Classical and Quantum Fluctuation Theorems for Heat Exchange

Christopher Jarzynski

Theoretical Division, T-13, MS B213, Los Alamos National Laboratory, Los Alamos, NM 87545[*]

Daniel K. Wójcik

Center for Nonlinear Science, School of Physics, Georgia Institute of Technology,
837 State Street, Atlanta, GA 30332-0430, USA and
Department of Neurophysiology, Nencki Institute of Experimental Biology, 3 Pasteur Str., 02-093 Warsaw, Poland[†]

(Dated: April 21, 2004)

The statistics of heat exchange between two classical or quantum finite systems initially prepared at different temperatures are shown to obey a fluctuation theorem.

The *fluctuation theorem* (FT) refers to a collection of theoretical predictions [1, 2, 3, 4, 5, 6, 7], recently confirmed experimentally [8], pertaining to a system evolving under non-equilibrium conditions. These results are roughly summarized by the equation

$$\ln \frac{p(+\Sigma)}{p(-\Sigma)} = \Sigma, \qquad (1)$$

where $p(\Sigma)$ denotes the probability that an amount of entropy Σ is generated during a specified time interval. Both *transient* and *steady state* versions of the FT have been obtained. The definition of "entropy generated" (Σ) depends on the dynamics used to model the evolution of the system under consideration. However, for a variety of physical situations, and a variety of equations of motion (both deterministic and stochastic) used to model them, the FT has been established under reasonable definitions of entropy generation. Moreover, the FT is related [9] to a set of *free energy relations* (see e.g. [10, 11]) which connect equilibrium free energy differences to non-equilibrium work values, and which have recently been confirmed experimentally [12].

The situations modeled in Refs. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] all involve an externally driven system, in the presence of a heat reservoir. The purpose of this paper is to point out that a similar result can be derived in a different setting. Namely, we will obtain a symmetry relation constraining the statistics of heat exchange between two bodies initially prepared at different temperatures. We will present both classical and quantum derivations, and will use the term *exchange fluctuation theorem* (XFT) to refer to these results.

In what follows, the XFT (Eq. (2)) will be stated and derived. A corollary result related to the Second Law of Thermodynamics will then be presented (Eq. (17)).

Consider two finite bodies, A and B, separately prepared in equilibrium states at temperatures T_A and T_B, respectively, then placed in thermal contact with one another for a time τ, and then separated again. Let Q denote the net heat transfer from A to B during the interval of contact, i.e. the amount of energy lost by A and gained by B. Now imagine repeating this experiment many times, always initializing the two bodies at the specified temperatures, and let $p_\tau(Q)$ denote the observed distribution of values of Q over the ensemble of repetitions. Then we claim that this distribution satisfies

$$\ln \frac{p_\tau(+Q)}{p_\tau(-Q)} = \Delta\beta \cdot Q, \qquad (2)$$

where $\Delta\beta = T_B^{-1} - T_A^{-1}$ is the difference between the inverse temperatures at which the bodies are prepared.

In the quantum case we must define Q through an experimental procedure: starting with the two systems initially prepared at different temperatures, we first measure the energy of each system, then we allow them to weakly interact over a time τ, and finally we again measure the energy of each system. We then interpret heat transfer in terms of the changes in these measured energies (Eq. (14)). This approach is similar in spirit to that taken by [13, 14, 15], who considered related problems. For an alternative approach see e.g. [16].

Eq. (2) clearly resembles the usual FT, Eq. (1). Indeed, if we invoke macroscopic thermodynamics to argue that $-Q/T_A$ is the entropy change of A, and $+Q/T_B$ is that of B, then the net entropy generated by the exchange of heat is given by $\Sigma = \Delta\beta \cdot Q$, and Eq. (2) becomes Eq. (1). However, this argument works only if the heat transferred is very small in comparison with the internal energy of either body, whereas the validity of Eq. (2) does *not* require this assumption. Therefore, we will leave Eq. (2) as a statistical statement about heat exchange, rather than trying to force it to be a statement about entropy generation *per se*.

To derive Eq. (2) from classical equations of motion, let $\mathbf{z}_A$ denote the phase space coordinates specifying the microstate of body A (e.g. the positions and momenta of all its degrees of freedom); and let $H^A(\mathbf{z}_A)$ be a Hamiltonian whose value defines the internal energy of A, as a function of its microstate. Similarly for $H^B(\mathbf{z}_B)$. Let

$h^{\text{int}}(\mathbf{z}_A, \mathbf{z}_B)$ denote a small interaction term, turned "on" at $t = 0$, and "off" at $t = \tau$, coupling the two bodies. Let $\mathbf{y} = (\mathbf{z}_A, \mathbf{z}_B)$ specify a point in the *full* phase space of all participating degrees of freedom. During any realization of the process in which we are interested, the microscopic evolution of the two bodies is described by a trajectory $\mathbf{y}(t)$, evolving from $t = 0$ to $t = \tau$ under Hamilton's equations, as derived from the Hamiltonian

$$\mathcal{H}(\mathbf{y}) = H^A(\mathbf{z}_A) + H^B(\mathbf{z}_B) + h^{\text{int}}(\mathbf{y}). \tag{3}$$

We now further assume *time-reversal invariance*:

$$H^i(\mathbf{z}_i) = H^i(\mathbf{z}_i^*), \quad h^{\text{int}}(\mathbf{y}) = h^{\text{int}}(\mathbf{y}^*), \tag{4}$$

where $i = A, B$ and the asterisk (*) denotes the time-reversal operation, usually the reversal of momenta: $(\mathbf{q}, \mathbf{p})^* = (\mathbf{q}, -\mathbf{p})$. This assumption has the crucial consequence that microscopic realizations of the process come in pairs related by time-reversal: for any trajectory $\mathbf{y}(t)$ which is a solution of Hamilton's equations, its time-reversed twin, $\overline{\mathbf{y}}(t) = \mathbf{y}^*(\tau - t)$, is also a solution. For future reference let $\mathbf{y}^0$ and $\mathbf{y}^\tau$ denote the initial and final conditions of the "forward" realization [19], $\mathbf{y}(t)$; hence the "reverse" realization, $\overline{\mathbf{y}}(t)$, evolves from $\overline{\mathbf{y}}^0 = \mathbf{y}^{\tau*}$ to $\overline{\mathbf{y}}^\tau = \mathbf{y}^{0*}$, as illustrated in Figure 1.
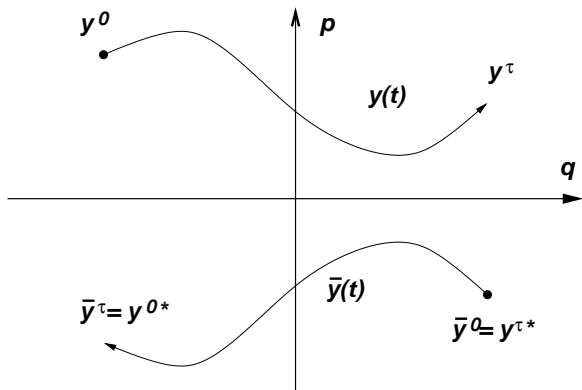

FIG. 1: Twin trajectories $y(t)$ and $\bar{y}(t) = y^*(\tau - t)$ related by time reversal.

By our assumption regarding the equilibrium preparation of the two bodies, the probability distribution for sampling initial conditions $\mathbf{y}^0$ is given by:

$$P(\mathbf{y}^0) = \frac{1}{Z_A Z_B} e^{-H^A(\mathbf{z}_A^0)/T_A} e^{-H^B(\mathbf{z}_B^0)/T_B}, \tag{5}$$

where the Z's are partition functions. Given a trajectory $\mathbf{y}(t)$ and its time-reversed twin $\overline{\mathbf{y}}(t)$, the ratio of probabilities of sampling their respective initial conditions is then:

$$\frac{P(\mathbf{y}^0)}{P(\overline{\mathbf{y}}^0)} = e^{\Delta E_B/T_B} e^{\Delta E_A/T_A}, \tag{6}$$

where $\Delta E_A = H^A(\mathbf{z}_A^\tau) - H^A(\mathbf{z}_A^0) = H^A(\overline{\mathbf{z}}_A^0) - H^A(\mathbf{z}_A^0)$, and similarly for ΔE_B. The quantities ΔE_A and ΔE_B

represent the net change in the internal energies of the two bodies, over the course of the realization described by $\mathbf{y}(t)$. If we neglect the small amount of work performed in switching on and off the interaction term h^{int}, then the net change in the energy of one system is compensated by an opposite change in the energy of the other, i. e. $\Delta E_B \approx -\Delta E_A$, and it is natural to view these changes as representing a quantity of heat transferred from A to B: $Q := \Delta E_B \approx -\Delta E_A$. Hence,

$$\frac{P(\mathbf{y}^0)}{P(\overline{\mathbf{y}}^0)} = e^{\Delta \beta \cdot \hat{Q}(\mathbf{y}^0)}, \tag{7}$$

where the function $\hat{Q}(\mathbf{y})$ denotes the value of Q during a realization evolving from initial conditions $\mathbf{y}$. Note that

$$\hat{Q}(\overline{\mathbf{y}}^0) = -\hat{Q}(\mathbf{y}^0), \tag{8}$$

that is, the heat transfer during the forward realization is opposite to that during the reverse realization.

Combining Eqs. (7) and (8) we get:

$$
\begin{aligned}
p_\tau(Q) &= \int d\mathbf{y}^0 P(\mathbf{y}^0) \delta[Q - \hat{Q}(\mathbf{y}^0)] \\
&= e^{\Delta \beta \cdot Q} \int d\overline{\mathbf{y}}^0 P(\overline{\mathbf{y}}^0) \delta[Q + \hat{Q}(\overline{\mathbf{y}}^0)] \\
&= e^{\Delta \beta \cdot Q} p_\tau(-Q),
\end{aligned} \tag{9}
$$

which is equivalent to Eq. (2). Here the change in the variables of integration between the first and second lines is justified by the invariance of the Liouville measure under time evolution ($d\mathbf{y}^0 = d\mathbf{y}^\tau$), as well as under time reversal ($d\mathbf{y}^\tau = d\mathbf{y}^{\tau*} = d\overline{\mathbf{y}}^0$).

These formal manipulations can be understood intuitively. $p_\tau(Q)$ is a sum of contributions from all realizations for which the heat transfer takes on a specified value, Q; and $p_\tau(-Q)$ is a sum over those for which the heat transfer is $-Q$. But these two sets of realizations are in one-to-one correspondence; for every trajectory $\mathbf{y}(t)$ belonging to the former set, its twin $\overline{\mathbf{y}}(t)$ belongs to the latter (Eq. 8). Moreover, the ratio of initial condition sampling probabilities for such a twinned pair of realizations is $e^{\Delta \beta \cdot Q}$ (Eq. (7)). Therefore, when we add the sampling probabilities $P(\mathbf{y}^0)$ from the first set of realizations to get $p_\tau(Q)$, and $P(\overline{\mathbf{y}}^0)$ from the second set to get $p_\tau(-Q)$, the ratio of the sums is $e^{\Delta \beta \cdot Q}$.

The above derivation, based on comparing the sampling probabilities for pairs of twinned trajectories, is similar to that carried out by Evans and Searles [2] for the transient FT. Note also that this derivation is valid for arbitrary times τ; there are no hidden assumptions that the temperatures of the two systems remain constant, or even well-defined after $t = 0$.

The sole approximation that we have made is the neglect of the interaction term, h^{int}. In reality, a finite amount of work is required to turn on this interaction initially, δw_{on}, and then to turn it off finally, δw_{off}.

The resulting balance of energy reads: $\Delta E_A + \Delta E_B = \delta w_{\mathrm{on}} + \delta w_{\mathrm{off}}$, hence $\delta w = \delta w_{\mathrm{on}} + \delta w_{\mathrm{off}}$ enters as a correction to the approximation $\Delta E_B \approx -\Delta E_A$ used earlier. The validity of our final result thus requires that the work performed in coupling and later uncoupling the systems ($|\delta w|$) be much smaller than the typical energy change in either system ($|\Delta E_A|$, $|\Delta E_B|$). Whether or not this condition is met depends, of course, on details of the two systems, on the strength of the interaction term ($\delta w \sim h^{\mathrm{int}}$), and on the duration τ. It will be interesting to investigate this issue in the context of specific models.

We proceed now to the proof of the quantum version of our theorem. We assume that systems A and B have equilibrated to temperatures T_A, T_B before the experiment, and are thus described by density matrices $\rho_i = \exp(-\beta_i H^i)/Z_i$, where $i = A, B$. At time $t = 0^-$ we separate the systems from the reservoirs and measure their energies. As a result, each system i is projected onto a pure state $|n_i\rangle$ with probability $e^{-\beta_i E^i_{n_i}}/Z_i$, and the combined system is described by the product state $|n_A n_B\rangle$. We then allow the systems to interact through a weak coupling term h^{int}. Thus the Hamiltonian takes the form $\mathcal{H} = H^A \otimes I^B + I^A \otimes H^B + h^{\mathrm{int}}$.

Let us now assume, as in the classical case (Eq. (4)), that the system and both its subsystems are time-reversal invariant. In quantum mechanics the time-reversal invariance of a system is expressed by the condition

$$\Theta H = H\Theta, \qquad (10)$$

where H is the system Hamiltonian, and Θ is the quantum time-reversal operator [17, 18]. This operator reverses linear and angular momentum while keeping position unchanged, and is *anti-linear*:

$$\Theta\Big(\alpha_1 |\psi\rangle + \alpha_2 |\phi\rangle\Big) = \alpha_1^\dagger \Theta|\psi\rangle + \alpha_2^\dagger \Theta|\phi\rangle, \qquad (11)$$

where the dagger denotes complex conjugation. When dealing with such operators, the Dirac bra-ket notation, invented to deal with linear operators, becomes cumbersome: the expression $\langle\phi|\Theta|\psi\rangle$ is ambiguous until we specify whether Θ is acting to the right or to the left. To avoid this inconvenience we will use the standard product in Hilbert space $(|\phi\rangle, |\psi\rangle)$, rather than the more abbreviated Dirac bra-ket, $\langle\phi|\psi\rangle$, to denote the inner product between two wave functions. From Eq. (10) it follows that, for every eigenstate $|n\rangle$ of H there corresponds a time-reversed eigenstate $\Theta|n\rangle$ with the same energy; these two states are either linearly independent, or else identical apart from an overall phase. Moreover, since Θ preserves wave function normalization, it is not just anti-linear but also *anti-unitary*: $(\Theta|\phi\rangle, \Theta|\psi\rangle) = (|\psi\rangle, |\phi\rangle)$. We will make use of these properties in the analysis below.

Having turned on the interaction term at $t = 0$, we allow the systems to evolve for a time τ. The combined system then reaches a state $|\Psi\rangle$, obtained from the initial state $|n_A n_B\rangle$ by evolution under Schrödinger's equation. We now separate the two systems – that is, we turn off the interaction term – and once again measure their energies. The state $|\Psi\rangle$ is thus projected onto a product state $|m_A m_B\rangle$. As before, we make no assumptions regarding τ, in particular the systems have not necessarily equilibrated.

Letting $P_\tau(|n\rangle \to |m\rangle)$ denote the probability of observing a transition from $|n\rangle \equiv |n_A n_B\rangle$ to $|m\rangle \equiv |m_A m_B\rangle$, we have

$$P_\tau(|n\rangle \to |m\rangle) = |(|m\rangle, U_\tau |n\rangle)|^2 \frac{e^{-\beta_A E^A_{n_A} - \beta_B E^B_{n_B}}}{Z_A Z_B},$$

where $U_\tau = e^{-i\tau\mathcal{H}}$ is the quantum evolution operator, and $\hbar = 1$. The second factor on the right is the probability for sampling the initial state $|n\rangle$; the first factor is the transition probability from $|n\rangle$ to $|m\rangle$. Similarly, the probability of observing the time-reversed transition from $\Theta|m\rangle$ to $\Theta|n\rangle$ is

$$P_\tau(\Theta|m\rangle \to \Theta|n\rangle) = |(\Theta|n\rangle, U_\tau \Theta|m\rangle)|^2 \frac{e^{-\beta_A E^A_{m_A} - \beta_B E^B_{m_B}}}{Z_A Z_B}.$$

But, since Θ is anti-unitary, and $U_\tau \Theta = \Theta U_{-\tau}$ [20], we have

$$\begin{aligned}(\Theta|n\rangle, U_\tau \Theta|m\rangle) &= (\Theta|n\rangle, \Theta U_{-\tau}|m\rangle) = (U_{-\tau}|m\rangle, |n\rangle) \\ &= (|m\rangle, U_\tau |n\rangle),\end{aligned}$$

therefore

$$\frac{P_\tau(|n\rangle \to |m\rangle)}{P_\tau(\Theta|m\rangle \to \Theta|n\rangle)} = e^{-\beta_A(E^A_{n_A} - E^A_{m_A})} e^{-\beta_B(E^B_{n_B} - E^B_{m_B})}. \qquad (12)$$

Since we assumed that the interaction is weak, we expect the energy of the total system to be almost preserved:

$$E^A_n + E^B_n \approx E^A_m + E^B_m. \qquad (13)$$

It follows that the energy changes in the two systems are approximately equal

$$Q_{n \to m} := E^B_m - E^B_n \approx E^A_n - E^A_m. \qquad (14)$$

We interpret Q as the heat exchange between the systems A and B. Thus,

$$\frac{P_\tau(|n\rangle \to |m\rangle)}{P_\tau(\Theta|m\rangle \to \Theta|n\rangle)} \approx e^{\Delta\beta \cdot Q_{n \to m}}. \qquad (15)$$

Since every eigenstate has a corresponding time-reversed twin, the net probability of the heat transfer Q in time τ is

$$\begin{aligned}p_\tau(Q) &= \sum_{n,m} P_\tau(|n\rangle \to |m\rangle)\delta(Q - Q_{n \to m}) \\ &= e^{\Delta\beta \cdot Q} \sum_{\Theta n, \Theta m} P_\tau(\Theta|m\rangle \to \Theta|n\rangle)\delta(Q + Q_{\Theta m \to \Theta n}) \\ &= e^{\Delta\beta \cdot Q} p_\tau(-Q). \qquad (16)\end{aligned}$$

This result is true for the quantities as we have defined them. We can rewrite Eq. (16) in the form of Eq. (2) if we further assume a sufficiently dense spectrum, so that $p_\tau(Q)$ can be replaced by a locally smooth function.

At the level of macroscopic thermodynamics (and in the absence of external work), the passage of heat from a colder to a hotter body constitutes a violation of the Second Law. From Eq. (2), we can derive an upper bound on the probability of observing such a "violation", of at least some finite magnitude, as follows. Assume that $T_A > T_B$, i.e. $\Delta\beta > 0$. The probability that the heat transfer from A to B will fall below a specified value q is given by $\int_{-\infty}^{q} p_\tau(Q)dQ$. Using Eq. (2) to replace $p_\tau(Q)$ by $p_\tau(-Q)\exp(\Delta\beta \cdot Q)$, and then invoking the inequality chain

$$\int_{-\infty}^{q} p_\tau(-Q)e^{\Delta\beta \cdot Q}dQ \le e^{\Delta\beta \cdot q}\int_{-\infty}^{q} p_\tau(-Q)dQ \le e^{\Delta\beta \cdot q},$$

we get

$$\int_{-\infty}^{q} p_\tau(Q)\, dQ \le e^{\Delta\beta \cdot q}. \tag{17}$$

Choosing $q < 0$, this result tells us that the probability of observing a net heat transfer in the "wrong" direction ($Q < 0$), from B (cold) to A (hot), of at least some magnitude $|q|$, dies exponentially (or faster) with that magnitude. Eq. (2) also implies that the average of $\exp(-\Delta\beta \cdot Q)$, over the ensemble of realizations for any time τ, is unity:

$$\overline{e^{-\Delta\beta \cdot Q}} \equiv \int dQ p_\tau(Q)e^{-\Delta\beta \cdot Q} = 1. \tag{18}$$

In conclusion, a result analogous to the FT for entropy generation (Eq. (1)), and valid for arbitrary times τ, has been derived for the statistics of heat exchange between finite classical or quantum systems separately prepared in equilibrium (Eqs. (2)). In our derivation we invoke statistical mechanics to describe the initial preparation of the systems, then treat their evolution during the interval of contact dynamically. We also assume a negligible energy of interaction between the two systems, and a time-reversal invariant Hamiltonian. In the quantum case, an additional source of randomness arises from the fact that the initial quantum state of the system does not uniquely determine the outcome of the final energy measurements. Nevertheless, this does not spoil our result.

We finally mention that a similar theorem can be derived for particle exchange between two reservoirs, driven by a difference in initial chemical potentials (unpublished).

———————

[*] Electronic address: chrisj@lanl.gov
[†] Electronic address: danek@cns.physics.gatech.edu
[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).
[3] G. Gallavotti and E. G. D. Cohen, J. Stat. Phys. **80**, 931 (1995).
[4] J. Kurchan, J. Phys. A-Math. Gen. **31**, 3719 (1998).
[5] J. L. Lebowitz and H. Spohn, J. Stat. Phys. **95**, 333 (1999).
[6] C. Maes, J. Stat. Phys. **95**, 367 (1999).
[7] D. J. Evans and D. Searles, Adv. in Phys. **51**, 1529 (2002).
[8] G. M. Wang, E. M. Sevick, E. Mittag, D. J. Searles, and D. J. Evans, Phys. Rev. Let. **89**, 050601 (2002).
[9] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[10] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[11] G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998).
[12] J. Liphardt, S. Dumont, S. Smith, I. Tinoco, and C. Bustamante, Science **296**, 1832 (2002).
[13] J. Kurchan (2000), cond-mat/0007360.
[14] S. Mukamel, Phys. Rev. Lett. **90**, 170604 (2003).
[15] T. Monnai and S. Tasaki (2003), cond-mat/0308337.
[16] W. De Roeck and C. Maes (2003), cond-mat/0309498.
[17] E. Merzbacher, *Quantum Mechanics* (J. Wiley & Sons, New York, 2001), 3rd ed.
[18] L. Ballentine, *Quantum Mechanics: A Modern Development* (World Scientific, Singapore, 1998), 2nd ed.
[19] Of course, for any pair of realizations related by time-reversal, the designation of which is the forward realization, and which is the reverse, is arbitrary.
[20] This follows from Eq.10 and the anti-linearity of Θ.