Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach

C. Jarzynski*

Theoretical Astrophysics, T-6, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 18 June 1997)

In has recently been shown that the Helmholtz free-energy difference between two equilibrium configurations of a system may be obtained from an ensemble of *finite-time* (nonequilibrium) measurements of the work performed in switching an external parameter of the system. Here this result is established, as an identity, within the master equation formalism. Examples are discussed and numerical illustrations provided. [S1063-651X(97)10710-3]

PACS number(s): 05.20.−y, 02.50.Ga, 05.70.Ln, 82.20.Fd

INTRODUCTION

Consider some finite classical system that depends on an external parameter λ. For instance, the system might be a lattice of coupled classical spins and λ may denote the strength of an externally applied magnetic field, or the system might be a gas of particles and λ a parameter specifying the volume of a box enclosing the gas. Now suppose that, after allowing the system to come to equilibrium with a heat reservoir at temperature T, we change, or ''switch,'' the external parameter, infinitely slowly, from an initial (say, $\lambda = 0$) to a final ($\lambda = 1$) value. The system will remain in quasistatic equilibrium with the reservoir throughout the switching process and the total *work* performed on the system will equal the Helmholtz free-energy difference between the initial and final configurations [1]:

$$W_\infty = \Delta F \equiv F_1 - F_0. \tag{1}$$

Here F_λ denotes the equilibrium free energy of the system at temperature T, for a fixed value of λ. The subscript on W reminds us that this result is valid for infinitely slow switching of the parameter.

Now, what happens if, after allowing the system and reservoir to equilibrate, we switch the value of λ at a *finite* rate? In this case the system will lag behind quasistatic equilibrium with the reservoir and the total work will depend on the microscopic initial conditions of system and reservoir. Thus an *ensemble* of such switching measurements, each prepared by first allowing the system to equilibrate with the reservoir, will yield a distribution of values of W. Let $\rho(W, t_s)$ denote this distribution, where t_s is the ''switching time'' over which the value of λ is changed from 0 to 1. (Without loss of generality, assume a uniform switching rate $\dot{\lambda} = t_s^{-1}$.) In other words, $\rho(W, t_s) dW$ is the probability that the work performed in switching λ from 0 to 1, over a time t_s, will fall between W and $W + dW$. In the limit $t_s \to \infty$, we get $W = \Delta F$ [Eq. (1)] and so $\rho \to \delta(W - \Delta F)$ in this limit. For finite t_s, however, (i) the distribution ρ acquires a finite

width, reflecting the fluctuations in W from one switching measurement to the next, and (ii) its centroid shifts to the right,

$$\overline{W} \equiv \int dW \rho(W, t_s) W \geqslant \Delta F, \tag{2}$$

as a result of dissipation (see Fig. 1) [2,3].

Equation (1) gives the free-energy difference ΔF in terms of the work performed during a single infinite-time (quasistatic) process. By contrast, Eq. (2) gives an upper bound on ΔF, from an ensemble of finite-time (hence nonequilibrium) repetitions of the switching process. Recently [4], it was shown that one can in fact extract the value of ΔF itself, not just an upper bound, from the information contained in $\rho(W, t_s)$. Specifically, the following result was shown to be valid *for any switching time* t_s:

$$\overline{\exp(-\beta W)} \equiv \int dW \rho(W, t_s) \exp(-\beta W) = \exp(-\beta \Delta F), \tag{3}$$

where $\beta^{-1} \equiv k_B T$. This result gives the value of an equilibrium quantity ΔF in terms of an ensemble of finite-time, nonequilibrium measurements. As discussed in Ref. [4] and in Sec. IV below, the inequality $\overline{W} \geqslant \Delta F$ [Eq. (2)] follows immediately from Eq. (3).

Equation (3) was derived by treating the system of interest and reservoir, coupled together, as a large, isolated, classical Hamiltonian system. The Hamiltonian governing the motion in the full phase space was taken to be a sum of three terms: one for the system of interest (H_λ), one for the reservoir, and a final term h_{int} coupling the two. The magnitude of the interaction term h_{int} was explicitly assumed to be negligible in comparison to the other two terms. With these assumptions, Eq. (3) follows from the properties of Hamiltonian evolution (see Ref. [4] for details).

The purpose of the present paper is to rederive the same result within a different framework. Instead of considering deterministic evolution in the full phase space, containing both the system of interest and reservoir degrees of freedom, we will treat only the evolution of the system of interest itself, described by a trajectory $\mathbf{z}(t)$. This evolution will be stochastic rather than deterministic and will be governed by a
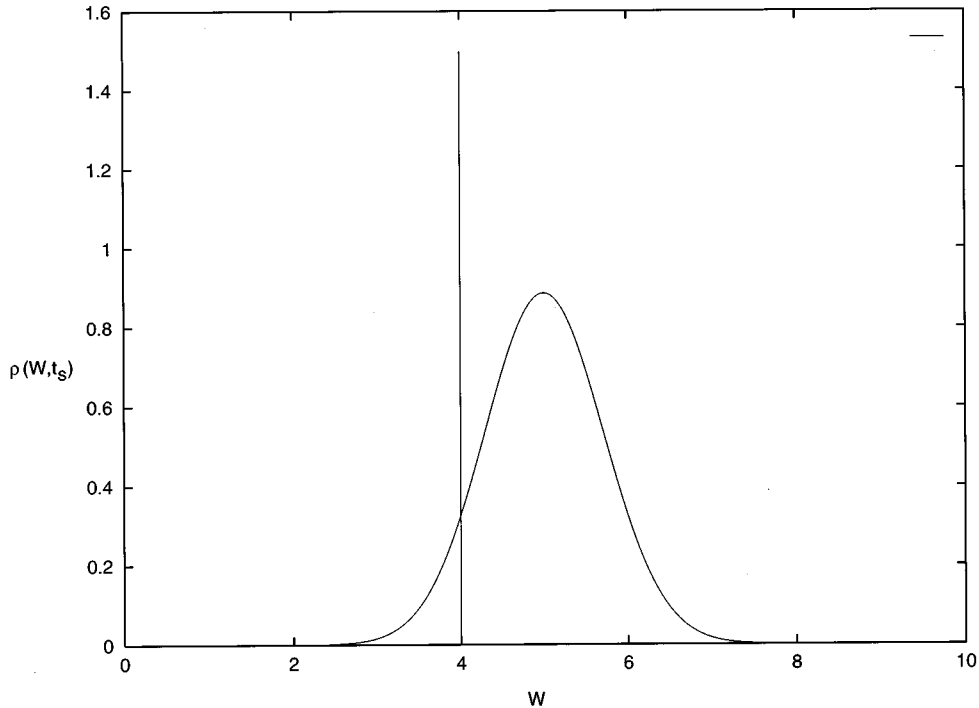
---

*Electronic address: chrisj@t6-serv.lanl.gov

FIG. 1. Distribution of values of work $\rho(W,t_s)$ performed during an ensemble of independent switching measurements at a given switching time t_s. The vertical line represents a δ function at $W=\Delta F$ and corresponds to $t_s \rightarrow \infty$; in that limit, the work performed during a single switching process is exactly equal to ΔF. The smooth distribution represents $\rho(W,t_s)$ for a finite value of t_s. In this case the ensemble-average work exceeds the free-energy difference, $\overline{W} > \Delta F$, since energy is dissipated in a finite-time (irreversible) process.

master equation. We will assume that this stochastic evolution is Markovian and that it satisfies detailed balance.

In presenting this alternative derivation, we are motivated by several factors. First, master equations are a common tool in statistical physics; therefore, it is reassuring to see that Eq. (3) follows in a natural way from the master-equation approach. Second, in this derivation there is no need to explicitly assume weak coupling between the system and reservoir since a reservoir *per se* does not enter into the analysis; given the assumptions stated below, Eq. (3) is *identically* true. Finally, one might come away from Ref. [4] with the feeling that the validity of Eq. (3) depends directly on the properties of Hamiltonian evolution, in particular Liouville's theorem. The treatment herein dispels this notion: Hamilton's equations appear nowhere in the derivation. This point is particularly relevant in the context of numerical simulations, where the evolution of a thermostated system is often realized with the use of non-Hamiltonian equations of motion.

The plan of this paper is as follows. In Sec. I we establish notation and terminology and specify precisely our assumptions regarding Markovian evolution and detailed balance. In Sec. II we derive our central result. In Sec. III we consider specific examples of stochastic processes for which the result is valid. In Sec. IV we briefly discuss the possible utility of Eq. (3) to the numerical computation of free-energy differences. In Sec. V we illustrate our central result with numerical simulations. We conclude with a few remarks in Sec. VI. Two appendixes provide derivations of results used in the main text.

I. PRELIMINARIES

We begin by specifying precisely what we mean by the terms *work* and *free energy* throughout this paper. We as-

sume that there exists a phase space of variables (e.g., the positions and momenta of constituent particles) such that the instantaneous microscopic state of our system is completely described by the values of these variables. Let $\mathbf{z}$ denote a point in this phase space. The evolution of our system with time is then described by a trajectory $\mathbf{z}(t)$. The kind of evolution that will most interest us is not the evolution of an isolated system, but rather that of a system in contact with a heat bath. Hence the trajectory $\mathbf{z}(t)$ will in general be *stochastic*, reflecting the "random" influence of the heat bath.

Next, assume the existence of a parameter-dependent Hamiltonian $H_\lambda(\mathbf{z})$. The Hamiltonian is just a function that, for a fixed value of λ, gives the total energy of the system of interest, in terms of its instantaneous state ($\mathbf{z}$). The value of λ thus parametrizes the external forces acting on the system (arising from, e.g., external fields and confining potentials). When the system is isolated, H_λ happens to generate the time evolution of the system, through Hamilton's equations, but we will not make use of this property in deriving our central result.

Given $H_\lambda(\mathbf{z})$, we now define

$$Z_\lambda(\beta) \equiv \int d\mathbf{z} \, \exp[-\beta H_\lambda(\mathbf{z})], \tag{4a}$$

$$F_\lambda(\beta) \equiv -\beta^{-1} \ln Z_\lambda(\beta), \tag{4b}$$

where β is a real, positive constant. (We are being a bit cavalier with units here: Z_λ should be divided by a constant to make it dimensionless. However, since this constant only shifts the value of F_λ by a fixed amount and therefore does not affect the free-energy difference $F_1 - F_0$, we ignore it.)

$Z_\lambda(\beta)$ and $F_\lambda(\beta)$ are of course the *partition function* and *free energy* of the system, respectively, and β^{-1} is the temperature, in units of energy ($\beta \equiv 1/k_B T$). However, in following the derivation presented below, it may be most convenient to view these quantities abstractly rather than in connection with their physical significance: β is just some positive constant and $Z_\lambda(\beta)$ and $F_\lambda(\beta)$ are the functions of β defined by Eq. (4). In what follows, we will never compare free energies or partition functions at different temperatures, hence the dependence of F_λ and Z_λ on β will be left implicit. The central quantity of interest in this paper will be the free-energy difference $\Delta F \equiv F_1 - F_0$ for a fixed value of β.

As mentioned, the time evolution of the system of interest is described entirely by a (stochastic) phase-space trajectory $\mathbf{z}(t)$. In general, this evolution will depend on the externally imposed time dependence of λ, describing the changing external fields to which the system is subject. Let us consider the evolution of the system from an initial time $t = 0$ to a final time $t = t_s$ over which the value of λ is switched from 0 to 1 at a uniform rate $\lambda(t) = t/t_s$. Given this time dependence of λ and given the trajectory $\mathbf{z}(t)$ describing the evolution of the system, the total *work* performed on the system is the time integral of $\dot{\lambda} \partial H_\lambda / \partial \lambda$ along the trajectory

$$W \equiv \int_0^{t_s} dt \, \dot{\lambda} \frac{\partial H_\lambda}{\partial \lambda} (\mathbf{z}(t)), \tag{5}$$

where $\dot{\lambda} = d\lambda/dt = t_s^{-1}$. For the evolution of an isolated Hamiltonian system, this reduces to $W = H_1(\mathbf{z}(t_s)) - H_0(\mathbf{z}(0))$, by Hamilton's equations; in this case the work performed on the system is just the change in its energy. For a system in contact with a heat bath, however, this no longer holds since there is a constant exchange of energy with the bath.

We will assume that the evolution of our system in phase space is a *Markov process* [5]. This means that the stochastic evolution $\mathbf{z}(t)$ is entirely characterized by the transition probability function $P(\mathbf{z}',t|\mathbf{z},t+\Delta t)$. This gives the probability distribution for finding the system in a state $\mathbf{z}$ at time $t + \Delta t$, given that at an earlier time t it was known to be at $\mathbf{z}'$. Taking the derivative of P with respect to Δt and evaluating at $\Delta t \to 0^+$, we get a function

$$R(\mathbf{z}',\mathbf{z};t) \equiv \frac{\partial}{\partial(\Delta t)} P(\mathbf{z}',t|\mathbf{z},t+\Delta t)|_{\Delta t \to 0^+}, \tag{6}$$

which gives the instantaneous transition rate from $\mathbf{z}'$ to $\mathbf{z}$, at time t. The dependence of R on time arises through whatever external parameters of the system and reservoir are available and time dependent. In our case we assume only one such parameter λ, characterizing external forces; therefore, we write

$$R(\mathbf{z}',\mathbf{z};t) \to R_\lambda(\mathbf{z}',\mathbf{z}). \tag{7}$$

In other words, the instantaneous transition rate R from $\mathbf{z}'$ to $\mathbf{z}$ depends on t only through the value of $\lambda(t)$.

Let us now shift our focus from the description of a single system to that of an *ensemble* of systems, each evolving according to the stochastic Markov process just described. This ensemble represents infinitely many independent real-

izations of the switching process. If $f(\mathbf{z},t)$ denotes the time dependent distribution of this ensemble in phase space, then this distribution obeys

$$\frac{\partial f}{\partial t}(\mathbf{z},t) = \int d\mathbf{z}' f(\mathbf{z}',t) R_\lambda(\mathbf{z}',\mathbf{z}), \tag{8a}$$

where $\lambda = \lambda(t)$. We will abbreviate this as

$$\frac{\partial f}{\partial t} = \hat{R}_\lambda f, \tag{8b}$$

where $\hat{R}_\lambda$ is a linear operator acting on the space of phase-space densities f. Equation (8) is our master equation; when the time dependence of λ is known and an initial distribution f_0 is specified, Eq. (8) uniquely determines the subsequent evolution of the phase-space density f.

In addition to the Markov assumption [Eq. (8)], we will impose another assumption on our stochastic process: detailed balance. If λ is held fixed, $\mathbf{z}(t)$ becomes a *stationary* Markov process, describing the evolution of a system in contact with a heat reservoir, when the external forces acting on the system are time independent. Under such evolution, the canonical distribution in phase space (corresponding to the fixed value of λ) ought to be invariant. By Eq. (8), this is equivalent to the statement that $\hat{R}_\lambda$ annihilates the canonical distribution:

$$\hat{R}_\lambda \exp[-\beta H_\lambda(\mathbf{z})] = 0. \tag{9}$$

This places a condition on the linear operator $\hat{R}_\lambda$. We assume that our stochastic process satisfies this condition and will refer to this assumption as *detailed balance*. (Usually, detailed balance is expressed as the somewhat stronger assumption

$$\frac{R_\lambda(\mathbf{z}',\mathbf{z})}{R_\lambda(\mathbf{z},\mathbf{z}')} = \frac{\exp[-\beta H_\lambda(\mathbf{z})]}{\exp[-\beta H_\lambda(\mathbf{z}')]}. \tag{10}$$

Equation (9) follows immediately from Eq. (10). Just set the products of the cross terms equal and then integrate over $\mathbf{z}'$. The distinction between Eqs. (9) and (10) is of little importance in the present context, so for simplicity we refer to Eq. (9) as detailed balance.)

Having been led to assume detailed balance by considering the behavior of the system when λ is fixed, it may seem natural to make another assumption. Namely, if our stochastic process is meant to describe a system in contact with a reservoir, then for fixed λ we expect the system to thermalize: After an initial relaxation time, the system samples its phase space canonically; equivalently, an arbitrary initial ensemble $f_0(\mathbf{z})$ relaxes to a canonical distribution in phase space. We may express this assumption formally as

$$\lim_{t \to \infty} \hat{U}_\lambda(t) f_0(\mathbf{z}) = \frac{1}{Z_\lambda} \exp[-\beta H_\lambda(\mathbf{z})] \tag{11}$$

for any normalized $f_0(\mathbf{z})$. The operator $\hat{U}_\lambda(t) \equiv \exp(\hat{R}_\lambda t)$ appearing on the left-hand side is just the evolution operator corresponding to the equation of motion $\partial f/\partial t = \hat{R}_\lambda f$, for

fixed λ. Equation (11) says that any initial distribution $f_0(\mathbf{z})$ relaxes to a canonical distribution and then stays there. We will refer to this as the assumption of *thermalization*. Note that while thermalization [Eq. (11)] implies detailed balance [Eq. (9)], the converse is not true. Since it will turn out that the proof of Eq. (3) requires only the weaker assumption of detailed balance and not the stronger thermalization assumption, we will assume in what follows that Eq. (9) holds, *but not necessarily Eq. (11)*, unless explicitly stated.

Physically, we expect both thermalization and detailed balance to hold as long as our master equation describes a system in contact with a genuine heat reservoir. However, one can easily imagine situations in which the evolution satisfies Eq. (9) but not Eq. (11). A specific example, which we will analyze in Sec. III, is that of an isolated Hamiltonian system. Note that if thermalization does not hold, then the validity of Eq. (1), $W_\infty = \Delta F$, is not guaranteed since that result assumes that during an infinitely slow switching process the system remains in quasistatic equilibrium with a reservoir. Equation (3) nevertheless remains valid, provided detailed balance is satisfied.

At this point we are ready to proceed with a proof of Eq. (3). Concrete examples of stochastic processes satisfying the assumptions made in this section will be discussed in Sec. III, and numerically simulated in Sec. V.

II. DERIVATION

The overbar appearing on the left-hand side of Eq. (3) denotes an average over an infinite ensemble of independent realizations (repetitions) of the switching process. Each realization is described by a trajectory $\mathbf{z}(t)$, $0 \leq t \leq t_s$, specifying the evolution of the complete set of phase-space variables as the external parameter λ is switched from 0 to 1. The entire ensemble is then described by a time-dependent phase-space density: At any time t, $f(\mathbf{z},t)$ represents a snapshot of the distribution of trajectories in phase space. Since we assume that the system equilibrates with the reservoir (with λ held fixed at 0) before the start of each realization, we have a canonical distribution of initial conditions [6]

$$f(\mathbf{z},0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z})]. \tag{12}$$

During the switching process, however, the ensemble does not (in general) remain in instantaneous canonical equilibrium. In other words, although the distribution $f = Z_\lambda^{-1} \exp(-\beta H_\lambda)$ is a (stationary) solution of Eq. (8) for λ *fixed*, it is not, in general, a solution of Eq. (8) when λ depends on time. Thus, for $t > 0$, a snapshot of the trajectories will reveal a distribution $f(\mathbf{z},t)$ that lags behind the canonical distribution corresponding to $\lambda(t)$. The amount of lag present by the time a given value of λ is reached will depend on how rapidly or slowly we performed the switching on the way to that value.

For every trajectory $\mathbf{z}(t)$ in our ensemble, we can compute the total work W performed on the system [Eq. (5)]. Our task is now to evaluate the ensemble average of $\exp(-\beta W)$. To do this, let us first define, for a given trajectory $\mathbf{z}(t)$, a function $w(t)$ that is the "work accumulated" up to time t:

$$w(t) = \int_0^t dt' \dot{\lambda} \frac{\partial H_\lambda}{\partial \lambda}(\mathbf{z}(t')). \tag{13}$$

Thus $W = w(t_s)$. Now consider all those trajectories in the ensemble that happen to pass through the phase-space point $\mathbf{z}$ at time t and let $Q(\mathbf{z},t)$ denote the average value of $\exp[-\beta w(t)]$ over this particular subset of trajectories. Finally, define

$$g(\mathbf{z},t) = f(\mathbf{z},t) Q(\mathbf{z},t). \tag{14}$$

Note that $g(\mathbf{z},0) = f(\mathbf{z},0)$ since $w(0) = 0$ for every trajectory. Given these definitions, the ensemble average of $\exp(-\beta W)$ may be expressed as

$$\overline{\exp(-\beta W)} = \int d\mathbf{z}\, g(\mathbf{z},t_s). \tag{15}$$

We will now solve for $g(\mathbf{z},t_s)$.

The function $g(\mathbf{z},t)$ obeys

$$\frac{\partial g}{\partial t} = \left(\hat{R}_\lambda - \beta \dot{\lambda} \frac{\partial H_\lambda}{\partial \lambda} \right) g, \tag{16}$$

with $\lambda = \lambda(t)$. To see this, imagine for a moment that each trajectory $\mathbf{z}(t)$ in the ensemble represents a "particle" moving about in phase space. Furthermore, assume that the "mass" of each particle is time dependent and is given by $\mu(t) = \exp[-\beta w(t)]$. Each particle thus begins with mass unity $\mu(0) = 1$. The function $Q(\mathbf{z},t)$ is then the average mass of those particles that are found at a point $\mathbf{z}$ at time t and $g(\mathbf{z},t)$ represents the time-dependent *mass density* in phase space (after normalization of the number density to unity $\int f d\mathbf{z} = 1$). This mass density is time dependent for two reasons. First, (a) the mass of each particle changes with time at a rate proportional to its instantaneous mass:

$$\dot{\mu}(t) = -\beta \dot{w}(t) \mu(t) = -\beta \dot{\lambda} \frac{\partial H_\lambda}{\partial \lambda}(\mathbf{z}(t)) \mu(t). \tag{17}$$

This contributes a term

$$\left. \frac{\partial g}{\partial t} \right|_A = -\beta \dot{\lambda} \frac{\partial H_\lambda}{\partial \lambda}(\mathbf{z}) g(\mathbf{z},t) \tag{18}$$

to $\partial g / \partial t$. Second, (b) the mass density evolves due to the flow of particles, described by our master equation $\partial f / \partial t = \hat{R}_\lambda f$. This flow of particles contributes a term

$$\left. \frac{\partial g}{\partial t} \right|_B = \hat{R}_\lambda g. \tag{19}$$

Adding these contributions (a) and (b) gives Eq. (16). An alternative derivation of this evolution equation, based on a path-integral formulation, is given in Appendix A.

Given the initial conditions $g(\mathbf{z},0) = f(\mathbf{z},0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z})]$, Eq. (16) is solved by

$$g(\mathbf{z},t) = Z_0^{-1} \exp[-\beta H_\lambda(\mathbf{z})] = \frac{Z_\lambda}{Z_0} f_\lambda^C(\mathbf{z}), \tag{20}$$

where $\lambda = \lambda(t)$ and f_λ^C denotes the canonical distribution in phase space for a given value of λ. This result is easily verified with the help of Eq. (9). Then, using Eq. (15), we finally arrive at

$$\overline{\exp(-\beta W)} = Z_0^{-1} \int d\mathbf{z} \, \exp[-\beta H_1(\mathbf{z})]$$

$$= \frac{Z_1}{Z_0} = \exp(-\beta \Delta F). \tag{21}$$

Q.E.D. A different proof of Eq. (3) has been discovered by Crooks [7].

It is worthwhile to draw attention to a curious feature of the evolution of our ''mass density'' $g(\mathbf{z},t)$. Recall that the evolution of $f(\mathbf{z},t)$ depends nontrivially on the rate at which we switch λ: At finite switching rates, $f(\mathbf{z},t)$ lags behind the instantaneous equilibrium distribution. By contrast, the time dependence of $g(\mathbf{z},t)$ is very simple: The mass density is determined uniquely by the instantaneous value of λ; see Eq. (20). Thus no matter how slowly or rapidly we switch λ from 0 to 1, $g(\mathbf{z},t)$ will always evolve through exactly the same continuous sequence of ''canonical'' mass densities, specified by Eq. (20); g *never develops a lag* in the sense that f does.

Equation (20) also implies the result

$$\overline{\exp[-\beta w(t)]} = \frac{Z_\lambda}{Z_0} = \exp[-\beta(F_\lambda - F_0)], \tag{22}$$

$\lambda = \lambda(t)$, which is in effect a restatement of our central result, Eq. (3).

III. EXAMPLES

We have shown that Eq. (3) is true for Markovian processes satisfying detailed balance. Let us consider a few examples of such processes.

A. Hamiltonian evolution

As a first case, we take deterministic evolution under the Hamiltonian H_λ, as λ is varied from 0 to 1:

$$\dot{\mathbf{z}} = \{\mathbf{z}, H_\lambda\}, \tag{23}$$

where $\{ , \}$ denotes the Poisson bracket. Ordinarily, one would not call this a stochastic process, but of course it may be viewed as a special case of such. This evolution is Markovian and $\hat{R}_\lambda$ is just the Poisson bracket operator

$$\hat{R}_\lambda f = \{H_\lambda, f\}. \tag{24}$$

It immediately follows that detailed balance [Eq. (9)] is satisfied. The central result of this paper then tells us that if we (i) start with a canonical distribution of initial conditions $f(\mathbf{z},0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z})]$ and (ii) allow trajectories to evolve deterministically from these initial conditions [Eq. (23)] as λ is varied from 0 to 1, then this ensemble of trajectories will satisfy Eq. (3), regardless of how slowly or quickly we perform the switching. This result was proven more directly in Ref. [4].

Since evolution under Eq. (23) describes an isolated system (no heat bath), the thermalization assumption [Eq. (11)] is not met and the work performed in the limit of infinitely slow switching will in general not equal the free-energy difference: $W_\infty \neq \Delta F$. Equation (3) nevertheless remains true, even in this limit.

This last statement is easily illustrated with a one-dimensional harmonic oscillator. Let the Hamiltonian be

$$H_\lambda(x,p) = \frac{p^2}{2} + \omega_\lambda^2 \frac{x^2}{2}. \tag{25}$$

For a given temperature, the partition function is given by $Z_\lambda = 2\pi/\beta\omega_\lambda$ and the free-energy difference is

$$\Delta F = -\beta^{-1} \ln \frac{Z_1}{Z_0} = \beta^{-1} \ln \frac{\omega_1}{\omega_0}. \tag{26}$$

Let us now imagine a trajectory $\mathbf{z}(t)$ evolving under this Hamiltonian, as λ is changed infinitely slowly from 0 to 1; assume for specificity that $\omega_1 > \omega_0$. Since H_λ/ω_λ is an adiabatic invariant for the harmonic oscillator [8], we have $E_1 = (\omega_1/\omega_0)E_0$, where E_0 (E_1) is the initial (final) energy of the oscillator. The work performed on an isolated system is equal to the change in its energy, so we get $W_\infty = [(\omega_1/\omega_0) - 1]E_0$. A canonical distribution of initial energies E_0 then leads, after some algebra, to the following ensemble distribution of values of W:

$$\lim_{t_s \to \infty} \rho(W,t_s) = \frac{\omega_0 \beta}{\omega_1 - \omega_0} \exp\left(-\frac{\omega_0 \beta W}{\omega_1 - \omega_0}\right) \theta(W), \tag{27}$$

where θ denotes the unit step function. It is straightforward to verify that this distribution satisfies Eq. (3).

As another example, consider a single particle bouncing around inside a three-dimensional cavity with hard walls, where the shape of the cavity is a function of λ. Let $\mathcal{V}_\lambda$ denote the volume of the cavity. The free-energy difference is then $\Delta F = \beta^{-1} \ln(\mathcal{V}_0/\mathcal{V}_1)$. Assume that, when λ is held fixed, the motion of the particle is ergodic over the energy shell (the five-dimensional surface of constant energy in six-dimensional phase space); also assume that $\mathcal{V}_1 \lesssim \mathcal{V}_0$. Now, imagine that we allow the particle to evolve as λ is switched infinitely slowly from 0 to 1. For this system the quantity $H_\lambda^{3/2}\mathcal{V}_\lambda$ is an adiabatic invariant [9]; therefore the total work performed on a particle with initial energy E_0 is $W_\infty = [(\mathcal{V}_0/\mathcal{V}_1)^{2/3} - 1]E_0$. Taking an ensemble of such particles, defined by a canonical distribution of initial conditions (at $\lambda = 0$), we get

$$\lim_{t_s \to \infty} \rho(W,t_s) = \left(\frac{4\beta^3 W}{\pi r^3}\right)^{1/2} \exp\left(-\frac{\beta W}{r}\right) \theta(W), \tag{28}$$

where $r = (\mathcal{V}_0/\mathcal{V}_1)^{2/3} - 1$. As with the example of the harmonic oscillator, it is straightforward to show that this distribution ρ satisfies Eq. (3).

B. Langevin evolution

Next, let us consider modifying Hamilton's equations by adding both a frictional and a stochastic force

$$\dot{x} = p, \tag{29a}$$

$$\dot{p} = -\partial_x V_\lambda - \gamma p + \widetilde{F}(t). \tag{29b}$$

We have assumed a one-degree-of-freedom system and a Hamiltonian $H_\lambda = p^2/2 + V_\lambda(x)$. Let us furthermore take the stochastic force $\widetilde{F}(t)$ to represent white noise, with an auto-correlation function given by

$$\langle \widetilde{F}(t_1)\widetilde{F}(t_2)\rangle = D_P \delta(t_2 - t_1). \tag{30}$$

Finally, let us impose a fluctuation-dissipation relation between the frictional and stochastic forces

$$\gamma = \beta D_P/2. \tag{31}$$

An ensemble of trajectories evolving under the stochastic process just defined satisfies the Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \{H_\lambda, f\} + \gamma \frac{\partial}{\partial p}(pf) + \frac{D_P}{2}\frac{\partial^2 f}{\partial p^2}. \tag{32}$$

Since Eq. (32) is of the form $\partial f/\partial t = \hat{R}_\lambda f$, this process is Markovian. Furthermore, inspection reveals that $\hat{R}_\lambda$ satisfies Eq. (9), i.e., detailed balance holds. Thus the conditions for the validity of Eq. (3) are met: Given an ensemble of systems evolving under this stochastic process, as λ is changed from 0 to 1 over a time t_s, and given an initial distribution $f(\mathbf{z},0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z})]$, we are guaranteed that this ensemble will satisfy $\overline{\exp(-\beta W)} = \exp(-\beta \Delta F)$ for any switching time t_s.

Equation (29) is a *Langevin* set of equations. If λ is held fixed, an ensemble of trajectories evolving under these stochastic equations of motion will relax to a canonical distribution on phase space. In other words, the evolution specified by Eqs. (29)–(31) satisfies not only detailed balance, but the stronger assumption of thermalization as well. If we start with a canonical distribution, at $\lambda = 0$, and then switch the value of the external parameter infinitely slowly to $\lambda = 1$, then the ensemble of trajectories will remain in quasistatic equilibrium over the course of the switching process:

$$f(\mathbf{z},t) = Z_\lambda^{-1}\exp[-\beta H_\lambda(\mathbf{z})], \quad \lambda = \lambda(t), \quad t_s \to \infty. \tag{33}$$

The creeping value of λ thus drags $f(\mathbf{z},t)$ through a continuous sequence of canonical distributions. Furthermore, in this limit, the "work accumulated" will be the same function of time for every trajectory in the ensemble: $w(t) = F_\lambda - F_0$; each trajectory represents a system evolving in quasistatic equilibrium with the reservoir. Thus the quantity $Q(\mathbf{z},t)$ defined earlier is given, in this quasistatic limit, by

$$Q(\mathbf{z},t) = \exp[-\beta(F_\lambda - F_0)] = Z_\lambda/Z_0. \tag{34}$$

It then follows that the function $g \equiv fQ$ [Eq. (14)] satisfies $g(\mathbf{z},t) = Z_0^{-1}\exp[-\beta H_\lambda(\mathbf{z})]$ [Eq. (20)], from which we get $\overline{\exp(-\beta W)} = \exp(-\beta \Delta F)$. Of course, when we switch λ from 0 to 1 over a *finite* time t_s, neither Eq. (33) nor Eq. (34) will in general hold; nevertheless, Eq. (20), and therefore our central result, will remain valid since Eq. (16) governs the

time evolution of g regardless of how slowly or quickly we switch the external parameter.

C. Isothermal molecular dynamics

The Langevin equations above provide a simple method for numerically simulating the evolution of a thermostated system (i.e., a system in contact with a heat bath), without explicitly simulating the many degrees of freedom of the bath. The term $\widetilde{F}(t)$ may be implemented by generating a small, random momentum "kick" at each time step in the numerical integration of the equations of motion. This method works both for static Hamiltonians and, as in the case considered in the present paper, for Hamiltonians made time dependent through the variation of an external parameter. In the past decade or so, methods for using explicitly *deterministic* equations of motion to simulate thermostated systems have proven very useful [10]. These generally go under the name of *isothermal molecular dynamics* (IMD). Typically, the heat bath is represented by one or more "extra" degrees of freedom. Then, in the extended phase space that includes both the system of interest and the extra degree(s) of freedom, the evolution is governed by a set of deterministic (but non-Hamiltonian) equations of motion. These are tailored so that when the Hamiltonian describing the system of interest is static, the variables representing the system of interest explore phase space canonically, at least to a good approximation.

An example of an IMD scheme is *Nosé-Hoover dynamics* [11,12], represented by the equations of motion

$$\{\dot{q} = p, \ \dot{p} = -\nabla V_\lambda - \zeta p\}_n, \tag{35a}$$

$$\dot{\zeta} = (K/K_0 - 1)/\tau^2. \tag{35b}$$

We have again assumed a kinetic-plus-potential Hamiltonian; the index n runs over all D degrees of freedom of the system, $K = \sum_n p_n^2/2$ is the total kinetic energy of the system, $K_0 = D/2\beta$ is the thermal average of K, and the parameter τ acts as a relaxation time. Let us imagine a trajectory evolving in the extended phase space, $(\mathbf{z},\zeta)$ space, under these equations of motion, as λ is switched from 0 to 1. The work W performed on the system of interest is defined, as before, by Eq. (5). While this expression does not explicitly contain the bath variable ζ, W is nevertheless a function of the full set of initial conditions $(\mathbf{z}_0,\zeta_0)$ since the evolution of the system of interest ($\mathbf{z}$) is coupled to that of the heat bath (ζ).

In Ref. [4], it was shown that if one started with a distribution of initial conditions in the extended phase space given by

$$f(\mathbf{z},\zeta,0) \propto \exp\{-[\beta H_0(\mathbf{z}) + \zeta^2\tau^2/2]\} \tag{36}$$

and if one then propagated an ensemble of trajectories from these initial conditions, under the Nosé-Hoover (NH) equations of motion, as λ was switched from 0 to 1 over a time t_s and computed the work W for each trajectory, then Eq. (3) would hold identically for this ensemble. This was shown by inspection.

Our purpose here is to use the central result of the present paper to establish simple criteria for determining whether or not Eq. (3) is valid for a particular implementation of IMD.

In general, the heat bath is represented by N variables, ζ_1 to ζ_N. Let $\mathbf{y}=(\mathbf{z},\zeta_1,\ldots,\zeta_N)$ denote a point in the $(2D+N)$-dimensional extended phase space, where D is the number of degrees of freedom of the system of interest. Then the thermostating scheme in question is defined by a set of deterministic equations of motion in this space:

$$\dot{\mathbf{y}}=\mathbf{K}_\lambda(\mathbf{y}), \tag{37}$$

of which Eq. (35) is an example. An ensemble of trajectories evolving under these equations of motion may be described by a distribution $f(\mathbf{y},t)$, which satisfies the continuity equation

$$\frac{\partial f}{\partial t}=-\frac{\partial}{\partial\mathbf{y}}\cdot(\mathbf{K}_\lambda f\,)\equiv\hat{R}_\lambda f. \tag{38}$$

Now suppose that we can find a function $q(\zeta_1,\ldots,\zeta_N)$ such that the distribution

$$f_\lambda^C\propto\exp\{-\beta[H_\lambda(\mathbf{z})+q(\zeta_1,\ldots,\zeta_N)]\} \tag{39}$$

is stationary under the evolution defined by Eq. (37) when λ is held fixed. That is,

$$\hat{R}_\lambda f_\lambda^C=0. \tag{40}$$

(For the NH equations, $q=\zeta^2\tau^2/2\beta$ satisfies this condition.) We will allow q to depend on β and any other *constant* parameters, but not on λ. The distribution $f_\lambda^C(\mathbf{y})$ may be viewed as the canonical distribution in the extended phase space since it is invariant when λ is held fixed. Now define a function

$$\mathcal{H}_\lambda(\mathbf{y})\equiv H_\lambda(\mathbf{z})+q(\zeta_1,\ldots,\zeta_N), \tag{41}$$

which we may think of as an *extended Hamiltonian*. [This is not meant to imply that $\mathcal{H}_\lambda$ generates Eq. (37) as a ''real'' Hamiltonian would; in general those equations are non-Hamiltonian.] By Eq. (40)

$$\hat{R}_\lambda\exp[-\beta\mathcal{H}_\lambda(\mathbf{y})]=0. \tag{42}$$

Now forget for a moment the division between the system of interest and the heat bath and treat the entire extended phase space as a phase space for some system of interest, for which a parameter-dependent energy function $\mathcal{H}_\lambda(\mathbf{y})$ is defined. The evolution given by Eq. (37) then satisfies the two conditions listed in Sec. II: (i) it is Markovian [Eq. (38)], and (ii) it satisfies detailed balance [Eq. (42)]. Thus our central result $\overline{\exp(-\beta W)}=\exp(-\beta\Delta F)$ is identically true for an ensemble of trajectories evolving under Eq. (37) from an initial distribution $f_0\propto\exp(-\beta\mathcal{H}_0)$, *provided we replace $H_\lambda(\mathbf{z})$ by $\mathcal{H}_\lambda(\mathbf{y})$ in computing W and ΔF*. However, it is easily verified that we will get exactly the same values for both W and ΔF using $\mathcal{H}_\lambda(\mathbf{y})$, as we would obtain with $H_\lambda(\mathbf{z})$. In the case of W, this is because only the first term of $\mathcal{H}_\lambda$ (namely, H_λ) depends on λ; for ΔF, it follows from the fact that the partition function for the extended Hamiltonian factorizes into an integral over the $\mathbf{z}$ variables and an integral over the ζ variables and only the former depends on λ. [See the defini-

tions of work and free energy, Eqs. (4) and (5).] These considerations lead to the following simple conclusion.

Suppose we have a system described by a Hamiltonian $H_\lambda(\mathbf{z})$ and we wish to compute the free-energy difference $\Delta F=F_1-F_0$. Suppose furthermore that we are given a scheme for IMD; that is, we have a set of ''heat bath'' variables $(\zeta_1,\ldots,\zeta_N)$, along with equations of motion in the extended phase space, as per Eq. (37). Then, if a function q of the heat bath variables can be found such that the distribution $f_\lambda^C\propto\exp[-\beta(H_\lambda+q)]$ is stationary under the equations of motion (with λ fixed), then we can compute ΔF as follows. Let an ensemble of trajectories evolve from an initial distribution $f_0\propto\exp[-\beta(H_0+q)]$, as λ is switched from 0 to 1 over a finite switching time t_s. Compute the work W for each trajectory, as per Eq. (5), and then compute the ensemble average of $\exp(-\beta W)$. This average will equal $\exp(-\beta\Delta F)$.

D. Monte Carlo evolution

As a final example, again, motivated by computer simulations, let us consider the evolution of a *Monte Carlo* (MC) trajectory, as λ is switched from 0 to 1. In this case both the trajectory and the parameter λ evolved in discrete steps, rather than continuously:

$$\mathbf{z}(t)\rightarrow\mathbf{z}_0,\mathbf{z}_1,\ldots,\mathbf{z}_N, \tag{43}$$

$$\lambda(t)\rightarrow\lambda_0,\lambda_1,\ldots,\lambda_N\ ,\quad\lambda_n=n/N. \tag{44}$$

The initial point in phase space $\mathbf{z}_0$ is sampled from a canonical distribution, with the value of λ fixed at $\lambda_0=0$. One then imagines that λ changes abruptly from $\lambda_0=0$ to $\lambda_1=1/N$; as a result, a quantity of work $\delta W_1=H_{\lambda_1}(\mathbf{z}_0)-H_{\lambda_0}(\mathbf{z}_0)$ is performed on the system. Then, the system jumps to the next point in phase space $\mathbf{z}_1$ generated from $\mathbf{z}_0$ by a MC algorithm appropriate to the Hamiltonian H_{λ_1} (see Appendix B). One then continues to alternate between discrete changes in λ and discrete MC jumps in phase space, until the entire ''trajectory'' $(\mathbf{z}_0,\ldots,\mathbf{z}_N)$ is obtained and the value of λ is 1. The total work performed during this discrete switching process is

$$W=\sum_{n=1}^{N}\delta W_n=\sum_{n=1}^{N}[H_{\lambda_n}(\mathbf{z}_{n-1})-H_{\lambda_{n-1}}(\mathbf{z}_{n-1})]. \tag{45}$$

Note that ''time'' does not enter into this scheme. The quasistatic limit is obtained by letting the number of steps become arbitrarily large: $N\rightarrow\infty$.

Let us now imagine that we generate an infinite ensemble of MC trajectories, each of length N. We do this by implementing the above procedure repeatedly, each time feeding in a different string of random numbers to generate the initial conditions $\mathbf{z}_0$ and the subsequent MC steps. Given such an ensemble, we compute the work W performed on each trajectory [Eq. (45)] and then the ensemble average of $\exp(-\beta W)$. As shown in Appendix B, this average will equal $\exp(-\beta\Delta F)$. This should come as no surprise: A trajectory generated by the MC algorithm is a *Markov chain*, with detailed balance built into the individual steps. This evolution thus satisfies, in discretized form, the two assumptions required for the validity of Eq. (3).

Let us now consider the two limiting cases $N=1$ and $N\rightarrow\infty$. The latter is the MC equivalent of the quasistatic limit. In this case our ensemble proceeds through a discrete, infinitesimally spaced sequence of canonical equilibrium distributions, and the work performed on each trajectory is ΔF, as per Eq. (1). The result $\overline{\exp(-\beta W)}=\exp(-\beta\Delta F)$ then follows automatically. In the opposite limit $N=1$, the work W performed on a particular trajectory is given by

$$W=H_1(\mathbf{z}_0)-H_0(\mathbf{z}_0)\equiv\Delta H(\mathbf{z}_0) \quad (N=1). \qquad (46)$$

Equation (3) then reduces to

$$\langle\exp(-\beta\Delta H)\rangle_0=\exp(-\beta\Delta F), \qquad (47)$$

where $\langle\ \rangle_0$ denotes a canonical average with respect to $\lambda=0$. This result is a well-known identity for the free-energy difference ΔF; see Eq. (48) below.

IV. FREE-ENERGY COMPUTATIONS

While Eq. (3) is interesting in its own right, it may additionally prove useful in the numerical computation of free-energy differences. The field of free-energy computations is decades old, with diverse applications, and a very large body of literature exists on the subject [13]. In this section, without attempting a survey of the field, we discuss a few points relevant to the possible application of Eq. (3) to free-energy computations. These comments expand on ones made in Ref. [4].

Most methods of computing free-energy differences are variants of either the *thermodynamic integration* (TI) or *thermodynamic perturbation* (TP) methods. (For an exception to this statement, see the work of Holian, Posch, and Hoover [14], where two different expressions for ΔF, based on time-integrated heat transfer, are derived within the framework of isothermal molecular dynamics.) TP is based on the identity [15]

$$\Delta F=-\beta^{-1}\ln\langle\exp(-\beta\Delta H)\rangle_0, \quad (\Delta H\equiv H_1-H_0), \qquad (48)$$

where $\langle\ \rangle_\lambda$ denotes a canonical average with respect to a fixed value of λ. Using a method such as Monte Carlo, one samples N points in phase space from the canonical distribution corresponding to $\lambda=0$ and then one takes the average of $\exp(-\beta\Delta H)$ over these N points. In principle, the method is exact for $N\rightarrow\infty$; in practice, unless the canonical distributions corresponding to H_0 and H_1 overlap to a significant degree, the average of $\exp(-\beta\Delta H)$ will be dominated by points in phase space that are visited extremely rarely during the canonical sampling, so numerical convergence with N will be prohibitively slow. One way to get around this problem is to break up the λ interval $[0,1]$ into small subintervals and then use Eq. (48) to compute the free-energy difference corresponding to each subinterval. Other, more sophisticated methods of extracting the best efficiency from Eq. (48) have been developed over the years [16].

TI is based on the identity [17]

$$\Delta F=\int_0^1 d\lambda\left\langle\frac{\partial H_\lambda}{\partial\lambda}\right\rangle_\lambda. \qquad (49)$$

The integral on the right-hand side may be evaluated by sampling n_s points in phase space from each of M different canonical distributions (corresponding to equally spaced values of λ from 0 to 1), for a grand total of $N=n_s M$ points sampled. The average of $\partial H_\lambda/\partial\lambda$ is then computed at each value of λ, and from these M averages the integral is obtained. With the exception of a few very simple systems (e.g., ideal gases and harmonic oscillators), the standard way to obtain a canonical distribution is to first allow the system to ''age,'' that is, to relax to an equilibrium statistical state, under some MC or IMD scheme. In implementing the numerical evaluation of $\int_0^1 d\lambda\langle\partial H_\lambda/\partial\lambda\rangle_\lambda$, it is often too time consuming to age the system independently at each of the M selected values of λ. Instead, the final point sampled at one λ value may be used to generate the initial point sampled at the next value of λ. Of course, this means that at each λ (except $\lambda=0$) we are actually sampling from a slightly out-of-equilibrium distribution, which leads to a systematic error (in fact, an overestimate) in the evaluation of the integral.

A limiting case of this procedure arises when we take $n_s=1$, i.e., exactly one point is sampled at each value of λ. This is the *slow growth* method [18]. As stressed by Reinhardt and co-workers [3,19,20], if we view the chain of points $(\mathbf{z}_0,\ldots,\mathbf{z}_N)$ thus generated as a *trajectory* evolving in phase space (as λ is switched from 0 to 1), then the integral appearing on the right-hand side of Eq. (49) represents the work W performed on the system over the course of the switching process. This picture is a compelling one, as it attaches a very physical interpretation to the numerical evaluation of Eq. (49): Instead of computing an integral, we are simulating the evolution of a thermostated system, with the idea that, in the limit of infinitely slow switching, the work performed on the system will equal the free-energy difference ΔF. For a finite switching rate, the above-mentioned systematic error inherent in the slow growth method is simply a manifestation of the inequality $\overline{W}\geq\Delta F$; see Eq. (2) and Refs. [3,20,21]. This interpretation of free-energy computations is referred to as *adiabatic switching* (or *finite-time variation*) and indeed may be viewed as a separate method, distinct from thermodynamic integration.

In the context of adiabatic switching, Eq. (3) says that if we run an ensemble of finite-time (or finite-N, for Monte Carlo) simulations of the switching process, using, for instance, one of the methods described in Sec. III, then the average of $\exp(-\beta W)$ over this ensemble of simulations will equal $\exp(-\beta\Delta F)$. Assuming perfect numerical accuracy and an infinite ensemble of simulations, this is an *exact* statement. Another way of putting it is as follows: For a single finite-time switching simulation, the value of $\exp(-\beta W)$ provides an *unbiased* estimate of $\exp(-\beta\Delta F)$. By contrast, the value of W gives a biased estimate of ΔF (i.e., $\overline{W}\geq\Delta F$). Indeed, the former statement implies the latter in the same way that Eq. (48) implies the Gibbs-Bogoliubov-Feynman inequality [22]: $\langle\Delta H\rangle_0\geq\Delta F$. Let us consider this for a moment.

Given a real function of a real variable $y(x)$, it is easy to show that if $d^2y/dx^2\geq0$ for all x, then

$$\frac{1}{N}\sum_{i=1}^N y(x_i)\geq y\left(\frac{1}{N}\sum_{i=1}^N x_i\right) \qquad (50)$$
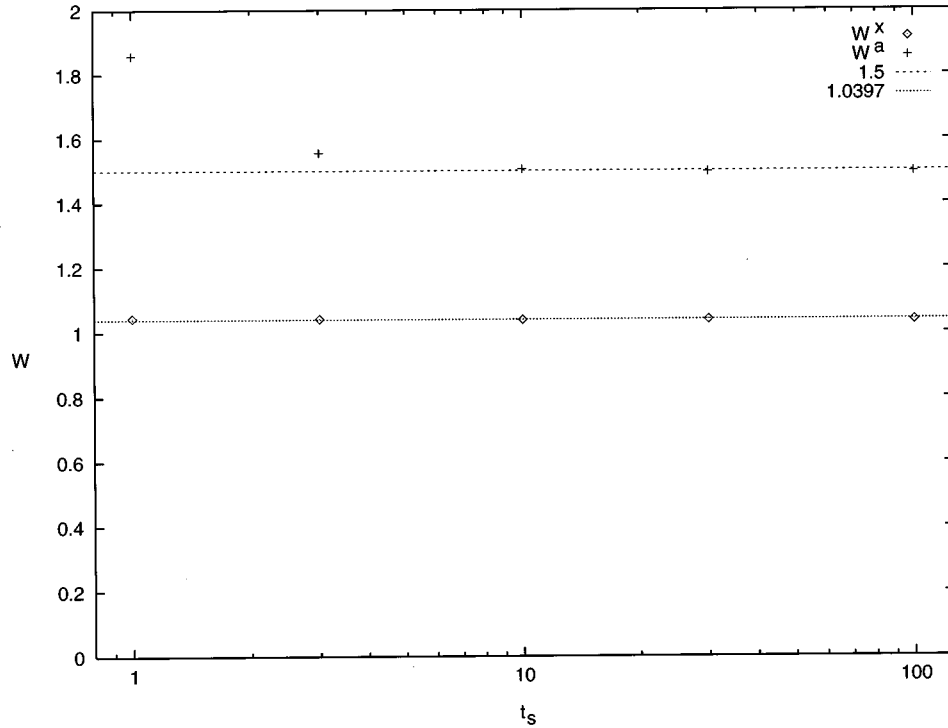
FIG. 2. Simulations of an isolated harmonic oscillator whose natural frequency is switched from $\omega_0 = 1.0$ to $\omega_1 = 2.0$ over a switching time t_s. At each of five values of t_s, 10^5 simulations were carried out. The upper and lower sets of points show the ordinary averages (W^a) and the exponential averages (W^x) of the work, respectively. The dashed line is at $W = 1.5$, the dotted line at $W = \Delta F = 1.0397$.

for any set of points $\{x_1, \ldots, x_N\}$. If these points are the result of random sampling from some ensemble, then Eq. (50), in the limit $N \to \infty$, may be rewritten as

$$\overline{y(x)} \geq y(\overline{x}), \tag{51}$$

where the overbar denotes an ensemble average. Applying this result to $y(x) = \exp x$, with $x = -\beta W$, we get

$$\overline{\exp(-\beta W)} \geq \exp(-\beta \overline{W}). \tag{52}$$

This is identically true for any distribution $\rho(W)$. We may now combine this result with Eq. (3) to get $\overline{W} \geq \Delta F$.

Now, what if we perform a finite number N_s of identical switching simulations? Let W_i denote the work performed on the system during the ith simulation and let

$$W^a \equiv \frac{1}{N_s} \sum_{i=1}^{N_s} W_i \tag{53}$$

be the average over these values. We may view the W_i's as numbers sampled randomly from a distribution $\rho(W)$ satisfying Eq. (3). Then the expectation value of W^a provides a rigorous upper bound on ΔF:

$$\langle\langle W^a \rangle\rangle = \overline{W} \equiv \int dW \rho(W) W \geq \Delta F. \tag{54}$$

The double angular brackets, denoting *expectation value*, specifically mean an average over all possible sets of N_s simulations. Now, Eq. (3) suggests that rather than W^a, we consider the following quantity as our best estimate of ΔF:

$$W^x \equiv -\beta^{-1} \ln \frac{1}{N_s} \sum_{i=1}^{N_s} \exp(-\beta W_i). \tag{55}$$

For $N_s = 1$, W^x and W^a are identical and the expectation value of either is $\overline{W}$. For $N_s \to \infty$, by contrast, W^x converges to ΔF, whereas W^a converges to $\overline{W}$. For intermediate values of N_s, the following inequality chain holds:

$$\Delta F \leq \langle\langle W^x \rangle\rangle \leq \langle\langle W^a \rangle\rangle. \tag{56}$$

[Both inequalities are derived by combining Eqs. (3) and (50) with the definitions of W^a and W^x.] In other words, as an estimate of ΔF, the "exponential average" W^x is statistically less biased than the ordinary average W^a for $N_s > 1$.

On the face of it, the last statement seems to imply that if we perform more than one switching simulation, then we are better off using W^x rather than W^a as our best guess (or upper bound) for ΔF. In practice, however, Eq. (3) may be subject to the same disease as the TP identity, Eq. (48). Namely, if the values of W obtained from repetitions of the switching simulation typically differ from one another by much more than β^{-1}, then the average of $\exp(-\beta W)$ will be dominated by values of W that are very rarely sampled [23]. Thus the convergence of W^x to ΔF, in the limit $N_s \to \infty$, may be much slower than the convergence of W^a to $\overline{W}$, in the same limit. In other words, for a finite number of switching simulations, W^x may be subject to considerably larger *statistical* fluctuations than W^a, even though its *systematic* error (expectation value minus ΔF) is, by Eq. (56), smaller.
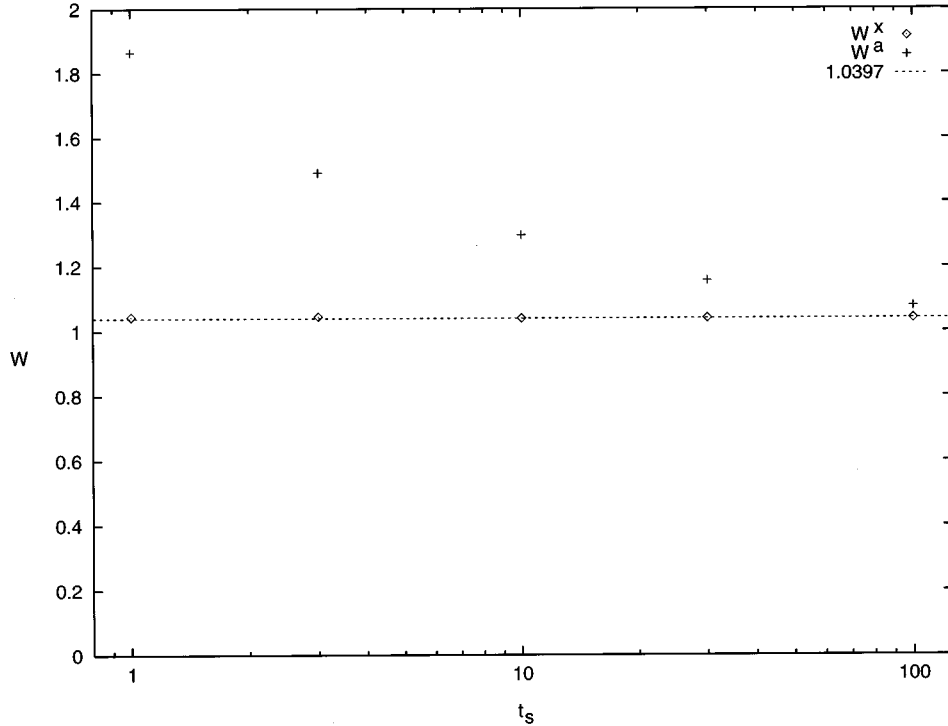
FIG. 3. Same as Fig. 2, except that the harmonic oscillator is now subject to a frictional and a stochastic force, as per Eq. (29). The dashed line gives the free-energy difference $\Delta F = 1.0397$.

The preceding comments point to the following tentative conclusion. If one runs a set of N_s switching simulations, with the goal of computing ΔF, and if the spread in the values of work W obtained is not much larger than β^{-1}, then the exponential average W^x defined by Eq. (55) should provide a better estimate of (or tighter upper bound on) ΔF than the ordinary average W^a. This conclusion is supported by a calculation by Hunter, as described in Ref. [4]; see also the numerical illustrations in the following section.

Of course, a more detailed study of the possible utility of Eq. (3) to free-energy computations should be made. In particular, it is not ruled out that there exist methods around the limitation mentioned in the previous paragraphs [24].

V. NUMERICAL RESULTS

In this section we illustrate our central result with numerical experiments. The first four sets of simulations involve a harmonic oscillator whose natural frequency is switched from $\omega_0 = 1.0$ to $\omega_1 = 2.0$. The evolution is implemented using, in turn, each of the four examples discussed in Sec. III. Then, we present results involving a more complicated system: a gas of interacting particles inside an externally pumped piston. All of these cases satisfy the condition discussed at the end of Sec. IV, namely, the spread in the values of W is not much greater than β^{-1}. (Otherwise, the convergence of W^x to ΔF, in the limit of many simulations, would be poor.)

For the harmonic-oscillator simulations, we take the Hamiltonian given by Eq. (25), with $\omega_\lambda = 1.0 + \lambda$ and $\lambda(t) = t/t_s$ as throughout this paper. Also, we take $\beta^{-1} = 1.5$. Thus the free-energy difference is $\Delta F = \beta^{-1} \ln(\omega_1/\omega_0) = 1.0397$.

In the first set of simulations, the oscillator is isolated (it evolves under Hamilton's equations, with λ time dependent), although the initial conditions are sampled from a canonical ensemble corresponding to $\lambda = 0$:

$$f(x,p,0) = \frac{\beta \omega_0}{2\pi} \exp[-\beta(p^2 + \omega_0^2 x^2)/2]. \qquad (57)$$

Five different values of the switching time were chosen: $t_s = 1.0, 3.0, 10.0, 30.0,$ and 100.0, and for each t_s a total of $N_s = 10^5$ simulations were carried out. Figure 2 shows the average value of work obtained at each switching time W^a as well as the exponential average W^x. [See Eqs. (53) and (55).] Since Hamiltonian evolution satisfies detailed balance, but not thermalization (as defined in Sec. I), we do not expect the work performed in the limit $t_s \to \infty$ to equal the free-energy difference ΔF. Rather, we expect $W_\infty = [(\omega_1/\omega_0) - 1]E_0$ for a single trajectory (see Sec. III A) and therefore for a canonical distribution of initial energies

$$\lim_{t_s \to \infty} \overline{W} = [(\omega_1/\omega_0) - 1]\beta^{-1} = 1.5. \qquad (58)$$

The values of W^a shown in Fig. 2 are consistent with this expectation. Equation (3), meanwhile, predicts

$$\lim_{N_s \to \infty} W^x = \Delta F \qquad (59)$$

for *any* value of t_s. Again, the numerical results are consistent with the prediction: The values of W^x shown in Fig. 2 all fall very close to ΔF.

For the second set of simulations, we added a frictional and a stochastic force, as described by Eqs. (29)–(31), with
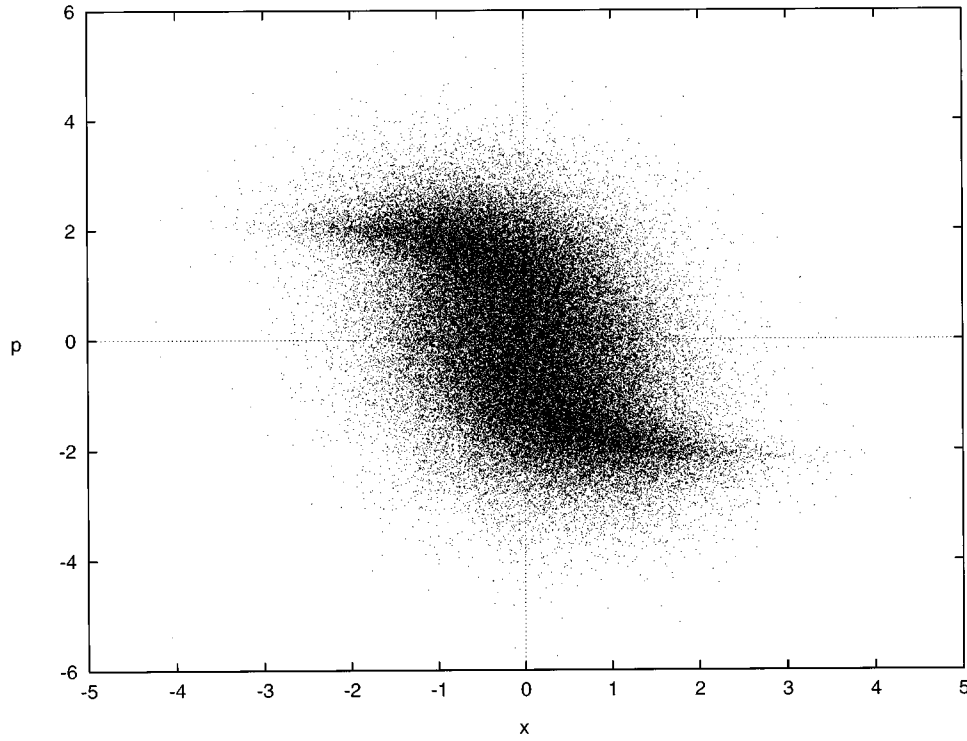
FIG. 4. In these simulations, the harmonic oscillator is thermostated with the IMD scheme described in the text. 10^5 simulations at $t_s = 1.0$ were performed and the dots in this figure show the final locations in phase space of these trajectories.

$D_P = 0.6$ and $\beta^{-1} = 1.5$. The evolution now represents that of an oscillator coupled to a heat bath. The stochastic force was implemented by generating a random momentum kick (sampled from a Gaussian distribution) at each time step $dt = 0.01$ in the numerical integration. As with the Hamiltonian evolution, 10^5 simulations were performed at each of the five switching times and the results for W^a and W^x are plotted in Fig. 3. Here we *do* expect that the work W will approach $\Delta F = 1.0397$ as $t_s \to \infty$, and the results for W^a support this. At the same time, the exponential average W^x falls very close to ΔF for each switching time, as predicted by Eq. (3).

The next simulations again involved a thermostated harmonic oscillator, only this time isothermal molecular dynamics was used to implement the coupling to the heat bath. The particular IMD scheme used was developed by Hoover and Holian [25] and involves two heat bath variables ζ and ξ. The equations of motion in the extended phase space are

$$\dot{x} = p, \tag{60a}$$

$$\dot{p} = -\omega_\lambda^2 x - \zeta p - \beta \xi p^3, \tag{60b}$$

$$\dot{\zeta} = (\beta p^2 - 1)/\tau^2, \tag{60c}$$

$$\dot{\xi} = (\beta^2 p^4 - 3\beta p^2)/\tau^2, \tag{60d}$$

where τ is a relaxation time whose value was set to unity. A total of $N_s = 10^5$ switching simulations were performed, with a switching time $t_s = 1.0$. At the start of each simulation, initial conditions were sampled from the distribution

$$f(x, p, \zeta, \xi, 0) \propto \exp[-\beta(p^2 + \omega_0^2 x^2)/2 - \tau^2(\zeta^2 + \xi^2)/2]. \tag{61}$$

It is easily verified by inspection that, if the value of λ were held fixed at 0, then this distribution would be invariant under Eq. (60). This is therefore the canonical distribution corresponding to this IMD scheme and the function q defined in Sec. III C is given by

$$q(\zeta, \xi) = \tau^2(\zeta^2 + \xi^2)/2\beta. \tag{62}$$

This set of simulations was used to illustrate Eq. (20), evaluated at $t = t_s$. (See the discussion at the end of Sec. II.) Figure 4 is a scatter plot showing the distribution of final values in phase space $(x(t_s), p(t_s))$, for the 10^5 trajectories. Figure 5 shows several contour lines of this distribution, after smearing each point with a Gaussian of variance 0.04 in both the x and p directions. Thus the lines shown are actually contours of the function

$$\tilde{f}(x, p) = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_\epsilon(x - x_i(t_s))\delta_\epsilon(p - p_i(t_s)), \quad \epsilon = 0.04, \tag{63}$$

where δ_ϵ is a normalized Gaussian of variance ϵ, and $(x_i(t), p_i(t))$ gives the phase space evolution of the ith trajectory in the ensemble of simulations. As can be seen from both figures, this distribution does not correspond to the canonical distribution for H_1. Indeed, its skewness illustrates the lag that develops between the evolving phase-space density and the instantaneous canonical distribution.

Next, the solid lines in Fig. 6 show contours of the function $g(x, p, t_s)$ defined in Sec. II, obtained from the same set
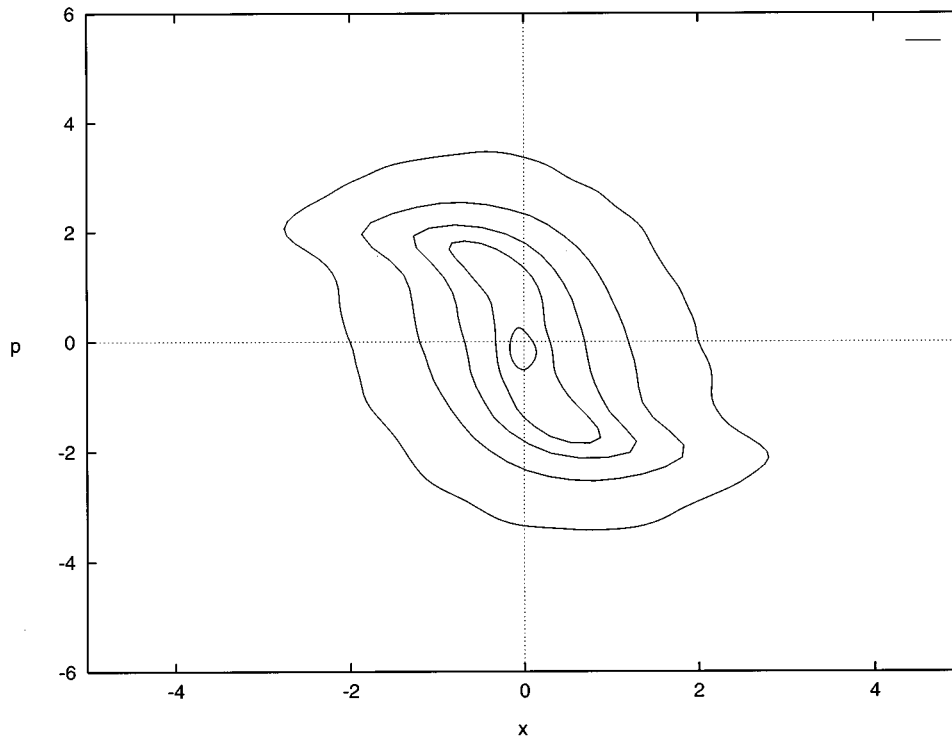
FIG. 5. Contour plot of the distribution $f(x,p,t_s)$, constructed from the data shown in Fig. 4, with Gaussian smoothing.

of simulations. Again, Gaussian smoothing was used, so the solid lines are contours of the function

$$\widetilde{g}(x,p,t_s) = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_\epsilon(x - x_i(t_s)) \delta_\epsilon(p - p_i(t_s))$$

$$\times \exp(-\beta W_i), \quad \epsilon = 0.04, \tag{64}$$

where W_i is the total work performed on the system during the ith simulation. The dashed lines in Fig. 6 show the corresponding contours of the *predicted* mass density $g(x,p,t_s)$ [Eq. (20)], with the same Gaussian smoothing function folded in. The agreement between the two sets of contours is very good. This shows that, indeed, when one assigns a weight $\exp(-\beta W_i)$ to each of the points in the scatter plot, Fig. 4, then the resulting weighted distribution is canonical in the sense of Eq. (20).

In the final set of simulations involving the harmonic oscillator, we used Monte Carlo evolution, with the Metropolis algorithm. Here, the duration of a simulation is characterized by the number of MC steps N rather than by a switching time t_s. Ten different values of N were considered, $N = 5,10,20,50,\ldots,5000$, and for each a total of 10^5 simulations were performed. Figure 7 shows W^a and W^x for each value of N; as before, the results agree nicely with Eqs. (1)–(3). Figure 8 shows $\rho_N(W)$, the distribution of values of W obtained from the 10^5 simulations, for each of the ten values of N. Although the distributions ρ_N are quite different, the integral $\int dW \rho_N(W) \exp(-\beta W)$ [i.e., $\exp(-\beta W^x)$] is independent of N, as shown by the values of W^x in Fig. 7.

As a final example, we take a system more complicated than a harmonic oscillator, namely, a gas of $n_p = 50$ interacting particles inside a piston that is taken through one cycle of pumping. Specifically, the particles are confined within a two-dimensional box with hard walls, whose initial dimensions are 1×1; over the course of the switching process, one of the wall first moves inward until the area enclosed by the box is three-quarters of its initial value and then back out again. This pumping of the piston is cosinusoidal: the x and y dimensions of the box are given by

$$L_x(\lambda) = 1.0, \tag{65}$$

$$L_y(\lambda) = 0.875 + 0.125\cos(2\pi\lambda), \tag{66}$$

where $\lambda = t/t_s$ and the total switching time is $t_s = 10.0$.

In their interactions with one another and with the walls, the particles act as hard disks of radius $R = 0.005$; between collisions each particle moves freely. (Thus, over a switching time $t_s = 10.0$, a typical particle suffers several collisions with other particles.) Work is performed on the gas each time a particle bounces off the moving wall.

Molecular dynamics was used for the evolution, i.e., continuous trajectories for the particles were computed as functions of time. However, at each time step in the integration of the equations of motion, a single particle was randomly selected and a random kick, a discrete change in the momentum of the chosen particle, was generated. The kick was then either accepted or rejected according to the Metropolis algorithm [26], corresponding to a temperature $\beta^{-1} = 0.5$. This
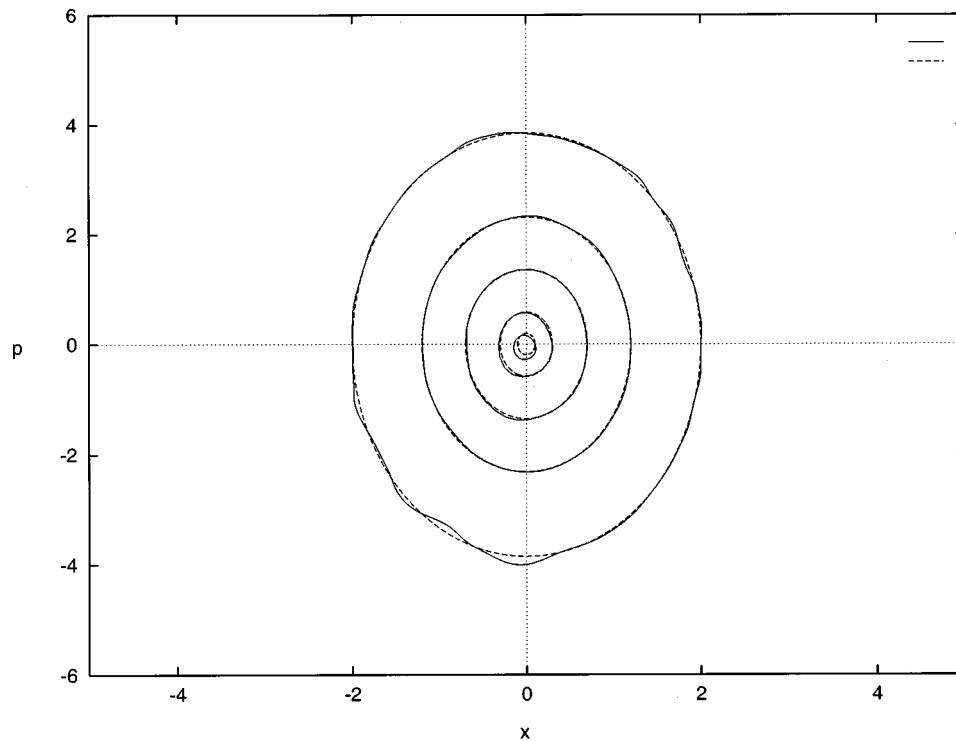
FIG. 6. The solid line shows contours of the function $g(x,p,t_s)$, constructed from the data shown in Fig. 4; the dashed line shows contours of the theoretical prediction for $g(x,p,t_s)$ [Eq. (20)]. Both are smoothed with Gaussians.

simulates coupling to a heat reservoir: If all the walls of the box were fixed, the gas would relax to thermal equilibrium from any initial conditions.

Figure 9 shows results obtained from $N_s = 10^4$ such simulations. For each simulation, the initial conditions of the gas were chosen from a canonical ensemble ($\beta^{-1} = 0.5$) and the work performed on the gas, as a function of time, was computed. The horizontal axis shows time. The solid line gives the work performed on the gas up to time t, i.e., the work accumulated $w(t)$, averaged over all N_s simulations. Let
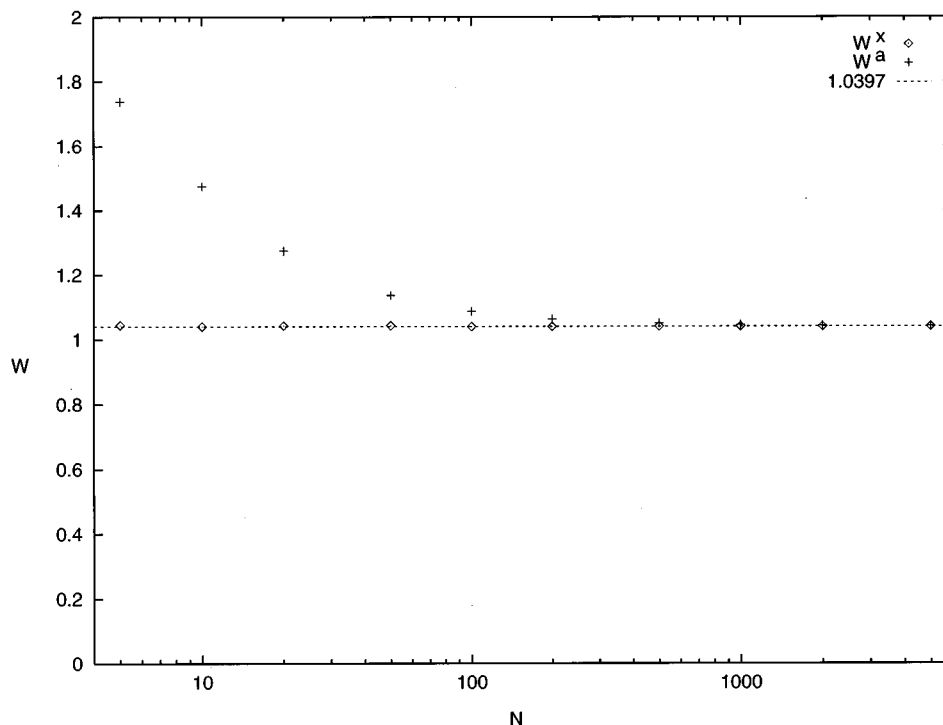

FIG. 7. Similar to Fig. 3, except here the evolution of the thermostated oscillator is implemented using Monte Carlo rather than Langevin evolution. The duration of a simulation is now characterized by the number of MC steps N rather than a switching time t_s. For each of ten values of N, 10^5 simulations were carried out.
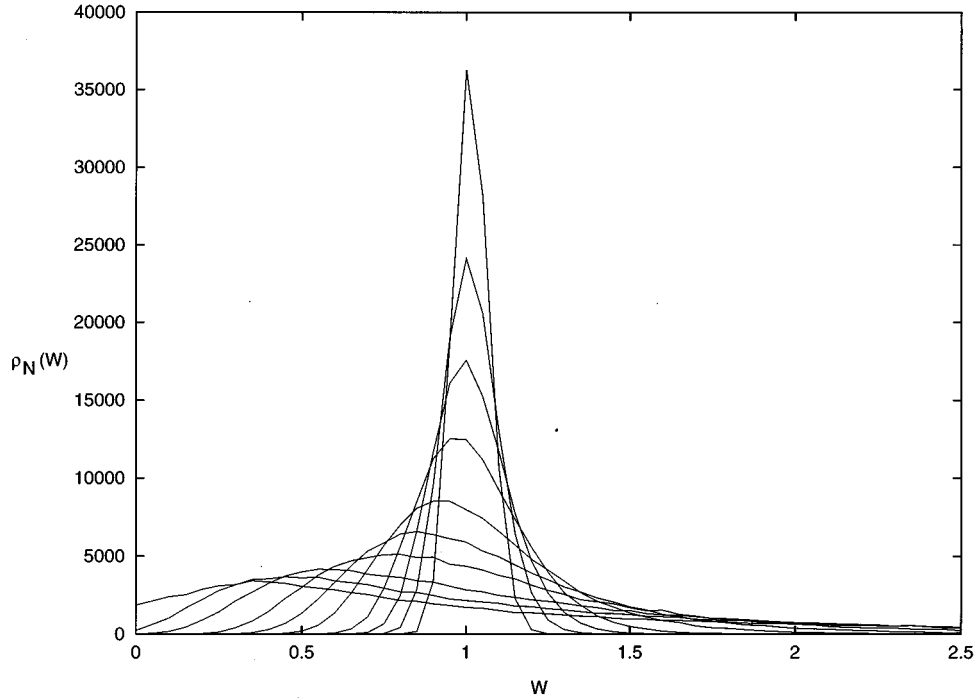
FIG. 8. For each of the ten sets of MC simulations (see Fig. 7), the distribution of values of work $\rho_N(W)$ was obtained. This figure shows these ten distributions, from $N=5$ (lowest peak) to $N=5000$ (highest). (Although the peak moves toward the right with increasing N, the actual average work performed goes down; see the values of W^a in Fig. 7.)

$w^a(t) \equiv (1/N_s)\Sigma_{i=1}^{N_s} w_i(t)$ denote this average, where $w_i(t)$ is the work accumulated during the ith simulation. The dashed line gives the exponential average

$$w^x(t) \equiv -\beta^{-1} \ln \frac{1}{N_s} \sum_{i=1}^{N_s} \exp[-\beta w_i(t)] \qquad (67)$$

of the work accumulated. Both $w^a(t)$ and $w^x(t)$ were computed from the same set of 10^4 trajectories.

Since the piston returns to its initial position at the end of the switching process, the final free-energy difference is zero: $\Delta F \equiv F_1 - F_0 = 0$. We see that the dashed line indeed returns to zero at $t = t_s$, with very good accuracy. By contrast, the average work performed on the system (upper line) ends at $W^a \equiv w^a(t_s) = 1.534$. This represents dissipated energy: the gas ''heats up'' when pumped at a finite rate.

At intermediate times, we expect

$$w^x(t) = F_\lambda - F_0, \quad \lambda = \lambda(t) \qquad (68)$$

(in the limit of infinitely many switching simulations), by Eq. (22). If the gas were truly ideal, then the free-energy difference would be given by

$$F_\lambda - F_0 = n_p \beta^{-1} \ln(A_0/A_\lambda) \quad \text{(ideal gas)}, \qquad (69)$$

where A_λ denotes the area enclosed by the box. However, since the particles do interact with one another, as hard disks, this expression for $F_\lambda - F_0$ is not exact. Nevertheless, the size of each particle is small enough ($R=0.005$) that Eq. (69) ought to represent an excellent approximation. The dotted line in Fig. 9 shows this approximation to $F_\lambda - F_0$. This line is very close to the exponential average (dashed line), in

confirmation of our central result, in the form given by Eq. (22). Note that the dotted line represents the work that *would have been* performed on the gas, in the limit of infinitely slow switching ($t_s \to \infty$). Thus for the dashed line, we have effectively extracted this *quasistatic* behavior, from an ensemble of *finite-time* switching simulations.

VI. SUMMARY AND DISCUSSION

The central goal of this paper has been to establish the (exact) validity of the result $\exp(-\beta W) = \exp(-\beta \Delta F)$, within the framework of the master-equation approach. This result is unusual in that it expresses *in the form of an equality* [rather than an inequality, e.g., Eq. (2)] the relationship between the work W performed on an out-of-equilibrium system (more precisely, on an *ensemble* of systems driven out of equilibrium by varying an external parameter) and the free-energy difference ΔF between two equilibrium states of the system. A few comments are now in order.

In classical statistical mechanics, the equilibrium ''state'' of a system is described by a canonical distribution in phase space: $f^C(\mathbf{z}) = Z^{-1} \exp[-\beta H(\mathbf{z})]$. Its free energy is then given by

$$F = \langle H \rangle - \beta^{-1} S = -\beta^{-1} \ln Z, \qquad (70)$$

where $\langle H \rangle = \int f^C H$ and $S = -k_B \int f^C \ln f^C$. Thus the free energy F (like the entropy S) is a quantity associated with a *statistical ensemble* of microscopic states of the system. When the system depends on some external parameter, then so does the canonical ensemble, as in turn does the free-energy. The quantity ΔF of central interest throughout this
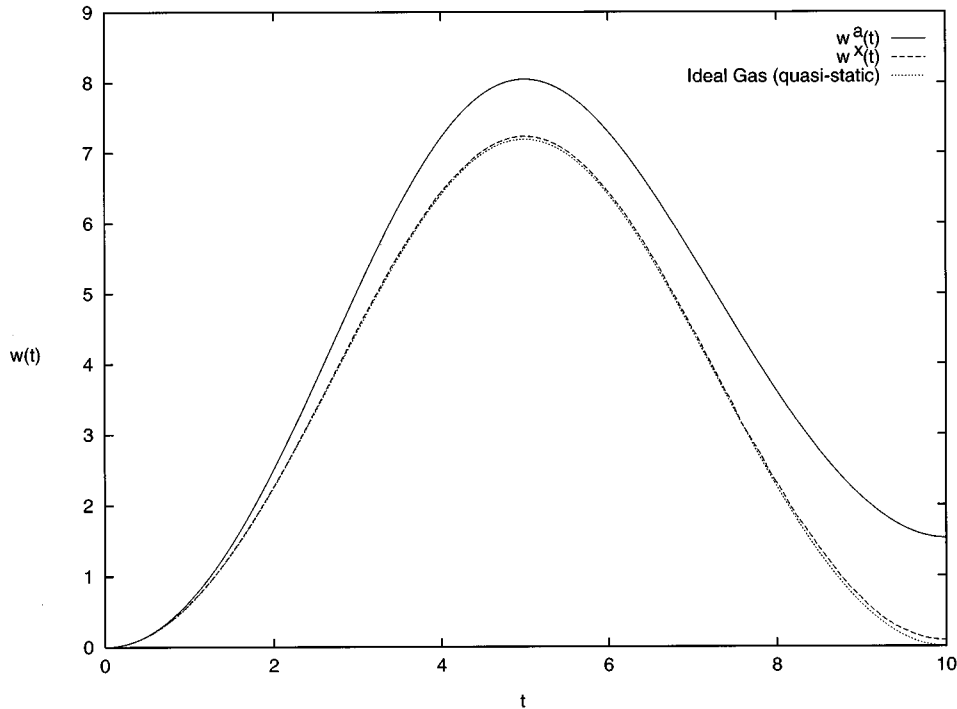
FIG. 9. Simulations of a gas of interacting hard disks inside a piston that goes through one cycle of pumping (first in, then out), over a switching time t_s. The evolution is MD, but MC ''kicks'' are included to provide a thermostat. A total of 10^4 simulations were performed. The solid line gives the average and the dashed line the exponential average of the work as a function of time, $w^a(t)$ and $w^x(t)$, respectively. The dotted line gives the theoretical prediction for $F_\lambda - F_0$ [Eq. (69)], with $\lambda = \lambda(t)$, for the case of an ideal gas.

paper has been the free-energy difference (at constant temperature) between two such equilibrium ensembles $f_{\lambda=0}^C(\mathbf{z})$ and $f_{\lambda=1}^C(\mathbf{z})$.

In deriving Eq. (3) we introduced a time-dependent phase-space density $f(\mathbf{z},t)$, describing the evolution of our ensemble of trajectories. Note that, although the initial phase-space density coincides with the canonical distribution used to compute the free energy F_0,

$$f(\mathbf{z},0) = f_{\lambda=0}^C(\mathbf{z}), \tag{71}$$

the final density does *not* (typically) coincide with the distribution for which F_1 is defined:

$$f(\mathbf{z},t_s) \neq f_{\lambda=1}^C(\mathbf{z}). \tag{72}$$

This is due to the lag that develops between the ensemble of trajectories and an instantaneous canonical distribution in phase space. Thus ΔF is not the free-energy difference between the initial and final states of the system [27], but rather, as stated in the preceding paragraph, between two different canonical ensembles, $f_{\lambda=0}^C$ and $f_{\lambda=1}^C$, only the former of which reflects the actual distribution of microscopic states of the system at any time during the switching process.

Another way of putting this is as follows. Suppose we are interested in the free-energy difference between two equilibrium statistical states of a system, A and B (corresponding to $f_{\lambda=0}^C$ and $f_{\lambda=1}^C$, respectively). Ordinarily, we would compute or measure ΔF by *reversibly* carrying the system from A to B, i.e., by switching λ infinitely slowly. Equation (3) tells us

that even if we switch *irreversibly*, so that the system ends up in some nonequilibrium statistical state B^*, we can still extract ΔF $(\equiv F^B - F^A)$ from an ensemble of such measurements.

Of course, if we are dealing with a system that satisfies the thermalization assumption, then, at the end of the switching process, we can always, at no cost in work, hold the value of λ fixed and allow our ensemble to relax, for an additional time t_{rel}, to thermal equilibrium: $f(\mathbf{z},t_s+t_{rel}) = f_{\lambda=1}^C(\mathbf{z})$. In this case the ΔF that appears on the right-hand side of Eq. (3) *does* equal the free-energy difference between the initial and final statistical states of the system.

Equation (3) was derived, as an identity, under the assumptions of Markovian evolution and detailed balance, as spelled out in Sec. I. This derivation is complementary to the one presented in Ref. [4], in which the degrees of freedom of the heat reservoir were treated explicitly. Neither of the assumptions of Sec. I was assumed in Ref. [4], but the coupling between the system and reservoir was taken to be weak, so the result there was an approximate one, with small corrections expected from the small but finite interaction Hamiltonian. Of course, in a real physical system, neither the Markov assumption nor detailed balance will be met exactly, so the derivation presented herein is strictly speaking valid only for a particular class of *models* of physical reality. Nevertheless, because the result is exact for these models and because the Markov and detailed balance assumptions are often very good approximations for physical systems, the result is a useful one. Furthermore, as illustrated in Sec. III, models of thermostated systems that are commonly used in

theoretical and numerical studies *do* satisfy these assumptions; Eq. (3) is therefore exactly valid for these models.

It would be very interesting, of course, to find a physical system on which a laboratory (as opposed to numerical) experiment testing the validity of Eq. (3) would be feasible. As mentioned in Ref. [4], such a system would almost certainly have to be microscopic, or at most mesoscopic in size.

Finally, from both a theoretical and a computational point of view, it would be worthwhile to consider possible extensions or generalizations of Eq. (3). In particular, are analogous results valid for ensembles other than the canonical ensemble (fixed N,V,T) considered here, e.g., microcanonical, grand canonical, and isothermal isobaric? Presumably, the role of the Helmholtz free energy F would then be played by other thermodynamic potentials, for instance, the Gibbs free energy in the case of the isothermal-isobaric ensemble.

ACKNOWLEDGMENTS

APPENDIX A

In this appendix we present a derivation of Eq. (16) different from the one given in Sec. II. For a given stochastic trajectory $\mathbf{z}(t)$, the work W is given by a path integral along that trajectory [Eq. (5)]. We are interested in evaluating the average of $\exp(-\beta W)$ over an ensemble of trajectories, obtained by sampling initial conditions from a canonical ensemble and then evolving stochastically from each of these initial conditions. The quantity $\overline{\exp(-\beta W)}$ thus constitutes a sum over all paths, with each path $\mathbf{z}(t)$ in our ensemble weighted by the factor $\exp(-\beta W)$. We may write this as

$$\overline{\exp(-\beta W)}=\int dm\,\mathcal{P}(\mathbf{z}(t))\exp(-\beta W), \qquad (A1)$$

where dm denotes a measure in the space of paths $\mathbf{z}(t)$, $\mathcal{P}(\mathbf{z}(t))$ denotes the probability density (with respect to this measure) of choosing $\mathbf{z}(t)$ by sampling randomly from the ensemble of trajectories, and $W=W(\mathbf{z}(t))$ [as per Eq. (5)].

Let us now divide the time interval $[0,t_s]$ into N time steps of duration $\delta t=t_s/N$ and let us denote a particular trajectory $\mathbf{z}(t)$ by its phase-space locations $\mathbf{z}_n\equiv\mathbf{z}(t_n)$ at times $t_n\equiv n\delta t$, $0\leq n\leq N$. Thus

$$\mathbf{z}(t)\rightarrow(\mathbf{z}_0,\mathbf{z}_1,\ldots,\mathbf{z}_N). \qquad (A2)$$

The limit $N\rightarrow\infty$ (with t_s fixed) is implied. Choosing a Euclidean measure in path space,

$$\int dm=\int d\mathbf{z}_0\int d\mathbf{z}_1\cdots\int d\mathbf{z}_N, \qquad (A3)$$

the probability density for a particular path is

$$\mathcal{P}(\mathbf{z}(t))=p(\mathbf{z}_0)P_{\lambda_1}^{\delta t}(\mathbf{z}_0|\mathbf{z}_1)P_{\lambda_2}^{\delta t}(\mathbf{z}_1|\mathbf{z}_2)\cdots P_{\lambda_N}^{\delta t}(\mathbf{z}_{N-1}|\mathbf{z}_N). \qquad (A4)$$

Here $p(\mathbf{z}_0)=Z_0^{-1}\exp[-\beta H_0(\mathbf{z}_0)]$ is the probability distribution for the initial condition $\mathbf{z}_0$, $P_\lambda^{\delta t}(\mathbf{z}'|\mathbf{z})$ is the transition probability from $\mathbf{z}'$ to $\mathbf{z}$ (in time δt) as a function of λ, and $\lambda_n\equiv n/N$. It is the Markov assumption that allows this factorization. The work W may be expressed as

$$W(\mathbf{z}(t))=\sum_{n=1}^{N}\delta H_n(\mathbf{z}_{n-1}), \qquad (A5)$$

where $\delta H_n\equiv H_{\lambda_n}-H_{\lambda_{n-1}}$. [In writing Eqs. (A4) and (A5) we implicitly assume that $\lambda(t)$ evolves in N discrete steps $\delta\lambda=1/N$ that occur at times $t_0,t_1,\ldots,t_{N-1}$. This "staircase" evolution becomes $\lambda(t)=t/t_s$ in the limit $N\rightarrow\infty$.]

Combining Eqs. (A1)–(A5), we arrive at

$$\overline{\exp(-\beta W)}=\left[\prod_{n=0}^{N}\int d\mathbf{z}_n\right]p(\mathbf{z}_0)e^{-\beta\delta H_1(\mathbf{z}_0)}$$
$$\times P_{\lambda_1}^{\delta t}(\mathbf{z}_0|\mathbf{z}_1)\cdots e^{-\beta\delta H_N(\mathbf{z}_{N-1})}P_{\lambda_N}^{\delta t}(\mathbf{z}_{N-1}|\mathbf{z}_N). \qquad (A6)$$

Let us now introduce

$$g_M(\mathbf{z})=\left[\prod_{n=0}^{M-1}\int d\mathbf{z}_n\right]p(\mathbf{z}_0)e^{-\beta\delta H_1(\mathbf{z}_0)}$$
$$\times P_{\lambda_1}^{\delta t}(\mathbf{z}_0|\mathbf{z}_1)\cdots e^{-\beta\delta H_M(\mathbf{z}_{M-1})}P_{\lambda_M}^{\delta t}(\mathbf{z}_{M-1}|\mathbf{z}), \qquad (A7)$$

where $1\leq M\leq N$. This is the discretized version of the function $g(\mathbf{z},t)$ introduced in the main body of the text:

$$g_M(\mathbf{z})=g(\mathbf{z},t_M). \qquad (A8)$$

In particular, note that $\overline{\exp(-\beta W)}=\int d\mathbf{z}\,g_N(\mathbf{z})$. This set of functions g_M satisfies the recursion relation

$$g_{M+1}(\mathbf{z})=\int d\mathbf{z}_M g_M(\mathbf{z}_M)e^{-\beta\delta H_{M+1}(\mathbf{z}_M)}P_{\lambda_{M+1}}^{\delta t}(\mathbf{z}_M|\mathbf{z}). \qquad (A9)$$

Now, to first order in δt, we have

$$e^{-\beta\delta H_{M+1}(\mathbf{z}_M)}=1-\beta\delta H_{M+1}(\mathbf{z}_M), \qquad (A10)$$

$$P_{\lambda_{M+1}}^{\delta t}(\mathbf{z}_M,\mathbf{z})=\delta(\mathbf{z}_M-\mathbf{z})+\delta t R_{\lambda_{M+1}}(\mathbf{z}_M,\mathbf{z}). \qquad (A11)$$

Combining this with our recursion relation gives (to leading order)

$$\frac{1}{\delta t}[g_{M+1}(\mathbf{z})-g_M(\mathbf{z})]=-\beta g_M(\mathbf{z})\delta H_{M+1}(\mathbf{z})/\delta t$$
$$+\int d\mathbf{z}_M g_M(\mathbf{z}_M)R_{\lambda_{M+1}}(\mathbf{z}_M,\mathbf{z}), \qquad (A12)$$

which becomes Eq. (16) in the limit $N\rightarrow\infty$.

APPENDIX B

Here we prove the assertion (made in Sec. III) that Eq. (3) is identically true when the switching process is carried out using the Monte Carlo method. Some of the steps in the proof will be similar to those in Appendix A, but the assumption $N\rightarrow\infty$ will not be made here.

As mentioned in Sec. III, a trajectory $(\mathbf{z}_0,\ldots,\mathbf{z}_N)$ is obtained by alternating discrete changes in the value of λ with random jumps in phase space generated by the MC algorithm. This algorithm, parametrized by the value of λ, takes as input a point $\mathbf{z}$ and output a point $\mathbf{z}'$. Let $P_\lambda(\mathbf{z}|\mathbf{z}')$ denote the probability of generating an output $\mathbf{z}'$ from an input $\mathbf{z}$, for a given value of λ. Detailed balance is built into the algorithm

$$\int d\mathbf{z}\, e^{-\beta H_\lambda(\mathbf{z})} P_\lambda(\mathbf{z}|\mathbf{z}') = e^{-\beta H_\lambda(\mathbf{z}')} \quad (B1)$$

for any λ. (This may be accomplished by, e.g., the Metropolis method [26].) Thus a canonical distribution of input $\mathbf{z}$ gives a canonical distribution of output $\mathbf{z}'$.

The probability of obtaining a particular trajectory $(\mathbf{z}_0,\ldots,\mathbf{z}_N)$ over the course of the entire switching process is then

$$\mathcal{P}(\mathbf{z}_0,\ldots,\mathbf{z}_N) = \frac{1}{Z_0} e^{-\beta H_0(\mathbf{z}_0)} P_{\lambda_1}(\mathbf{z}_0|\mathbf{z}_1)\cdots P_{\lambda_N}(\mathbf{z}_{N-1}|\mathbf{z}_N). \quad (B2)$$

Combining this with Eq. (45) for the work, we get

$$\overline{\exp(-\beta W)} = \left[\prod_{n=0}^{N} \int d\mathbf{z}_n \right] \frac{1}{Z_0} e^{-\beta H_0(\mathbf{z}_0)} e^{-\beta\delta H_1(\mathbf{z}_0)}$$
$$\times P_{\lambda_1}(\mathbf{z}_0|\mathbf{z}_1)\cdots e^{-\beta\delta H_N(\mathbf{z}_{N-1})}$$
$$\times P_{\lambda_N}(\mathbf{z}_{N-1}|\mathbf{z}_N), \quad (B3)$$

where $\delta H_n \equiv H_{\lambda_n} - H_{\lambda_{n-1}}$. Now notice that $\exp[-\beta H_0(\mathbf{z}_0)]$ can be combined with $\exp[-\beta\delta H_1(\mathbf{z}_0)]$ to give

$\exp[-\beta H_{\lambda_1}(\mathbf{z}_0)]$. The only other factor in the integrand that depends on $\mathbf{z}_0$ is $P_{\lambda_1}(\mathbf{z}_0|\mathbf{z}_1)$. Performing the integral $\int d\mathbf{z}_0$, we get

$$\int d\mathbf{z}_0 P(\mathbf{z}_0|\mathbf{z}_1) \exp[-\beta H_{\lambda_1}(\mathbf{z}_0)] = \exp[-\beta H_{\lambda_1}(\mathbf{z}_1)], \quad (B4)$$

using Eq. (B1). This takes care of the first of the $N+1$ integrals appearing in Eq. (B3). We now repeat this process, first combining $\exp[-\beta H_{\lambda_1}(\mathbf{z}_1)]$ (obtained from the $d\mathbf{z}_0$ integration) with $\exp[-\beta\delta H_2(\mathbf{z}_1)]$ to get $\exp[-\beta H_{\lambda_2}(\mathbf{z}_1)]$, then integrating over $\mathbf{z}_1$, and so forth. At the end of this process of ''rolling up'' the factors and integrating, we are left with

$$\overline{\exp(-\beta W)} = \int d\mathbf{z}_N \frac{1}{Z_0} \exp[-\beta H_1(\mathbf{z}_N)]$$
$$= \frac{Z_1}{Z_0} = \exp(-\beta\Delta F). \quad (B5)$$

Q.E.D. Note that Eq. (20) for $g(\mathbf{z},t)$, derived within the framework of continuous-time evolution, also has a Monte Carlo counterpart. Namely, for $1\leq M\leq N$, let us define

$$g_M(\mathbf{z}) = \prod_{n=0}^{M-1} \int d\mathbf{z}_n \mathcal{P}_M(\mathbf{z}_0,\ldots,\mathbf{z}_{M-1},\mathbf{z}) \exp(-\beta w_M), \quad (B6)$$

where $\mathcal{P}_M$ gives the probability of sampling a particular sequence of phase space points in the first M Monte Carlo steps and w_M is the work accumulated during those steps. Thus, in terms of our ensemble of MC trajectories, $g_M(\mathbf{z})$ is the weighted phase-space density after M steps, where the weight assigned to each trajectory is $\exp(-\beta w_M)$. Then, writing an explicit expression for $\mathcal{P}_M$ in the form given by Eq. (B2) and rolling up factors and integrating as above, it follows easily that

$$g_M(\mathbf{z}) = \frac{1}{Z_0} \exp[-\beta H_{\lambda_M}(\mathbf{z})] = \frac{Z_{\lambda_M}}{Z_0} f_{\lambda_M}^C(\mathbf{z}). \quad (B7)$$

[1] L. D. Landau and E. M. Lifshitz, *Statistical Physics*, 3rd ed. (Pergamon, Oxford, 1990), Pt. 1, Sec. 15.

[2] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Ref. [1]), Sec. 20.

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

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

[5] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics* (Springer-Verlag, Berlin, 1985), Chap. 2.

[6] We assume these initial conditions regardless of whether or not the thermalization assumption mentioned above is met. Thus, even if, for instance, the system is isolated during the switching process, we still assume that it was earlier allowed

to thermalize with some heat reservoir.

[7] Gaun E. Crooks (unpublished).

[8] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison-Wesley, Reading, MA, 1980), Secs. 10-5 and 11-7.

[9] R. Brown, E. Ott, and C. Grebogi, J. Stat. Phys. **49**, 511 (1987).

[10] W. G.Hoover, *Computational Statistical Mechanics* (Elsevier, Amsterdam, 1991).

[11] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[12] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[13] For reviews and reference books see T. P. Straatsma and J. A. McCammon, Annu. Rev. Phys. Chem. **43**, 407 (1992); M. Karplus and G. A. Petsko, Nature (London) **347**, 631 (1990); D. L. Beveridge and F. M. DiCapua, Annu. Rev. Biophys. Biophys.

Chem. **18**, 431 (1989); C. L. Brooks III, M. Karplus, and B. M. Pettitt, Adv. Chem. Phys. **71**, 1 (1988); *Simulations of Liquids and Solids*, edited by D. Frenkel, I. R. McDonald, and G. Ciccotti (North-Holland, Amsterdam, 1986), Chap. 2; D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, Boston, 1996), Chap. 7.

[14] B. L. Holian, H. A. Posch, and W. G. Hoover, Phys. Rev. E **47**, 3852 (1993).

[15] R. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).

[16] See, for instance, C. H. Bennett, J. Comput. Phys. **22**, 245 (1976); G. M. Torrie and J. P. Valleau, *ibid* J. Comput. Phys. **23**, 187 (1977).

[17] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).

[18] See, for instance, T. P. Straatsma, H. J. C. Berendsen, and J. P. M. Postma, J. Chem. Phys. **85**, 6720 (1986); U. C. Singh *et al.*, J. Am. Chem. Soc. **109**, 1607 (1987).

[19] M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).

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

[21] L.-W. Tsao, S.-Y. Sheu, and C.-Y. Mou, J. Chem. Phys. **101**, 2302 (1994).

[22] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987), Sec. 5.5.

[23] This is easy to see: Typically sampled values of W are those within a standard deviation or so of the maximum of $\rho(W)$, while the values of W that are most important in determining $\overline{\exp(-\beta W)}$ are those near the maximum of $\rho(W)\exp(-\beta W)$. As a general rule, the two sets will overlap significantly only if the function $\exp(-\beta W)$ does not change much over one standard deviation in W.

[24] Stimulating correspondence with D. Frenkel on this point is gratefully acknowledged.

[25] W. G. Hoover and B. L. Holian, Phys. Lett. A **211**, 253 (1996).

[26] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953).

[27] Indeed, ''the free energy of the system at the end of the switching process'' is not a well-defined quantity since the statistical state of the system does not correspond to canonical equilibrium.