Journal of Statistical Physics, Vol. 96, Nos. 1/2, 1999

Microscopic Analysis of Clausius–Duhem Processes

C. Jarzynski[1]

Received November 16, 1998; final March 23, 1999

Given a thermodynamic process which carries a system from one equilibrium state to another, we construct a quantity whose average, over an ensemble of microscopic realizations of the process, depends only on these end states, even if at intermediate times the system is *out* of equilibrium. This result (1) can be used to express the entropy difference between two equilibrium states in terms of an *irreversible* process connecting them, (2) leads to two statistical statements of the Clausius–Duhem inequality, and (3) can be generalized to situations in which the system begins and/or ends in nonequilibrium states.

KEY WORDS: Irreversible processes.

The Clausius–Duhem inequality of classical thermodynamics—a statement of the Second Law—applies to thermodynamic processes during which a system evolves from one equilibrium state (A) to another (B). It asserts that the integrated heat absorbed by the system, divided by the temperature at which that heat is absorbed, is bounded from above by the net change in the entropy of the system:

$$\int_A^B \frac{dQ}{T} \leqslant \Delta S \equiv S^B - S^A \tag{1}$$

By "thermodynamic process", we have in mind a situation in which the system is brought into thermal contact with a sequence of heat reservoirs at different temperatures, one at a time, while one or more external parameters of the system are varied with time (see Fig. 1); the denominator in Eq. (1) denotes the temperature of the reservoir from which the system absorbs a quantity of heat dQ. In general, the process is irreversible: it is

[1] Theoretical Division, T-6 and T-13, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545; e-mail: chrisj@lanl.gov.

0022-4715/99/0700-0415$16.00/0

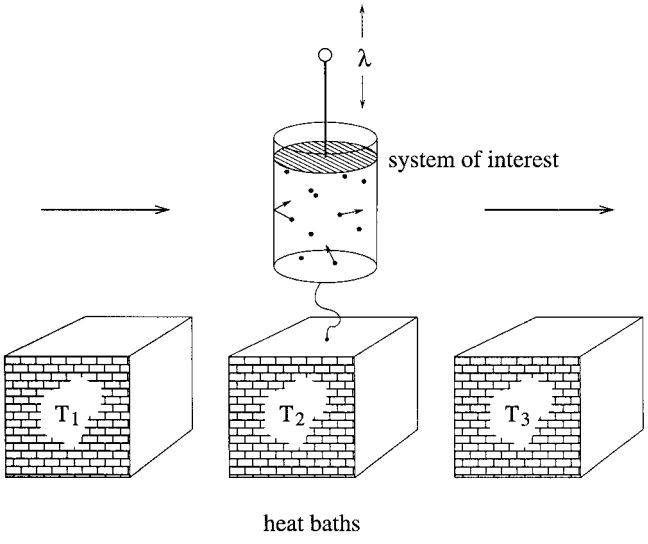

Fig. 1. A schematic representation of the sort of thermodynamic process considered in this paper. The system of interest here is a gas inside a container, closed off at one end by a movable piston. The position of the piston is our externally controlled parameter, λ. The three "heat baths" are simply objects with heat capacities much greater than that of the gas. The system of interest is brought into thermal contact with these baths, one at a time, e.g., as depicted by the filament connecting the bath at temperature T_2 to the container of gas. At the same time, λ is varied externally.

carried out over a finite time with a finite number of reservoirs, and the system evolves, from A to B, through a sequence of *non-equilibrium* intermediate states.

The aim of the present paper is a classical, microscopic analysis of such "Clausius–Duhem" processes, explicitly accounting for all degrees of freedom involved. This analysis will follow a statistical approach: we will consider an *ensemble* of microscopic realizations of the thermodynamic process. Each realization (described by a trajectory specifying the evolution of all the degrees of freedom which make up the system of interest and reservoirs) represents a possible "microscopic history," consistent with the macroscopic prescription for carrying out the thermodynamic process.

A statistical ensemble of realizations implies *fluctuations*—from one realization to another—of various quantities of physical interest. In taking a microscopic approach, therefore, we will be interested in the statistical *distribution* (over the ensemble of microscopic histories) of values of quantities such as $\int_A^B dQ/T$. Working with the assumption that the system of interest begins and ends in equilibrium states ($A \to B$), we will construct a

quantity whose average, over the ensemble of realizations, *depends only on those two states*, and not on the intermediate evolution of the system. This result, Eq. (9) below—the central result of this paper—is valid even if the system is driven far from equilibrium at intermediate times. We will argue that this result can be viewed as the generalization—to irreversible processes of the well-known identity which relates ΔS to an arbitrary *reversible* process from A to B (Eq. (10)). As we will show, the central result of this paper allows one to express ΔS in terms of an arbitrary *irreversible* process from A to B (Eq. (12)). We will also show that Eq. (9) immediately implies two inequalities which are closely related to the Clausius–Duhem inequality. In particular, Eq. (16) places a tight upper bound on the frequency with which finite-size violations of the Clausius–Duhem inequality do occasionally occur. We will finally generalize these results to processes in which the system of interest begins and/or ends in nonequilibrium statistical states.

In the spirit (and level of rigor) of Gibbs,[1] we will work with the following quite general picture. The "system of interest" is a closed, finite system with a single external parameter, λ. Additionally, we assume the presence of N other closed, finite systems, which will play the role of heat reservoirs. We will refer to these as "baths," and assume that they have been prepared at temperatures T_1, T_2,..., T_N. Our thermodynamic process then consists of a sequence of steps during which the system of interest is placed in thermal contact with the baths, one at a time, while the value of λ is varied along a pre-determined path $\lambda(t)$, over a time interval $0 \leqslant t \leqslant \tau$. Let $\lambda^A \equiv \lambda(0)$ and $\lambda^B \equiv \lambda(\tau)$ denote the initial and final parameter values, and let $n(t)$ identify the bath with which the system is in contact at time t. The functions $\lambda(t)$ and $n(t)$ embody the macroscopic instructions (the "protocol") which specify the thermodynamic process.

For any parameter value λ and temperature T, there exists an *equilibrium state* (λ, T) of the system of interest, described at the microscopic level by a canonical distribution in the phase space of the system. Through most of the paper (up to Eq. (24)), we will restrict our attention to processes for which the system begins $(t=0)$ and ends $(t=\tau)$ in such equilibrium states, denoted by $A \equiv (\lambda^A, T^A)$ and $B \equiv (\lambda^B, T^B)$, respectively. In other words, we assume that the system of interest is prepared in equilibrium: over the ensemble of realizations, the microscopic initial conditions of the system are distributed canonically; and that, at the end of the process, the system is once again described (statistically) by an equilibrium distribution: the microscopic final conditions of the system of interest are distributed canonically. Strictly speaking, these assumptions are mathematically suspect: there exists (to my knowledge) no rigorous proof that a canonical ensemble can be achieved in a finite time, with finite resources.

Thus, one cannot say with certainty that there exist physically realizable methods for preparing a system in a canonical ensemble; or that, given a system initially in equilibrium, there exist processes which carry that system to a different equilibrium state in a finite time. However, it is a widely held prejudice that such processes do exist in Nature, and that the canonical ensemble is the appropriate statistical representation for a closed system in thermal equilibrium. We will therefore adopt the point of view that it is a legitimate exercise to assume initial and final equilibrium, and to explore the consequences of these (seemingly reasonable) assumptions, without concerning ourselves here with the separate problem of establishing the validity of these assumptions from first principles. Later, as mentioned, we will generalize to nonequilibrium initial and final states.

We assume Hamiltonian evolution at the microscopic level. Let $\mathbf{z}$ denote a point in the phase space of the system of interest—specifying, e.g., the positions and momenta of all its constituent particles—and let $\mathbf{z}_n$ denote a point in the phase space of the nth bath. Let $\mathbf{y} = (\mathbf{z}, \mathbf{z}_1, ..., \mathbf{z}_N)$ then specify the instantaneous state of all degrees of freedom involved. Evolution in the full phase space ($\mathbf{y}$-space) is governed by a time-dependent Hamiltonian

$$\mathscr{H}(\mathbf{y}, t) = H_{\lambda(t)}(\mathbf{z}) + \sum_{n=1}^{N} H_n^b(\mathbf{z}_n) + \sum_{n=1}^{N} \delta_{n, n(t)} h_n^{\text{int}}(\mathbf{z}, \mathbf{z}_n) \tag{2}$$

Here, $H_\lambda(\mathbf{z})$ is a Hamiltonian describing the system of interest, for a parameter value λ; H_n^b is the Hamiltonian for the nth bath; and h_n^{int} couples the system of interest to that bath.[2] $\mathscr{H}(\mathbf{y}, t)$ is (essentially) the most general classical Hamiltonian describing a system of interest, with a specified external-parametric time dependence $\lambda(t)$, coupled to N other systems, one at a time. Once the initial conditions are fully specified, Hamilton's equations uniquely determine the evolution of all degrees of freedom.

Let us consider a single realization of the thermodynamic process, described by a trajectory $\mathbf{y}(t)$ evolving under Hamilton's equations from some initial conditions $\mathbf{y}^0 \equiv \mathbf{y}(0)$, and ending at $\mathbf{y}^\tau \equiv \mathbf{y}(\tau)$. Let E^0 and E^τ denote, respectively, the initial and final internal energies of the system of interest, and let ΔE_n denote the net change in the internal energy of the nth bath:

$$E^0 = H_{\lambda^A}(\mathbf{z}^0), \qquad E^\tau = H_{\lambda^B}(\mathbf{z}^\tau), \qquad \Delta E_n = H_n^b(\mathbf{z}_n^\tau) - H_n^b(\mathbf{z}_n^0) \tag{3}$$

[2] As usual, we take the interaction energies h_n^{int} to be negligible in comparison with the other terms in $\mathscr{H}$. While this assumption does not enter explicit calculations, it allows us to view the total energy of the system of interest and N baths as simply the sum of the internal energies of these $N+1$ objects. The terms appearing in the First Law, $dE = dW + dQ$, are then unambiguously defined.

Now define

$$\Sigma \equiv \frac{E^\tau}{T^B} - \frac{E^0}{T^A} + \sum_{n=1}^{N} \frac{\Delta E_n}{T_n} \tag{4}$$

where T^A and T^B are the temperatures corresponding to the initial and final statistical states of the system of interest, and the T_n's are the temperatures at which the baths are prepared. Working with units in which Boltzmann's constant $k_B = 1$, let us now compute $\langle \exp(-\Sigma) \rangle$, where *angular brackets signify an average over the statistical ensemble of realizations*, i.e., ever the ensemble of trajectories $\mathbf{y}(t)$. For a given realization, Σ happens to depend only on the initial and final points of $\mathbf{y}(t)$: $\Sigma = \Sigma(\mathbf{y}^0, \mathbf{y}^\tau)$. Since evolution in $\mathbf{y}$-space is deterministic, we can formally express the final conditions as a function of the initial conditions, $\mathbf{y}^\tau = \mathbf{y}^\tau(\mathbf{y}^0)$, and then compute the desired average over realizations, by integrating over the distribution of initial conditions, $f(\mathbf{y}^0)$:

$$\langle \exp(-\Sigma) \rangle = \int d\mathbf{y}^0 \, f(\mathbf{y}^0) \exp[-\Sigma(\mathbf{y}^0, \mathbf{y}^\tau(\mathbf{y}^0))] \tag{5}$$

By our assumptions regarding initial conditions, $f(\mathbf{y}^0)$ is a product of canonical distributions, hence we have

$$\begin{aligned} \langle \exp(-\Sigma) \rangle &= \frac{1}{\mathcal{N} Z^A} \int d\mathbf{y}^0 \exp\left[-\frac{E^0}{T^A} - \sum_{n=1}^{N} \frac{H_n^b(\mathbf{z}_n^0)}{T_n} \right] \exp(-\Sigma) \\ &= \frac{1}{\mathcal{N} Z^A} \int d\mathbf{y}^\tau \exp\left[-\frac{E^\tau}{T^B} - \sum_{n=1}^{N} \frac{H_n^b(\mathbf{z}_n^\tau)}{T_n} \right] \\ &= \frac{Z^B}{Z^A} = \exp\left(-\frac{F^B}{T^B} + \frac{F^A}{T^A} \right) \end{aligned} \tag{6}$$

Here, Z^A and Z^B denote partition functions associated with equilibrium states of the system of interest, and F^A and F^B denote free energies:

$$F^i = -T^i \ln Z^i = -T^i \ln \int d\mathbf{z} \exp[-H_{\lambda^i}(\mathbf{z})/T^i], \qquad i = A, B \tag{7}$$

$\mathcal{N}$ is the product of the partition functions for the N baths, each corresponding to the temperature at which that bath was prepared: $\mathcal{N} = \prod_{n=1}^{N} \int d\mathbf{z}_n \exp[-H_n^b(\mathbf{z}_n)/T_n]$. In going from the first to the second line in

Eq. (6), we have: (1) used Eq. (4) to rewrite the integrand as an explicit function of y^τ alone, and (2) changed the variables of integration from $\mathbf{y}^0$ to $\boldsymbol{y}^\tau$. By Liouville's theorem, the Jacobian $|\partial \mathbf{y}^\tau / \partial \mathbf{y}^0| = 1$.

At this point, Eq. (6) is just a statement about energy exchange among a number of finite, closed systems, one of which has been singled out as being "of interest." Its validity does not depend on the relative sizes of these objects. To establish contact with more familiar results, let us now imagine the limiting case in which the heat capacities of the N baths become arbitrarily greater than that of the system of interest. The baths then assume the role of "infinite" heat reservoirs, and we can rewrite the term $\sum_n \Delta E_n / T_n$ in Eq. (4) as $-\int_A^B dQ/T$:

$$\Sigma = \frac{E^\tau}{T^B} - \frac{E^0}{T^A} - \int_A^B \frac{dQ}{T} \tag{8}$$

(During the time interval over which the system of interest is coupled to the n'th bath, the heat absorbed by the system is equal to the energy lost by that bath, so the value of $\int dQ/T$ over that interval of time is simply $-\Delta E_n / T_n$.) Eq. (6) now becomes:

$$\left\langle \exp\left[-\Delta\left(\frac{E}{T}\right) + \int_A^B \frac{dQ}{T} \right] \right\rangle = \exp\left[-\Delta\left(\frac{F}{T}\right) \right] \tag{9}$$

using the shorthand notation $\Delta(E/T) \equiv E^\tau/T^B - E^0/T^A$ and $\Delta(F/T) \equiv F^B/T^B - F^A/T^A$. This is the central result of this paper. We add here that a result equivalent to Eq. (9) has been derived independently by Gavin E. Crooks[2]—using stochastic, Markovian dynamics to model the evolution of the system of interest—and that this result has been shown to be closely related to the Fluctuation Theorem for non-equilibrium steady states.[3, 4, 7]

Since internal energies are quantities associated with specific microscopic states (i.e., points in phase space), the exact values of E^0, E^τ and $\int_A^B dQ/T (= -\sum_n \Delta E_n / T_n)$ differ from one realization of the thermodynamic process to the next. By contrast, the free energies F^A and F^B are associated with *canonical ensembles* of microstates of the system of interest. Thus, $\Sigma = \Delta(E/T) - \int_A^B dQ/T$ is a linear combination of quantities (E^0, E^τ, the ΔE_n's) which vary in value from one realization to the next, and Eq. (9) makes an assertion regarding the statistical distribution of values of Σ: it claims that the average of $\exp(-\Sigma)$, over the ensemble of realizations, is equal to $\exp[-\Delta(F/T)]$, which depends only on the equilibrium states A and B, and not on the sequence of (nonequilibrium) statistical states

through which the system evolves in getting from A to B! Now, for macroscopic, *reversible* processes, it is well known that[3]

$$\int_A^B \frac{d\hat{Q}}{T} = \Delta\hat{S} \equiv \hat{S}^B - \hat{S}^A \qquad \text{(REVERSIBLE)}, \tag{10}$$

regardless of the path (through equilibrium state space) taken from A to B. Since Eq. (10) can be written, at the macroscopic level, as $\hat{\Sigma} = \Delta(\hat{F}/T)$ (because $\hat{F} = \hat{E} - \hat{S}T$), our central result $[\langle e^{-\Sigma}\rangle = e^{-\Delta(F/T)}]$ may be viewed as the microscopic extension of Eq. (10) to *irreversible* processes. Moreover, and somewhat surprisingly, this result is valid *regardless of how far the system is driven away from equilibrium* between the initial and final times: no matter how violent the process, Eq. (9) will hold, provided the system begins in A and ends in B.

For a given equilibrium state, the microscopic expressions for free energy, average internal energy, and entropy satisfy $F = \bar{E} - ST$, where the overbar denotes an equilibrium (canonical) average. Thus, the quantity which we have called $\Delta(F/T)$ can be written as:

$$\Delta\left(\frac{F}{T}\right) = \frac{\bar{E}^B}{T^B} - \frac{\bar{E}^A}{T^A} - \Delta S = \frac{\langle E^\tau\rangle}{T^B} - \frac{\langle E^0\rangle}{T^A} - \Delta S \tag{11}$$

where the second equality follows from our assumptions regarding the initial and final distributions of microstates. Combining this with Eq. (9) gives:

$$\Delta S = \frac{\langle E^\tau\rangle}{T^B} - \frac{\langle E^0\rangle}{T^A} + \ln\left\langle \exp\left[-\Delta\left(\frac{E}{T}\right) + \int_A^B \frac{dQ}{T}\right]\right\rangle \tag{12}$$

This result expresses the entropy difference $\Delta S = S^B - S^A$ in terms of an arbitrary—in general *irreversible*—thermodynamic process from A to B. In principle, by repeatedly measuring E^0, E^τ, and $\int_A^B dQ/T$ for independent realizations of such a process, we can construct the averages appearing in Eq. (12), and therefore compute the value of ΔS.[4]

[3] Since quantities such as heat, entropy, etc., appear in both thermodynamics and statistical mechanics, and since these are (formally) separate theories, it is useful to distinguish between the two. Here and below, we use a carat (e.g. $d\hat{Q}$) to denote that a certain quantity is to be understood in the macroscopic (thermodynamic) context, rather than in the microscopic (statistical) context.

[4] Note that Eq. (12) generally involves averaging over infinitely many finite-time realizations, in contrast to Eq. (10), which gives ΔS in terms of a single realization of infinite duration.

We now derive, as a byproduct of Eq. (9), two inequalities, one old and one new (Eqs. (14) and (16) below), which are closely related to the Clausius–Duhem inequality. By the convexity of the function e^x, Eq. (9) implies

$$-\frac{\langle E^\tau\rangle}{T^B}+\frac{\langle E^0\rangle}{T^A}+\left\langle\int_A^B\frac{dQ}{T}\right\rangle\leqslant-\frac{F^B}{T^B}+\frac{F^A}{T^A} \tag{13}$$

Once again invoking the identity $F=\bar{E}-ST$, along with our assumptions of initial and final equilibria, $\langle E^0\rangle=\bar{E}^A$ and $\langle E^\tau\rangle=\bar{E}^B$, we can rewrite Eq. (13) as:

$$\left\langle\int_A^B\frac{dQ}{T}\right\rangle\leqslant\Delta S \tag{14}$$

This result says, effectively, that the Clausius–Duhem inequality is satisfied "on average," where the average is taken over an ensemble of microscopic realizations of a given thermodynamic process. This still leaves open the possibility that there exist individual realizations for which the inequality is violated. We will now use Eq. (9) to investigate the frequency of occurrence of such violations.

Macroscopically, the Clausius–Duhem inequality can be written as $\Delta(\hat{E}/T)-\int_A^B d\hat{Q}/T\geqslant\Delta(\hat{F}/T)$, or simply $\hat{\Sigma}\geqslant\Delta(\hat{F}/T)$. Thus, thermodynamics tells us that we will "never" observe a value of $\hat{\Sigma}$ below $\Delta(\hat{F}/T)$. To investigate the *microscopic* validity of this statement, let $p(\Sigma)$ denote the distribution of values of Σ corresponding to the statistical ensemble of microscopic realizations of a given thermodynamic process. Then the probability of observing a value of Σ no greater than some fixed value Σ_0 is just: $\mathrm{Prob}[\Sigma\leqslant\Sigma_0]=\int_{-\infty}^{\Sigma_0}d\Sigma\ p(\Sigma)$. But Eq. (9) tells us that $\int_{-\infty}^{+\infty}pe^{-\Sigma}=e^{-\Delta(F/T)}$. When we combine this with the inequality chain

$$\int_{-\infty}^{+\infty}pe^{-\Sigma}\geqslant\int_{-\infty}^{\Sigma_0}pe^{-\Sigma}\geqslant e^{-\Sigma_0}\int_{-\infty}^{\Sigma_0}p \tag{15}$$

and take $\Sigma_0=\Delta(F/T)-\Gamma_0$, where $\Gamma_0>0$, we obtain

$$\mathrm{Prob}[\Sigma\leqslant\Delta(F/T)-\Gamma_0]\leqslant\exp(-\Gamma_0/k_B) \tag{16}$$

where we have explicitly put in the Boltzmann constant k_B. Thus, *the probability of observing a violation of the Clausius–Duhem inequality, by an amount no less than* Γ_0, *is bounded from above by* $e^{-\Gamma_0/k_B}$. A macroscopic violation would be one for which $\Gamma_0/k_B\gg1$, hence such violations are extremely rare: the Clausius–Duhem inequality is "never" violated by a macroscopic amount.[5]

[5] It is interesting to note the similarity between Eq. (16)—which pertains to a nonequilibrium thermodynamic process—and the Einstein–Boltzmann expression for microscopic fluctuations of a system in equilibrium; see e.g. Ref. 5, Eq. (112.1).

The previous two paragraphs are by no means intended as a first-principles derivation of the Second Law, as they assume—reasonably, but without proof—canonical distributions. Indeed, it has long been known (see, e.g. Ref. 1) that canonical ensembles imply inequalities such as Eq. (14), stating that the Second Law is not violated "on average." (However, I believe that the upper bound given by Eq. (16) is a new result.) The aim here is rather to reveal the close connection between the central result of this paper (Eq. (9)) and the Clausius–Duhem inequality.

While Eq. (9) is valid for any thermodynamic process which carries a system from A to B, it is instructive to ponder limiting cases of such processes. We will now consider three examples, for which we will be able to verify Eq. (9) directly (without invoking Liouville's theorem), by solving explicitly for Σ.

The first example involves bringing a system—initially at a temperature T^A—into contact with a reservoir at temperature T^B, and allowing the system to relax to the temperature of the reservoir. Let us therefore imagine that, at time $t=0$, we start with the system in the equilibrium state $A=(\lambda, T^A)$. Then, at $t=0^+$ ("immediately after $t=0$"), we place the system in thermal contact with a reservoir at temperature T^B, and let the two equilibrate. We assume the reservoir to have the usual property of "infinite" heat capacity, so that the system of interest relaxes to the equilibrium state $B=(\lambda, T^B)$. In this situation, we get $\int_A^B dQ/T=(1/T^B)\times\int_A^B dQ=(E^\tau-E^0)/T^B$—where the initial and final energies are given by $E^0=H_\lambda(\mathbf{z}^0)$ and $E^\tau=H_\lambda(\mathbf{z}^\tau)$—from which it follows that

$$\Sigma=\frac{E^\tau}{T^B}-\frac{E^0}{T^A}-\frac{E^\tau-E^0}{T^B}=\frac{H_\lambda(\mathbf{z}^0)}{T^B}-\frac{H_\lambda(\mathbf{z}^0)}{T^A} \tag{17}$$

We now average over realizations by integrating over the distribution of initial conditions to get:

$$\langle\exp(-\Sigma)\rangle=\frac{1}{Z^A}\int d\mathbf{z}^0\exp\left[-\frac{H_\lambda(\mathbf{z}^0)}{T^A}\right]\exp(-\Sigma) \tag{18}$$

$$=\frac{1}{Z^A}\int d\mathbf{z}^0\exp\left[-\frac{H_\lambda(\mathbf{z}^0)}{T^B}\right] \tag{19}$$

$$=\frac{Z^B}{Z^A}=\exp\left[-\Delta\left(\frac{F}{T}\right)\right] \tag{20}$$

in agreement with Eq. (9).

The second example involves making a sudden change in the value of the external parameter, $\lambda^A \to \lambda^B$, and then letting the system (assumed in contact at all times with a reservoir at temperature T) relax to the equilibrium state corresponding to the new parameter value. Thus, we begin ($t=0$) with the system in equilibrium state $A=(\lambda^A, T)$, coupled to a reservoir at temperature T; a moment later ($t=0^+$), we instantaneously change the parameter value from λ^A to λ^B, and then we allow the system to relax to the equilibrium state $B=(\lambda^B, T)$. The initial energy of the system is given by $E^0 = H_{\lambda^A}(\mathbf{z}^0)$; the energy just after λ is switched to λ^B is given by $E^{0+} = H_{\lambda^B}(\mathbf{z}^0)$; and the final energy is $E^\tau = H_{\lambda^B}(\mathbf{z}^\tau)$. Then $\int_A^B dQ/T = (E^\tau - E^{0+})/T$, and

$$\Sigma = \frac{E^\tau}{T} - \frac{E^0}{T} - \frac{E^\tau - E^{0+}}{T} = \frac{H_{\lambda^B}(\mathbf{z}^0)}{T} - \frac{H_{\lambda^A}(\mathbf{z}^0)}{T} \tag{21}$$

from which we again get

$$\begin{aligned} \langle \exp(-\Sigma) \rangle &= \frac{1}{Z^A} \int d\mathbf{z}^0 \exp\left[-\frac{H_{\lambda^A}(\mathbf{z}^0)}{T} \right] \exp(-\Sigma) \\ &= \exp\left[-\Delta\left(\frac{F}{T}\right) \right] \end{aligned} \tag{22}$$

The third example combines these two, and gives us a specific prescription for carrying a system from one arbitrary equilibrium state to another. We start with the system in state $A=(\lambda^A, T^A)$. Then we instantaneously switch the parameter value to λ^B, after which we place the system in contact with a reservoir at temperature T^B, and allow it to relax to the state $B=(\lambda^B, T^B)$. Following steps as above, we get $\Sigma = H_{\lambda^B}(\mathbf{z}^0)/T^B - H_{\lambda^A}(\mathbf{z}^0)T^A$, from which it once more immediately follows that $\langle e^{-\Sigma} \rangle = e^{-\Delta(F/T)}$.

In each of these examples, a convenient cancellation of terms gave us the value of Σ explicitly in terms of known functions of the initial conditions of the system of interest ($\mathbf{z}^0$). For a more general thermodynamic process, in which more reservoirs are involved and λ changes at a finite rate, this is not the case; if we wanted to compute Σ for a realization launched from a known set of initial conditions, we would need to actually integrate the equations of motion (in the *full* phase space) to get E^τ and $\int_A^B dQ/T$.

Let us now suppose that we prepare the system in state $A=(\lambda^A, T)$, and we switch the parameter at an arbitrary finite rate from λ^A to λ^B

(driving the system out of equilibrium), while keeping the system in thermal contact with a reservoir at temperature T; at the end we hold λ fixed at λ^B and allow the system to relax to $B=(\lambda^B, T)$. In this situation, we have

$$\int_A^B \frac{dQ}{T}=\frac{1}{T}\int_A^B dQ=\frac{1}{T}(E^\tau-E^0-W) \tag{23}$$

where W is the external work performed on the system by driving the parameter. We thus have $\Sigma=W/T$; and Eq. (9) reduces to the following relationship between the work performed (during realizations of this nonequilibrium process) and the free energy difference $\Delta F\equiv F^B-F^A$:

$$\left\langle \exp\left(-\frac{W}{T}\right)\right\rangle=\exp\left(-\frac{\Delta F}{T}\right) \tag{24}$$

Note the conditions for the validity of this result: there is only one heat reservoir[6] and its temperature must equal that at which the system is initially prepared. Eq. (24) has recently been derived in a number of ways, and confirmed in numerical experiments.[6]

We now generalize our analysis to the situation in which the system of interest begins and ends in nonequilibrium statistical states. For instance, we might prepare the system by heating it at one end and cooling at another, until a steady-state thermal gradient is achieved. (The N baths, however, are prepared in equilibrium, as before.) Whatever the method of preparation, let $\rho^0(\mathbf{z})$ denote the statistical distribution of initial conditions (of the system of interest) achieved by that preparation. Similarly, let $\rho^\tau(\mathbf{z})$ denote the distribution of final conditions; this will of course depend on the sequence of steps defining the thermodynamic process.

As before, a realization of the process is described by a microscopic trajectory $\mathbf{y}(t)$, determined by the initial conditions, $\mathbf{y}^0$. For a given realization, let us define

$$\Gamma\equiv -\ln\rho^\tau(\mathbf{z}^\tau)+\ln\rho^0(\mathbf{z}^0)+\sum_{n=1}^N\frac{\Delta E_n}{T_n} \tag{25}$$

and let us compute the average of $\exp(-\Gamma)$ over our ensemble of realizations:

$$\langle\exp(-\Gamma)\rangle=\frac{1}{\mathcal{N}}\int d\mathbf{y}^0\,\rho^0(\mathbf{z}^0)\exp\left[-\sum_{n=1}^N\frac{H_n^b(\mathbf{z}_n^0)}{T_n}\right]\exp(-\Gamma)=1 \tag{26}$$

following steps like those leading to Eq. (6).

[6] Or, if there are more, they share a common temperature.

Since we were able to derive Eq. (14) from Eq. (9), it is natural to wonder whether an interesting inequality can similarly be obtained from Eq. (26). The convexity of e^x in this case gives us $-\langle \Gamma \rangle \leqslant 0$, or

$$\langle \ln \rho^{\tau}(\mathbf{z}^{\tau}) \rangle - \langle \ln \rho^{0}(\mathbf{z}^{0}) \rangle + \left\langle \int dQ/T \right\rangle \leqslant 0 \tag{27}$$

Now note that $-\langle \ln \rho^{0}(\mathbf{z}^{0}) \rangle = -\int \rho^{0} \ln \rho^{0} = S_G[\rho^{0}]$, where the integration is over the phase space of the system of interest, and $S_G[\rho^{0}]$ represents the statistical (Gibbs) entropy associated with the initial statistical state of the system. Similarly, $-\langle \ln \rho^{\tau}(\mathbf{z}^{\tau}) \rangle = S_G[\rho^{\tau}]$. Eq. (27) then reads

$$\left\langle \int dQ/T \right\rangle \leqslant S_G[\rho^{\tau}] - S_G[\rho^{0}] \equiv \Delta S_G \tag{28}$$

That is, *the expectation value of* $\int dQ/T$ *is bounded from above by the net change in the statistical entropy*, S_G, characterizing the initial and final states of the system of interest. Note that in the case of an isolated system (no heat baths), this inequality reduces to a trivial result, as both sides are identically zero.

Eq. (28) is equivalent to the statement: $\langle \Gamma \rangle \geqslant 0$. Following a line of reasoning like the one leading to Eq. (16), we can use Eq. (26) to place an upper bound on the probability of observing a value of Γ no greater than $-\Gamma_0$:

$$\text{Prob}[\Gamma \leqslant -\Gamma_0] \leqslant e^{-\Gamma_0/k_B} \tag{29}$$

Thus, we will "never" observe a "macroscopically negative" value of Γ. Eqs. (26), (28) and (29) together constitute a generalization of Eqs. (9), (14) and (16), to situations in which the system of interest begins and ends in states not necessarily corresponding to thermal equilibrium.

A macroscopic, reversible process between two equilibrium states A and B has the property that $\Delta(\hat{E}/T) - \int_A^B d\hat{Q}/T = \Delta(\hat{F}/T)$ (Eq. (10)). The central result of this paper is a microscopic, statistical generalization of this result to irreversible processes between two such states: $\langle e^{-\Delta(E/T) + \int_A^B dQ/T} \rangle = e^{-\Delta(F/T)}$ (Eq. (9)), where the average is taken over an ensemble of realizations of the process. In both cases, the right side of the equation depends only on the states A and B, and not on the (equilibrium or non-equilibrium) path connecting them. We have used Eq. (9) to derive an expression for the entropy difference between two equilibrium states, in

terms of an arbitrary (generally irreversible) thermodynamic process connecting them (Eq. (12)). We have also shown that Eq. (9) leads to statistical statements of the Clausius–Duhem inequality, in particular placing an upper bound on the probability for observing violations of the Clausius–Duhem inequality above an arbitrary threshold Γ_0 (Eq. (16)). Finally, we have extended this analysis to processes which begin and/or end out of equilibrium.

ACKNOWLEDGMENTS

I would like to thank G. Crooks, E. Lieb, C. Maes, and J. Percus for stimulating conversations and correspondence regarding central issues in this paper, and C. Zalka for simplifying the derivation of Eq. (16). This research was partially supported by the Polish-American Maria Sklodowska-Curie Joint Fund II, under project PAA/NSF-96-253.

REFERENCES

1. J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (New York, 1902, Charles Scribner's Sons), Chapter XIII.
2. G. Crooks, "The Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences," cond-mat/9901352, available from the Los Alamos preprint archives (http://xxx.lanl.gov).
3. D. J. Evans, E. G. D. Cohen, and G. P. Morriss, *Phys. Rev. Lett.* **71**:2401 (1993).
4. G. Gallavotti and E. G. D. Cohen, *J. Stat. Phys.* **80**:931 (1995); *Phys. Rev. Lett.* **74**:2694 (1995).
5. L. D. Landau and E. M. Lifshitz, *Statistical Physics*, 3rd ed., Part 1, Chapter XII (Pergamon Press, Oxford, 1980).
6. C. Jarzynski, *Phys. Rev. Lett.* **78**:2690 (1997); C. Jarzynski, *Phys. Rev. E* **56**:5018 (1997); G. E. Crooks, *J. Stat. Phys.* **90**:1481 (1998); C. Jarzynski, *Acta Phys. Pol. B* **29**:1609 (1998); R. M. Neal, "Annealed Importance Sampling," Technical Report No. 9805 (revised), Dept. of Statistics, Toronto (1998).
7. D. J. Evans and D. J. Searles, *Phys. Rev. E* **50**:1645 (1994).