MOLECULAR PHYSICS, 20 May 2003, VOL. 101, NO. 10, 1551–1554

Taylor & Francis
Taylor & Francis Group

RESEARCH NOTE

A non-equilibrium free energy theorem for deterministic systems

DENIS J. EVANS*

Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia

(*Received 4 November 2002; revised version accepted 9 December 2002*)

Jarzynski and Crooks have recently shown that equilibrium free energy differences can be computed from *non-equilibrium* thermodynamic path integrals. In the present paper we give a new derivation of this extraordinary relation. Our derivation which is valid for time reversible deterministic systems highlights the close relationship between the non-equilibrium free energy theorems and the fluctuation theorem.

1. Introduction

The fluctuation theorems (FTs) [1–9] give formulae for the logarithm of the probability ratio that the time averaged dissipative flux takes a value B to minus that value, $-B$, in non-equilibrium systems. A subset of these theorems, known as transient fluctuation theorems (TFTs), compute these finite time probability ratios for systems which start at $t=0$, from some known initial distribution—usually an equilibrium distribution. A TFT has recently been successfully tested in laboratory experiments employing optical tweezers [10]. In the present paper we derive a TFT for non-equilibrium transitions between two equilibrium states. The resulting formulae are not new, having been derived earlier by Jarzynski [11] and Crooks [12]. However, almost all of the work carried out by Crooks and Jarzynski was for stochastic systems. Our derivation is applicable to realistically thermostatted, time reversible, deterministic systems.

These equilibrium-to-equilibrium TFTs relate the distribution of thermodynamic work done along all possible time reversible, non-equilibrium paths connecting the equilibrium systems, to differences in the free energy of the two equilibrium states. Thus we term these relations non-equilibrium free energy theorems (NEFETs).

2. Derivation

Consider two N-particle equilibrium systems with coordinates and peculiar momenta, $\{\mathbf{q}_1, \mathbf{q}_2, \ldots \mathbf{q}_N, \mathbf{p}_1, \ldots \mathbf{p}_N\} \equiv (\mathbf{q}, \mathbf{p}) \equiv \mathbf{\Gamma}$. The systems are described by Hamiltonians $H_1(\mathbf{\Gamma})$, $H_2(\mathbf{\Gamma})$. The systems are of volume V and are assumed to include a heat bath maintained at a temperature T. Thus we can characterize the phase space distributions of the two systems by the appropriate canonical distributions, $f_1(\mathbf{\Gamma})$, $f_2(\mathbf{\Gamma})$

$$f_i(\mathbf{\Gamma}) \sim \frac{\exp[-\beta H_i(\mathbf{\Gamma})]}{\int \mathrm{d}\Gamma \exp[-\beta H_i(\mathbf{\Gamma})]}, \quad i = 1, 2, \tag{1}$$

with corresponding Helmholtz free energies,

$$A_i = -k_\mathrm{B} T \ln\left[\int \mathrm{d}\Gamma \exp[-\beta H_i(\mathbf{\Gamma})]\right]. \tag{2}$$

Consider a transformation from H_1 to H_2. We call this the forward (F) direction for the transformation and we denote the reverse direction by the symbol R. Consider, for example,

$$H(\mathbf{\Gamma}, t) = H_1(\mathbf{\Gamma})(1 - \lambda(t)) + H_2(\mathbf{\Gamma}) \quad \lambda(t), \quad 0 < t < \tau \tag{3}$$

with

$$\dot{\lambda}_{\mathrm{F,R}} = \pm\frac{1}{\tau}, \ 0 < t < \tau. \tag{4}$$

The choice of the actual pathway in the transformation from H_1 to H_2 is, as we shall see, extraordinarily general. It need not be the simple linear pathway as in (3) and (4). Thus, equations (3) and (4) are simply a convenient example of such a pathway. The equations of motion for the system in the time interval $(0, \tau)$ are assumed to be [13],

$$\begin{aligned} \dot{\mathbf{q}}_i &= \frac{\partial H(\mathbf{\Gamma}, t)}{\partial \mathbf{p_i}}, \\ \dot{\mathbf{p}}_i &= -\frac{\partial H(\mathbf{\Gamma}, t)}{\partial \mathbf{q_i}} - S_i \alpha(\mathbf{\Gamma}) \mathbf{p_i}, \\ \dot{\lambda}_{\mathrm{F,R}} &= \pm\frac{1}{\tau}, \end{aligned} \tag{5}$$

where α is the thermostat multiplier [13] which in this case is applied to fix the kinetic temperature of the

* e-mail: evans@rsc.anu.edu.au

Molecular Physics ISSN 0026–8976 print/ISSN 1362–3028 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0026897031000085173

thermostatting walls at a temperature T,

$$\sum_{i=1}^{N} S_i \frac{p_i^2}{2m} = \sum_{i=1}^{N_w} \frac{p_i^2}{2m} = \frac{3N_w k_B T}{2}. \tag{6}$$

For simplicity we assume all particles have the same mass m. Boltzmann's constant is denoted as k_B. S_i is a switch that controls which particles are thermostatted. We assume particles labelled i from 1 to N_w comprise thermostatting walls. The remaining particles $N_w + 1 \leqslant i \leqslant N$, comprise the system of interest. We can also consider homogeneously thermostatted systems in which all the particles are thermostatted, that is $N_w = N$. For wall-thermostatted systems it is natural to assume that the wall particles are unaltered in the transformation $H_1 \to H_2$.

We define a work function W as [9],

$$\begin{aligned}\beta \Delta W(\tau) &= \beta(W_2 - W_1) \\ &= \beta[H(\tau) - H(0)] - \int_0^\tau \mathrm{d}s\, \Lambda(s)\end{aligned} \tag{7}$$

where the phase space compression factor is defined as

$$\Lambda \equiv \frac{\partial}{\partial \boldsymbol{\Gamma}} \cdot \dot{\boldsymbol{\Gamma}}, \tag{8}$$

and $\beta \equiv 1/k_B T$. The Liouville equation for the N-particle phase space distribution function $f(\boldsymbol{\Gamma}, t)$ of the system can be written as [13],

$$\frac{\mathrm{d}f(\boldsymbol{\Gamma}, t)}{\mathrm{d}t} = -f(\boldsymbol{\Gamma}, t)\Lambda(\boldsymbol{\Gamma}) = 3N_w \alpha(t) f(\boldsymbol{\Gamma}, t) + \mathrm{O}(1), \tag{9}$$

where we used the equations of motion to evaluate the phase space compression factor.

We seek an expression for the probability ratio that in the transition $(1 \to 2)$ (i.e. in the forward direction) the work function takes a value B compared with the probability that starting from system 2, the change in the work function for the reverse process $(2 \to 1)$, takes a value $-B$.

From figure 1 we can see that since the Jacobean of the time reversal map $M^{\mathrm{T}}(M^{\mathrm{T}}(\mathbf{q}, \mathbf{p}, \dot{\lambda}) = (\mathbf{q}, -\mathbf{p}, -\dot{\lambda}))$ is unity, the volume elements $\mathrm{d}\Gamma_0^T(\tau), \mathrm{d}\Gamma_0(\tau)$ have the same measure. Since the equations of motion are time reversible,

$$\mathrm{d}\Gamma_0(\tau)/\mathrm{d}\Gamma_0(0) = \mathrm{d}\Gamma_0^{\mathrm{T}}(\tau)/\mathrm{d}\Gamma_0^{\mathrm{T}}(0). \tag{10}$$

Clearly also, the work function will take on opposite values for the forward and reverse trajectories. We have drawn figure 1 as though there is only one contiguous region in system 1 for which $\Delta W(t) = B, \mathrm{d}B$. However, this will not usually be so. Usually there will be multiply disconnected regions within which trajectories originate with the required path integral values.

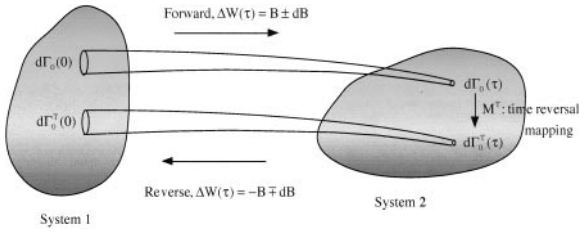

Figure 1. A trajectory bundle within the phase space for system 1 which has the specified value for the change in the work function, $B - \mathrm{d}B < \Delta W(\tau) < B + \mathrm{d}B$. The figure also shows the conjugate bundle of antitrajectories which necessarily have the corresponding negative values of the change in the work function. In practice there may be numerous non-contiguous trajectory bundles which each have the same value of the change in the work function.

Obviously for finite τ, the intermediate states are not in equilibrium. We assume that no matter how far from equilibrium the trajectories may be in mid transition, they nevertheless must originate and terminate in equilibrium systems. This places a constraint on the transformation $H_1 \to H_2$ (3, 4), at least near both end points. Thus we can compute the required probability ratio,

$$\begin{aligned}&\frac{\mathrm{Pr_F}(\Delta W = B)}{\mathrm{Pr_R}(\Delta W = -B)} \\ &= \frac{\displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[-\beta H_1(\boldsymbol{\Gamma}_0(0))] \Big/ \int \mathrm{d}\Gamma \exp[-\beta H_1]}{\displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0^{\mathrm{T}}(\tau) \exp[-\beta H_2(\boldsymbol{\Gamma}_0^{\mathrm{T}}(\tau))] \Big/ \int \mathrm{d}\Gamma \exp[-\beta H_2]} \\ &= \frac{\mathrm{e}^{\beta A_1} \displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) - B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[-\beta H_1(\boldsymbol{\Gamma}_0(0))]}{\mathrm{e}^{\beta A_2} \displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(\tau) \exp[-\beta H_2(\boldsymbol{\Gamma}_0(\tau))]} \\ &= \frac{\mathrm{e}^{-\beta \Delta A} \displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[-\beta H_1(\boldsymbol{\Gamma}_0(0))]}{\left[\displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \times \exp\Big[-\int_0^\tau \mathrm{d}s 3N_w \alpha(s) \mathrm{d}s - \beta\{H_1(\boldsymbol{\Gamma}_0(0)) + \Delta W(\tau) - \int_0^\tau \mathrm{d}s 3N_w k_B T \alpha(s)\, \mathrm{d}s\}\Big]\right]} \\ &= \frac{\mathrm{e}^{-\beta \Delta A} \displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[-\beta H_1(\boldsymbol{\Gamma}_0(0))]}{\displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[-\beta\{H_1(\boldsymbol{\Gamma}_0(0)) + \Delta W(\tau)\}]} \\ &= \frac{\mathrm{e}^{-\beta \Delta A} \displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0) \exp[\beta \Delta W(\tau)]}{\displaystyle\sum_{\{\boldsymbol{\Gamma}_0 | \Delta W(\tau) = B, \mathrm{d}B\}} \mathrm{d}\Gamma_0(0)} \\ &= \mathrm{e}^{-\beta \Delta A} \mathrm{e}^{\beta B}.\end{aligned} \tag{11}$$

All the sums in (11) are computed over contiguous trajectory bundles each of which is centred on $\mathbf{\Gamma}_0(0)$ or $\mathbf{\Gamma}_0(\tau)$ and have volumes $d\Gamma_0(0)$, $d\Gamma_0(\tau)$ respectively.

The first line of (11) assumes the two end states are in thermal equilibrium and are describable by the canonical probability distributions. It also assumes that the system (5) is time reversible and that the starting phases for the reverse pathways can be obtained from a time reversal map applied to the end phase of the conjugate, forward pathway. We use (2) and (10) to obtain the second line. We also use the fact that all Hamiltonians considered here are invariant under the time reversal mapping. The third line uses (9) to obtain the relationship between $d\Gamma_0(\tau)$ and $d\Gamma_0(0)$ and also uses the equations of motion to relate $H_2(\mathbf{\Gamma}_0(\tau))$ to $H_1(\mathbf{\Gamma}_0(0))$. Line 4 involves simple algebraic manipulation, as does line 5.

This non-equilibrium work relation was first derived by Crooks [12] for stochastic transitions. We refer to equation (11) as a non-equilibrium free energy theorem (NEFET). It shows how equilibrium free energy differences, in this case the Helmholtz difference $\Delta A \equiv A_2 - A_1$, can be computed by *non-equilibrium* thermodynamic path integrals. Although the paths may be far from equilibrium, it is essential that near both end points sufficient time must be allowed for the establishment of the two equilibrium end states.

From (11) a simple algebraic rearrangement shows that,

$$\int_{-\infty}^{+\infty} dB \Pr_F(\Delta W = B) e^{-\beta B} = e^{-\beta \Delta A} \int_{-\infty}^{+\infty} dB \Pr_R(\Delta W = -B) \quad (12)$$

thus,

$$\langle e^{-\beta \Delta W} \rangle_F = e^{-\beta \Delta A}, \quad (13)$$

where the subscript F denotes the fact that the change in the work function is computed relative to the 'forward direction' (i.e. $1 \to 2$) thus, $\Delta W = W_2 - W_1$. This NEFET (13) was first derived by Jarzynski [11]. Its relationship to the stochastic TFT was first clarified by Crooks [12].

3. Discussion

It is extraordinary that differences in an equilibrium thermodynamic state function can be computed from sets of *non-equilibrium* thermodynamic path integrals. These differences are independent of the actual non-equilibrium pathways. The two equilibrium thermodynamic states could be connected by pathways other than the linear Hamiltonian transformation given by (3) and (4). In fact an S-shaped pathway would be more efficient than the linear pathway given in (3) and (4). Provided the pathways are continuous and allow the construction of time reversible reaction paths, the final expressions for the NEFETs ((11) and (13)) are unchanged. The NEFETs therefore generalize the concept of path independent state functions, outside the domain of purely equilibrium pathways.

Some comments are required regarding the thermostats. If the NEFETs are meant to describe experimental systems then we need to employ (as above) wall thermostats. While it is true that the Gaussian isokinetic equations are 'unnatural', the Gaussian thermostats can, as we have argued before [9], be embedded in walls that are arbitrarily remote from the physical system of interest. If this is the case, then it is clear that there is no way that the system of interest can 'know' whether the thermostatting is due to a Gaussian isokinetic thermostat, a Nosé–Hoover thermostat [13], or whether (in those remote walls) there is simply some material with a very large heat capacity. In this way the Gaussian isokinetic thermostat is a convenient but ultimately irrelevant mathematical device.

On the other hand if the NEFETs are to be used in a computer simulation to calculate free energy differences, then an homogeneous Gaussian thermostat [13] provides an efficient and easy way to allow the thermostatted transition to occur.

Finally we point out that these NEFETs can easily be generalized to handle other transitions (isoenergetic, isobaric, etc.) In fact the two equilibrium end states do not have to have common values for any thermodynamic properties.

We wish to acknowledge D. J. Searles and E. M. Sevick for their useful comments. We also thank that Australian Research Council for financial support.

References

[1] Evans, D. J., Cohen, E. G. D., and Morriss, G. P., 1993, *Phys. Rev. Lett.*, **71,** 2401.
[2] Evans, D. J., and Searles, D. J., 1994, *Phys. Rev.* E, **50,** 1645; 1995, *ibid.*, E **52,** 5839; 1996, *ibid.*, E **53,** 5808.
[3] Searles, D. J., and Evans, D. J., 1999, *Phys. Rev.* E, **60,** 159; 2000, *J. chem. Phys.*, **112,** 9727; 2000, *ibid.*, **113,** 3503; 2000.
[4] Searles, D. J., and Evans, D. J., 2001, *Int. J. Thermophys.*, **22,** 123; Ayton, G., Evans, D. J., and Searles, D. J., 2001, *J. chem. Phys.*, **115,** 2033.
[5] Evans, D. J., Searles, D. J., and Mittag, E., 2001, *Phys. Rev.* E, **63,** 051105; Mittag, E., Searles, D. J., and Evans, D. J., 2002, *J. chem. Phys.*, **116,** 6879.
[6] Gallavotti, G., and Cohen, E. G. D., 1995, *J. statist. Phys.*, **80,** 931; 1995, *Phys. Rev. Lett.*, **74,** 2694.
[7] Kurchan, J., 1998, *J. Phys.* A, **31,** 3719.

[8] LEBOWITZ, J.L., and SPOHN, H., 1999, *J. statist. Phys.*, **95,** 333.
[9] EVANS, D. J., and SEARLES, D. J., 2002, *Adv. Phys.*, **51,** 1529.
[10] WANG, G. M., SEVICK, E., MITTAG, E., SEARLES, D. J., and EVANS, D. J., 2002, *Phys. Rev. Lett.*, **89,** 050601.
[11] JARZYNSKI, C., 1997, *Phys. Rev. Lett.*, **78,** 2690; 1997, *Phys. Rev.* E. **56,** 5018.
[12] CROOKS, G. E., 1998, *J. statist. Phys.*, **90,** 1481; 1999, *Phys. Rev.* E, **60,** 2721; 2000, *ibid.*, **61,** 236; CROOKS, G. E., and CHANDLER, D., 2001, *Phys. Rev.* E, **64,** 026109.
[13] EVANS, D. J., and MORRISS, G. P., 1990, *Statistical Mechanics of Nonequilibrium Liquids* (London: Academic Press), downloadable at: http://rsc.anu.edu.au/~evans/evansmorrissbook.htm