Nonequilibrium Equality for Free Energy Differences

C. Jarzynski*
Institute for Nuclear Theory, University of Washington, Seattle, Washington 98195
(Received 7 June 1996)

An expression is derived for the equilibrium free energy difference between two configurations of a system, in terms of an ensemble of *finite-time* measurements of the work performed in parametrically switching from one configuration to the other. Two well-known identities emerge as limiting cases of this result. [S0031-9007(97)02845-7]

PACS numbers: 05.70.Ln, 87.10.+e, 82.20.Wt

Consider a finite classical system in contact with a heat reservoir. A central concept in thermodynamics is that of the *work* performed on such a system, when some external parameters of the system are made to change with time. (These parameters may represent, for instance, the strength of an external field, or the volume of space within which the system is confined, or, more abstractly, some particle-particle interactions which are turned on or off during the course of a molecular dynamics simulation.) When the parameters are changed *infinitely slowly* along some path γ from an initial point A to a final point B in parameter space, then the total work W performed on the system is equal to the Helmholtz free energy difference ΔF between the initial and final configurations [1]: $W = \Delta F \equiv F^B - F^A$. [Here F^A (F^B) refers to the equilibrium free energy of the system, with the parameters held fixed at A (B).] By contrast, when the parameters are switched along γ at a *finite* rate, then W will depend on the microscopic initial conditions of the system and reservoir, and will, on average, exceed ΔF:

$$\overline{W} \geq \Delta F\,. \tag{1}$$

The overbar denotes an average over an *ensemble* of measurements of W, where each measurement is made after first allowing the system and reservoir to equilibrate at temperature T, with the parameters fixed at A. (The path γ from A to B, and the rate at which the parameters are switched along this path, remain unchanged from one measurement to the next.) Note that the right side of Eq. (1) still refers to the *equilibrium* free energy difference $F^B - F^A$. The difference $\overline{W} - \Delta F$ is just the dissipated work, $W_{\rm diss}$, associated with the increase of entropy during an irreversible process.

Equation (1) is an inequality. By contrast, the new result derived in this paper is the following *equality*:

$$\overline{\exp(-\beta W)} = \exp(-\beta \Delta F)\,, \tag{2a}$$

or, equivalently,

$$\Delta F = -\beta^{-1} \ln \overline{\exp(-\beta W)}\,, \tag{2b}$$

where $\beta \equiv 1/k_B T$. This result, which is independent of both the path γ from A to B, and the rate at which the parameters are switched along the path, is surprising: It says that we can extract equilibrium information (ΔF) from the ensemble of *nonequilibrium* (finite-time) measurements described above.

Before proceeding with the proof of Eq. (2), we establish notation and then relate Eq. (2) to two well-known equilibrium identities for ΔF. Since we have fixed our attention on a particular path γ in parameter space, it will be convenient to henceforth view the system as parametrized by a single quantity λ, which increases from 0 to 1 as we travel from A to B along γ. Let $\mathbf{z} \equiv (\mathbf{q}, \mathbf{p})$ denote a point in the phase space of the system, and let $H_\lambda(\mathbf{z})$ denote the Hamiltonian for the system, parametrized by the value of λ. Next, let Z_λ denote the partition function, let $\langle \cdots \rangle_\lambda$ denote a canonical average, and let $F_\lambda = -\beta^{-1} \ln Z_\lambda$ denote the free energy, all with respect to the Hamiltonian H_λ and the temperature T. We are interested in the following scenario, which we will refer to as "the switching process": The system evolves, in contact with a heat reservoir, as the value of λ is switched from 0 to 1, over a total switching time t_s. Without loss of generality, assume a constant switching rate, $\dot{\lambda} = t_s^{-1}$. For a given realization of the switching process, the evolution of the system is described by a (effectively stochastic) trajectory $\mathbf{z}(t)$, and the work performed on the system is the time integral of $\dot{\lambda} \partial H_\lambda / \partial \lambda$ along this trajectory:

$$W = \int_0^{t_s} dt\, \dot{\lambda}\, \frac{\partial H_\lambda}{\partial \lambda}\,(\mathbf{z}(t))\,. \tag{3}$$

Now imagine an *ensemble* of realizations of the switching process (with γ and t_s fixed), with initial conditions for the system and reservoir generated from a canonical ensemble at temperature T. Then W may be computed for each trajectory $\mathbf{z}(t)$ in the ensemble, and the overbars appearing in Eqs. (1) and (2) indicate an average over the distribution of values of W thus obtained.

In the limiting cases of infinitely slow and infinitely fast switching of the external parameters, we know explicitly the ensemble distribution of values of W, and thus can readily check the validity of our central result. In the slow limit ($t_s \to \infty$), the system is in quasistatic equilibrium with the reservoir throughout the switching process, hence $W = \int_0^1 d\lambda \langle \partial H_\lambda / \partial \lambda \rangle_\lambda$ for every trajectory in the

 0031-9007/97/78(14)/2690(4)$10.00 © 1997 The American Physical Society

ensemble. Equation (2b) then reduces to

$$\Delta F = \int_0^1 d\lambda \left\langle \frac{\partial H_\lambda}{\partial \lambda} \right\rangle_\lambda . \tag{4}$$

In the opposite limit ($t_s \to 0$), the switching is instantaneous, and the work is simply $W = H_1 - H_0 \equiv \Delta H$, evaluated at the initial conditions [2]. Since we have a canonical distribution of initial conditions, Eq. (2b) becomes, in this case,

$$\Delta F = -\beta^{-1} \ln\langle \exp -\beta \Delta H \rangle_0 . \tag{5}$$

These two results, Eqs. (4) and (5), are well-established identities [3,4]. Note that both give ΔF in terms of equilibrium (canonical) averages. By contrast, for finite t_s, our ensemble of trajectories lags behind the equilibrium distribution in phase space as H_λ changes with time. In this sense, Eq. (2) is the explicitly *non*equilibrium results.

To prove our central result, it is instructive to first consider what happens when there is *no* reservoir during the switching process. The evolution of the system is then described by a deterministic trajectory $\mathbf{z}(t)$ which evolves under $H_\lambda(\mathbf{z})$, as λ changes from 0 to 1 over a time t_s. Consider an ensemble of such trajectories, defined by a canonical distribution of initial conditions at temperature T. (This corresponds to allowing the system to equilibrate with a reservoir, and then decoupling the two, before each realization of the switching process.) The ensemble of trajectories is described by a phase space density $f(\mathbf{z}, t)$ which satisfies $f(\mathbf{z}, 0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z})]$, and which evolves under the Liouville equation, $\partial f/\partial t + \{f, H_\lambda\} = 0$, with $\lambda = \lambda(t) = t/t_s$. Here, $\{\cdot, \cdot\}$ denotes the Poisson bracket. Since the evolution is deterministic, a particular trajectory in this ensemble is uniquely specified by a single point: There is exactly one trajectory which passes through a given $\mathbf{z}$ at time t. This means we can define a "work accumulated" function $w(\mathbf{z}, t)$, as follows. For the trajectory which passes through the point $\mathbf{z}$ at time t, $w(\mathbf{z}, t)$ is the work performed on that trajectory (the time integral of $\dot{\lambda} \partial H_\lambda/\partial \lambda$) up to time t. Since the total work W is just the work accumulated up to time t_s [Eq. (3)], the ensemble average $\overline{\exp(-\beta W)}$ may be expressed as

$$\overline{\exp(-\beta W)} = \int d\mathbf{z}\, f(\mathbf{z}, t_s) \exp[-\beta w(\mathbf{z}, t_s)] . \tag{6}$$

Now, the work done on an isolated system is equal to the change in its energy. Thus, $w(\mathbf{z}, t) = H_\lambda(\mathbf{z}) - H_0(\mathbf{z}_0)$, where $\mathbf{z}_0 = \mathbf{z}_0(\mathbf{z}, t)$ is the initial condition for the trajectory which passes through $\mathbf{z}$ at time t; and $\lambda = \lambda(t)$. Furthermore, Liouville's theorem tells us that phase space density is conserved along any trajectory, hence $f(\mathbf{z}, t) = f(\mathbf{z}_0, 0) = Z_0^{-1} \exp[-\beta H_0(\mathbf{z}_0)]$. Combining these results gives

$$f(\mathbf{z}, t) \exp[-\beta w(\mathbf{z}, t)] = Z_0^{-1} \exp[-\beta H_\lambda(\mathbf{z})] . \tag{7}$$

Equation (6) then becomes

$$\overline{\exp(-\beta W)} = Z_0^{-1} \int d\mathbf{z} \exp[-\beta H_1(\mathbf{z})] = Z_1/Z_0 . \tag{8}$$

Since $\Delta F = -\beta^{-1} \ln(Z_1/Z_0)$, we have established the validity of Eq. (2) for the case in which the system is isolated during the switching process.

Now consider the situation in which the system remains coupled to the reservoir. We assume that the system of interest and the reservoir together constitute a larger, *isolated* Hamiltonian system. Let $\mathbf{z}'$ denote a point in the phase space of the reservoir, let $\mathcal{H}(\mathbf{z}')$ be the Hamiltonian for the reservoir alone, and let $\mathbf{y} = (\mathbf{z}, \mathbf{z}')$ denote a point in the full phase space of system and reservoir. Motion in the full phase space is deterministic, and governed by a Hamiltonian $G_\lambda(\mathbf{y}) = H_\lambda(\mathbf{z}) + \mathcal{H}(\mathbf{z}') + h_{\text{int}}(\mathbf{z}, \mathbf{z}')$, where the interaction term h_{int} couples the system of interest to the reservoir. Let Y_λ be the partition function for G_λ. We explicitly assume the reservoir to be large enough, and the interaction energy h_{int} small enough [5], that when λ is held fixed the system of interest samples its phase space according to the Boltzmann factor $e^{-\beta H_\lambda(\mathbf{z})}$. Now imagine that, at $t = 0$, we populate the *full* phase space with a canonical distribution of initial conditions [6], using the Boltzmann factor $e^{-\beta G_0(\mathbf{y})}$. From this ensemble of initial conditions, an ensemble of trajectories $\mathbf{y}(t)$ evolves deterministically under G_λ, as λ switches from 0 to 1. Since the system of interest and reservoir together constitute an isolated Hamiltonian system, the work W performed on the system of interest is equal to the change in the *total* energy of the system and reservoir: $W = G_1(\mathbf{y}(t_s)) - G_0(\mathbf{y}(0))$. Therefore, applying the analysis of the previous paragraph to the situation considered here, with $\mathbf{y}$, G_λ, and Y_λ replacing $\mathbf{z}$, H_λ, and Z_λ, respectively, we get

$$\overline{\exp(-\beta W)} = Y_1/Y_0 . \tag{9}$$

The right side of Eq. (9) depends only on the initial and final Hamiltonians $G_0(\mathbf{y})$ and $G_1(\mathbf{y})$, and on the temperature T, which means that $\overline{\exp(-\beta W)}$ is independent of the switching time t_s. But we already know that $\overline{\exp(-\beta W)} = \exp(-\beta \Delta F)$ in the limit $t_s \to \infty$, since $W = \Delta F$ for every member of the ensemble, in that limiting case. We therefore conclude that

$$\overline{\exp(-\beta W)} = \exp(-\beta \Delta F) \tag{10}$$

for *all* values of t_s (and all paths γ).

Equation (9), which tells us that $\overline{\exp(-\beta W)}$ is independent of both γ and t_s, is identically true, given the formulation of the problem. However, in going from Eq. (9) to Eq. (10), we invoke a result from quasiequilibrium statistical mechanics, which relies on the assumption of weak coupling (small h_{int}). Equation (2), therefore, is valid *for sufficiently weak coupling* between the system and reservoir. This may be seen more directly by writing an explicit expression for the ratio Y_1/Y_0: only if h_{int} may be

neglected does this ratio immediately reduce to Z_1/Z_0 [$= \exp(-\beta\Delta F)$].

Note that the inequality $\overline{W} \geq \Delta F$ [Eq. (1)] follows directly from the equality $\overline{\exp(-\beta W)} = \exp(-\beta\Delta F)$ [Eq. (2a)], by application of the mathematical identity $\overline{\exp x} \geq \exp\overline{x}$ [7]. This establishes $\overline{W} \geq \Delta F$ directly from a microscopic, Hamiltonian basis rather than by invoking the increase of entropy. [In the limit $t_s \to 0$, we have $\overline{W} = \langle \Delta H \rangle_0$, and Eq. (1) reduces to the Gibbs-Bogoliubov-Feynman bound [7], $\langle \Delta H \rangle_0 \geq \Delta F$.]

It is also worthwhile to point out that the right side of Eq. (2b) may be expanded as a sum of cumulants [see Eq. (9) of Ref. [4]]:

$$\Delta F = \sum_{n=1}^{\infty} (-\beta)^{n-1} \frac{\omega_n}{n!}, \tag{11}$$

where ω_n is the nth cumulant of the ensemble distribution of values of W. If this distribution happens to be Gaussian (as may be expected for sufficiently slow switching), then only the first two terms survive, and we have

$$\Delta F = \overline{W} - \beta\sigma^2/2, \tag{12}$$

where $\sigma^2 \equiv \overline{W^2} - \overline{W}^2$. The dissipated work W_{diss} ($= \overline{W} - \Delta F$) is then related to the fluctuations in W by $W_{\text{diss}} = \beta\sigma^2/2$. This fluctuation-dissipation relation has been obtained previously by Hermans [8].

The central result of this paper, Eq. (2), makes a concrete prediction regarding the outcome of an ensemble of measurements, which, in principle, is subject to experimental verification. In practice, however, the *applicability* of Eq. (2) may be severely limited by the following considerations. If the fluctuations in W from one measurement to the next are much larger than k_BT (i.e., if $\sigma \gg \beta^{-1}$), then the ensemble average of $\exp(-\beta W)$ will be dominated by values of W many standard deviations below $\overline{W}$. Since such values of the work represent statistically very rare events, it would require an unreasonably large number of measurements of W to determine $\overline{\exp(-\beta W)}$ with accuracy. Therefore, given a specific system of interest, switching path γ, and switching time t_s, the fluctuations in the work W must not be much greater than k_BT, if we are to have any hope of verifying Eq. (2) experimentally. This condition pretty much rules out macroscopic systems of interest. In recent years, however, the direct manipulation of *nanoscale* objects—and the measurement of forces thereon [9]—has become feasible. Such systems may offer the best chance for experimentally testing the new result of this paper.

So far, we have assumed that our system is coupled to a *physical* heat reservoir. It is interesting, however, to discuss this problem within the context of numerical simulations. On a computer, a heat reservoir must be "mocked up." One way to accomplish this is with a Nosé-Hoover (NH) thermostat [10]. In its simplest form, this method replaces the reservoir with a single variable ζ; motion in the extended phase space $(\mathbf{z}, \zeta)$ is governed by the NH equations,

$$\{\dot{q} = p/m, \quad \dot{p} = -\nabla\Phi_\lambda - \zeta p\}_n, \tag{13}$$

$$\dot{\zeta} = (K/K_0 - 1)/\tau^2. \tag{14}$$

[We have assumed $H_\lambda = p^2/2m + \Phi_\lambda(\mathbf{q})$. The index n runs over all D degrees of freedom of the system, $K = p^2/2m$ is the total kinetic energy of the system, $K_0 = \beta^{-1}D/2$ is the thermal average of K, and the parameter τ acts as a relaxation time.] For λ fixed, a trajectory $\mathbf{z}(t)$ generated by these equations of motion samples phase space according to the Boltzmann factor $\exp[-\beta H_\lambda(\mathbf{z})]$, provided that the evolution is sufficiently chaotic.

It is interesting to ask, does Eq. (2) remain valid if the system evolves under the NH equations, rather than under the influence of a physical reservoir? Let us consider an ensemble of initial conditions described by the density,

$$f(\mathbf{z}, \zeta, 0) = cZ_0^{-1} \exp[-\beta Q_0(\mathbf{z}, \zeta)], \tag{15}$$

where $Q_\lambda(\mathbf{z}, \zeta) \equiv H_\lambda(\mathbf{z}) + D\zeta^2\tau^2/2\beta$, and $c = (D\tau^2/2\pi)^{1/2}$ is a normalization factor. [The distribution $cZ_\lambda^{-1} \times \exp(-\beta Q_\lambda)$ is stationary under the NH equations when λ is held fixed, and may be viewed as the "canonical" distribution in the extended phase space.] Allowing these initial conditions to evolve under the NH equations, as λ changes from 0 to 1, we obtain an ensemble of trajectories described by a time-dependent density $f(\mathbf{z}, \zeta, t)$. As before, the work performed on each member of the ensemble is defined to be the time integral of $\dot{\lambda}\partial H_\lambda/\partial\lambda$. We now introduce a work accumulated function $w(\mathbf{z}, \zeta, t)$, analogous to $w(\mathbf{z}, t)$ introduced earlier. It is straightforward to establish that

$$f(\mathbf{z}, \zeta, t) = f(\mathbf{z}_0, \zeta_0, 0) \exp\left[D \int_0^t \zeta(t')\, dt' \right], \tag{16}$$

$$w(\mathbf{z}, \zeta, t) = Q_\lambda(\mathbf{z}, \zeta) - Q_0(\mathbf{z}_0, \zeta_0) + \beta^{-1}D \int_0^t \zeta(t')\, dt', \tag{17}$$

where $(\mathbf{z}_0, \zeta_0)$ are the initial conditions associated with the trajectory which passes through $(\mathbf{z}, \zeta)$ at time t, and the integral $\int_0^t \zeta\, dt'$ is performed along this trajectory. Then, repeating the steps leading to Eq. (8), we again get $\overline{\exp(-\beta W)} = \exp(-\beta\Delta F)$, where the overbar now denotes an average over our ensemble of NH trajectories. Thus, Eq. (2) remains valid [given the canonical initial distribution specified by Eq. (15)] when the system is coupled to a Nosé-Hoover thermostat. This result is identically true: No weak coupling assumption is necessary, nor do we need to assume that the evolution is chaotic.

It may similarly be established that Eq. (2) is identically valid when the thermostat is numerically implemented using the Metropolis Monte Carlo algorithm (see, e.g., Ref. [11]) rather than Nosé-Hoover dynamics.

Numerical simulations of this sort are often used to compute free energy differences of physical, chemical, or biological interest [12,13]. Typically, a number of simulations of slow switching from one configuration to another are performed, and the work W obtained from each simulation is treated as an estimate of the free energy difference ΔF. This estimate contains both statistical errors (W differs from one simulation to the next) and systematic errors [any finite-rate simulation has a bias, as per Eq. (1)]. Statistical errors are removed by averaging over many simulations, but the systematic error remains, thus the average of W really represents an upper bound on ΔF. (Reversing direction, a lower bound is established as well [11].) Now, Eq. (2a) tells us that if we use $e^{-\beta W}$ as an estimate for $e^{-\beta \Delta F}$ (rather than W for ΔF), then *this* estimate is *un*biased: There are only statistical errors. We can take advantage of this fact by using the "exponential average," $W^x \equiv -\beta^{-1} \ln \overline{\exp(-\beta W)}$, rather than the ordinary average $\overline{W}$, as an estimate of ΔF; the overbar now denotes an average over a finite number N_s of simulations. It is easily shown that the systematic error in W^x is smaller than that in $\overline{W}$, and vanishes as $N_s \to \infty$ [Eq. (2b)]. The upshot is that, if we perform more than one simulation of the switching process, then W^x will provide a tighter upper bound on ΔF than $\overline{W}$.

Hunter [14] has performed a "back-of-the-envelope" test of this idea. Reference [15] details the results of six switching simulations in which a threonine dipeptide is converted to an alanine dipeptide, and vice versa. Using this data, Hunter computed W^x for both sets of simulations, obtaining the following bounds: $-5.4 \leq \Delta F \leq -4.5$. This compares favorably with the bounds obtained by computing $\overline{W}$: $-6.2 \leq \Delta F \leq -3.8$.

In the time since the original submission of this paper, numerical simulations by Tams (manuscript in preparation) have provided a nice illustration of Eq. (2).

To summarize, the central result of this paper gives the equilibrium free energy difference ΔF between two configurations A and B of a classical system, in terms of an ensemble of finite-time measurements of the work performed on the system as it is switched from A to B. The derivation relies on the usual assumption of weak coupling between system and reservoir, but otherwise follows directly from Hamilton's equations. Two well-known equilibrium identities for ΔF, Eqs. (4) and (5), emerge as limiting cases of this more general, nonequilibrium result. Practical considerations, in all likelihood, limit the applicability of Eq. (2) to systems of no more than a moderate number of degrees of freedom. Finally, the equality may be useful when numerical simulations of thermostatted systems are used to compute free energy differences.

It is a pleasure to acknowledge that numerous stimulating discussions—including those with G. F. Bertsch, A. Bulgac, M. DenNijs, P. DeVries, G. J. Hogenson, J. Hunter III, D. B. Kaplan, W. P. Reinhardt, T. Schaefer, D. Thouless, and R. Venugopalan—contributed to the derivation and understanding of the results presented in this paper. This work was supported by the Department of Energy under Grant No. DE-FG06-90ER40561.

*Present address: Theoretical Astrophysics, T-6, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.
Electronic address: chrisj@t6–serv.lanl.gov

[1] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon Press, Oxford, 1990), 3rd ed., Part 1, Sect. 15. We use the term "configuration" to denote a fixed set of parameter values.
[2] However, if the system is confined within a box with perfectly hard walls, and if changing λ corresponds to moving these walls, then Eq. (2) remains true for any finite t_s, but does not reduce to Eq. (5) in the limit $t_s \to 0$.
[3] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[4] R. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[5] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965), p. 95.
[6] That is, before each realization of the switching process, we bring the coupled system and reservoir to equilibrium at temperature T (for instance, by temporarily coupling them to a much larger "super-reservoir").
[7] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University, New York, 1987), Sect. 5.5.
[8] J. Hermans, J. Chem. Phys. **95**, 9029 (1991).
[9] See, e.g., N. Agraït, G. Rubio, and S. Vieira, Phys. Rev. Lett. **74**, 3995 (1995).
[10] S. Nosé, J. Chem. Phys. **81**, 511 (1984); W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[11] J. E. Hunter III, W. P. Reinhardt, and T. F. Davis, J. Chem. Phys. **99**, 6856 (1993).
[12] See, e.g., L.-W. Tsao, S.-Y. Sheu, and C.-Y. Mou, J. Chem. Phys. **101**, 2302 (1994); T. C. Beutler and W. F. van Gunsteren, J. Chem. Phys. **101**, 1417 (1994); B. L. Holian, H. A. Posch, and W. G. Hoover, Phys. Rev. E **47**, 3852 (1993); M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).
[13] For reviews, 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, G. Ciccotti (North-Holland, Amsterdam, 1986).
[14] J. Hunter III (private communication).
[15] M. J. Mitchell and J. A. McCammon, J. Comput. Chem. **12**, 271 (1991).