IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Nonequilibrium work theorem for a system strongly coupled to a thermal environment

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

J. Stat. Mech. (2004) P09005

(http://iopscience.iop.org/1742-5468/2004/09/P09005)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 200.133.215.4
The article was downloaded on 25/02/2011 at 15:48

Please note that terms and conditions apply.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Nonequilibrium work theorem for a system strongly coupled to a thermal environment

Chris Jarzynski

Theoretical Division, Los Alamos National Laboratory, NM 87545, USA
E-mail: chrisj@lanl.gov

Received 6 August 2004
Accepted 30 August 2004
Published 21 September 2004
Online at stacks.iop.org/JSTAT/2004/P09005
doi:10.1088/1742-5468/2004/09/P09005

Abstract. In a recent paper (2004 *J. Stat. Mech.: Theor. Exp.* P07006), Cohen and Mauzerall (CM) have argued that the derivation of the nonequilibrium work relation given in a previous work of Jarzynski (1997 *Phys. Rev. Lett.* **78** 2690) is flawed. Here I attempt to answer their criticisms, both by presenting a detailed version of that derivation and by addressing specific objections raised by CM. The derivation presented here is in fact somewhat stronger than the one I gave in 1997, as it does not rely on the assumption of a weak coupling term connecting the system of interest and its thermal environment.

Keywords: exact results, fluctuations (theory)

JSTAT (2004) P09005

©2004 IOP Publishing Ltd PII: S1742-5468(04)85580-8 1742-5468/04/P09005+13$30.00

Contents

1. Derivation 3
2. Response to specific points 9
Acknowledgments 11
Appendix 11
References 12

JSTAT (2004) P09005

In a recent paper [1], Cohen and Mauzerall (CM) have raised questions about the validity of the nonequilibrium work relation, equation (1) below, which relates the external work W performed on a system during a nonequilibrium process to the free energy difference ΔF between two equilibrium states of the system. This prediction has been derived by various means in e.g. [2]–[11], and has been confirmed in an experiment performed by Liphardt *et al* [12], along lines suggested by Hummer and Szabo [7], involving the forced unfolding and refolding of a single strand of RNA [13]. Two recent papers review theoretical, computational, and experimental aspects of the nonequilibrium work relation and related results [14, 15]. While most of the criticisms of [1] are aimed at a derivation of equation (1) that I published in 1997 [2], CM also suggest that other derivations (in particular that given by Crooks in [4]) are relevant only for near-equilibrium processes, and that the experiment of [12] cannot be viewed as a confirmation of equation (1).

Some of the issues raised by CM pertain to aspects of equation (1) that are counter-intuitive, or at least sufficiently unusual to arouse justifiable scepticism. For instance, while the right-hand side of equation (1) is familiar enough—it is a ratio of partition functions—the left-hand side is not. The quantity inside the angular brackets is constructed by combining two values—the work W performed on a system that is driven out of equilibrium, and the temperature T of the initial equilibrium state of the system—into something that looks like a Boltzmann factor, namely $\mathrm{e}^{-\beta W}$ (where $1/\beta = k_{\mathrm{B}}T$). To CM this is an ad hoc, unjustified construction, and I partially agree with them; I can think of no good *a priori* reason to consider this particular quantity rather than some other. However, as long as β and W are both well-defined, then so is the value of $\mathrm{e}^{-\beta W}$, and it is a perfectly legitimate exercise to investigate its properties. Any justification for embarking on such an investigation can only come *a posteriori*: if it leads us to an interesting and potentially useful result, then that is sufficient reason for studying it in the first place!

My aim here is to address the arguments of CM, in particular their assertion that the derivation presented in [2] is flawed. To do so, I will first present a detailed version of that derivation (section 1), with the goal of establishing the nonequilibrium work relation as a mathematical identity, within the context of a Hamiltonian model for the evolution of the system of interest and its thermal environment. Following that, I will discuss several specific points raised by CM (section 2).

1. Derivation

The nonequilibrium work relation (or 'Jarzynski equality', in CM) can be stated as follows. Imagine a finite system that has been prepared in a state of equilibrium with a thermal environment at temperature T, and suppose that we subject this system to a thermodynamic process, by externally varying a work parameter of the system, λ, from an initial value A to a final value B. In doing so we both drive the system out of equilibrium, and perform some amount of work, W, on it. The precise value of W depends of course on the specific motions of the microscopic degrees of freedom that constitute the system, and these motions are in turn influenced by the degrees of freedom of the environment. Therefore let us imagine that we carry out this process infinitely many times. During each of these repetitions, or *realizations*, of the process, we begin with the system and environment in a state of equilibrium, and we always vary the work parameter λ in precisely the same manner from A to B. After each realization we note down the amount of work W performed on the system during that realization, and in the end we construct the distribution of work values, $\rho(W)$, observed over this set of realizations of the process. The nonequilibrium work relation states that this distribution satisfies a strong constraint, namely

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle \equiv \int \mathrm{d}W\, \rho(W)\, \mathrm{e}^{-\beta W} = \mathrm{e}^{-\beta \Delta F}, \tag{1}$$

which remains valid regardless of how slowly or quickly we varied the work parameter during the process. Here, $1/\beta = k_{\mathrm{B}}T$, and ΔF is the free energy difference between the equilibrium states associated with the initial and final values of the work parameter. To be precise, let Z_λ denote the partition function (defined by equation (21) below) corresponding to the equilibrium state of the system of interest, when the work parameter is held fixed at the value λ, and the system is in equilibrium with the environment. Then the free energy of that equilibrium state is given by the usual formula, $F_\lambda = -\beta^{-1} \ln Z_\lambda$, and the quantity ΔF appearing in equation (1) is *defined* to be

$$\Delta F \equiv F_B - F_A = -\beta^{-1} \ln \frac{Z_B}{Z_A}. \tag{2}$$

A special case of equation (1), pertaining to the situation in which an external perturbation to the system is turned on and then off (hence $\Delta F = 0$), was derived earlier by Bochkov and Kuzovlev [16].

I will now give a detailed version of the derivation of the nonequilibrium work relation found in [2], and will begin by spelling out the assumptions behind this derivation.

First, let us treat the system and its thermal environment as a set of classical degrees of freedom that are well isolated from the rest of the universe, and described by a Hamiltonian

$$\mathcal{H}(\Gamma; \lambda) = H(x; \lambda) + H_{\mathrm{E}}(y) + h_{\mathrm{int}}(x, y). \tag{3}$$

Here, x denotes a point in the phase space of the system of interest; y is a point in the (typically vastly larger) phase space of the thermal environment; $\Gamma = (x, y)$ is a point in the combined phase space of system and environment; and λ is an externally controlled work parameter. The terms on the right side of equation (3) correspond to the bare Hamiltonian for the system of interest (H), the bare Hamiltonian for the environment (H_{E}), and the energy of interaction between the system and the environment (h_{int}).

JSTAT (2004) P09005

Let us now imagine that we prepare the system and environment in an initial state of thermal equilibrium at a temperature T, with the work parameter held fixed at an initial value $\lambda = A$. To be specific, suppose that this preparation is accomplished by placing the combined system and environment in weak thermal contact with a much larger 'super-environment' at temperature T, and then removing the super-environment after an appropriate equilibration time. As a result of this preparation, the system and environment find themselves in a microstate $\Gamma_0 = (x_0, y_0)$ that can effectively be viewed as being sampled randomly from the canonical distribution in the full phase space:

$$p(\Gamma_0) = \frac{1}{Y_A} \exp[-\beta \mathcal{H}(\Gamma_0; A)]. \tag{4}$$

The normalization factor Y_A is a particular case ($\lambda = A$) of the quantity

$$Y_\lambda \equiv \int \mathrm{d}\Gamma \, \exp[-\beta \mathcal{H}(\Gamma; \lambda)]. \tag{5}$$

This is the classical partition function for the equilibrium state of the system and environment, when the work parameter is held fixed at a value λ.

Having prepared the initial state of equilibrium and removed the super-environment, we allow the system and environment to evolve over time as we vary the work parameter from $\lambda = A$ at $t = 0$ to $\lambda = B$ at $t = \tau$, according to some arbitrary but pre-determined schedule. The microscopic history of system and environment during this process is described by a trajectory $\{\Gamma_t\}$ evolving under Hamilton's equations in the full phase space. Here I use the notation $\{\Gamma_t\}$ to denote the *entire trajectory*, that is the microscopic history of the system and environment from $t = 0$ to τ. By contrast, the notation Γ_t (without the braces) denotes simply the microstate of the system and environment at a specific time t. Similarly, $\{\lambda_t\}$ will refer to the schedule for varying the work parameter from A to B, and λ_t to the value of the work parameter at a particular time t. The schedule $\{\lambda_t\}$ specifies how we act on the system during the process in question, whereas $\{\Gamma_t\}$ specifies how the system and environment respond, at the microscopic level, during a given realization of the process.

Let us interpret $H(x; \lambda)$ as the *internal energy* of the system of interest[1]. The net change in this quantity during a single realization of the process is equal, identically, to

$$H(x_\tau; B) - H(x_0; A) = \int_0^\tau \mathrm{d}t \, \dot{\lambda} \frac{\partial H}{\partial \lambda}(x_t, \lambda_t) + \int_0^\tau \mathrm{d}t \, \dot{x} \frac{\partial H}{\partial x}(x_t, \lambda_t). \tag{6}$$

It is natural to interpret the first integral on the right side as the *external work* (or *mechanical work* in [1]) performed on the system, and the second term as the *heat* absorbed by the system (see for instance the discussions in [4, 17, 18]). Equation (6) can then be viewed as a statement of the first law of thermodynamics, which asserts that the change

[1] This is a reasonable interpretation when the coupling between the system and environment is negligibly weak. However, when this is not the case, then the system and the environment effectively share the energy in the term h_{int}, and it is not completely obvious whether this energy should be viewed as belonging to the system or to the environment. In that situation, other considerations (not discussed here) suggest that the quantity $H^*(x; \lambda)$ (equation (20)) should replace $H(x; \lambda)$ in equations (7) and (8). This will change the definition of heat, equation (8), but will have no effect on the definition of work, equation (7), since $\partial H^*/\partial\lambda = \partial H/\partial\lambda$. Hence whether we use H or H^* in equations (7) and (8) has no bearing on the derivation of equation (1) presented in the present section.

JSTAT (2004) P09005

in the internal energy of the system is due to two contributions: the work performed on the system,

$$W \equiv \int_0^\tau \mathrm{d}t\, \dot{\lambda} \frac{\partial H}{\partial \lambda}(x_t, \lambda_t), \tag{7}$$

and the heat absorbed by the system,

$$Q \equiv \int_0^\tau \mathrm{d}t\, \dot{x} \frac{\partial H}{\partial x}(x_t, \lambda_t). \tag{8}$$

Equations (7) and (8) define W and Q in terms of the microscopic history of the system alone, $\{x_t\}$. In effect, while Nature integrates Hamilton's equations in the full phase space, we observers need to monitor only the degrees of freedom of the system of interest in order to deduce how much work was performed on the system, and how much heat was absorbed by it, during a given realization of the process. Note, however, that for a realization described by a trajectory $\{\Gamma_t\}$ in the full phase space, we have[2]

$$\mathcal{H}(\Gamma_\tau; B) - \mathcal{H}(\Gamma_0; A) = \int_0^\tau \mathrm{d}t\, \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{H}(\Gamma_t; \lambda_t) \tag{9}$$

$$= \int_0^\tau \mathrm{d}t\, \dot{\lambda} \frac{\partial \mathcal{H}}{\partial \lambda}(\Gamma_t; \lambda_t) \tag{10}$$

$$= \int_0^\tau \mathrm{d}t\, \dot{\lambda} \frac{\partial H}{\partial \lambda}(x_t; \lambda_t), \tag{11}$$

that is,

$$W = \mathcal{H}(\Gamma_\tau; B) - \mathcal{H}(\Gamma_0; A). \tag{12}$$

This tells us that the work performed on the system of interest is equal to the net change in the Hamiltonian of the combined system and environment.

The distinction between equations (7) and (12) is an important one: the former *defines* the external work performed on the system, in terms of its microscopic evolution; the latter states that the quantity W thus defined is equal to the net change in the combined energy of the system and environment (under the assumption of Hamiltonian evolution for the combined system and environment).

The preceding paragraphs have specified a *model* used to represent a system that is both in contact with a thermal environment, and subject to an externally controlled work parameter. Two strong but commonly made assumptions underlie this model: first, that quantum effects can be ignored; second, that the system and environment can be treated as being isolated from all other degrees of freedom. I will now show that equation (1) follows as a direct consequence of this model.

So far the discussion has focused on a single realization of the thermodynamic process in question. Now imagine that we carry out the process very many—in principle, infinitely many—times. We always prepare the system and environment in equilibrium as described

[2] The second line follows from the first by a well-known property of Hamilton's equations, namely that the total time derivative of the Hamiltonian along a trajectory that is a solution of Hamilton's equations is equal to the partial derivative of the Hamiltonian function with respect to time (see for instance [19]). To get to the third line we use the assumption that the full Hamiltonian $\mathcal{H}$ depends on the work parameter λ only through the term $H(x; \lambda)$, equation (3).

JSTAT (2004) P09005

above, and we always use the same schedule $\{\lambda_t\}$ to vary the work parameter from A to B. In other words, we act on the system in precisely the same way, over and over again. Nevertheless, the microscopic history of the system and environment, $\{\Gamma_t\}$, will differ from one realization to the next, simply because the initial microstate Γ_0 differs from one realization to the next (see equation (4)). Over the course of each realization we observe the evolution of the system's degrees of freedom, $\{x_t\}$, and from that empirical data we compute the value of W using equation (7). Finally, from the set of work values collected over these realizations, we construct the average of $\mathrm{e}^{-\beta W}$. Formally,

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathrm{e}^{-\beta W_n}, \tag{13}$$

where N denotes the number of realizations, and W_n is the work performed on the system during the nth realization. As above, $1/\beta = k_{\mathrm{B}}T$, where T is the initial temperature at which the system and environment are prepared, which is 'the only known temperature available', as CM correctly point at the bottom of page 4 of [1].

The preceding paragraph describes how to construct the desired average from a series of measurements. Let us now analyse this quantity theoretically: for a given Hamiltonian $\mathcal{H}$, initial temperature T, and schedule $\{\lambda_t\}$, what value will we obtain for the average defined by equation (13)? Note that this problem is fully specified: the probability distribution for the initial conditions Γ_0 is given by equation (4); the subsequent evolution in the full phase space, $\{\Gamma_t\}$, is determined by Hamilton's equations; and the quantities β and W are precisely defined. Thus the question that we have just posed has a unique answer within the context of our model, and it remains only to do the mathematics.

To carry out the analysis, let us introduce a function $\tilde{W}(\Gamma_0)$, which is the work performed on the system for a realization launched from initial conditions Γ_0 in the full phase space. (The initial conditions Γ_0 uniquely determine a trajectory $\{\Gamma_t\}$ in the full phase space, and from such a trajectory the value of W can be obtained through either equation (7) or (12).) Then the average we wish to evaluate can be written as

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \int \mathrm{d}\Gamma_0\, p(\Gamma_0)\, \exp[-\beta \tilde{W}(\Gamma_0)], \tag{14}$$

with $p(\Gamma_0)$ as given by equation (4). If we now combine equation (4) with (12), then after a single cancellation we get

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \int \mathrm{d}\Gamma_0\, \frac{1}{Y_A}\, \exp[-\beta \mathcal{H}(\Gamma_\tau(\Gamma_0); B)], \tag{15}$$

where the notation stresses the fact that Hamiltonian evolution is deterministic: $\Gamma_\tau(\Gamma_0)$ represents the final microstate in the full phase space, for the realization launched from the initial microstate Γ_0. Since there is a one-to-one correspondence between the initial conditions Γ_0 and the final conditions Γ_τ, we can perform a change of variables in the above integral:

$$\int \mathrm{d}\Gamma_0 \cdots = \int \mathrm{d}\Gamma_\tau \left| \frac{\partial \Gamma_\tau}{\partial \Gamma_0} \right|^{-1} \cdots, \tag{16}$$

JSTAT (2004) P09005

where $|\partial\Gamma_\tau/\partial\Gamma_0|$ is the Jacobian associated with this change of variables. By Liouville's theorem, this Jacobian is equal to unity[3]. We therefore have

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \int \mathrm{d}\Gamma_\tau \, \frac{1}{Y_A} \, \exp[-\beta \mathcal{H}(\Gamma_\tau; B)] = \frac{Y_B}{Y_A}, \tag{17}$$

by equation (5). Since the ratio Y_B/Y_A depends only on the parameter values $\lambda = A$ and B, and on the temperature T, equation (17) already establishes a strong result. Namely, even though the distribution of work values $\rho(W)$ generally depends on the specific protocol for varying λ from A to B, the quantity $\int \mathrm{d}W \, \rho(W) \, \mathrm{e}^{-\beta W}$ does not!

To this point, no approximations have been made in the analysis. Now, the right side of equation (17) is a ratio of partition functions of the combined system *and environment* (see equation (5)). We want to replace this by some expression pertaining to the system itself. In [2] this was accomplished by explicitly assuming the interaction energy h_{int} to be negligible in comparison with the other two terms in the Hamiltonian $\mathcal{H}$. (In this situation the contributions to Y_A and Y_B from the environmental degrees of freedom cancel one another:

$$\frac{Y_B}{Y_A} = \frac{\int \mathrm{d}\Gamma \, \exp[-\beta \mathcal{H}(\Gamma; B)]}{\int \mathrm{d}\Gamma \, \exp[-\beta \mathcal{H}(\Gamma; A)]} \approx \frac{\int \mathrm{d}x \, \exp[-\beta H(x; B)]}{\int \mathrm{d}x \, \exp[-\beta H(x; A)]} = \frac{Z_B}{Z_A} = \mathrm{e}^{-\beta \Delta F}, \tag{18}$$

where $Z_\lambda = \int \mathrm{d}x \, \exp[-\beta H(x; \lambda)]$ is the appropriate expression for the partition function of the system of interest when its coupling to the environment is negligibly weak.) In many situations of physical interest, however, h_{int} is *not* negligibly small. In this case the evaluation of the ratio Y_B/Y_A requires a bit more effort. As a starting point, let us recall that when h_{int} is finite, the equilibrium distribution of the system of interest is given by the following modification of the familiar Boltzmann–Gibbs formula:

$$p_{\mathrm{S}}(x; \lambda) \propto \exp[-\beta H^*(x; \lambda)], \tag{19}$$

where

$$H^*(x; \lambda) \equiv H(x; \lambda) - \beta^{-1} \ln \frac{\int \mathrm{d}y \exp\Big[-\beta[H_{\mathrm{E}}(y) + h_{\mathrm{int}}(x, y)]\Big]}{\int \mathrm{d}y \exp[-\beta H_{\mathrm{E}}(y)]} \tag{20}$$

is a *potential of mean force* (PMF) associated with the phase space variables of the system of interest [20]. The notation p_{S} indicates the probability distribution of the system of interest; this is obtained from the probability distribution in the full phase space, equation (4), by integrating over the environmental degrees of freedom [21][4]. For the equilibrium distribution given by equation (19), the partition function (normalization factor) is

$$Z_\lambda = \int \mathrm{d}x \, \exp[-\beta H^*(x; \lambda)]. \tag{21}$$

[3] Imagine an infinitesimal cell of volume ϵ centred around the point Γ_0 in the full phase space. If we propagate a trajectory for a time τ from each set of initial conditions found within this cell, then we will end up with an infinitesimal cell of final conditions, of volume ϵ', containing the point Γ_τ. The Jacobian, by definition, is the ratio of the volumes of these two cells, ϵ'/ϵ. But Liouville's theorem tells us that Hamiltonian dynamics gives rise to no phase space expansion or contraction; hence $\epsilon' = \epsilon$.

[4] Section 2 of this article contains a concise and very clear discussion of the PMF formalism in the context of non-negligible coupling between the system and environment.

JSTAT (2004) P09005

With these definitions, we have

$$Y_\lambda = Z_\lambda \cdot \int \mathrm{d}y \, \exp[-\beta H_{\mathrm{E}}(y)], \tag{22}$$

which immediately implies

$$\frac{Y_B}{Y_A} = \frac{Z_B}{Z_A}. \tag{23}$$

This is not an approximation, but follows identically from equations (5), (20), and (21) along with the definition of $\mathcal{H}$, equation (3). Combining this result with equations (2) and (17), we finally arrive at the nonequilibrium work relation:

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \frac{Y_B}{Y_A} = \frac{Z_B}{Z_A} = \mathrm{e}^{-\beta \Delta F}, \tag{24}$$

without resorting to a weak coupling assumption. For physical situations in which h_{int} happens to be negligible, we recover the situation discussed in [2].

Note that the quantity ΔF has been defined *mathematically*, in terms of a ratio of partition functions (equations (2), (21)). In the appendix, I briefly argue that it is reasonable to view this quantity as a *physical* free energy difference.

It is important to stress that the derivation which has just been presented is *exact*: equation (24) is a mathematical equality, given the model specified above. The key feature of this model is that the system and environment are treated as an isolated collection of classical degrees of freedom evolving under Hamilton's equations. While this model represents the traditional approach of classical statistical mechanics, there is a subtlety associated with it, even if we agree to ignore quantum effects. This subtlety arises from the fact that it is in practice impossible to completely isolate a system and its immediate thermal environment; unavoidable interactions with the rest of the universe introduce effectively random perturbations to the evolution of the trajectory $\{\Gamma_t\}$. (Moreover these perturbations are correlated with $\{\Gamma_t\}$.) No matter how weak these perturbations are, their effect on a given trajectory becomes magnified exponentially with time if the evolution of $\{\Gamma_t\}$ is chaotic, as is typically the case for a realistic thermal environment. With this in mind, are we really justified in invoking Liouville's theorem in going from equation (15) to (17)? This theorem reflects a delicate balance between sets of initial and final conditions of trajectories evolving under Hamilton's equations, a balance that might well be upset by the addition of even the tiniest amount of randomness into the equations of motion. Questions such as this make it all the more important that equation (1) be tested in actual laboratory experiments.

The above derivation relies explicitly on the assumption that the initial equilibrium state is represented by a canonical distribution in the full phase space (equation (4)). This assumption was justified by the somewhat artificial construct of a super-environment. However, the validity of equation (1) might not depend as strongly on a literal interpretation of equation (4) as the derivation suggests. For instance, in the pulling experiment of [12], the microscopic system of interest—a single strand of RNA, two micron-size beads, and the DNA handles used to attach the RNA to the beads—is immersed in a macroscopic bath of water molecules. As long as that bath is prepared at a well-defined temperature, one intuitively expects the behaviour of the biomolecule, immersed deep within the aqueous solution, to not depend in any significant way on

JSTAT (2004) P09005

whether the combined system and environment are prepared *exactly* in the canonical distribution given by equation (4). Thus we would expect the same outcome (apart from extremely small corrections) if the initial conditions Γ_0 were instead sampled from a microcanonical distribution, $p \propto \delta(E - \mathcal{H})$; this is a variant of the usual expectation of *equivalence of ensembles* [22]. For a quantitative discussion of this issue, see section II.B of [15].

2. Response to specific points

A central claim of [1] is that the heat exchange between the system and environment has 'not been properly taken into account' (page 5 of [1]). While it is true that the derivation given above never makes explicit use of the quantity Q, this does not imply that Q is assumed to be zero, or that in some other way the heat has been mishandled in the analysis. It simply means that—within the context of the model—one can evaluate the average of $\exp(-\beta W)$ without mentioning Q in the calculation.

As an example of their claim that heat has not been treated properly, CM consider the situation in which the system of interest is out of equilibrium at the moment when the work parameter reaches its final value (page 5 of [1]). It is worthwhile to discuss this example in some detail. To begin, recall that ΔF should always be understood as the free energy difference between *the two equilibrium states associated with the initial and final values of the work parameter*, rather than as the free energy difference between the initial and final states of the system of interest. (Indeed, if the system is out of equilibrium at the end of the process, then its final free energy might not be well-defined.) To be absolutely precise, for any process during which the work parameter is varied from A to B, the quantity ΔF appearing in equation (1) is defined by (see equations (2), (21))

$$\Delta F = \frac{\int \mathrm{d}x \, \exp[-\beta H^*(x; B)]}{\int \mathrm{d}x \, \exp[-\beta H^*(x; A)]}, \tag{25}$$

regardless of whether or not the system is in equilibrium at the end of the process.

Now let us consider a two-stage schedule for varying the work parameter. During the first stage ($0 \le t \le \tau_1$), λ is changed in some arbitrary way from A to B; during the second stage ($\tau_1 \le t \le \tau_2$), λ is held fixed at the value B. Let us assume that the system is out of equilibrium at the time τ_1, but that the second ('relaxation') stage is sufficiently long for the system to relax to equilibrium. Now imagine two observers, one of whom monitors the behaviour of the system only during the first stage, while the other monitors the behaviour during both stages. Thus according to the first observer the process ends at time τ_1 (with the system out of equilibrium), whereas the second observer contends that the process ends at a later time τ_2 (with the system in equilibrium). As before, we imagine infinitely many repetitions of the process.

For every realization, the two observers agree on the precise amount of work performed on the system during the process, even though they disagree as to when the process ends. This follows from equation (7): since λ is fixed for $t > \tau_1$, there is no contribution to W from the relaxation stage. Therefore, when the two observers independently construct the average $\langle \exp(-\beta W) \rangle$ after many realizations of the process, they both arrive at the same number for the left side of equation (1). Moreover, since the two observers agree that the work parameter begins at A (at $t = 0$) and ends at B (at $t = \tau_1$ or τ_2), they also agree on

JSTAT (2004) P09005

the value of ΔF, as defined by equation (25). Hence when using their data to assess the validity of equation (1) the two observers will be comparing exactly the same numbers. In other words, whether or not we choose to include a relaxation stage—during which we hold λ fixed at its final value, so as to let the system come to equilibrium—has no bearing whatsoever on the validity of equation (1). This is a simple consequence of the definitions of the quantities W and ΔF.

Although no work is done on the system during the relaxation stage, there is typically a certain amount of heat exchanged between system and environment during this stage. Thus the observed values of Q generally do depend on whether the process is defined to end at time τ_1 or time τ_2. But this in no way affects the validity of equation (1), since that prediction concerns work, as defined by equation (7), and not heat.

Another point raised by CM concerns the factor $\beta = 1/k_{\rm B}T$, where T denotes the initial temperature of the system and environment. Once the system is driven away from equilibrium, it might not have a well-defined temperature, and even if it did there would be no guarantee that it would be equal to the initial temperature T. Therefore, in the expression $\mathrm{e}^{-\beta W}$, a value that pertains to a system out of equilibrium (W) is divided by a temperature (T) that does not meaningfully represent the state of that system, except at $t = 0$. CM assert, with some justification, that such a pairing appears arbitrary and without foundation (see e.g. page 4 of [1]). Note the nature of this criticism: CM do not claim that the quantity $\mathrm{e}^{-\beta W}$ is somehow inherently 'illegal' or ill-defined, but rather that it is ad hoc. However, as discussed briefly in the introduction above, as long as β and W are well-defined, it is perfectly acceptable to inquire about the average of $\mathrm{e}^{-\beta W}$ over different realizations of the process. As the detailed calculation of section 1 reveals in the context of a particular model—and as has been shown by a number of other derivations using significantly different models [3]–[11]—this average works out to be equal to $\mathrm{e}^{-\beta\Delta F}$. It might be surprising that a construction as intuitively unnatural as $\mathrm{e}^{-\beta W}$ should lead to such a simple result, but this does not automatically invalidate the result. Indeed, the fact that this quantity seems unnatural might simply reflect our limited intuition regarding nonequilibrium processes [23].

In section 2 of their paper, CM consider two factors that play an important role in determining whether a process is reversible or irreversible. The first is the rate of heat transfer between the system and environment, $\dot{c}$, which is related to the strength of the coupling between them; the second is the rate at which work is performed, $\dot{w} = \dot{\lambda}\,\partial H/\partial\lambda$. CM discuss several cases illustrating how the balance between $\dot{c}$ and $\dot{w}$ affects the reversibility or irreversibility of the process. Their discussion is physically motivated and certainly seems correct, but it does not bear on the validity of the derivation of the nonequilibrium work relation given in [2]. As shown in detail in section 1 above, that derivation is based on very general considerations involving Hamilton's equations, Liuoville's theorem, and the use of the canonical ensemble to represent the initial equilibrium state of the system and environment. These considerations remain valid independently of the values of $\dot{c}$ and $\dot{w}$. To put it another way: the analysis presented in section 1 above depends neither on the rate at which the work parameter is varied, nor on the strength of the coupling between the system and its environment; hence it is as valid for irreversible processes as it is for reversible ones.

In section 5 of [1] CM discuss the example of an isolated harmonic oscillator whose frequency is externally switched from an initial value ω_0 to a final value ω_1 (where $\omega_1 > \omega_0$),

JSTAT (2004) P09005

over a switching time t_s. In [3] I had shown that in the limit of infinitely slow switching, and assuming a canonical distribution of initial conditions for the oscillator, one can solve exactly for the distribution of work values:

$$\lim_{t_s\to\infty}\rho(W)=\frac{\omega_0\beta}{\omega_1-\omega_0}\exp\left(\frac{-\omega_0\beta W}{\omega_1-\omega_0}\right)\theta(W), \tag{26}$$

where $\theta(\cdot)$ is the unit step function. It is easy to verify that this distribution satisfies the nonequilibrium work relation, equation (1) above. The derivation of equation (26) makes use of an adiabatic invariant and therefore is valid only in the limit $t_s\to\infty$. This does not, however, imply that the nonequilibrium work relation fails for finite values of t_s, only that for finite switching times it is not easy to obtain an exact expression for $\rho(W)$. (Exactly solvable models are hard to come by! See, however, [24].) Therefore in [3] the analytical calculation of $\rho(W)$ for infinite switching times (equation (26) above) was supplemented by numerical experiments carried out for five different finite values of t_s. The results of these experiments were in excellent agreement with equation (1), as shown by the diamonds in figure 2 of [3]. It is difficult to reconcile these results with CM's statement that the nonequilibrium work relation 'is critically dependent on adiabatic invariance for finite t_s, even for the harmonic oscillator' (page 14), particularly since in [3] adiabatic invariance was only invoked in the limit $t_s\to\infty$.

Of course a single example does not establish universal validity. General proofs of equation (1) for Hamiltonian systems were given in [2] and [3], and the harmonic oscillator was meant to serve only as a simple illustration.

Acknowledgments

It is a pleasure to acknowledge useful discussions and correspondence with A Adib, C Bustamante, E G D Cohen, G E Crooks, J Jarzynski, J Liphardt, and F Ritort. This research was supported by the Department of Energy, under contract W-7405-ENG-36.

Appendix

Equation (22) can be written explicitly as

$$\int \mathrm{d}\Gamma\, \mathrm{e}^{-\beta\mathcal{H}(\Gamma;\lambda)}=\int \mathrm{d}x\, \mathrm{e}^{-\beta H^*(x;\lambda)}\cdot\int \mathrm{d}y\, \mathrm{e}^{-\beta H_\mathrm{E}(y)}. \tag{A.1}$$

Thus the partition function for the combined system and environment factorizes nicely as the product of two partition functions, one for the system of interest (which includes all the effects of the interaction energy) and the other for the environment. If we take the natural logarithm of both sides and multiply by $-\beta^{-1}$, we can rewrite the above result as

$$\mathcal{F}_\lambda=F_\lambda+F_\mathrm{E}^0, \tag{A.2}$$

where $\mathcal{F}_\lambda=-\beta^{-1}\ln Y_\lambda$ can be viewed as the equilibrium free energy of the combined system and environment, and $F_\mathrm{E}^0=-\beta^{-1}\ln\int \mathrm{d}y\, \mathrm{e}^{-\beta H_\mathrm{E}(y)}$ as that of the bare environment. Note that F_E^0 is a macroscopic quantity, describing a macroscopic thermal environment, whereas the characteristic magnitude of F_λ is determined by the size of the system

JSTAT (2004) P09005

of interest; e.g. F_λ (and therefore ΔF) is microscopic for a single-molecule pulling experiment.

Given the definitions in section 1, it is easy to show that the quantity $F_\lambda = -\beta^{-1} \ln Z_\lambda$ satisfies $\partial F_\lambda / \partial \lambda = \langle \partial H^* / \partial \lambda \rangle_\lambda^{\rm eq} = \langle \partial H / \partial \lambda \rangle_\lambda^{\rm eq}$, or equivalently

$$\Delta F = \int_A^B {\rm d}\lambda \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda^{\rm eq}. \tag{A.3}$$

Here $\langle \cdots \rangle_\lambda^{\rm eq} = \int {\rm d}x\, p_{\rm S}(x;\lambda) \cdots$ denotes an equilibrium average at a fixed value of the work parameter. (The derivation of equation (A.3), not reproduced here, is just a few lines long, and essentially identical to the derivation of the well known *thermodynamic integration* identity; see e.g. [25].)

Now suppose we carry out a *reversible* process: the system passes through a continuous sequence of equilibrium states as we slowly vary λ from A to B. Thus, at any time t during this process, the system of interest is sampling its phase space according to the equilibrium distribution ($p_{\rm S}$) corresponding to the current value of work parameter, λ_t. In this situation we can replace the value of $\partial H / \partial \lambda$ appearing in the definition of work (equation (7)), by its equilibrium average:

$$W \to \int_0^\tau {\rm d}t\, \dot\lambda \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda_t}^{\rm eq} = \int_A^B {\rm d}\lambda \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda^{\rm eq} = \Delta F, \tag{A.4}$$

invoking equation (A.3). The arrow denotes that we are considering the special case of a reversible, quasi-static process. By these arguments, then, $W = \Delta F$ for any reversible process during which λ is changed, quasi-statically, from A to B. This suggests that we are justified in interpreting ΔF, defined mathematically in terms of the modified Boltzmann distribution, equation (19), as a physical equilibrium free energy difference.

References

[1] Cohen E G D and Mauzerall D, 2004 *J. Stat. Mech.: Theor. Exp.* P07006 [cond-mat/0406128]
[2] Jarzynski C, 1997 *Phys. Rev. Lett.* **78** 2690
[3] Jarzynski C, 1997 *Phys. Rev.* E **56** 5018
[4] Crooks G E, 1998 *J. Stat. Phys.* **90** 1481
[5] Crooks G E, 1999 *Phys. Rev.* E **60** 2721
[6] Crooks G E, 2000 *Phys. Rev.* E **61** 2361
[7] Hummer G and Szabo A, 2001 *Proc. Nat. Acad. Sci.* **98** 3658
[8] Jarzynski C, 2002 *Dynamics of Dissipation* ed P Garbaczewski and R Olkiewicz (Berlin: Springer)
[9] Sun S X, 2003 *J. Chem. Phys.* **118** 5769
[10] Evans D J, 2003 *Mol. Phys.* **101** 1551
[11] Mukamel S, 2003 *Phys. Rev. Lett.* **90** 170604
[12] Liphardt J *et al*, 2002 *Science* **296** 1832
[13] Schurr J M and Fujimoto B S, 2003 *J. Phys. Chem.* B **107** 14007 contains a detailed discussion of the relation between this experiment and the theoretical prediction
[14] Ritort F, 2003 *Poincaré Seminar* **2** 195 [cond-mat/0401311]
[15] Park S and Schulten K, 2004 *J. Chem. Phys.* **120** 5946
[16] Bochkov G N and Kuzovlev Y E, 1977 *Zh. Eksp. Teor. Fiz.* **72** 238
Bochkov G N and Kuzovlev Y E, 1977 *Sov. Phys. JETP* **45** 125 (translation)
Bochkov G N and Kuzovlev Y E, 1981 *Physica* A **106** 443
Bochkov G N and Kuzovlev Y E, 1981 *Physica* A **106** 480
[17] Sekimoto K, 1998 *Prog. Theor. Phys. Suppl.* **130** 17
[18] Jarzynski C, 1998 *Acta Phys. Polon.* B **29** 1609
[19] Goldstein H, 1980 *Classical Mechanics* 2nd edn (Reading, MA: Addison-Wesley) chapter 8

JSTAT (2004) P09005

[20] Kirkwood J G, 1935 *J. Chem. Phys.* **3** 300
[21] Roux B, *Implicit solvent models*, 2001 *Computational Biochemistry and Biophysics* ed O Becker, A D MacKerell, B Roux and M Watanabe (New York: Dekker)
[22] Ruelle D, 1999 *Statistical Mechanics: Rigorous Results* (London: Imperial College Press) section 1.1
[23] Adib A B, 2004 personal communication
[24] Mazonka O and Jarzynski C, 1999 *Preprint* cond-mat/9912121
[25] Frenkel D and Smit B, 2002 *Understanding Molecular Simulation: From Algorithms to Applications* (San Diego, CA: Academic) section 7.1

JSTAT (2004) P09005