29), it does allow us to argue heuristically that whenever a set of mappings succeeds in increasing the value of C, the convergence of the Bennett estimator is improved. We will illustrate this point in Sec. V.

V. EXAMPLES

A. Cavity expansion

As a first example, we estimate the free energy cost associated with growing a hard-sphere solute in a fluid. Consider a system composed of n_p point particles inside a cubic container of volume L^3, centered at the origin with periodic boundaries. The particles are excluded from a spherical region of radius R, also centered at the origin. The particles interact with one another via the WCA pairwise interaction potential[1] which is denoted by $V(\mathbf{r}_k, \mathbf{r}_l)$. The energy of the system at a

microstate $\mathbf{z} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{n_p})$ is given by

$$H_R(\mathbf{z}) = \Theta(\mathbf{z}, R) + \sum_{k=1}^{n_p-1} \sum_{l>k}^{n_p} V(\mathbf{r}_k, \mathbf{r}_l), \tag{49}$$

where $\Theta(\mathbf{z}, R) = 0$ whenever $|\mathbf{r}_k| > R$ for all $k = 1, \cdots n_p$, that is when there are no particles inside the spherical cavity; and $\Theta(\mathbf{z}, R) = \infty$ otherwise. The function $\Theta(\mathbf{z}, R)$ ensures that particles are excluded from the spherical region around the origin. We wish to compute the free energy cost, ΔF, associated with increasing the radius of the cavity from R_A to R_B (see Fig. 2).

A hypothetical estimate of ΔF using *unescorted* nonequilibrium simulations [Eq. (5)] involves "growing out" the spherical cavity in discrete increments, as follows. Starting with a microstate $\mathbf{z}_0$ sampled from equilibrium at $R = R_A$, the radius of the sphere is increased by an amount δR_0. If all n_p fluid particles remain outside the enlarged sphere, then $\delta W_0 = 0$; but if one or more particles now finds itself inside the sphere ($r_k < R_A + \delta R_0$) then $\delta W_0 = \infty$. One or more Monte Carlo steps are then taken, after which the radius is again increased by some amount, δR_1, and δW_1 is determined in the same fashion as δW_0. In principle this continues until the radius of the sphere is R_B, and then the work is tallied for the entire trajectory: $W = \Sigma_i \delta W_i$. In practice the trajectory can be terminated as soon as $\delta W_i = \infty$ at some step i, since this implies $W = \infty$. For this procedure, Eq. (1) can be rewritten as

$$P = e^{-\beta \Delta F}, \tag{50}$$

where P is the probability of generating a trajectory for which $W = 0$; that is, a trajectory in which the sphere is successfully grown out to radius R_B, without overtaking any fluid particles along the way. The quantity P is estimated directly, by generating a number of trajectories and counting the "successes" ($W = 0$). For a sufficiently dense fluid, however, a successful trajectory is a rare event ($P \ll 1$), and this approach converges poorly. Note also that this approach does not give the correct free energy difference in the reverse case of a shrinking sphere (from $R = R_B$ to $R = R_A$), since $W = 0$ for every trajectory in that situation.

For the hypothetical procedure just described, Eq. (50) implies that the probability to generate a successful trajectory does not depend on the number of increments used to grow the cavity from R_A to R_B. Therefore the most computationally efficient implementation is to grow the sphere out in a single step, which corresponds to the free energy perturbation method (FEP) (Refs. 1 and 2). In this case P is just the
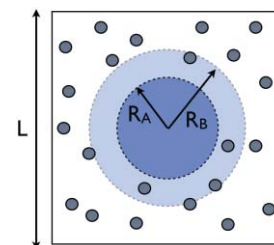

FIG. 2. A schematic of the cavity expansion problem.

probability to observe no particles in the region $R_A < r < R_B$, for an equilibrium simulation at cavity radius R_A.

To improve convergence by means of escorted simulations [Eq. (10)], we constructed mapping functions M_i that move the fluid particles out of the way of the growing sphere, to prevent infinite values of δW_i. Specifically, as the cavity radius R is increased from R_i to R_{i+1}, the location of the nth particle, $\mathbf{r}_n$, is mapped to $\mathbf{r}'_n = m_i(\mathbf{r}_n)$,[13] where

$$m_i(\mathbf{r}_n) = \left[1 + \frac{(R_{i+1}^3 - R_i^3)(L^3 - 8r_n^3)}{(L^3 - 8R_i^3)r_n^3}\right]^{1/3}\mathbf{r}_n$$

$$\text{if} \quad \mathrm{r}_n \leq \mathrm{L}/2, \tag{51}$$

and $m_i(\mathbf{r}_n) = \mathbf{r}_n$ if $r_n > L/2$. The notation $m_i : \mathbf{r}_n \rightarrow \mathbf{r}'_n$ denotes a single-particle mapping; the full mapping $M_i : \mathbf{z} \rightarrow \mathbf{z}'$ is obtained by applying m_i to all n_p fluid particles. To picture the effect of this mapping, let $\mathcal{S}_i$ denote the region of space defined by the conditions $R_i \leq r \leq L/2$, that is a spherical shell of inner radius R_i and outer radius $L/2$ (just touching the sides of the cubic container). Under the mapping $m_i : \mathbf{r} \rightarrow \mathbf{r}'$, the shell $\mathcal{S}_i$ is compressed uniformly onto the shell $\mathcal{S}_{i+1}$, leaving the eight corners of the box $r > L/2$ untouched.[30] In this manner, the particles that would otherwise have found themselves inside the enlarged sphere are pushed outside of it, resulting in a finite contribution to the work [Eq. (13)],

$$\delta W_i = \sum_{k=1}^{n_p-1}\sum_{l>k}^{n_p}\left[V(\mathbf{r}'_k, \mathbf{r}'_l) - V(\mathbf{r}_k, \mathbf{r}_l)\right] - n_0\beta^{-1}\ln\gamma, \tag{52}$$

where $n_0 = n_0(\mathbf{z})$ is the number of particles found within the shell $R_i \leq r \leq L/2$ (before the mapping is applied), and $\gamma = (L^3 - 8R_{i+1}^3)/(L^3 - 8R_i^3) < 1$ is the ratio of shell volumes, $|\mathcal{S}_{i+1}|/|\mathcal{S}_i|$. The first term on the right side of Eq. (52) gives the net change in the energy of the system associated with the escorted switch $[\mathbf{z}_i, R_i] \overset{M_i}{\Rightarrow} [\mathbf{z}'_i, R_{i+1}]$, while the second is the Jacobian term $-\beta^{-1}\ln J_i(\mathbf{z}_i)$.

Unlike the unescorted approach or free energy perturbation, the escorted approach with the mapping given by Eq. (51) is applicable in both the forward (growing spherical cavity) and reverse (shrinking cavity) directions. In the reverse direction, as the solute radius is decreased from R_{i+1} to R_i, the shell $\mathcal{S}_{i+1}$ is uniformly expanded onto the shell $\mathcal{S}_i$. The corresponding increment in work is given by a formula similar to Eq. (52). As a result, one can combine work values from forward and reverse escorted simulations using BAR, Eq. (38).

We have performed both forward and reverse simulations of this system using $n_p = 1000$ WCA particles, with $L = 10.42\sigma$, $R_A = 2.0\sigma$, $R_B = 2.05\sigma$, and $T^* \equiv k_BT/\epsilon = 1$, where the WCA parameters σ and ϵ set the units of length and energy, respectively. Minimum image convention and periodic boundary conditions were used.[1]

Figure 3 shows a running estimate of $P = \exp(-\beta\Delta F)$ obtained from escorted simulations in which the solute radius was switched from R_A to R_B in $N = 10$ steps, with each increment in R alternating with one Monte Carlo sweep. Using a total of $N_s = 50\,000$ independent escorted trajectories, estimates of ΔF and the figures of merit were obtained, and are summarized in Table I (the value of C and $\Delta F_{\mathrm{BAR}}^{\mathrm{est}}$ were
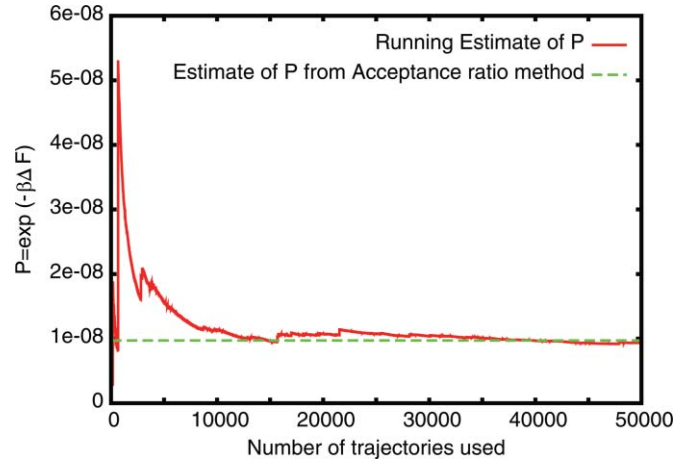

FIG. 3. Running estimate of $P = \exp(-\beta\Delta F)$ from escorted free energy simulations, plotted as a function of the number of trajectories used to obtain the estimate.

estimated using $n_F = n_R = N_s = 50\,000$ trajectories). Statistical error bars were computed using the bootstrap method.[31] While an analytical expression for ΔF is not available for this example, the agreement between the estimates obtained by growing the solute (F), shrinking it (R), and applying BAR gives us confidence in the result, $\Delta F \approx 18.4\,k_BT$.

As an additional consistency check, in Fig. 4 we verify that the escorted simulations satisfy the fluctuation theorem Eq. (35). We do this by following steps analogous to those in Sec. III of Ref. 19 to obtain a restatement of Eq. (35),

$$L_2(W) - L_1(W) \equiv \left[\ln P_R(-W) + \beta\frac{W}{2}\right]$$

$$- \left[\ln P_F(W) - \beta\frac{W}{2}\right] = \beta\Delta F. \tag{53}$$

In Fig. 4 we plot L_1, L_2, and $L_2 - L_1$ as functions of W. The flatness of the difference $L_2 - L_1$ over the region for which we have good statistics is in agreement with Eq. (53), and provides a useful and stringent consistency check,[1,32] which gives us further confidence in our estimates.

While the highly accurate estimates listed in Table I were generated using $N_s = 50\,000$ escorted trajectories, we found that we were able to obtain estimates of ΔF with error bars around $1 k_BT$ using only $N_s = 100$ realizations for the

TABLE I. Estimates and figures of merit. Here $\Delta F_F^{\mathrm{est}}$ denotes the estimate of $\Delta F \equiv F_B - F_A$ from the forward process ($R_A \rightarrow R_B$) and $\Delta F_R^{\mathrm{est}}$ denotes the estimate of $-\Delta F$ from the reverse process ($R_A \leftarrow R_B$). $\Delta F_{\mathrm{BAR}}^{\mathrm{est}}$ denotes the estimate of ΔF obtained from Bennett's acceptance ratio method.

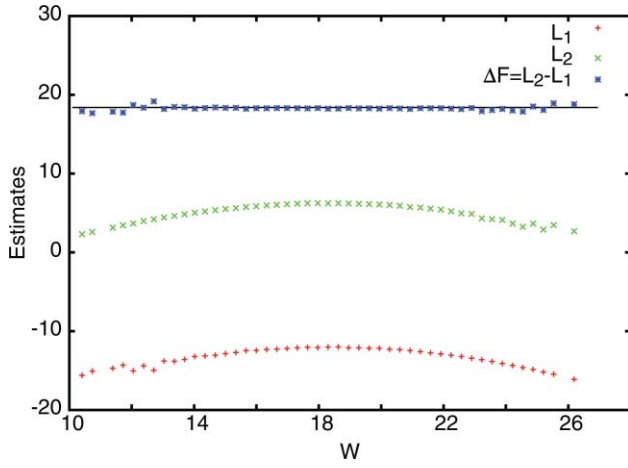|	
$\langle W \rangle_F$	22.288 ± 0.012
$\langle W \rangle_R$	-14.458 ± 0.013
$\langle W \rangle_F + \langle W \rangle_R$	7.830 ± 0.018
$\Delta F_F^{\mathrm{est}}$	18.487 ± 0.085
$\Delta F_R^{\mathrm{est}}$	-18.334 ± 0.078
$\Delta F_{\mathrm{BAR}}^{\mathrm{est}}$	18.456 ± 0.011
C	0.120 ± 0.001

FIG. 4. Graphical verification of the fluctuation theorem and estimation of ΔF. The horizontal line indicates the estimate of ΔF obtained from the acceptance ratio method (Table I).

unidirectional estimators, and $N_s = 10$ realizations for the bidirectional estimator (data not shown).

To compare the escorted method with unescorted FEP, we first sampled $N_s = 100\,000$ independent configurations from the canonical ensemble with cavity radius $R = R_A$, by generating a single, long equilibrium Monte Carlo trajectory and sampling one configuration per ten Monte Carlo sweeps. This involved a total computational time approximately equal to that of generating 50 000 escorted trajectories. Among these 10^5 configurations we did not observe a single one in which the region $R_A \leq r \leq R_B$ was spontaneously devoid of particles ($W = 0$), in other words we were unable to obtain an estimate of ΔF using free energy perturbation. This is consistent with the result $P \approx e^{-18.4} \approx 10^{-8}$ (Fig. 3, Table I), which suggests that roughly 10^8 independent configurations are needed to observe one for which $W = 0$.

For a more efficient implementation of FEP, we divided the interval $[R_A, R_B]$ into ten stages (subintervals), and then used FEP to estimate the free energy change for each stage, keeping the total computational time fixed. This provided a final estimate of ΔF with error bars comparable to those of the unidirectional escorted estimators in Table I, but still considerably larger than those of the bidirectional estimates (data not shown).[33]

B. Dipole fluid

As our second example, we consider n_p point Lennard-Jones dipoles in a cubic container of size L with periodic boundaries, and we compute the free energy cost associated with introducing a uniform electric field in the container. The energy of the system in an external electric field $\mathbf{E} = E\hat{\mathbf{e}}_z$, where $\hat{\mathbf{e}}_z$ denotes a unit vector along the z axis, is given by

$$H_{E,\gamma}(\mathbf{z}) = -\sum_{k=1}^{n_p} \mathbf{p}_k \cdot \mathbf{E} + \sum_{k=1}^{n_p-1} \sum_{l>k}^{n_p} V_{LJ}(\mathbf{r}_k, \mathbf{r}_l)$$
$$-\gamma \frac{\mathbf{p}_k \cdot \mathbf{p}_l}{|\mathbf{r}_k - \mathbf{r}_l|^4}, \tag{54}$$

where $\mathbf{z} = \{\mathbf{r}_1, \mathbf{p}_1, \ldots, \mathbf{r}_{n_p}, \mathbf{p}_{n_p}\}$, $\mathbf{p}_k$ denotes the dipole moment vector of the kth particle, and $V_{LJ}(\mathbf{r}_k, \mathbf{r}_l)$ denotes the Lennard-Jones pairwise interaction potential. The parameter γ controls the strength of the dipole–dipole interaction. We set $|\mathbf{p}_k| = 1$ for all k. In spherical polar coordinates, $\mathbf{p}_k = (1, \theta_k, \phi_k)$, and the measure on $\mathbf{z}$ space is hence $d\mathbf{z} = \Pi_{k=1}^{n_p} d\mathbf{r}_k d\cos(\theta_k) d\phi_k$.

Taking the electric field to be the external parameter, we wish to compute the free energy difference between the ensembles corresponding to $E = 0$ and $E = E_f$ at some temperature β^{-1} by performing nonequilibrium switching simulations. Our first task is to construct a mapping function that escorts the system along a near equilibrium path as E is switched. Following Eq. (43), we consider the energy function $\bar{H}_E(\mathbf{z}) \equiv H_{E,0}(\mathbf{z})$ [i.e., $\gamma = 0$ in Eq. (54)], which describes a system of noninteracting Lennard-Jones dipoles in a field of strength E. The change in free energy as the field is switched from E_i to E_{i+1} can be solved analytically and is given by

$$\bar{F}_{E_{i+1}} - \bar{F}_{E_i} = -n_p \frac{1}{\beta} \ln\left[\frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)} \frac{E_i}{E_{i+1}}\right]. \tag{55}$$

We now use this result to solve for a perfect set of mappings for this system of noninteracting dipoles.

Let $m_i : \zeta \equiv \cos(\theta) \to \zeta'$ denote a mapping that acts on the $\zeta = \cos(\theta)$ degree of freedom of a dipole when the external field is switched from E_i to E_{i+1}. The full mapping M_i is obtained by applying the mapping m_i to all n_p particles. We look for the perfect mapping M_i that transforms the canonical distribution corresponding to $\bar{H}_{E_i}(\mathbf{z})$ to the canonical distribution corresponding to $\bar{H}_{E_{i+1}}(\mathbf{z}')$. The following equation for the perfect single particle mapping m_i can be obtained from Eq. (40) by using Eqs. (54) and (55) and by noting that $\mathbf{p}_k \cdot \mathbf{E} = E\zeta_k$:

$$E_{i+1}m_i(\zeta) - E_i\zeta - \frac{1}{\beta} \ln \frac{dm_i(\zeta)}{d\zeta}$$
$$= -\frac{1}{\beta} \ln \frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)} \frac{E_i}{E_{i+1}}. \tag{56}$$

This differential equation has the solution

$$m_i(\zeta) = \frac{1}{\beta E_{i+1}} \ln\left[\frac{\sinh(\beta E_{i+1})}{\sinh(\beta E_i)}(e^{\beta E_i \zeta} - e^{\beta E_i}) + e^{\beta E_{i+1}}\right]. \tag{57}$$

While Eq. (57) is a perfect mapping only when there are no dipole–dipole interactions ($\gamma = 0$) we expect this mapping to work reasonably well for small values of γ. We will use the term *simple mapping* in reference to Eq. (57).

We also constructed a set of mapping functions using mean field[34] arguments as follows. In the absence of long range order, mean field theory suggests that the interacting dipole-fluid system ($\gamma \neq 0$) in an electric field of strength E can be approximated by a system of noninteracting dipoles ($\gamma = 0$) in an effective field of strength E'. We obtained approximate values for this effective electric field by first numerically evaluating the single-dipole distribution $P(\zeta), \zeta = \cos(\theta)$, at $E = E_f$. The thermal distribution of ζ for a noninteracting dipole in a field of strength E'_f, is $P_0(\zeta) \propto \exp(\beta E'_f \zeta)$. Hence E'_f can be estimated by fitting

P_0 to the numerically obtained distribution $P(\zeta)$. For all other values of E, we calculate the effective fields by linear scaling, $E' = E E'_f / E_f$. Again, using Eq. (40) with $\bar{H}_E(\mathbf{z}) = H_{E',0}(\mathbf{z})$ we obtain a new set of mapping functions. In particular, when the E field is switched from E_i to E_{i+1}, the $\zeta_k = \cos(\theta_k)$ degree of freedom of the kth dipole is transformed according to Eq. (58)

$$m_i(\zeta_k) = \frac{1}{\beta E'_{i+1}} \ln \left[\frac{\sinh(\beta E'_{i+1})}{\sinh(\beta E'_i)} (e^{\beta E'_i \zeta_k} - e^{\beta E'_i}) + e^{\beta E'_{i+1}} \right].$$

$$(58)$$

We will refer to Eq. (58) as a *mean field mapping*. Since the single-dipole distributions for the interacting system at field strength E are (by construction) closely approximated by the single-particle distributions for the noninteracting system at E', we expect the mean field mappings to perform better than the simple mappings of Eq. (57).

We performed numerical simulations with $n_p = 800$ particles. The parameters σ, ϵ of the Lennard-Jones potential set the length and the energy scale of the system, and we took $L = 10\sigma$ and $T^* = k_B T / \epsilon = 1$. Minimum image convention and periodic boundary conditions[1] were used. We performed $N_s = 10^4$ forward and reverse simulations to estimate the free energy difference between the ensembles corresponding to $E = 0$ and $E = 1$, switching the field strength in $N = 10$ equal increments. Ten Monte Carlo sweeps were performed between these updates in E. We obtained estimates of ΔF using: (1) unescorted switching simulations [Eq. (1)], (2) escorted simulations with the simple mappings [Eq. (57)], and (3) escorted simulations with the mean field mappings [Eq. (58)]. For the latter, the effective fields were obtained as described in the previous paragraph. In particular, we found $E'_f \approx 1.5 E_f$ and therefore we took $E'_i = 1.5 E_i$ in Eq. (58).

Figure 5 shows the work distributions $P_F(W)$ and $P_R(-W)$ for these sets of simulations, and reveals a progression from virtually no overlap for the unescorted simulations, to some overlap for the simulations with the simple mappings, to nearly perfect overlap when using the mean field mappings. This trend is in agreement with the expectations mentioned above, and provides direct evidence that the mappings we have constructed substantially reduce the lag and dissipation. The first three rows of Table II quantify these observations. In particular, row 3 gives the distance between the
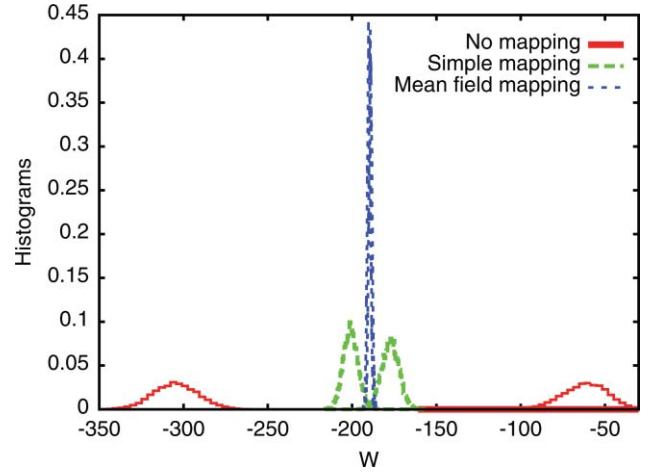

FIG. 5. Work histograms obtained from forward and reverse simulations performed at $\gamma = 0.1$. The degree of overlap between $P_F(W)$ and $P_R(-W)$ provides an indication of the efficiency of the free energy estimate. For unescorted simulations (no mapping) we see no overlap, reflecting considerable dissipation and poor efficiency (Table II). With the mapping given by Eq. (57) the overlap is much improved, and with the mean field mapping, Eq. (58) the forward and reverse distributions are nearly identical.

means of $P_F(W)$ and $P_R(-W)$, and shows that this hysteresis proceeds from nearly $250\,k_B T$ to about $24\,k_B T$ to less than $1 k_B T$ in the three cases. Rows 4 to 6 illustrate the effect of this trend on the efficiency and accuracy of the free energy estimates. The estimates of ΔF (that is, $\Delta F_F^{\mathrm{est}}$, $-\Delta F_R^{\mathrm{est}}$, and $\Delta F_{\mathrm{BAR}}^{\mathrm{est}}$) obtained from the unescorted simulations differ substantially from one another, indicating a high degree of bias. The estimates corresponding to the simple mappings are markedly better, though they still suggest a degree of bias on the order of $1 k_B T$. Finally, the simulations with the mean field mappings are in agreement to within about $0.05 k_B T$, indicating excellent accuracy and efficiency. These findings are also in agreement with the values of the overlap integral C, shown in row 7. This was too low to be estimated using the unescorted simulations, and approaches its maximal value of $1/2$ when using the mean field mappings. Using escorted simulations with the mean field mappings, with the acceptance ratio method (BAR), we found that we were able to generate estimates of ΔF with error bars on the order of $0.2 k_B T$, with about $N_s \sim 1/C^2 \sim 10$ (data not shown).

TABLE II. Estimates and figures of merit for $\gamma = 0.1$. Note that the simulations with the mapping are much more efficient than those without. The forward and reverse work histograms obtained from the simulations without any mappings were so far apart that a reliable estimate of C could not be obtained.

	No mapping	Mapping	Mean field mapping
$\langle W \rangle_F$	-60.409 ± 0.126	-177.074 ± 0.039	-189.079 ± 0.010
$\langle W \rangle_R$	302.958 ± 0.132	200.607 ± 0.045	189.971 ± 0.010
$\langle W \rangle_F + \langle W \rangle_R$	242.549 ± 0.182	23.533 ± 0.060	0.892 ± 0.014
$\Delta F_F^{\mathrm{est}}$	-114.189 ± 3.913	-187.612 ± 0.405	-189.552 ± 0.011
$\Delta F_R^{\mathrm{est}}$	262.232 ± 0.711	191.877 ± 0.310	189.502 ± 0.0140
$\Delta F_{\mathrm{BAR}}^{\mathrm{est}}$	-128.215 ± 3.324	-189.599 ± 0.110	-189.530 ± 0.008
C	~ 0	0.011 ± 0.001	0.407 ± 0.001

VI. SUMMARY

Nonequilibrium fast switching estimates of free energy differences often perform poorly due to dissipation (see Fig. 1). The strategy developed here seeks to address this issue. By modifying the dynamics with additional terms that serve to escort the system along a near equilibrium trajectory and consequently reduce dissipation, we obtain efficient fast switching estimators [Eq. (19)] for the free energy difference. The success of the strategy depends crucially on the choice of the mapping functions M_i: the more effectively these reduce the dissipation, the more efficient the resulting estimator of ΔF.

The examples presented in Sec. V illustrate this point. For the hard sphere solute, we used a simple mapping function that uniformly compresses the solvent, vacating the region into which the hard sphere expands [Eq. (51)]. With this escorting function we were able to estimate ΔF directly from single-stage switching simulations, which would not have been feasible without escorting. In the example of the Lennard-Jones dipole fluid, we used a reference system of noninteracting dipoles to construct a reasonable set of mapping functions [Eq. (57)], and then we further refined these mappings using mean field arguments [Eq. (58)]. Figure 5 and Table II illustrate the correlation between reduced dissipation and increased computational efficiency. Because mean field theory often provides an good description of many-body systems, we speculate that this approach will prove effective for more complex problems of physical interest.

We have also discussed figures of merit, specifically the dissipation in the forward and reverse processes, and the overlap integral C [Eqs. (44), (45), (47)]. For the two examples in Sec. V, we found that these quantities indeed track the effectiveness of the mapping functions. This suggests that these figures of merit might be useful to iteratively improve the performance of the mapping functions.

Finally, the efficiency of our method might further be improved by applying it in combination with other methods, such as biased or umbrella sampling algorithms (see, e.g., Refs. 35–38).

ACKNOWLEDGMENTS

APPENDIX: FIGURE OF MERIT FOR BAR ESTIMATOR

Here we derive a relation between N_s and C for the bidirectional estimator, Eq. (38). The Bennett estimator, Eq. (38) can be rewritten as a ratio of two free energy perturbation identities[39]

$$\frac{\langle P_H(W)/P_F(W)\rangle_{P_F(W)}}{\langle P_H(W)/P_R(-W)\rangle_{P_R(-W)}} = 1, \tag{A1}$$

where $\langle\ldots\rangle_{P_F(W)}$ denotes an average over W values sampled from $P_F(W)$, $\langle\ldots\rangle_{P_R(-W)}$ denotes an

average over W values sampled from $P_R(-W)$, $P_H(W) \equiv C^{-1}(P_F(W)P_R(-W)/P_F(W) + P_R(-W))$ with $C = \int dW(P_F(W)P_R(-W)/P_F(W) + P_R(-W))$ is the normalized harmonic mean distribution. As the averages in the numerator and the denominator are over different ensembles, let us separately consider the number of the realizations required for each to converge.

The dominant contributions to the average in the numerator come from work values that are typically sampled from the harmonic mean distribution P_H.[6] The probability that these dominant values are observed in the forward process can be given by $P = \int_{\text{Typical}} dW\, P_F(W) = \int_{\text{Typical}} dW\, P_H P_F(W)/P_H$, where $\int_{\text{Typical}}$ denotes that the integration is performed over the range of W values that are typically sampled from the harmonic mean distribution ($P_H(W)$).

Following Ref. 6, we now write

$$P \sim \int_{\text{Typical}} dW\, P_H e^{\ln \frac{P_F}{P_H}} \sim e^{\left\langle \ln \frac{P_F}{P_H}\right\rangle_H}$$

$$\times \int_{\text{Typical}} dW\, P_H \sim e^{\left\langle \ln \frac{P_F}{P_H}\right\rangle_H}. \tag{A2}$$

The number of realizations N_s required for adequate sampling can be roughly given by $N_s \sim P^{-1} \sim \exp D[P_H||P_F]$, where we have used $-\langle \ln \frac{P_F}{P_H}\rangle_H = D[P_H||P_F]$. The relative entropy $D[P_H||P_F]$ satisfies the following inequality

$$D[P_H||P_F] = \int \frac{1}{C}\frac{P_R P_F}{P_R + P_F}\ln\frac{P_R}{C(P_F + P_R)}$$

$$\leq \ln\int \frac{1}{4C^2}\frac{4P_R^2 P_F}{(P_R + P_F)^2}$$

$$\leq \ln\frac{1}{4C^2}\int P_R$$

$$= -2\ln 2C, \tag{A3}$$

where we have used the Jensen's inequality[27] for concave functions together with the identity $4P_F P_R \leq (P_F + P_R)^2$. Finally, using Eq. (A3), the number of realizations required to obtain a reliable estimate of ΔF using Bennett's method is bounded by

$$N_s \leq \frac{1}{C^2}. \tag{A4}$$

We have not included the numerical factors in the above relation as it is already an approximate equation.

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, ed. (Academic, San Diego, 2002).

[2] C. Chipot and A. Pohorille, *Free Energy Calculations* (Springer, Berlin, 2007).

[3] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[4] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997).

[5] D. A. Kofke, Mol. Phys. **104**, 3701 (2006), and references therein.

[6] C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).

[7] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A **100**, 12564 (2003).

[8] D. A. Pearlman and P. Kollman, J. Chem. Phys. **91**, 7831 (1989).

[9] R. Wood, J. Phys. Chem. **95**, 4838 (1991).

[10] J. Hermans, J. Phys. Chem. **95**, 9029 (1991).

[11] S. Vaikuntanathan and C. Jarzynski, EPL **87**, 60005 (2009).

[12] S. Vaikuntanathan and C. Jarzysnki, Phys. Rev. Lett. **100**, 190601 (2008).

[13] C. Jarzynski, Phys. Rev. E **65**, 046122 (2002).

[14] A. M. Hahn and H. Then, Phys. Rev. E **79**, 011113 (2009).

[15] M. A. Miller and W. P. Reinhardt, J. Chem. Phys. **113**, 7035 (2000).

[16] G. Crooks, Phys. Rev. E **61**, 2361 (2000).

[17] G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998).

[18] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[19] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[20] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).

[21] As is usually the case with Monte Carlo simulations, we do not include momenta in the microstate.

[22] N. G.V. Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, New York, 2007).

[23] W. P. Reinhardt and J. E. Hunter III, J. Chem. Phys. **97**, 1599 (1992).

[24] J. E. Hunter III, W. P. Reinhardt, and T. F. Davis, J. Chem. Phys **99**, 6856 (1993).

[25] D. M. Zuckerman and T. B. Woolf, Phys. Rev. Lett. **89**, 180602 (2002).

[26] E. H. Feng and G. E. Crooks, Phys. Rev. Lett. **101**, 090602 (2008).

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006).

[28] A. M. Hahn and H. Then, Phys. Rev. E **81**, 041117 (2010).

[29] For $C \ll 1$, the upper and lower bounds in Eq. (48) can be orders of magnitude apart. Nevertheless, Eq. (48) can serve as a good consistency check for the quality of the estimates. For example an estimate of ΔF using Bennett's method from a data set of size $N_s \sim 10^6$ is reliable if $C \sim 0.001$.

[30] An even better mapping would uniformly compress the entire region $r > R_i$, including the eight corners, onto the region $r > R_{i+1}$. However, due to the geometric mismatch between the spherical inner surface and cubic outer surface of these regions, such a mapping is not represented by a simple formula such as Eq. (51), and would need to be constructed numerically.

[31] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans* (Society for Industrial Mathematics, Philadelphia, Pennsylvania, 1982).

[32] A. Pohorille, C. Jarzynski, and C. Chipot, J. Phys. Chem. B **114**, 10235 (2010).

[33] Of course, even after dividing the problem into stages, one can apply escorting by separately treating each stage as a switching simulation with one step, $N = 1$, and using the mappings given by Eq. (51). We found that this further reduces the error bars by nearly a factor of six.

[34] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987).

[35] F. M. Ytreberg and D. M. Zuckerman, J. Chem. Phys. **120**, 10876 (2004).

[36] D. Wu and D. A. Kofke, J. Chem. Phys. **122**, 204104 (2005).

[37] S. X. Sun, J. Chem. Phys. **118**, 5759 (2003).

[38] D. D. L. Minh, J. Chem. Phys. **130** (2009).

[39] R. J. Radmer and P. Kollman, J. Comp. Chem. **18**, 902 (1997).