Journal of Statistical Physics, Vol. 98, Nos. 1/2, 2000

Hamiltonian Derivation of a Detailed Fluctuation Theorem

C. Jarzynski[1]

Received June 21, 1999; final August 13, 1999

We analyze the microscopic evolution of a system undergoing a far-from-equilibrium thermodynamic process. Explicitly accounting for the degrees of freedom of participating heat reservoirs, we derive a hybrid result, similar in form to both the fluctuation theorem and a statement of detailed balance. We relate this result to the steady-state fluctuation theorem and to a free energy relation valid far from equilibrium.

KEY WORDS: Fluctuation theorem; irreversible processes.

I. INTRODUCTION

The *fluctuation theorem* refers collectively to a number of theoretical results in the field of nonequilibrium statistical mechanics, which are striking because they are valid *far* from thermal equilibrium. Following the original numerical discovery by Evans, Cohen, and Morriss,(1) a *transient* fluctuation theorem (applicable to systems driven away from an initial state of equilibrium), and a *steady*-*state* fluctuation theorem (for systems in a non-equilibrium steady state), were derived by Evans and Searles,(2) and by Gallavotti and Cohen,(3) respectively, for systems evolving under deterministic but non-Hamiltonian equations of motion. These results have stimulated considerable research,(4–19) in which the fluctuation theorem has been generalized (in particular to stochastic evolution) and related to linear response theory, specific examples have been studied, and the relation between the transient and steady-state fluctuation theorems has been discussed.

[1] Theoretical Division, T-13, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545; e-mail: chrisj@lanl.gov.

The steady-state version of the fluctuation theorem can be written as:

$$\lim_{\tau \to \infty} \frac{1}{\tau} \ln \frac{p_\tau(+\bar{\sigma})}{p_\tau(-\bar{\sigma})} = \frac{\bar{\sigma}}{k_B} \tag{1}$$

where k_B is the Boltzmann constant, and $p_\tau(\bar{\sigma})$ is the probability distribution of observing an average entropy production rate $\bar{\sigma}$ over a time interval of length τ. The distribution is defined with respect to an ensemble of trajectory segments of duration τ, sampled while the system in question evolves in a nonequilibrium steady state.

Physically, maintaining a system in a nonequilibrium steady state requires the participation of one or more heat reservoirs, for instance to absorb the heat generated by shear forces, or to maintain boundaries of the system at different temperatures. Derivations of the fluctuation theorem which have appeared in the literature (whether pertaining to systems driven away from equilibrium, or to those in a nonequilbrium steady state) have discussed a variety of thermostating schemes, both deterministic[(1–14)] and stochastic,[(15–19)] to model the presence of reservoirs. Many of these schemes originated as numerical strategies for simulating the microscopic evolution of a system in thermal contact with a heat reservoir, without simulating the huge number of degrees of freedom making up the reservoir itself. The very fact that the fluctuation theorem seems to be independent of the thermostating scheme lends it considerable support. Indeed, Maes[(19)] has argued on quite general grounds that the fluctuation theorem can be understood as a Gibbs property of space-time histories; see also ref. 20 for illustrative examples.

Recently, Crooks[(18)] has shown that the fluctuation theorem is closely related to another set of results[(21–27)]—also valid far from equilibrium—which relate the free energy difference between two equilibrium states of a system, to the external work performed on the system during a *non*-equilibrium process from one state to the other.

The purpose of the present paper is to derive a result similar to the fluctuation theorem, *by explicitly including the degrees of freedom of heat reservoirs in the analysis*, and assuming Hamiltonian evolution at the microscopic level. This approach corresponds closely to the situation present in a laboratory experiment, where the "thermostating" is precisely the result of interactions between the system and the innumerable degrees of freedom which constitute its environment. We will argue that, when all microscopic degrees of freedom are taken into account, then there emerges a "detailed fluctuation theorem" (Eq. (4) below), valid for finite times τ, and expressed without reference to steady states.

A Hamiltonian treatment of nonequilibrium processes, similar in spirit to that taken here, was carried out by Bochkov and Kuzovlev[28, 29] (though with less emphasis on a distinction between the "system of interest" and its "environment"). The central result derived below, however, is new, as is the connection to the fluctuation theorem. More recently, Eckmann, Pillet, and Rey-Bellet[30] have introduced and studied an exactly solvable, Hamiltonian model of a system (a chain of nonlinear oscillators) coupled to two heat reservoirs at different temperatures. It would be very interesting to establish the precise relation between the results which they obtain for their model—especially in connection with the nonequilibrium steady state—and the approach taken in Section IV of the present paper.

In the following two sections, the central result is stated and derived. While this result does not explicitly refer to a nonequilibrium steady state, we argue in Section IV that, under appropriate circumstances, it leads to the steady-state fluctuation theorem. In Section V we show that the nonequilibrium free energy relation of refs. 21–27 also follows from the central result of the present paper. We end with a discussion in Section VI.

II. STATEMENT OF CENTRAL RESULT

Suppose we have the following ingredients at our disposal:

1. a finite, classical *system of interest*, ψ,
2. a number of *heat reservoirs*, $\theta_1, \theta_2,..., \theta_N$,
3. and, possibly, a *work parameter*, λ.

The work parameter is some degree of freedom which we control externally, for instance an external field, and which interacts directly with ψ (but not with the reservoirs). The reservoirs are also finite, classical systems, prepared ahead of time at various temperatures. We suppose that we can establish or break thermal contact between our system of interest and any of the reservoirs, as we choose. Finally, we assume that, at the most fundamental level of description, the system and reservoirs are composed of a large number of microscopic degrees of freedom, and that the collection of these obeys Hamiltonian evolution.

Our ability to directly manipulate λ, and to make or break contact with the reservoirs, allows us to subject our system of interest to a variety of thermodynamic processes. We will take the word *process* to be synonymous with an explicit prescription spelling out "what we do to the system" at the macroscopic level, using the tools at our disposal (the work parameter and heat reservoirs). More precisely, a process Π is defined by a set of instructions specifying:

1. the temperatures $(T_1,..., T_N)$ at which to prepare the reservoirs,
2. when to establish and/or break thermal contact between ψ and any of the θ's, and
3. the time-dependence of the work parameter, $\lambda(t)$.

We will refer to items 2 and 3 as the *protocol*. We will restrict ourselves to processes occurring over a finite interval of time, $[0, \tau]$, and will say that a process is *static* if the work parameter and thermal contacts are constant over the course of the process. Note that a process is defined without reference to the preparation of the system of interest itself.

Having introduced the notion of a process to specify "what we do to ψ" at the macroscopic level, let us now turn our attention to the microscopic response of the system of interest and reservoirs. A complete description of this response is provided by a trajectory $\Gamma(t)$, detailing the (Hamiltonian) evolution of all participating degrees of freedom. We will, however, be interested in a less complete descriptions consisting of: the microscopic history of ψ itself, and the net entropy generated, ΔS, over the course of the process. By the former, we mean a trajectory $\mathbf{z}(t)$ specifying the evolution of the microstate of the system of interest (the position and momentum of each constituent degree of freedom), from $t=0$ to $t=\tau$. By *entropy generated*, we mean the quantity

$$\Delta S \equiv -\sum_{n=1}^{N} \frac{Q_n}{T_n} \tag{2}$$

where Q_n denotes the net heat absorbed by ψ from the nth reservoir, over the course of the process. We justify the nomenclature by noting that, at the macroscopic level of description, $-Q_n/T_n$ corresponds to the net change in the entropy of the nth reservoir. Thus, ΔS can be viewed as the change in the entropy of the *environment* of ψ (the collection of reservoirs), which we abbreviate to "entropy generated." (See also the definition of the rate of entropy production introduced in ref. 30.)

Note that both $\mathbf{z}(t)$ and ΔS can be obtained from the complete microscopic description, $\Gamma(t)$: the former, by projecting out the reservoir degrees of freedom; the latter, by using the initial and final conditions $\Gamma(0)$ and $\Gamma(\tau)$ to compute the net change in the internal energy of each reservoir. (See Eq. (8) below, and commentary in Section VI).

Because the system of interest interacts with the reservoirs, the microscopic evolution of ψ itself is not deterministic. Rather, an initial microstate $\mathbf{z}_A$ determines a *statistical ensemble* of possible realizations, each characterized by a particular history $\mathbf{z}(t)$, and a particular value of entropy generated ΔS. This is the ensemble of realizations which we would obtain by endlessly repeating the same process, always initializing ψ in the microstate $\mathbf{z}_A$; the difference from one realization to the next arises solely from the different initial conditions sampled for the reservoirs. From this ensemble, let us imagine constructing the statistic

$$P(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \tag{3}$$

which is the joint probability distribution of obtaining a final microstate $\mathbf{z}(\tau) = \mathbf{z}_B$, and an entropy generated ΔS, conditional on the initial microstate $\mathbf{z}(0) = \mathbf{z}_A$. This joint, conditional probability distribution P will be the object of central interest in this paper. This is admittedly a somewhat peculiar quantity to investigate. No attempt will be made to motivate our interest in this statistic other than that "it works," in the sense that consideration of $P(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$ leads to the neat result expressed by Eq. (4) below.

Let us now introduce a final piece of notation. For an arbitrary process Π^+, let its *time-reversed* counterpart, Π^-, denote the process obtained by using the same set of reservoir temperatures, but carrying out the protocol of Π^+ in reverse order (reversing the time-dependence of both the work parameter and the thermal contacts established and broken). We will, quite arbitrarily, refer to Π^+ as the "forward" process and Π^- as the "reverse" process. When discussing the conditional probability $P(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$, computed for two processes Π^+ and Π^- related by time-reversal, the notation P_+ and P_- is used to distinguish between the two cases.

The central result of this paper then asserts that the probability distributions P_+ and P_- satisfy the following relation:

$$\frac{P_+(\mathbf{z}_B, +\Delta S \mid \mathbf{z}_A)}{P_-(\mathbf{z}_A^*, -\Delta S \mid \mathbf{z}_B^*)} = \exp(\Delta S/k_B) \tag{4}$$

where the asterisk (*) denotes a reversal of momenta: $(\mathbf{q}, \mathbf{p})^* = (\mathbf{q}, -\mathbf{p})$. To obtain some intuition for what this result says, imagine filming the evolution of the system, work parameter, and reservoirs during one realization of the process Π^+, as the microstate of ψ evolves from $\mathbf{z}_A$ to $\mathbf{z}_B$ and the entropy generated is ΔS. Now imagine running the film backward; you will then see a realization of the process Π^-, with ψ evolving from $\mathbf{z}_B^*$ to $\mathbf{z}_A^*$, and entropy generation $-\Delta S$. Equation (4) thus relates the conditional probability of observing one set of events ($\mathbf{z}_A \to \mathbf{z}_B$, $+\Delta S$) during a given process, to that of observing the time-reversal of those events ($\mathbf{z}_B^* \to \mathbf{z}_A^*$, $-\Delta S$) during the time-reversed process: it states that the ratio of these two probabilities is just the exponent of the entropy generated, ΔS, in units of k_B. For a static thermodynamic process, we have $\Pi^+ = \Pi^-$, and therefore we can drop the subscripts + and − appearing in Eq. (4).

Equation (4) is a hybrid result, akin both to the fluctuation theorem (through dependence on ΔS), and to a statement of detailed balance (because of the appearance of the initial and final microstates of ψ); for this reason we refer to it as a *detailed fluctuation theorem.*

Note that if $\Delta S > 0$, then the conditional probability appearing in the numerator is greater than that in the denominator; if $\Delta S < 0$, the opposite is true. This makes intuitive sense: of the two scenarios, the one which remains obedient to the second law by generating positive entropy is more likely than its disobedient twin, by a factor exponential in the entropy generated.

The proof of Eq. (4) will follow directly from the assumption that evolution in the full phase space (including the degrees of freedom of $\psi, \theta_1, ..., \theta_N$) is deterministic and Hamiltonian. For a process Π, the statistical ensemble of realizations corresponding to a particular initial microstate $\mathbf{z}_A$ for the system of interest is then defined operationally: it is the ensemble obtained by initializing ψ in the microstate $\mathbf{z}_A$, then sampling the initial conditions of the reservoirs $(\mathbf{y}_1^0, ..., \mathbf{y}_N^0)$ from canonical distributions at the specified temperatures $(T_1, ..., T_N)$. Given $\mathbf{z}_A$, the sampled values of $(\mathbf{y}_1^0, ..., \mathbf{y}_N^0)$ uniquely determine the subsequent evolution of all degrees of freedom, $\Gamma(t)$. The probability distribution of obtaining a given realization, conditional on a given microstate $\mathbf{z}_A$ for ψ, then reduces to that of sampling the appropriate initial microstates of the reservoirs.

III. DERIVATION

To carry out the derivation of Eq. (4), we begin by introducing notation and spelling out assumptions, starting with the classical approximation: all quantal effects are ignored.

The system of interest, ψ, is taken to have a finite number of degrees of freedom, and its instantaneous microstate is described by a point $\mathbf{z} = (\mathbf{q}, \mathbf{p})$ in the phase space of ψ, with the usual assignment of $\mathbf{q}$ to denote configurational variables, and $\mathbf{p}$ the associated momenta. At any instant in time, the internal energy of ψ is given by a Hamiltonian $H_\lambda^\psi(\mathbf{z})$, parametrized by the current value of the work parameter λ.

Next, assume that each heat reservoir θ_n is itself a classical system with a finite number of degrees of freedom, whose microstate is described by a point $\mathbf{y}_n$ in the phase space associated with that reservoir. We do *not* assume the reservoirs to be physically identical, so the dimensionalities of the $\mathbf{y}_n$'s may differ. The internal energy of the nth reservoir is given by a Hamiltonian $H_n^\theta(\mathbf{y}_n)$.

Finally, let $h_n^{\rm int}(\mathbf{z}, \mathbf{y}_n)$ denote a weak coupling term between the system of interest and the nth reservoir. As discussed below, thermal contact between ψ and θ_n can be established or broken by turning this term "on" and "off," using parameters $c_n(t)$.

For simplicity of presentation assume time-reversal invariance ("no magnetic fields") for these Hamiltonian terms:

$$H_\lambda^\psi(\mathbf{z}^*) = H_\lambda^\psi(\mathbf{z}), \qquad H_n^\theta(\mathbf{y}_n^*) = H_n^\theta(\mathbf{y}_n), \qquad h_n^{\rm int}(\mathbf{z}^*, \mathbf{y}_n^*) = h_n^{\rm int}(\mathbf{z}, \mathbf{y}_n) \tag{5}$$

The vector

$$\Gamma = (\mathbf{z}, \mathbf{Y}) = (\mathbf{z}, \mathbf{y}_1, ..., \mathbf{y}_N) \tag{6}$$

specifies the instantaneous state of all degrees of freedom involved, where $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_N)$ denotes the collective microstate of the N reservoirs. The evolution of $\Gamma(t)$ is taken to be deterministic, and governed by a Hamiltonian

$$\mathscr{H}(\Gamma, t) = H_{\lambda(t)}^\psi(\mathbf{z}) + \sum_{n=1}^N H_n^\theta(\mathbf{y}_n) + \sum_{n=1}^N c_n(t)\, h_n^{\rm int}(\mathbf{z}, \mathbf{y}_n) \tag{7}$$

Here $\lambda(t)$ denotes the time-dependence of the work parameter, and the $c_n(t)$'s take on values of 0 or 1, which can also change with time. At a given time t, if $c_n(t) = 1$, then this indicates that the system of interest is in thermal contact with the nth reservoir at that time; when $c_n(t) = 0$, ψ and θ_n are *not* in contact. (More generally, we could let the c_n's take on a continuous range of values, allowing the thermal contacts to be turned on and off smoothly rather than abruptly. This modification would have no effect on the analysis.)

The collection $\{\lambda, \vec{c}\} \equiv \{\lambda, c_1, ..., c_N\}$ represents the "tool-kit" available for externally manipulating the system of interest. The protocol for a given process, Π, is then just a particular prescription for doing so: it is synonymous with a specific set of functions of time, $\{\lambda(t), \vec{c}(t)\}$, instructing us exactly how to manipulate the work parameter and thermal contacts over a time interval $[0, \tau]$. This protocol uniquely specifies the time-dependence of the Hamiltonian function $\mathscr{H}(\Gamma, t)$, since that time-dependence enters only through $\lambda(t)$ and $\vec{c}(t)$. We will use $\mathscr{H}_\Pi(\Gamma, t)$ to denote the time-dependent Hamiltonian corresponding to a particular process Π. If the protocol for a process Π^+ is $\{\lambda(t), \vec{c}(t)\}$, then that of its time-reversed counterpart Π^- is given by $\{\lambda(\tau - t), \vec{c}(\tau - t)\}$.

While the time-dependence of the parameters $\{\lambda, \vec{c}\}$ is controlled externally, the participating *dynamical* degrees of freedom $\Gamma = (\mathbf{z}, \mathbf{y}_1, ..., \mathbf{y}_N)$ evolve under Hamilton's equations, as determined by $\mathscr{H}_\Pi(\Gamma, t)$. Thus, initial conditions $\Gamma(0)$ uniquely determine a trajectory $\Gamma(t)$, which chronicles the microscopic history of all degrees of freedom. From this trajectory, as mentioned earlier, we can extract both the microscopic history of the system of interest, $\mathbf{z}(t)$ (by projecting out the reservoir variables $\mathbf{Y}$), and the net heat absorbed by ψ from each of the reservoirs. The heat absorbed by ψ from a particular reservoir is equal to the net decrease in the internal energy of that reservoir:

$$Q_n = H_n^\theta(\mathbf{y}_n(0)) - H_n^\theta(\mathbf{y}_n(\tau)) \tag{8}$$

From the Q_n's we in turn compute the entropy generated (Eq. (2)).

We note that, if $\Gamma_+(t)$ is a microscopic realization of a process Π^+, then $\Gamma_-(t) \equiv \Gamma_+^*(\tau - t)$ is a realization of the reverse process Π^-. This follows from the assumption of time-reversal invariance: if $\Gamma_+(t)$ satisfies Hamilton's equations for the forward process, then $\Gamma_-(t)$ will do so for the reverse.

The reservoirs, as mentioned, are initially prepared at specified temperatures, $T_1,..., T_N$. We take this to imply that their initial microstates $\mathbf{y}_n^0 \equiv \mathbf{y}_n(0)$ are sampled from canonical ensembles. This defines the following probability distribution for the collection of initial reservoir conditions:

$$p(\mathbf{Y}^0) = \mathcal{N}^{-1} \prod_{n=1}^{N} \exp[-H_n^\theta(\mathbf{y}_n^0)/k_B T_n] \tag{9}$$

where $\mathcal{N}(T_1,..., T_N)$ is a product of partition functions.

Finally, for a process Π^+ and a set of initial conditions Γ in the full phase space, let

$$\hat{\Gamma}{}^t_+(\Gamma) \equiv (\hat{\mathbf{z}}{}^t_+(\Gamma), \hat{\mathbf{Y}}{}^t_+(\Gamma)) \tag{10}$$

denote the point in phase space reached after time t, and let $\Delta\hat{S}_+(\Gamma)$ denote the net entropy generated over the entire realization of the process (from $t=0$ to $t=\tau$). The carats emphasize that $\hat{\Gamma}{}^t_+$, $\hat{\mathbf{z}}{}^t_+$, $\hat{\mathbf{Y}}{}^t_+$, and $\Delta\hat{S}_+$, are viewed as *functions* of the initial conditions Γ. For the time-reversed process, we adopt the same notation, with an obvious change in subscript ($\hat{\Gamma}{}^t_-$, $\hat{\mathbf{z}}{}^t_-$, etc.)

All the pieces needed to derive Eq. (4) are now in place. We begin with a formal expression for the joint, conditional probability distribution in which we are interested:

$$P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) = \int d\mathbf{Y}\, p(\mathbf{Y})\, \delta[\mathbf{z}_B - \hat{\mathbf{z}}{}^\tau_+(\mathbf{z}_A, \mathbf{Y})] \cdot \delta[\Delta S - \Delta\hat{S}_+(\mathbf{z}_A, \mathbf{Y})] \tag{11}$$

Using the identity $\hat{\mathbf{z}}^0_+(\mathbf{z}, \mathbf{Y}) = \mathbf{z}$, we rewrite this as:

$$P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \\ = \int d\Gamma\, p(\mathbf{Y}) \cdot \delta[\mathbf{z}_A - \hat{\mathbf{z}}^0_+(\Gamma)]\, \delta[\mathbf{z}_B - \hat{\mathbf{z}}^\tau_+(\Gamma)] \cdot \delta[\Delta S - \Delta\hat{S}_+(\Gamma)] \quad (12)$$

where $\Gamma \equiv (\mathbf{z}, \mathbf{Y})$. Now let $\Gamma' = (\mathbf{z}', \mathbf{Y}') = \hat{\Gamma}^\tau_+(\Gamma)$ denote the *final* point in the full phase space, for a realization of Π^+ launched from initial conditions Γ. For a given $\Gamma = (\mathbf{z}, \mathbf{Y})$, we can rewrite $p(\mathbf{Y})$ as:

$$p(\mathbf{Y}) = \frac{p(\mathbf{Y})}{p(\mathbf{Y}')}\, p(\mathbf{Y}') = \exp[\Delta\hat{S}_+(\Gamma)/k_B]\; p(\mathbf{Y}') \quad (13)$$

using Eqs. (2), (8), and (9), which leads to

$$P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \\ = e^{\Delta S/k_B} \int d\Gamma\, p(\mathbf{Y}')\, \delta[\mathbf{z}_A - \hat{\mathbf{z}}^0_+(\Gamma)]\, \delta[\mathbf{z}_B - \hat{\mathbf{z}}^\tau_+(\Gamma)] \cdot \delta[\Delta S - \Delta\hat{S}_+(\Gamma)] \quad (14)$$

Here $p(\mathbf{Y}')$ is *not* to be interpreted as "the probability distribution of final reservoir conditions," but rather as the function p defined by Eq. (9), evaluated at $\mathbf{Y}' = \hat{\mathbf{Y}}^\tau_+(\Gamma)$. Since Γ' is reached from Γ by time evolution under the process Π^+, and since we have assumed time-reversal invariance (Eq. (5)), it follows that, if we reverse the final momenta and launch a realization of Π^- from initial conditions Γ'^*, then we will obtain the time-reversed image of the original realization: $\hat{\Gamma}^t_-(\Gamma'^*) = [\hat{\Gamma}^{\tau-t}_+(\Gamma)]^*$. From this it follows that

$$\hat{\mathbf{z}}^t_-(\Gamma'^*) = [\hat{\mathbf{z}}^{\tau-t}_+(\Gamma)]^*, \qquad \Delta\hat{S}_-(\Gamma'^*) = -\Delta\hat{S}_+(\Gamma) \quad (15)$$

This allows us to rewrite Eq. (14) as:

$$P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) = e^{\Delta S/k_B} \int d\Gamma\, p(\mathbf{Y}'^*)\, \delta[\mathbf{z}^*_A - \hat{\mathbf{z}}^\tau_-(\Gamma'^*)] \\ \times \delta[\mathbf{z}^*_B - \hat{\mathbf{z}}^0_-(\Gamma'^*)] \cdot \delta[\Delta S + \Delta\hat{S}_-(\Gamma'^*)] \quad (16)$$

where we have used the fact that $p(\mathbf{Y}') = p(\mathbf{Y}'^*)$ (Eqs. (5) and (9)). Finally, since the integrand is expressed in terms of Γ'^*, which is an invertible function of Γ (defined by time evolution, followed by a reversal of momenta), let us change the variables of integration from Γ to Γ'^*. The Jacobian for this change of variables is unity (by Liouville's theorem), so we simply replace $d\Gamma$ by $d\Gamma'^*$ in Eq. (16). But then we can drop the prime and asterisk altogether (since Γ'^* is just a variable of integration) to get:

$$P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$$

$$= e^{\Delta S/k_B} \int d\Gamma\, p(\mathbf{Y})\, \delta[\mathbf{z}_A^* - \hat{\mathbf{z}}_-^\tau(\Gamma)]\, \delta[\mathbf{z}_B^* - \hat{\mathbf{z}}_-^0(\Gamma)] \cdot \delta[\Delta S + \Delta \hat{S}_-(\Gamma)] \tag{17}$$

$$= e^{\Delta S/k_B} \int d\mathbf{Y}\, p(\mathbf{Y})\, \delta[\mathbf{z}_A^* - \hat{\mathbf{z}}_-^\tau(\mathbf{z}_B^*, \mathbf{Y})] \cdot \delta[\Delta S + \Delta \hat{S}_-(\mathbf{z}_B^*, \mathbf{Y})] \tag{18}$$

$$= e^{\Delta S/k_B} P_-(\mathbf{z}_A^*, -\Delta S \mid \mathbf{z}_B^*) \tag{19}$$

which is the desired result.

The origin of the exponential term in Eq. (4) can be understood informally, as follows. Given a "forward" realization $\Gamma_+(t)$, and its time-reversed image $\Gamma_-(t)$, $e^{\Delta S/k_B}$ is the probability distribution for sampling the reservoir initial conditions corresponding to the forward realization, relative to those corresponding to the reverse realization, from canonical distributions: $e^{\Delta S/k_B} = p(\mathbf{Y})/p(\mathbf{Y}'^*)$. The probability P_+ appearing in the numerator of Eq. (4) is a sum of contributions from all realizations for which ψ evolves from $\mathbf{z}_A$ to $\mathbf{z}_B$ and the entropy generated is $+\Delta S$; and similarly for P_-. The two sets of realizations are in one-to-one correspondence with each other: for every $\Gamma_+(t)$ contributing to P_+ there is a time-reversed realization $\Gamma_-(t)$ contributing to P_-. Since, for every such pair of realizations, the ratio of probability distributions for sampling the associated initial reservoir conditions is $e^{\Delta S/k_B}$, the ratio of the two sums (P_+ to P_-) is equal to this exponential.

We end this section by pointing out, that a result similar to Eq. (4) can be derived for the statistic

$$P(\mathbf{z}_1, \mathbf{z}_2,..., \mathbf{z}_M, \Delta S \mid \mathbf{z}_0), \qquad M \geqslant 1 \tag{20}$$

which is the joint probability distribution that ψ will evolve through the sequence of points $\mathbf{z}_1, \mathbf{z}_2,..., \mathbf{z}_M$, at times $t_1, t_2,..., t_M$, where $t_m = m\tau/M$, and that the entropy generated will be ΔS, given $\mathbf{z}(0) = \mathbf{z}_0$. Formally, for a process Π^+,

$$P_+(\mathbf{z}_1 \cdots \mathbf{z}_M, \Delta S \mid \mathbf{z}_0)$$

$$= \int d\mathbf{Y}\, p(\mathbf{Y})\, \delta[\Delta S - \Delta \hat{S}_+(\mathbf{z}_0, \mathbf{Y})] \prod_{m=1}^{M} \delta[\mathbf{z}_m - \hat{\mathbf{z}}_+^{t_m}(\mathbf{z}_0, \mathbf{Y})] \tag{21}$$

A calculation similar to the one presented above then gives:

$$\frac{P_+(\mathbf{z}_1, \mathbf{z}_2 \cdots \mathbf{z}_M, +\Delta S \mid \mathbf{z}_0)}{P_-(\mathbf{z}^*_{M-1} \cdots \mathbf{z}^*_0, -\Delta S \mid \mathbf{z}^*_M)} = \exp(\Delta S/k_B) \tag{22}$$

Note that the discrete trajectory implied in the denominator ($\mathbf{z}^*_M \to \cdots \to \mathbf{z}^*_0$) is the time-reversed image of the one in the numerator ($\mathbf{z}_0 \to \cdots \to \mathbf{z}_M$). Equation (4) is just a special case, $M=1$, of Eq. (22). The latter remains valid as well in the opposite limit, $M \to \infty$ (with τ fixed), in which case the entire history of ψ is specified. We then write, in suggestive notation,

$$\frac{P_+[\mathbf{z}_+(t), +\Delta S \mid \mathbf{z}_+(0)]}{P_-[\mathbf{z}_-(t), -\Delta S \mid \mathbf{z}_-(0)]} = \exp(\Delta S/k_B) \tag{23}$$

where $\mathbf{z}_-(t) = \mathbf{z}^*_+(\tau - t)$.

For an *isolated* system ($N=0$) perturbed by external forces of finite duration, we have $\Delta S = 0$, by definition. Thus, by Eq. (22), the conditional probability distribution of observing the (isolated) system evolve through a given sequence of points during the process Π^+, is equal to that of observing it to pass through the time-reversed sequence during Π^-. This probability distribution will be a product of δ-functions: either the unique trajectory launched from $\mathbf{z}_0$ goes through the sequence $\mathbf{z}_1,..., \mathbf{z}_M$, or it does not. In this case ($N=0$), Eq. (23) is essentially equivalent to Eq. (7) of ref. 28. (A technical point of difference is that Bochkov and Kuzovlev consider the *unconditional* probability of observing a given realization, *assuming the system of interest begins in equilibrium*, for both the forward and the reverse realization; the exponential factor which they obtain is a ratio of probabilities of sampling microstates of ψ itself from a given equilibrium distribution.)

For a single heat reservoir ($N=1$), Eq. (22) is similar to Eq. (9) of ref. 23. The main difference is that in Crooks' formulation the evolution of the system of interest is explicitly taken to be a Markov process, occurring in discrete steps. Here, by contrast, the microstates $\mathbf{z}_m$ represent "snapshots" of ψ taken at equally-spaced time intervals during continuous-time evolution, and in general this sequence of states cannot be viewed as a Markov chain.

Evans and Searles[2] have also derived the (transient) fluctuation theorem by comparing the probabilities of sampling initial conditions of pairs of finite-time trajectories, one the time-reversed image of the other. In their approach, the system of interest evolves under deterministic but non-Hamiltonian equations of motion, to model the presence of a heat reservoir. They find, as above, that the probability measure of a given trajectory, relative to that of its time-reversed twin, is the exponent of the entropy generated (where the latter is identified with phase space contraction).

IV. RELATION TO THE STEADY-STATE FLUCTUATION THEOREM

In terms of the *average entropy generation rate*, $\bar{\sigma} \equiv \Delta S/\tau$, Eq. (4) can be rewritten as

$$\frac{1}{\tau} \ln \frac{P_{+}(\mathbf{z}_B,\ +\bar{\sigma}\tau \mid \mathbf{z}_A)}{P_{-}(\mathbf{z}_A^*,\ -\bar{\sigma}\tau \mid \mathbf{z}_B^*)} = +\bar{\sigma}/k_B \tag{24}$$

which is reminiscent of the steady-state fluctuation theorem, Eq. (1). However, the correspondence is not exact: Eq. (1) applies explicitly to a nonequilibrium steady state, contains the limit $\tau \to \infty$, and exhibits no dependence on initial and final microstates of the system of interest, all in contrast to Eq. (24). In this section we pursue the relationship between the detailed fluctuation theorem of this paper and the steady-state fluctuation theorem.

Rather than aiming at complete generality, we will focus on a specific physical situation which might exhibit a nonequilibrium steady state in the appropriate limit, with the expectation that the line of reasoning applied here can serve as a model for other examples. In Fig. 1, the system of interest is a fluid composed of particles of "type A," inside a finite cylindrical container of length l. The reservoirs are fluids of "type B" particles, contained in two cylinders of length L abutting the ends of ψ. Let v_ψ denote the number of particles constituting the system of interest, and let $v_\theta = v_1 = v_2$ denote the number in either reservoir. Assume that the forces between particles are pairwise, unaffected by the barriers between the cylinders, and have a short interaction range $r < l$. Also assume that all particles scatter elastically off the container walls.

Given this set-up, one cannot expect the system to reach a nonequilibrium steady state, *except possibly in the limit of infinitely large reservoirs.* We will now argue, quantitatively though not rigorously, that if ψ indeed reaches a steady state in this limit, and if fluctuations in the entropy production in that state are characterized by finite correlation times, then Eq. (4) (or 24) implies the steady-state fluctuation theorem.

The system of interest and reservoirs depicted in Fig. 1 are governed by a Hamiltonian of the form given in Eq. (7), with

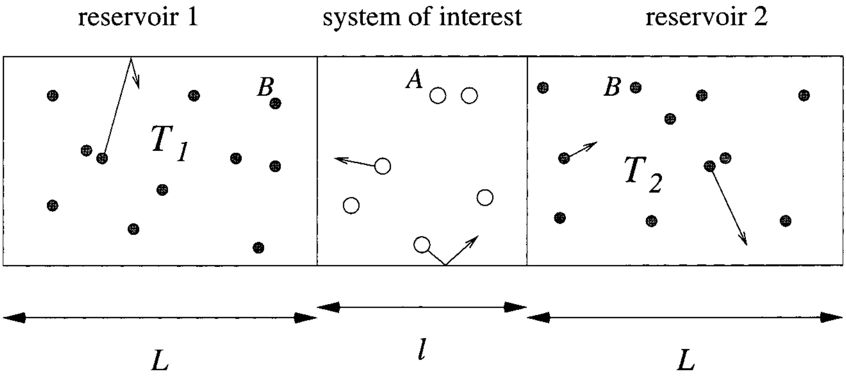

Fig. 1. Three interacting fluids. See text for details.

$$H^{\psi}(\mathbf{z}) = \sum_{i=1}^{\nu_\psi} \frac{\mathbf{p}^{[i]\,2}}{2m_A} + \sum_{i<j} V_{AA}(\mathbf{q}^{[i]}, \mathbf{q}^{[j]}) + \text{b.c.} \tag{25}$$

$$H_n^{\theta}(\mathbf{y}_n) = \sum_{i=1}^{\nu_\theta} \frac{\mathbf{p}_n^{[i]\,2}}{2m_B} + \sum_{i<j} V_{BB}(\mathbf{q}_n^{[i]}, \mathbf{q}_n^{[j]}) + \text{b.c.} \tag{26}$$

$$h_n^{\text{int}}(\mathbf{z}, \mathbf{y}_n) = \sum_{i=1}^{\nu_\psi} \sum_{j=1}^{\nu_\theta} V_{AB}(\mathbf{q}^{[i]}, \mathbf{q}_n^{[j]}) \tag{27}$$

where the index $n = 1, 2$ labels the reservoirs. Here, the microstates of the system of interest and reservoirs are denoted by

$$\mathbf{z} = (\mathbf{q}^{[1]}, \mathbf{p}^{[1]}, \ldots, \mathbf{q}^{[\nu_\psi]}, \mathbf{p}^{[\nu_\psi]}) \tag{28}$$

$$\mathbf{y}_n = (\mathbf{q}_n^{[1]}, \mathbf{p}_n^{[1]}, \ldots, \mathbf{q}_n^{[\nu_\theta]}, \mathbf{p}_n^{[\nu_\theta]}), \tag{29}$$

V_{xy} represents the short-range interaction potential between particles of type x and y, and "b.c." denotes boundary conditions, implying elastic reflection off the walls of the containers. There is no work parameter, and thermal contact between the system of interest and the reservoirs is always "on" ($c_n = 1$).

Let us now choose two temperatures T_1 and T_2 to be associated with the reservoirs θ_1 and θ_2. We can then subject the system of interest to a (static) thermodynamic process, by starting with ψ in some initial microstate $\mathbf{z}_A$, sampling the initial microstates $(\mathbf{y}_1^0, \mathbf{y}_2^0)$ of the reservoirs from canonical distributions at the chosen temperatures, and letting the entire system evolve for a time τ under the Hamiltonian $\mathscr{H} = H^{\psi} + \sum_n H_n^{\theta} + \sum_n h_n^{\text{int}}$. For this process, we can construct the statistic $P_\tau^{\omega}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$. This is the joint, conditional probability distribution defined in Section II, but with the dependence on the duration of the process, τ, and the size of the reservoirs, $\omega \equiv (L, \nu_\theta)$, explicitly stated.

The detailed fluctuation theorem, Eq. (4), then tells us that

$$\frac{P_\tau^\omega(\mathbf{z}_B, +\Delta S \mid \mathbf{z}_A)}{P_\tau^\omega(\mathbf{z}_A^*, -\Delta S \mid \mathbf{z}_B^*)} = \exp(\Delta S/k_B) \tag{30}$$

for this static process. Let us now change variables, from ΔS to $\bar{\sigma} = \Delta S/\tau$, by defining

$$p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A) \equiv P_\tau^\omega(\mathbf{z}_B, \bar{\sigma}\tau \mid \mathbf{z}_A) \cdot \tau, \tag{31}$$

the joint probability distribution of observing, after a time τ, a final microstate $\mathbf{z}_B$, and an average entropy generation rate $\bar{\sigma}$, conditional on an initial microstate $\mathbf{z}_A$.

We view $p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A)$ as a function of $\mathbf{z}_A$, $\mathbf{z}_B$, and $\bar{\sigma}$, parametrized by the values of τ, L, and ν_θ. Let us now *assume*, first, that

$$p_\tau^\Omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A) \equiv \lim_{\omega \to \infty} p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A) \qquad \text{exists (pointwise)} \tag{32}$$

where "$\omega \to \infty$" denotes the limit $L, \nu_\theta \to \infty$, with the particle density ν_θ/L held fixed. Thus, we assume that the dynamics of ψ converges to a well-defined limit, as we let the reservoirs become infinitely large. Equation (30) then implies that

$$\frac{1}{\tau} \ln \frac{p_\tau^\Omega(\mathbf{z}_B, +\bar{\sigma} \mid \mathbf{z}_A)}{p_\tau^\Omega(\mathbf{z}_A^*, -\bar{\sigma} \mid \mathbf{z}_B^*)} = \frac{\bar{\sigma}}{k_B} \tag{33}$$

for any finite value of τ.

[A bit of care is needed here, since, in the limit $\omega \to \infty$, the entropy generated becomes defined in terms of differences between infinite numbers (the initial and final energies of θ_1 and θ_2). The assumption expressed by Eq. (32) states that, for any *fixed* $\bar{\sigma}$, the quantity $p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A)$ converges to a particular value as the reservoirs become increasingly larger; roughly speaking, even though typical initial and final reservoir energies diverge in that limit, the energy differences Q_n do not. One can heuristically argue that this is a reasonable assumption, as follows. For a given particle density and temperature, there ought to be a characteristic "signal velocity" v with which information about the microstate of particles in one region of the reservoir gets propagated to other regions. If we choose $L \gg v\tau$, then we expect the particles at the far end of either reservoir to have negligible influence on the evolution of ψ, hence $p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A)$ will be unaffected by further increases in reservoir size, ω. When taking the limit $\tau \to \infty$ below, it will be understood that the limit $\omega \to \infty$ comes first.]

Let us next assume that, under the dynamics imposed by infinitely large reservoirs, ψ evolves to a statistical steady state. Let $f^S(\mathbf{z})$ denote the distribution of microstates, and $p_\tau^S(\bar{\sigma})$ the probability distribution of observing an average entropy production rate $\bar{\sigma}$ over a time interval of duration τ, in the steady state.[2] We further assume that, in the steady state, fluctuations in the entropy production are characterized by a finite correlation time t_c, so that the average entropy production rates measured over two adjacent time intervals of duration t_c can be treated as statistically independent. Finally, let $\bar{\sigma}^S$ denote the infinite-time average entropy production rate (equivalently, the *expectation value* of the average entropy production rate over any finite time interval) in the steady state:

$$\lim_{\tau \to \infty} p_\tau^S(\bar{\sigma}) = \delta(\bar{\sigma} - \bar{\sigma}^S) \tag{34}$$

Given these assumptions, we now want to justify replacing the numerator and denominator of Eq. (33) by $p_\tau^S(+\bar{\sigma})$ and $p_\tau^S(-\bar{\sigma})$, respectively, in the limit $\tau \to \infty$.

As a first step in this direction, we define

$$p_\tau(\mathbf{z}_B \mid \mathbf{z}_A) \equiv \int d\bar{\sigma} p_\tau^\Omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A) \tag{35}$$

and

$$p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B) \equiv p_\tau^\Omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A) / p_\tau(\mathbf{z}_B \mid \mathbf{z}_A) \tag{36}$$

(We will drop the superscript Ω henceforth, with the understanding that the limit of infinite reservoirs is assumed in the remainder of this section.) The former is the probability distribution of observing a microstate $\mathbf{z}_B$ at $t = \tau$, given $\mathbf{z}_A$ at $t = 0$. The latter is the distribution of average entropy production rates $\bar{\sigma}$ over the interval from $t = 0$ to $t = \tau$, conditional on an initial state $\mathbf{z}_A$ and a final state $\mathbf{z}_B$. Note that

$$\lim_{\tau \to \infty} p_\tau(\mathbf{z}_B \mid \mathbf{z}_A) = f^S(\mathbf{z}_B) \tag{37}$$

[2] Note the distinction between the use of the subscript τ in the statistic $p_\tau^S(\bar{\sigma})$, and its use in $p_\tau^\omega(\mathbf{z}_B, \bar{\sigma} \mid \mathbf{z}_A)$. In the latter, the time interval of duration τ is measured from the initial time, $t = 0$, at which the reservoir microstates are drawn from canonical distributions. In the former, the interval is measured starting at some moment *after the steady state has been achieved.* In both cases, the entropy generated is defined in terms of the net changes in internal energies of the reservoirs, over the interval in question.

by the assumption that ψ evolves to a stationary steady state.

Equation (33) now becomes

$$\frac{1}{\tau}\ln\frac{p_\tau(\mathbf{z}_B \mid \mathbf{z}_A)}{p_\tau(\mathbf{z}_A^* \mid \mathbf{z}_B^*)}+\frac{1}{\tau}\ln\frac{p_\tau(+\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)}{p_\tau(-\bar{\sigma} \mid \mathbf{z}_B^*, \mathbf{z}_A^*)}=\frac{\bar{\sigma}}{k_B} \tag{38}$$

which again is valid for any τ. By Eq. (37), the first term on the left will vanish as $\tau \to \infty$. It remains then to show that, in this limit, the numerator and denominator of the second term can be replaced by $p_\tau^S(+\bar{\sigma})$ and $p_\tau^S(-\bar{\sigma})$. It is tempting to argue that the dependence of $p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)$ on the specified initial and final microstates will vanish as $\tau \to \infty$, and therefore

$$p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B) \to p_\tau^S(\bar{\sigma}) \tag{39}$$

This is true in the sense that both sides of Eq. (39) converge to the same distribution of values of $\bar{\sigma}$, namely, $\delta(\bar{\sigma}-\bar{\sigma}^S)$. However, since that limiting distribution is singular, we cannot simply assume that, for instance, the ratio of $p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)$ to $p_\tau^S(\bar{\sigma})$ at a given value of $\bar{\sigma}$ converges to unity as $\tau \to \infty$. (In general it does not.) Justifying the above-mentioned replacement will therefore require some work.

Let $\underline{x}$ be a stochastic variable denoting the time-averaged entropy production rate during an interval of duration t_c, when the system is in the steady state. Thus, the value of $\underline{x}$ is a value of $\bar{\sigma}$ sampled randomly from the distribution $p_{t_c}^S(\bar{\sigma})$. Note that $\langle \underline{x} \rangle = \bar{\sigma}^S$, where angular brackets denote expectation value. For an interval of duration $\tau = Kt_c$ (where K is a positive integer) we then have, by our assumption of finite correlations,

$$p_\tau^S(\bar{\sigma}) = \langle \delta(\bar{\sigma}-\underline{X}) \rangle \tag{40}$$

where

$$\underline{X}=\frac{1}{K}\sum_{k=1}^{K} \underline{x}_k \tag{41}$$

and the $\underline{x}_k$'s denote independent samples of the same stochastic variable.

Equation (40) pertains to a system already in the steady state. Let us write down a similar equation for $p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)$. For τ sufficiently large, we can divide the interval $[0, \tau]$ into three segments: initial, intermediate, and final. During the initial segment, the system relaxes to the steady state, and the influence of the initial state $\mathbf{z}_A$ is felt statistically: the probability distribution of values of entropy produced during this segment depends on $\mathbf{z}_A$.

During the intermediate segment, the system is in the steady state, and its behavior is independent of either $\mathbf{z}_A$ or $\mathbf{z}_B$. During the final segment, the influence of the assumed final state $\mathbf{z}_B$ is felt statistically. The amounts of entropy generated during each of these segments are statistically independent of one another. For simplicity, let us assume that the initial and final segments are both of duration t_c, and that the total interval $\tau = Kt_c$ as above. Then we can write

$$p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B) = \langle \delta(\bar{\sigma} - \underline{Y}) \rangle \tag{42}$$

where

$$\underline{Y} = \frac{1}{K}\left(\underline{a} + \underline{b} + \sum_{k=1}^{K-2} \underline{x}_k\right) \tag{43}$$

where $\underline{a}$ and $\underline{b}$ are stochastic variables representing the average entropy production rate during the initial and final segments, respectively. The dependence of these on $\mathbf{z}_A$ and $\mathbf{z}_B$ is implicit.

Both $p_\tau^S(\bar{\sigma})$ and $p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)$ have been reduced to distributions of averages of K independently drawn samples. In the former case, all K samples are drawn from the same distribution (Eq. (41)), corresponding to the steady state. In the latter case, $K-2$ samples are drawn from that distribution, and the remaining two ($\underline{a}$ and $\underline{b}$) are drawn otherwise (Eq. (43)). How do these two distributions compare, in the limit $K \to \infty$ (in which one would expect the contributions of $\underline{a}$ and $\underline{b}$ in Eq. (43) to become negligible)? Introducing $\underline{c} \equiv \underline{a} + \underline{b}$, we can write

$$p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B) = \int dc\, \eta(c)\, p_{\tau'}^S(\bar{\sigma}') \tag{44}$$

where $\eta(c) = \langle \delta(c - \underline{c}) \rangle$ is the probability distribution of values of $\underline{c}$; $\tau' = \tau - 2t_c$; and

$$\bar{\sigma}' = \frac{K\bar{\sigma} - c}{K-2} \tag{45}$$

is the entropy generation rate implied for the intermediate segment (of duration τ'), if the rates during the initial and final segments sum to a value c, and the time-averaged rate for the entire interval $[0, \tau]$ is $\bar{\sigma}$. Thus, in Eq. (44) we are integrating over all possible ways of splitting a total entropy $\Delta S = \bar{\sigma}\tau$ into a sum of two terms: that generated during initial and final segments (ct_c), and the remnant ($\bar{\sigma}'\tau'$) during the intermediate, steady-state segment.

The distribution of averages of many independently drawn samples is governed by the *theory of large deviations*,[31] which predicts that, for a fixed value of $\bar{\sigma}$,

$$p_\tau^S(\bar{\sigma}) = p_{Kt_c}^S(\bar{\sigma}) \rightarrow \sqrt{\frac{KI_0''}{\pi}}\exp[-KI(\bar{\sigma})] \tag{46}$$

as τ, $K \rightarrow \infty$ (t_c fixed). Here, $I(x)$ is the *rate function* (or *entropy function*) associated with the stochastic variable $\underline{x}$; I_0'' is the second derivative of $I(x)$, evaluated at the minimum of that function, $x_{\mathrm{min}} = \bar{\sigma}^S$; and the normalization factor is obtained by steepest descent. [The rate function is defined up to an additive constant. For convenience we have set the value of I to zero at x_{min}. Equation (46) implies that, right around that minimum, $p_\tau^S(\bar{\sigma})$ tends toward a Gaussian of variance $(2KI_0'')^{-1}$. This is just the central limit theorem.] We can write down a similar result for $p_{\tau'}^S(\bar{\sigma}')$, replacing K by $K-2$ in Eq. (46), and $\bar{\sigma}$ by $\bar{\sigma}'$. Then taking the ratio of the two functions and considering the limit $\tau \rightarrow \infty$ (equivalently, $K \rightarrow \infty$) gives

$$\lim_{\tau \rightarrow \infty} \frac{p_{\tau'}^S(\bar{\sigma}')}{p_\tau^S(\bar{\sigma})} = \exp[2I(\bar{\sigma}) + (c - 2\bar{\sigma})\, I'(\bar{\sigma})] \tag{47}$$

where $I'(x) \equiv dI(x)/dx$, and $\bar{\sigma}'$ is given by Eq. (45).

Combining Eqs. (44) and (47), we finally get

$$\begin{aligned}\lim_{\tau \rightarrow \infty} \frac{p_\tau(\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)}{p_\tau^S(\bar{\sigma})} &= \int dc\, \eta(c) \exp[2I(\bar{\sigma}) + (c - 2\bar{\sigma})\, I'(\bar{\sigma})] \\ &\equiv R(\bar{\sigma}, \mathbf{z}_A, \mathbf{z}_B) \end{aligned} \tag{48}$$

where the dependence of R on $\mathbf{z}_A$ and $\mathbf{z}_B$ enters through the implicit dependence of $\underline{a}$ and $\underline{b}$ (hence, $\underline{c}$) on those microstates of ψ. We see that, indeed, the ratio of the two distributions does not generally converge to unity. However, it *does* converge (by the arguments just presented, and assuming the integral in Eq. (48) converges!) to a function which does not depend on τ. Therefore we get, for the second term on the left side of Eq. (38),

$$\frac{1}{\tau} \ln \frac{p_\tau(+\bar{\sigma} \mid \mathbf{z}_A, \mathbf{z}_B)}{p_\tau(-\bar{\sigma} \mid \mathbf{z}_B^*, \mathbf{z}_A^*)} \rightarrow \frac{1}{\tau} \ln \frac{p_\tau^S(+\bar{\sigma})}{p_\tau^S(-\bar{\sigma})} + \frac{1}{\tau} \ln \frac{R(+\bar{\sigma}, \mathbf{z}_A, \mathbf{z}_B)}{R(-\bar{\sigma}, \mathbf{z}_B^*, \mathbf{z}_A^*)} \tag{49}$$

as $\tau \to \infty$. Combining Eqs. (38) and (49) and dropping the terms which vanish in that limit, we finally get

$$\lim_{\tau \to \infty} \frac{1}{\tau} \ln \frac{p_\tau^S(+\bar{\sigma})}{p_\tau^S(-\bar{\sigma})} = +\bar{\sigma}/k_B \tag{50}$$

which is the steady-state fluctuation theorem (Eq. (1)).

While the arguments presented in this section lack mathematical rigor, they might point the way toward a proper derivation of Eq. (1) from Eq. (4). An interesting question is: what additional assumptions are required in order for the detailed fluctuation theorem to rigorously imply the steady-state fluctuation theorem? Clearly a carefully crafted assumption about the existence of a steady state in the limit of infinitely large reservoirs ($\omega \to \infty$) is a *sine qua non*. If we further assume exponential decay of the autocorrelation function of the instantaneous entropy generation rate, then perhaps the arguments advanced above could provide the backbone of a rigorous theorem.

A somewhat different approach has been taken by Eckmann, Pillet, and Rey-Bellet,[30] in their study of a chain of anharmonic oscillators coupled to two infinite heat reservoirs. They are able to project out the reservoir variables, and thus reduce the evolution of the oscillator chain (supplemented by a set of auxiliary variables) to a Markov diffusion process. By studying the generator of this diffusion process, they argue that their model exhibits entropy production in accordance with the steady-state fluctuation theorem.

V. RELATION TO FAR-FROM-EQUILIBRIUM FREE ENERGY RESULTS

Independently of the fluctuation theorem, another far-from-equilibrium result has been derived and generalized in recent years.[21–27] Consider a system initially in thermal equilbrium with a heat reservoir at temperature T. Now imagine externally changing a work parameter from an initial value (say, $\lambda = 0$) to a final value ($\lambda = 1$) over a finite time, while keeping the system in contact with the reservoir. Once the final value of the work parameter has been reached, hold the work parameter fixed and let the system and reservoir re-equilibrate. The system thus begins and ends in equilibrium states, corresponding to $\lambda = 0$ and $\lambda = 1$, but at intermediate times is driven out of equilibrium by the finite-rate variation of the work parameter. (The assumption that the system ends in equilibrium is not necessary, but makes for a more pleasant presentation.) Now imagine repeating this process infinitely many times, always following the same protocol for varying λ, and carefully measuring the external *work* W performed on the system during each realization. Then the distribution of values of W obtained from this ensemble of realizations obeys the following equality, *regardless of how gently or violently the parameter was switched* from 0 to 1:

$$\langle \exp(-\beta W) \rangle = \exp(-\beta \Delta F), \qquad \beta \equiv 1/k_B T \tag{51}$$

where $\langle \cdots \rangle$ denotes an average over the ensemble of realizations,

$$\Delta F = -\beta^{-1} \ln(Z_1/Z_0) \tag{52}$$

is the free energy difference between the initial and final equilibrium states of the system, and the Z's are the associated partition functions:

$$Z_\lambda = \int d\mathbf{z} \exp[-\beta H_\lambda^\psi(\mathbf{z})] \tag{53}$$

Equation (51) was originally derived using a Hamiltonian formulation,(21) but has also been shown to be valid under explicitly non-Hamiltonian evolution—including the Nosé–Hoover thermostating scheme,(21, 22) Markov-chain dynamics,(22–24) and Langevin evolution(22, 25)—and has been generalized to a wider class of thermodynamic processes.(18, 23, 26, 27)

[Equation (51) can be viewed as an extension, to irreversible processes, of the relation $W = \Delta F$, which holds for a *reversible*, isothermal process from one equilibrium state to another. Furthermore, it immediately implies the inequality $\langle W \rangle \geqslant \Delta F$, in agreement with the second law of thermodynamics, and places an exponentially decaying upper bound on the probability of observing finite-size violations of the second law:(26)

$$\text{Prob}(W \leqslant \Delta F - X) \leqslant e^{-X/k_B T} \tag{54}$$]

Let us now derive Eq. (51) from the detailed fluctuation theorem obtained in the present paper, following a line of reasoning similar to that presented by Crooks(23) for the case of Markov evolution.

Let Π^+ be a process involving a system of interest, ψ, a single reservoir, θ (prepared at temperature T), and a work parameter λ which is varied from $\lambda(0) = 0$ to $\lambda(\tau) = 1$. During a single realization of this process, the *work* W performed on ψ is given by

$$W = \Delta E - Q \tag{55}$$

where ΔE is the net change in the internal energy of ψ, and Q (Eq. (8)) is the heat absorbed by ψ from the reservoir. Following the notation of previous sections, let $\mathbf{z}_A$ and $\mathbf{z}_B$ denote the initial and final microstates of ψ, and ΔS the entropy generated, for a given realization. Since $Q = -T \cdot \Delta S$ (Eq. (2)), the value of W can be expressed as a function of $\mathbf{z}_A$, $\mathbf{z}_B$, and ΔS:

$$W(\mathbf{z}_A, \mathbf{z}_B, \Delta S) = H^{\psi}_{\lambda=1}(\mathbf{z}_B) - H^{\psi}_{\lambda=0}(\mathbf{z}_A) + T \cdot \Delta S \tag{56}$$

Assuming a canonical distribution of initial conditions, we can construct the ensemble average of $\exp(-\beta W)$ as follows:

$$\begin{aligned}\langle \exp(-\beta W) \rangle = \int d\mathbf{z}_A \frac{1}{Z_0} \exp[-\beta H^{\psi}_{\lambda=0}(\mathbf{z}_A)] \int d\mathbf{z}_B \\ \times \int d(\Delta S)\, P_+(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \exp(-\beta W)\end{aligned} \tag{57}$$

Invoking Eqs. (4) and (56) allows us to rewrite this as:

$$\begin{aligned}\langle \exp(-\beta W) \rangle = \frac{1}{Z_0} \int d\mathbf{z}_A \exp[-\beta H^{\psi}_{\lambda=1}(\mathbf{z}_B)] \int d\mathbf{z}_B \\ \times \int d(\Delta S)\, P_-(\mathbf{z}_A^*, -\Delta S \mid \mathbf{z}_B^*)\end{aligned} \tag{58}$$

$$= \frac{1}{Z_0} \int d\mathbf{z}_B \exp[-\beta H^{\psi}_{\lambda=1}(\mathbf{z}_B)] \tag{59}$$

$$= \frac{Z_1}{Z_0} = \exp(-\beta \Delta F) \tag{60}$$

as promised.

Note that Eq. (8) of ref. 28 can be viewed as a special case of Eq. (51) above, for the situation in which the initial and final Hamiltonian functions are the same: $H_{\lambda=0}(\mathbf{z}) = H_{\lambda=1}(\mathbf{z})$, hence the free energy difference is identically zero: $\Delta F = 0$.

VI. DISCUSSION

The motivation for this paper has been a belief that one ought to be able to derive the fluctuation theorem—or something like it—within the framework of traditional statistical mechanics; that is, by contemplating a system of interest interacting with a thermal environment. The central result of this exercise, Eq. (4), is a detailed fluctuation theorem, valid for finite times and made without reference to a steady state. Given the definitions and assumptions made in this paper, Eq. (4) is identically true. We now briefly address a number of issues related to this result.

The first involves the definition of ΔS. Specifying what is meant by the entropy generated during a single realization of a given process is inherently problematic, since entropy, in statistical mechanics, is ordinarily associated with an *ensemble* of microstates. Equation (2) must therefore be regarded as constituting a particular *choice* of definition of ΔS. This choice, however, is not entirely arbitrary, but rather (as indicated in Section II) guided by macroscopic thermodynamics, where the entropy increase of a thermal environment is identified with the heat, per unit temperature, absorbed by that environment.

There is another apparently troubling feature of the definition of entropy generated: to compute ΔS, it seems we must know the exact initial and final microstates of each reservoir (Eq. (8)). This is extremely unsatisfying, as it conflicts with the usual notion of a thermal environment as a huge collection of unmonitored (and uninteresting) degrees of freedom "out there." However, one can infer the Q_n values by monitoring the evolution of only those environmental degrees of freedom which are at any moment interacting (exchanging energy) with the system of interest. Therefore if, due to short-range interaction forces, the exchange of energy between ψ and the θ's occurs *locally*—say, at an interface—then ΔS can be computed by knowing only what goes on in the immediate vicinity of the system of interest (e.g., within boundary layers of width r in Fig. 1), *without* explicit knowledge of initial and final reservoir energies. Thus, while Eq. (8) is indisputably useful as a device in the derivation of the detailed fluctuation theorem, that final result is really a statement about ψ and the heat fluxes into and out of ψ, rather than one about ψ and initial and final reservoir energies.

Finally, there is a bit of arbitrariness even in the definition of the heat absorbed by a given reservoir (Eq. (8)), owing to the small but finite interaction energy h_{int}. As with Eq. (2), this definition represents a particular choice, but again this choice is consistent with macroscopic thermodynamics.

The derivation presented in Section III explicitly assumes that the reservoir degrees of freedom are initially sampled canonically (Eq. (9)). The result itself, however, might not depend as strongly on this assumption as the derivation suggests. For macroscopically large reservoirs, Eq. (4) ought to remain valid, at least to an excellent approximation, if the initial reservoir conditions are sampled from *micro*canonical, rather than canonical, distributions. The argument for this is similar to the usual one for the equivalence of microcanonical and canonical averages in the thermodynamic limit, and goes roughly as follows. Imagine constructing a joint, conditional probability distribution

$$\tilde{P}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \tag{61}$$

defined in the same way as $P(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$, but with reservoir initial conditions sampled from microcanonical distributions. In this case the temperatures T_n appearing in the definition of ΔS are the "microcanonical temperatures" of the heat reservoirs:

$$(k_B T)^{-1} = \frac{\partial}{\partial E} \ln \int d\mathbf{y}\, \delta[E - H^\theta(\mathbf{y})] \tag{62}$$

$\tilde{P}$ depends on the set of initial reservoir energies, just as P depends on the initial reservoir temperatures. Explicitly, we can write

$$\tilde{P} = \tilde{P}^{\varepsilon_1 \cdots \varepsilon_N}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A), \qquad P = P^{T_1 \cdots T_N}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \tag{63}$$

where ε_n denotes the known initial energy *per particle* (alternatively, per degree of freedom) of the nth reservoir: $\varepsilon_n = E_n / v_n$, where E_n is the initial energy of, and v_n the number of particles constituting, the nth reservoir. Since the v_n's are fixed, the initial microcanonical ensemble of each reservoir is uniquely specified by the value of ε_n. Now, $P^{T_1 \cdots T_N}$ can be expressed as a weighted average of $\tilde{P}^{\varepsilon_1 \cdots \varepsilon_N}$:

$$P^{T_1 \cdots T_N}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) = \left[\prod_{n=1}^{N} \int d\varepsilon_n\, w_n(\varepsilon_n; T_n) \right] \tilde{P}^{\varepsilon_1 \cdots \varepsilon_N}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \tag{64}$$

where $w_n(\varepsilon_n; T_n)$ is the statistical weight of the microcanonical ensemble at energy-per-particle ε_n, within the canonical ensemble at temperature T_n, for the nth reservoir.[3] In the thermodynamic limit of arbitrarily large reservoirs ($v_n \to \infty$, with extensive and intensive quantities scaled appropriately), $w_n(\varepsilon_n; T_n)$ becomes peaked arbitrarily sharply around the value $\varepsilon_n(T_n)$ whose corresponding "microcanonical temperature" is equal to T_n. Assuming as in Section IV that $\tilde{P}$ converges to a well-defined function of $\mathbf{z}_A$, $\mathbf{z}_B$, and ΔS, we will therefore get, in the thermodynamic limit,

$$P^{T_1 \cdots T_N}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) = \tilde{P}^{\varepsilon_1(T_1) \cdots \varepsilon_N(T_N)}(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A) \tag{65}$$

resulting in a microcanonical detailed fluctuation theorem.

[3] Formally, $w(\varepsilon; T) \propto e^{-v\varepsilon/k_B T} \int d\mathbf{y}\, \delta[v\varepsilon - H^\theta(\mathbf{y})]$, with the subscript n suppressed. The normalization is chosen so that $\int d\varepsilon\, w = 1$.

More generally, we expect Eq. (4) to remain valid, provided that each reservoir is prepared in what would be characterized at the macroscopic level as a state of thermodynamic equilibrium (regardless of whether the method of preparation truly yields a canonical distribution of microstates when carried out repeatedly). This reflects a strong prejudice that the canonical ensemble should be viewed primarily as a computational convenience, and ought not to be taken too seriously as characterizing the "true" statistical distribution of microstates of entire macroscopic bodies. Of course, all this presupposes macroscopically large heat reservoirs; for microscopic "reservoirs," the detailed fluctuation theorem still holds, but its validity then depends on a literal interpretation of Eq. (9).

Another issue involves the assumption of time-reversal invariance, Eq. (5). While this assumption was made for convenience, it is straightforward to generalize the analysis to the situation in which (possibly time-dependent) magnetic fields are present. In that case, given a process Π^+, its time-reversed counterpart Π^- is obtained by carrying out the protocol in reverse order, while also reversing all magnetic fields: $\mathbf{B}(t) \rightarrow -\mathbf{B}(\tau - t)$. With this modification, Eq. (4) remains valid (see also refs. 18, 28, and 29).

The derivation of the detailed fluctuation theorem presented in this paper is admittedly not as "clean" as in previous works,[(1–19)] where the (deterministic or stochastic) thermostating is accomplished without the explicit introduction of reservoir degrees of freedom. For instance, the present treatment forces us to consider the unphysical limit of infinite reservoirs; caveats need to be made about ignoring interaction energies; the time $t = 0$ is privileged (since that is when the reservoir microstates are sampled canonically); and so forth. There is some consolation, however, in knowing that these messy issues are likely to arise in any laboratory setting, where real thermal environments and finite times of observation are an inherent part of the game.

Nosé–Hoover-type thermostating schemes[(32)]—employing one or a handful of "reservoir" degrees of freedom (and non-Hamiltonian equations of motion)—might offer an interesting middle ground between the deterministic thermostats of refs. 1–14 and the approach taken here. In the original Nosé–Hoover scheme, the reservoir degree of freedom (ζ) is initially sampled from a Gaussian distribution. Therefore, given an appropriate definition of ΔS, it would be straightforward to define a joint, conditional probability distribution $P(\mathbf{z}_B, \Delta S \mid \mathbf{z}_A)$, as in this paper, basically replacing the canonical distribution of initial reservoir conditions in Eq. (11) by the Gaussian distribution of initial values of ζ. It would be interesting to see whether a detailed fluctuation theorem then follows. If so, this approach might lead (by arguments along the lines of Section IV) to a Nosé–Hoover steady-state fluctuation theorem. Presumably some chaoticity assumption would still be required to make this result rigorous, but at least there would be no need to worry about infinitely large reservoirs!

Finally, it would be very nice to come up with a laboratory experiment to test the main result derived in this paper. The system of interest in such an experiment would doubtless have to have very few degrees of freedom (ideally, only one), in order to collect data with sufficiently good statistics in a reasonable amount of time. Furthermore, one would want a process during which the typical entropy generated is not much greater than k_B; otherwise, prohibitively many realizations would be needed before observing a single one for which $\Delta S < 0$. In this respect, the fact that Eq. (4) is valid for *finite* durations τ is helpful.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge and thank F. Bonetto, G. E. Crooks, J. R. Dorfman, R. Mainieri, Y. Oono, J. Percus, L. Rey-Bellet, and D. J. Searles for stimulating conversations and correspondence. This research is supported by the Department of Energy, under contract W-7405-ENG-36.

REFERENCES

1. D. J. Evans, E. G. D. Cohen, and G. P. Morriss, *Phys. Rev. Lett.* **71**:2401 (1993).
2. D. J. Evans and D. J. Searles, *Phys. Rev. E* **50**:1645 (1994); *Phys. Rev. E* **52**:5839 (1995); *Phys. Rev. E* **53**:5808 (1996).
3. G. Gallavotti and E. G. D. Cohen, *J. Stat. Phys.* **80**:931 (1995); *Phys. Rev. Lett.* **74**:2694 (1995).
4. G. Gallavotti, *Phys. Rev. Lett.* **77**:4334 (1996).
5. E. G. D. Cohen, *Physica A* **240**:43 (1997).
6. F. Bonetto, G. Gallavotti, and P. L. Garrido, *Physica D* **105**:226 (1997).
7. G. Gallavotti, *Chaos* **8**:384 (1998).
8. G. Gallavotti, *Physica A* **263**:39 (1999).
9. F. Bonetto, N. I. Chernov, and J. L. Lebowitz, *Chaos* **8**:823 (1998).
10. D. Ruelle, *J. Stat. Phys.* **95**:393 (1999).
11. G. Ayton, D. J. Evans, and D. J. Searles, cond-mat/9901256.
12. D. J. Searles and D. J. Evans, cond-mat/9902021.
13. D. J. Searles and D. J. Evans, cond-mat/9906002.
14. E. G. D. Cohen and G. Gallavotti, *J. Stat. Phys.* **96**:1343 (1999).
15. J. Kurchan, *J. Phys. A* **31**:3719 (1998).
16. J. L. Lebowitz and H. Spohn, *J. Stat. Phys.* **95**:333 (1999).
17. D. J. Searles and D. J. Evans, *Phys. Rev. E* **60**:159 (1999).
18. G. E. Crooks, *Phys. Rev. E* **60**:2721 (1999); also "Path Ensemble Averages in Systems Driven Far From Equilibrium," unpublished preprint.
19. C. Maes, *J. Stat. Phys.* **95**:367 (1999).

20. C. Maes, F. Redig, and A. Van Moffaert, mp-arc/99-209.
21. C. Jarzynski, *Phys. Rev. Lett.* **78**:2690 (1997).
22. C. Jarzynski, *Phys. Rev. E* **56**:5018 (1997).
23. G. E. Crooks, *J. Stat. Phys.* **90**:1481 (1998).
24. R. M. Neal, "Annealed Importance Sampling," Technical Report No. 9805 (revised), Dept. of Statistics, Toronto (1998).
25. C. Jarzynski, *Acta Phys. Pol. B* **29**:1609 (1998).
26. C. Jarzynski, *J. Stat. Phys.* **96**:415 (1999).
27. T. Hatano, *Phys. Rev. E* **60**:R5017 (1999).
28. G. N. Bochkov and Yu. E. Kuzovlev, *Zh. Eksp. Teor. Fiz.* **72**:238 (1977) [*Sov. Phys.—JETP* **45**:125 (1977)].
29. G. N. Bochkov and Yu. E. Kuzovlev, *Physica A* **106**:443 (1981), *Physica A* **106**:480 (1981).
30. J.-P. Eckmann, C.-A. Pillet, and L. Rey-Bellet, *J. Stat. Phys.* **95**:305 (1999).
31. Y. Oono, *Prog. Theor. Phys. Suppl.* **99**:165 (1989), and references therein.
32. S. Nosé, *J. Chem. Phys.* **81**:511 (1984); W. G. Hoover, *Phys. Rev. A* **31**:1695 (1985).