Finite bath fluctuation theorem

Michele Campisi,* Peter Talkner, and Peter Hänggi
Institute of Physics, University of Augsburg, Universitätsstrasse 1, D-86153 Augsburg, Germany
(Received 30 April 2009; revised manuscript received 21 July 2009; published 28 September 2009)

We demonstrate that a finite bath fluctuation theorem of the Crooks type holds for systems that have been thermalized via weakly coupling them to a bath with energy independent finite specific heat. We show that this theorem reduces to the known canonical and microcanonical fluctuation theorems in the two respective limiting cases of infinite and vanishing specific heat of the bath. The result is elucidated by applying it to a two-dimensional hard disk colliding elastically with few other hard disks in a rectangular box with perfectly reflecting walls.

DOI: 10.1103/PhysRevE.80.031145 PACS number(s): 05.20.Gg, 05.70.Ln, 05.40.−a

I. INTRODUCTION

During the last decade a number of fluctuation theorems have been reported in the literatures, which have contributed a good deal to a better understanding of nonequilibrium thermodynamics [1–8]. These can be roughly divided in two categories: steady state fluctuation theorems and transient fluctuation theorems. The former apply to systems in nonequilibrium steady states and give information on the system fluctuations in the asymptotic regime of very large times (see [9–11] for reviews on this topic). The latter apply to systems that are temporarily driven out of equilibrium and give information about the fluctuations of work generated by the driving forces. The most representative example of the latter kind is the Crooks fluctuation theorem [5,6], that applies to systems that are initially in a canonical state. Although the canonical case is by far the most common case, one may need to study situations where the system is initially distributed according to some other statistics, instead. For example the system might be initially a microcanonical state of well defined energy. In this latter case it has been shown that a microcanonical fluctuation theorem of the type of Crooks also exists [12,13]. One naturally then wonders whether the same type of transient fluctuation theorem exists as well for yet other types of statistics.

In this work we focus on the probability distribution function (pdf) that describes the statistics of a subsystem of a total classical *ergodic* system with fixed energy. For the case where the interaction between the subsystem and the rest of the total system (which we will refer to as the *bath*) is weak, and the specific heat of the bath is independent of the energy (as for an ideal gas, or for a bath composed of hard spheres), the derivation of the pdf is a standard problem of statistical mechanics [14,15]. We make no assumptions regarding the size of the bath; in particular we do not assume that it is much larger or smaller than that of the system of interest as assumed in the canonical and microcanonical cases respectively. For this reason we refer to this type of bath as to a *finite heat bath*, and to the statistics of the subsystem as to the *finite bath statistics* [see Eq. (6) below]. For this statistics we show that a fluctuation theorem of the type of Crooks, i.e., a *finite bath fluctuation theorem* holds. This finite bath fluctuation theorem includes the Crooks canonical fluctuation theorem and the microcanonical fluctuation theorem, as the two limiting cases in which the bath specific heat goes to infinity and zero, respectively.

The present work is organized as follows. In Sec. II we review the derivation of finite bath statistics and recall some of its properties. In Sec. III we derive the corresponding finite bath fluctuation theorem, and show that it reduces to microcanonical and canonical fluctuation theorems in the limits of vanishing and infinite baths, respectively. In Sec. IV we apply the theory to a specific example [i.e., a two-dimensional (2D) hard disk elastically colliding with few other hard disks in a box] and test the validity of the finite bath fluctuation theorem, both analytically and numerically. Sec. V contains a discussion of the obtained results. The conclusions are drawn in Sec. VI.

II. FINITE BATH STATISTICS

Let us consider a finite classical Hamiltonian system of total N_{tot} particles and total energy E_{tot} composed of two *weakly interacting* subsystems: the "system of interest" (or simply the system) and the "bath." Assuming that the total system is ergodic, the pdf of the system is given in terms of density of states, $\Omega_B(E)$, of the bath and density of states of the total system, $\Omega_{tot}(E)$, as [16]:

$$\rho(\mathbf{z},\lambda) = \frac{\Omega_B(E_{tot} - H(\mathbf{z},\lambda))}{\Omega_{tot}(E_{tot})} \tag{1}$$

where $\mathbf{z}=(p_1,\ldots,p_s,q_1,\ldots,q_s)$ stands for the $2s$ dimensional phase space point of the system. Here we assume that the (sub)system Hamiltonian $H(\mathbf{z},\lambda)$ may depend on some externally controllable parameter λ (this could be for instance the volume of a vessel that contains the system, or an applied magnetic or electric field). For example, in the case of a bath composed of n hard spheres in three dimensions, it is $\Omega_B(E_B) \propto E_B^{3n/2-1}$ (see Appendix A), and one finds from Eq. (1) [14]:

$$\rho(\mathbf{z},\lambda) = \frac{[E_{tot} - H(\mathbf{z},\lambda)]_+^{3n/2-1}}{\int d\mathbf{z}[E_{tot} - H(\mathbf{z},\lambda)]_+^{3n/2-1}} \tag{2}$$

which is a known result of classical statistical mechanics [17]. The symbol $[x]_+$ is defined as $[x]_+ := x\theta(x)$, with $\theta(x)$

*michele.campisi@physik.uni-augsburg.de

denoting Heaviside step function. Note that, in this case, the specific heat of the bath $C(E_B)$, is energy independent and equal to $3n/2$ [18]. More generally one has the following theorem [19,20]:

Theorem 1 *The system pdf is given by*

$$\rho(\mathbf{z},\lambda)=\frac{[E_{tot}-H(\mathbf{z},\lambda)]_+^{C-1}}{\int d\mathbf{z}[E_{tot}-H(\mathbf{z},\lambda)]_+^{C-1}} \tag{3}$$

if and only if the specific heat of the bath C is energy independent.

Here C is the microcanonical specific heat of the bath, i.e.,

$$C(E_B):=\left(\frac{\partial}{\partial E_B}T_B(E_B)\right)^{-1} \tag{4}$$

where E_B is the energy, and $T_B(E_B):=\Phi_B(E_B)/\Omega_B(E_B)$ is the microcanonical temperature expressed in terms of the phase space volume $\Phi_B(E_B)$ of the bath, with energy below E_B, and the bath density of states $\Omega_B(E_B)=\partial\Phi_B(E_B)/\partial E_B$. In the following of this work we restrict ourselves to the case of energy independent, positive specific heat of the bath, abbreviated as $C(E_B):=C>0$.

Note that the pdfs in Eq. (3) are parametrized via the total system energy E_{tot}. It is however convenient to parametrize the pdfs via a property that pertains to the subsystem only, e.g., its average energy U. This is accomplished by writing $E_{tot}=U+CT$ (here CT represents the average energy of the bath), substituting this expression in Eq. (3), and imposing that $U=\int d\mathbf{z}H(\mathbf{z};\lambda)\rho(\mathbf{z},\lambda)$. This leads to solving the following equation for T, given the average energy U and λ

$$\frac{\int d\mathbf{z}H(\mathbf{z};\lambda)[1-[H(\mathbf{z};\lambda)-U]/(CT)]_+^{C-1}}{\int d\mathbf{z}[1-[H(\mathbf{z};\lambda)-U]/(CT)]_+^{C-1}}=U \tag{5}$$

We shall denote the value of T that satisfies Eq. (5) for given U and λ as $T(U,\lambda)$ [in Appendix B we prove that a solution $T(U,\lambda)$ always exists]. With this function at hand we can parametrize the pdfs in Eq. (3) via the subsystem average energy U and recast them in the form:

$$\rho_C(\mathbf{z};U,\lambda)=\frac{[1-[H(\mathbf{z};\lambda)-U]/[CT(U,\lambda)]]_+^{C-1}}{N_C(U,\lambda)} \tag{6}$$

where $N_C(U,\lambda)$ is the normalization:

$$N_C(U,\lambda)=\int d\mathbf{z}[1-[H(\mathbf{z};\lambda)-U]/[CT(U,\lambda)]]_+^{C-1}. \tag{7}$$

As discussed in Appendix B it is not always possible to invert $T(U,\lambda)$. For sake of simplicity, in this work we shall assume that $T(U,\lambda)$ is invertible with respect to the argument U. This means that we could also choose T as an independent parameter and express U as a function of T and λ. Thus we are free to choose between two possible parameterizations: a microcanonical-like parameterization (or U parameterization), and a canonical-like parameterization (or T parameterization) [21].

We shall refer to the numerator in Eq. (6) as to a "generalized Boltzmann factor." It is important to stress that a factor of the type $[1-(H(\mathbf{z};\lambda)-U)/(CT)]_+^{C-1}$ is a generalized Boltzmann factor only if $T=T(U,\lambda)$, in agreement with Eq. (5).

A. Remark

By expressing the specific heat C as $C=1/(1-q)>0$ one recognizes that the pdf in Eq. (6) is the Tsallis escort pdf of index q with $q<1$ [22]. Note that these *do not* exhibit heavy tails but rather have a *faster than exponential* decay with a finite cutoff occurring at the energy $U+CT=E_{tot}$. The physical meaning of this cutoff energy is that the system's energy cannot be larger than the total energy.

B. Properties

1. Equipartition

The following equipartition theorem holds for the finite bath statistics in Eq. (6) [22]:

$$\left\langle p_i\frac{\partial H}{\partial p_i}\right\rangle=T(U,\lambda) \tag{8}$$

where $\langle\cdot\rangle$ denotes average over ρ_C in Eq. (6), p_i is one of the momenta and repeated indices are not summed. Equation (8) says that $T(U,\lambda)$ can be interpreted as the temperature of the system.

2. Heat theorem

The finite bath statistics provides a *mechanical model of thermodynamics* [23], meaning that the temperature T, the external parameter λ, its conjugated generalized force f_λ, and the average energy U are related in such a way as to satisfy the *heat theorem* [24]:

$$\frac{dU+f_\lambda d\lambda}{T}=\text{exact differential} \tag{9}$$

where f_λ is defined in the usual way as:

$$f_\lambda=-\left\langle\frac{\partial H}{\partial\lambda}\right\rangle. \tag{10}$$

This property is an important one because it allows to determine the thermodynamic entropy associated with the finite bath statistics by finding the integral of the exact differential. This is given by [24]:

$$S_C(U,\lambda)=\ln N_C(U,\lambda). \tag{11}$$

3. Interpolation

The pdfs in Eq. (6) interpolate between canonical and microcanonical ensembles. Using the limits of infinite and null specific heat C, i.e.,

$$\lim_{C\to\infty}\left[1+\frac{x}{C}\right]_+^{C-1}=\lim_{C\to\infty}\left[1+\frac{x}{C}\right]_+^{C}=e^x; \tag{12}$$

$$\lim_{C\to\infty}\left[1+\frac{x}{C}\right]_+^C=\theta(x); \tag{13}$$

$$\lim_{C\to\infty}\left[1+\frac{x}{C}\right]_+^{C-1}=\delta(x) \tag{14}$$

one recovers the canonical and microcanonical pdfs [25]:

$$\lim_{C\to\infty}\rho_C(\mathbf{z};T,\lambda)=\frac{e^{-H(\mathbf{z};\lambda)/T}}{Z(T,\lambda)} \tag{15}$$

$$\lim_{C\to 0}\rho_C(\mathbf{z};U,\lambda)=\frac{\delta(U-H(\mathbf{z};\lambda))}{\Omega(U,\lambda)}, \tag{16}$$

respectively [20]. The microcanonical normalization $\Omega(U,\lambda)$ is the system density of states. Likewise one has, for the normalization, the following limits [20]:

$$\lim_{C\to\infty}N_C(T,\lambda)=\int d\mathbf{z}e^{-(H(\mathbf{z};\lambda)-U)/T}=e^{U/T}Z(T,\lambda) \tag{17}$$

$$\lim_{C\to 0}N_C(U,\lambda)=\int_{H(\mathbf{z};\lambda)\le U}d\mathbf{z}=\Phi(U,\lambda). \tag{18}$$

The quantity $\Phi(U,\lambda)$ is the volume of system phase space with energy below U. The density of states is related to Φ via a partial derivative $\Omega=\partial\Phi/\partial U$. By taking the logarithm one recovers canonical and microcanonical entropies; i.e.,

$$\lim_{C\to\infty}S_C(T,\lambda)=\frac{U}{T}+\ln Z(T,\lambda) \tag{19}$$

$$\lim_{C\to 0}S_C(U,\lambda)=\ln\Phi(U,\lambda). \tag{20}$$

III. FLUCTUATION THEOREM

Consider an ensemble of systems distributed according to Eq. (6). Assume the system being decoupled from its bath and that it is acted upon by an external force that changes the external parameter λ according to some prescribed protocol $\lambda(t)$ executed between times t_0 and t_f. The probability density that the external force does a certain work W on the system in that interval of time reads:

$$p_{t_f,t_0}^{C,U}(W):=N_{C,0}^{-1}(U)\int d\mathbf{z}_0\delta(H_f(\mathbf{z}_f)-H_0(\mathbf{z}_0)-W)\times\left[1-\frac{H_0(\mathbf{z}_0)-U}{CT_0(U)}\right]_+^{C-1} \tag{21}$$

where $\mathbf{z}_f=\mathbf{z}(t_f,t_0,\mathbf{z}_0)$ is the solution of Hamilton's equation with initial condition $\mathbf{z}_0$. For simplicity of notation we drop the variable λ in all quantities that depend on it, and replace it with a subscript 0 or f, depending on whether the quantity is taken at values of λ equal to $\lambda(t_0)$ or $\lambda(t_f)$, e.g., $H_0(\mathbf{z}):=H[\mathbf{z},\lambda(t_0)]$, $T_0(U):=T[U,\lambda(t_0)]$. By making the change of variables from $\mathbf{z}_0\to\mathbf{z}_f$ with a unitary Jacobian, one obtains

$$N_{C,0}(U)p_{t_f,t_0}^{C,U}(W)=\int d\mathbf{z}_f\delta(H_0(\mathbf{z}_0)-H_f(\mathbf{z}_f)+W)\times\left[1-\frac{H_f(\mathbf{z}_f)-(U+W)}{CT_0(U)}\right]_+^{C-1} \tag{22}$$

where now $\mathbf{z}_0=\mathbf{z}(t_0,t_f,\mathbf{z}_f)$, is the solution of Hamilton's equation with $\mathbf{z}_f$ as initial condition and time running backward. Note that the second term in the integrand *is not* a generalized Boltzmann factor because in general it does not satisfy Eq. (5). However for any δT one can rewrite the previous equation as:

$$N_{C,0}(U)p_{t_f,t_0}^{C,U}(W)=\left(\frac{T_0(U)+\delta T}{T_0(U)}\right)^{C-1}\times\int d\mathbf{z}_f\delta(H_0(\mathbf{z}_0)-H_f(\mathbf{z}_f)+W)\times\left[1-\frac{H_f(\mathbf{z}_f)-(U+W-C\delta T)}{C(T_0(U)+\delta T)}\right]_+^{C-1}. \tag{23}$$

We now choose δT as the solution of the following integral equation:

$$\frac{\int d\mathbf{z}H_f(\mathbf{z})B(\mathbf{z},U,W,\delta T)}{\int d\mathbf{z}B(\mathbf{z},U,W,\delta T)}=U+W-C\delta T \tag{24}$$

where, for convenience, we use the notation

$$B(\mathbf{z},U,W,\delta T):=\left[1-\frac{H_f(\mathbf{z})-(U+W-C\delta T)}{C(T_0(U)+\delta T)}\right]_+^{C-1}; \tag{25}$$

or, equivalently, as a solution of

$$T_0(U)+\delta T=T_f(U+W-C\delta T). \tag{26}$$

Then, we find

$$N_{C,0}(U)p_{t_f,t_0}^{C,U}(W)=\left[\frac{T_f(U+W-C\delta T)}{T_0(U)}\right]^{C-1}\times\int d\mathbf{z}_f\delta[H_0(\mathbf{z}_0)-H_f(\mathbf{z}_f)+W]\times\left[1-\frac{H_f(\mathbf{z}_f)-(U+W-C\delta T)}{T_f(U+W-C\delta T)}\right]_+^{C-1} \tag{27}$$

where the second term of the integrand is the Boltzmann factor of the pdf $\rho_C(\mathbf{z};U+W-C\delta T,\lambda(t_f))$. The integral is the product of $N_{C,f}(U+W-C\delta T)$ and the probability $p_{t_0,t_f}^{C,U+W-C\delta T}(-W)$ that the force performs the work $-W$ when the protocol is run backward and the system is initially in the state $\rho_C(\mathbf{z};U+W-C\delta T,\lambda(t_f))$.

Therefore the following fluctuation theorem is obtained:

$$\frac{p^{C,U}_{t_f,t_0}(W)}{p^{C,U_f}_{t_0,t_f}(-W)} = \left(\frac{T_f}{T_0}\right)^{C-1} \frac{N_{C,f}(U_f)}{N_{C,0}(U)}, \tag{28}$$

where

$$U_f := U + W - C\delta T \tag{29}$$

$$T_f := T_f(U_f). \tag{30}$$

Using Eqs. (11) and (28) can be rewritten in terms of entropy as:

$$\frac{p^{C,U}_{t_f,t_0}(W)}{p^{C,U_f}_{t_0,t_f}(-W)} = \left(\frac{T_f}{T_0}\right)^{C-1} \exp[\Delta S^{f,0}_C(U,W)] \tag{31}$$

where $\Delta S^{f,0}_C(U,W)=S_{C,f}(U_f)-S_{C,0}(U)$.

The finite bath fluctuation theorem of Eq. (31) allows to calculate the ratios of probability of work done on the system when it is driven arbitrarily away from equilibrium during the action of the forward and backward protocol, in terms of equilibrium properties such as entropy and temperature.

Recovering known special cases

1. Limit of microcanonical ensemble

In the limit $C\to 0$ Eq. (22) becomes [using the formula $\delta(ax)=a^{-1}\delta(x)$, and Eqs. (14) and (18)]

$$\Phi_0(U)p^{C,U}_{t_f,t_0}(W) = T_0(U)\int d\mathbf{z}_f \delta(H_0(\mathbf{z}_0) - H_f(\mathbf{z}_f) + W) \times \delta(H_f(\mathbf{z}_f) - (U+W)). \tag{32}$$

Using the microcanonical equipartition theorem [16] $T(U,\lambda)=\Phi(U,\lambda)/\Omega(U,\lambda)$, one recovers the microcanonical fluctuation theorem [12,13]:

$$\frac{p^{0,U}_{t_f,t_0}(W)}{p^{0,U+W}_{t_0,t_f}(-W)} = \frac{\Omega_f(U+W)}{\Omega_0(U)}. \tag{33}$$

Alternatively one can take the limit $C\to 0$ of Eq. (28) directly and obtain the expression $T_0(U)\Phi_f(U+W)/[T_f(U+W)\Phi_0(U)]$, which reduces to the previous one by virtue of the microcanonical equipartition theorem.

2. Limit of canonical ensemble

Likewise, using the T parameterization, it can be seen that, in the limit $C\to\infty$ Eq. (22) becomes

$$Z_0(T)p^{C,T}_{t_f,t_0}(W) = e^{W/T}\int d\mathbf{z}_f \delta(H_0(\mathbf{z}_0) - H_f(\mathbf{z}_f) + W)e^{-H_f(\mathbf{z}_f)/T} \tag{34}$$

One thus obtains the fluctuation theorem for the canonical ensemble of Crooks [5,6]

$$\frac{p^{\infty,T}_{t_f,t_0}(W)}{p^{\infty,T}_{t_0,t_f}(-W)} = \frac{Z_f(T)}{Z_0(T)}e^{W/T}. \tag{35}$$

IV. EXAMPLE: A 2D GAS OF HARD DISKS

In this section we illustrate the finite bath fluctuation theorem by applying it to a system composed of $n+1$ elastically colliding hard disks in a two-dimensional box with perfectly reflecting walls. One disk will be our system of interest, whereas the remaining n ones will form the bath. We assume that the disks do not have rotational degrees of freedom. As shown in the Appendix A, the specific heat is given in this case by $C=dn/2$ where d is the number of translational degrees of freedom of each disk. In this case $d=2$, hence $C=n$. Note the fact that C does not depend on energy.

A. Probability density function

The energy of the system of interest is simply its kinetic energy; i.e.,

$$H(p_x,p_y;M) = \frac{p_x^2+p_y^2}{2M}, \tag{36}$$

which fluctuates permanently due to the collisions with the bath's particles. According to Eq. (6), the probability that the disk has a given momentum (p_x,p_y) is given by

$$\rho_C(p_x,p_y;U,M) = N_C^{-1}(U,M)\left[1 - \frac{(p_x^2+p_y^2)/(2M) - U}{CT(U,M)}\right]_+^{C-1}. \tag{37}$$

We consider the mass of the disk M as an external parameter that can be changed at will in the course of time according to prespecified protocols. The function $T(U,M)$ has to be computed via Eq. (5). In general, the solution of Eq. (5) with a purely kinetic Hamiltonian with s translational degrees of freedom gives the usual equipartition of energy [22]: $T(U,M)=2U/s$. In the specific case of Eq. (36) $s=2$, hence

$$T(U,M) = U, \tag{38}$$

and

$$\rho_C(p_x,p_y;U,M) = N_C^{-1}(U,M)\left[1 - \left(\frac{p_x^2+p_y^2}{2M} - U\right)/(CU)\right]_+^{C-1}. \tag{39}$$

Using Eq. (7), with Eq. (39) gives

$$N_C(U,M) = 2\pi A[1+C^{-1}]^C MU \tag{40}$$

where A is the reduced volume (i.e., area in this two-dimensional case) of the box (see the Appendix A for the definition of reduced volume). From Eq. (39), one obtains the pdf of energy E of the disk:

$$p(E;U) = U^{-1}[1+C^{-1}]^{-C}\left[1 - \frac{(E-U)}{CU}\right]_+^{C-1}. \tag{41}$$

Interestingly, the energy pdf does not depend on the mass M. In Fig. 1 we compare Eq. (41), with the result of various numerical simulations with $C=1,2,3,4$. Note that for $C=1$ the distribution is flat, for $C=2$ it is linear, for $C=3$ it is quadratic etc. In view of theorem 1, the impressive agreement between theory and numerics corroborates the validity

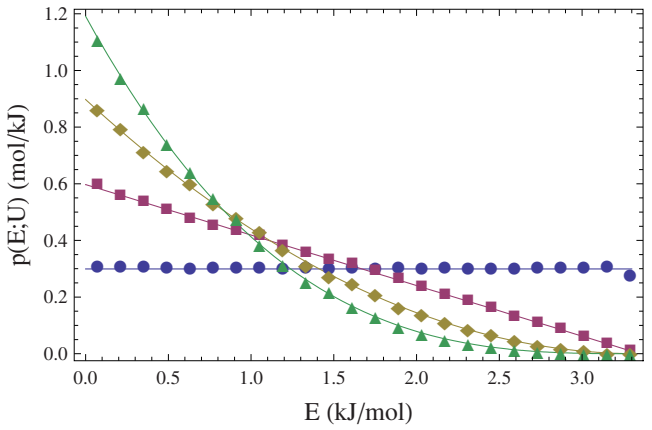

FIG. 1. (Color online) Energy pdf for a 2D hard disk of radius $r=1$ nm and mass $M=2$ amu, in a bath composed of 1(●), 2(■), 3(◆), 4(▲) other identical disks. The dots represent histograms of properly normalized relative frequencies from numerical simulations. All simulations were carried out for the same total energy $E_{tot}=3.3469$ kJ/mol, which corresponds to measured average energies of the disk of interest $U_1=1.67166$ kJ/mol, $U_2=1.11644$ kJ/mol, $U_3=0.835784$ kJ/mol, $U_4=0.671521$ kJ/mol. The solid lines represent the pdf predicted by the theory [Eq. (41)] for the measured average energies U_i, $i=1\ldots4$.

of the assumed ergodic hypothesis for this model system. Similar simulations have been reported in [26] for a one-dimensional harmonic oscillator coupled to a bath of n one-dimensional quartic oscillators. In that case the density of states of the bath is proportional to $E^{(3n-2)/4}$, and accordingly the specific heat, $C=(3n+2)/4$, is energy independent.

B. Analytical test of the finite bath fluctuation theorem

Consider a protocol $M(t)$ that changes the mass of the disk from the value $M_0=M(t_0)$ to $M_f=M(t_f)$. According to the general assumption of our derivation, the system is decoupled from the bath during the action of the protocol. We are interested in checking the validity of Eq. (28). To this end we need to compute the forward pdf of work, $p^{C,U}_{t_f,t_0}(W)$, the backward pdf of work $p^{C,U_f}_{t_0,t_f}(-W)$, and the starting average energy of the backward protocol U_f, given the starting average energy of the forward protocol U. Solving Eq. (26) with Eq. (38) [note that Eq. (38) does not depend on the value of M, hence $T_f(U)=T_0(U)=U$] we arrive at:

$$\delta T = W/(1+C) \tag{42}$$

hence from Eq. (29) we obtain

$$U_f = T_f = U + W/(1+C). \tag{43}$$

Using Eq. (40) with Eq. (43) we obtain the normalizations of the equilibrium pdfs with average energy and external parameters (U,M_0) and (U_f,M_f), respectively,

$$N_{C,0}(U) = 2\pi A[1+C^{-1}]^C M_0 U \tag{44}$$

$$N_{C,f}(U_f) = 2\pi A[1+C^{-1}]^C M_f\left(U+\frac{W}{1+C}\right). \tag{45}$$

Using Eqs. (43)–(45) we find

$$\left(\frac{T_f}{T_0}\right)^{C-1}\frac{N_{C,f}(U_f)}{N_{C,0}(U)} = \left(\frac{U_f}{U}\right)^C\frac{M_f}{M_0}. \tag{46}$$

From Eq. (21) we have:

$$p^{C,U}_{t_f,t_0}(W) = N^{-1}_{C,0}(U)A\int dp_x dp_y \delta\left(\frac{p_x^2+p_y^2}{2M_f}-\frac{p_x^2+p_y^2}{2M_0}-W\right) \times\left[1-\left(\frac{p_x^2+p_y^2}{2M_0}-U\right)/(CU)\right]_+^{C-1} \tag{47}$$

where we use the fact that the momentum (p_x,p_y) is a constant of motion. By applying the change of variable $E=(p_x^2+p_y^2)/(2M_0)$, and employing Eq. (44) we obtain

$$p^{C,U}_{t_f,t_0}(W) = U^{-1}[1+C^{-1}]^{-C}\frac{M_f}{|M_0-M_f|} \times\left[1-\left(\frac{M_f}{M_0-M_f}W-U\right)/(CU)\right]_+^{C-1}. \tag{48}$$

Similarly one finds the backward pdf of work

$$p^{C,U_f}_{t_0,t_f}(-W) = U_f^{-1}[1+C^{-1}]^{-C}\frac{M_0}{|M_f-M_0|} \times\left[1-\left(\frac{M_0}{M_0-M_f}W-U_f\right)/(CU_f)\right]_+^{C-1}. \tag{49}$$

Taking the ratio of Eq. (48) and Eq. (49) we obtain:

$$\frac{p^{C,U}_{t_f,t_0}(W)}{p^{C,U_f}_{t_0,t_f}(-W)} = \left(\frac{U_f}{U}\right)^C\frac{M_f}{M_0}. \tag{50}$$

By comparison with Eq. (46) we see that the finite bath fluctuation theorem of Eq. (28) is satisfied.

C. Numerical check of the finite bath fluctuation theorem

In order to check numerically the validity of Eq. (50) we simulated the forward work pdf $p^{C,U}_{t_f,t_0}(W)$ for a bath of $n=C$ 2D disks, a given value of U and a protocol that changes the mass of the disk from M_0 to $M_f=2M_0$. The pdf for the numerical work is calculated as follows. We first run a simulation of the motion of the disk with fixed U and M_0. We then construct a histogram that counts the number of occurrences of energy in the intervals $I_n=[E_n-\Delta E/2,E_n+\Delta E/2)$ for a certain ΔE (in our simulations, typically, $\Delta E=0.1$ kJ/mol, for a total of about 20 intervals and the histogram counts a total of about 10^5 events). This provides us with the starting statistics. At this point, we note that, independent of the functional form of $M(t)$, acting the protocol on a particle with energy E gives with probability 1 the work $W=E(M_0-M_f)/M_f$. The reason is that the time dependent system Hamiltonian $(p_x^2+p_y^2)/[2M(t)]$ generates the following equation of motion for the momenta: $\dot{p}_x=\dot{p}_y=0$. Hence $E(t_f)=(p_x^2+p_y^2)/[2M(t_f)]=E(t_0)M_0/M_f$ regardless of the details of the protocol. So we immediately obtain a count of work belonging to the intervals $J_n=[W_n-\Delta W,W_n+\Delta W)$, where $W_n=E_n(M_0-M_f)/M_f$ and $\Delta W=\Delta E(M_0-M_f)/M_f$. After proper

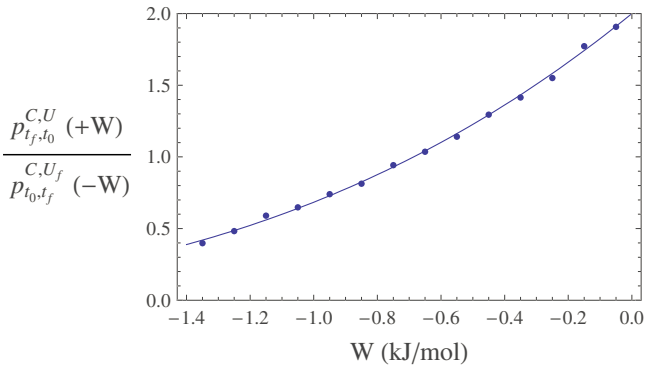

FIG. 2. (Color online) Comparison between the numerical values (dots) and the theoretical expression in Eq. (50) (continuous line) of $p^{C,U}_{t_f,t_0}(W)/p^{C,U_f}_{t_0,t_f}(-W)$ for a 2D hard disk of mass $M_0=2$ amu in a bath composed of three hard disks of the same mass. The initial energy is $U=0.831447$ kJ/mol and the protocol doubles the mass of the disk.

normalization, this yields a histogram, labeled

$$h^{C,U}_{t_f,t_0}(n)$$

that provides a numerical estimate for $p^{C,U}_{t_f,t_0}(W)$. Next, for each n, we simulate the motion of the disk with fixed parameters, $M_f=2M_0$ and $U_n=U+W_n/(C+1)$, and compute n different histograms for the backward probabilities $h^{C,U_n}_{t_0,t_f}(k)$ in the same way as the forward histogram was computed. By selecting the $k=n$ value from each of the backward histograms and collecting them to form the new histogram

$$h^{C,U_n}_{t_0,t_f}(n)$$

we obtain a numerical estimate for $p^{C,U_f}_{t_0,t_f}(-W)$. Finally, we compute the ratios $h^{C,U}_{t_f,t_0}(n)/h^{C,U_n}_{t_0,t_f}(n)$.

These ratios are depicted in Fig. 2 along with the theoretical values given by Eq. (50). The figure shows excellent agreement between analytical theory and numerical experiment. The visible differences are within the statistical errors. Note that, for the forward protocol, where the mass is increased by a factor 2, the work can only be negative and vice-versa for the backward protocol. Therefore, the graph shows only the negative values of nonequilibrium work W.

V. DISCUSSION

A. Physical meaning of δT

The basic quantity that enters the finite bath fluctuation theorem, and marks a distinction with the canonical fluctuation theorem of Crooks Eq. (35), is the quantity δT, defined formally as the solution of Eq. (26). This quantity enters in the definition of U_f and T_f. What is the physical meaning of these quantities? The hard sphere gas example turns useful in addressing this question. Calculations analogous to those leading to Eq. (42) show that for a gas of hard spheres with a total of s degrees of freedom, in contact with a bath with a specific heat C, it is

$$\delta T = W/C_{tot} \tag{51}$$

where C_{tot} is the total specific heat of the system+bath compound system: $C_{tot}:=s/2+C$. This δT is therefore the increment of temperature that would result if, after having injected the energy W in the system of interest this is brought back into contact with the bath and the compound system is let reach thermal equilibrium. Recall that during the forcing protocol we assumed that system and bath are decoupled. We shall refer to this process as to the *rethermalization*. After system and bath have rethermalized, the extra energy W, initially stored in the system, will be shared between system and bath according to the ratio of the respective specific heats. In particular the bath gets the energy $Q=C\delta T$, which is indeed the heat that flows from the system to the bath during rethermalization. Accordingly the system looses this amount of energy and its change in energy becomes $\Delta U=W-Q$, in agreement with the first law of thermodynamics. This means that U_f represents the average energy of the system after the rethermalization. To summarize: (a) the system is first in thermal contact with the bath. Its average energy is U_i and the temperature is T_i. (b) the system is decoupled from the bath and the forcing protocol is acted on it. As a result, the energy W is injected in the system with a certain probability density $p^{C,U}_{t_f,t_0}(W)$. (c) The system (carrying the extra energy W), and bath (still at temperature T_i) are now allowed to rethermalize. During rethermalization the heat $C\delta T$ flows in the bath, the system reaches the average energy U_f, and the new temperature T_f is reached in the compound system.

Remarkably, the temperature change δT vanishes in the canonical case: $\lim_{C\to\infty}\delta T=0$. However it is $\lim_{C\to\infty} C\delta T=W$, meaning that the whole extra energy W injected in the system, flows into the bath during rethermalization. However this does not affect its temperature (i.e., $T_i=T_f$), the specific heat being infinite in the canonical case. Therefore the term T_f/T_0 does not appear in the canonical fluctuation theorem of Crooks. In fact the latter gives information about the free-energy difference of two states with different parameter values, but *same* temperature. This is a much more fortunate situation as compared to the finite bath and microcanonical fluctuation theorems, in the sense that, in the canonical case, one should not bother to start the backward process from the "target" temperature T_f (which depends on W), but simply starts it from the same temperature as that of the forward process.

B. Implications for the second law of thermodynamics

From the canonical fluctuation theorem of Crooks, one obtains, after proper algebraic manipulations, and integration over W, the integral form of the fluctuation theorem, namely the Jarzynski equality $\langle e^{-\beta W}\rangle=e^{-\beta\Delta F}$ [3], which implies the second law in the form $\langle W\rangle\geq\Delta F$. A similar integral equation can be obtained for the finite bath fluctuation theorem too. It reads

$$\mathcal{N}\langle T_f^{C-1}e^{S_f(U_f)}\rangle = T_0^{C-1}e^{S_0(U_0)} \tag{52}$$

where

$$\mathcal{N} := \int p^{U_f}_{t_0,t_f}(W)dW \tag{53}$$

and $\langle\cdot\rangle$ denotes average over the normalized distribution $q_{t_0,t_f}(W) := p^{U_f}_{t_0,t_f}(W)/\mathcal{N}$. Equation (53) generalizes both the canonical Jarzynski equality and the microcanonical *entropy-from-work theorem* [13,27]. Note that, as for the entropy-from-work theorem, in general it is $\mathcal{N} \neq 1$ because the energy U_f in Eq. (53) is a function of W (see Eq. (29)). As pointed out already in [27], this prevents obtaining the second law directly from the integral form of the fluctuation theorem.

Nevertheless the validity of the second law of thermodynamics for a driving protocol acting on a system that is initially thermalized with a finite bath, can be proved directly without invoking the finite bath fluctuation theorem. To this end it is sufficient to recall the content of two theorems which have been recently reported in the literature [28–30]. According to these theorems, the second law of thermodynamics, in either the minimal work principle form, or the entropy increase form of Clausius, is obeyed whenever the initial phase space pdf $\rho(\mathbf{z})$ is a *decreasing* function of energy, namely $\rho(\mathbf{z}) \geq \rho(\mathbf{z}')$, for every $\mathbf{z}$, $\mathbf{z}'$ such that $H(\mathbf{z}) \leq H(\mathbf{z}')$. This condition is obeyed by the finite bath statistics, if the condition $C \geq 1$ is met (see Eq. (3)). In this regard we notice that this condition only is violated in the extremal case when the bath consists of a single degree of freedom (in which case it is $C=1/2$), or if there is no bath at all ($C=0$, microcanonical case).

The Crooks fluctuation theorem Eq. (35) can be seen as a statement according to which the probability of doing a certain negative work $-W$ during the backward protocol is *exponentially suppressed* with respect to the probability of doing the positive work W, in the forward protocol. For a cyclic protocol, this says that it is exponentially more probable to spend energy, rather harvesting it, in agreement with the Kelvin postulate (i.e., no energy extraction from a cyclic process). A similar situation occurs for the finite bath fluctuation theorem, with the exponential suppression being replaced by a power-law suppression. To exemplify this, consider again the gas of N hard spheres in d dimensions. Imagine the protocol consists of changing the volume of the box that contains the gas from V_0 to V_f. Straightforward calculations lead the following form of the finite bath fluctuation theorem

$$\frac{p^{C,U}_{t_f,t_0}(W)}{p^{C,U_f}_{t_0,t_f}(-W)} = \left(\frac{V_f}{V_i}\right)^{N/d}\left(1+\frac{W}{C_{tot}T_0}\right)^{C_{tot}-1} \tag{54}$$

where it is evident that the power-law term $[1+W/(C_{tot}T_0)]^{C_{tot}-1}$ becomes the exponential term appearing in the Crooks theorem Eq. (35) for very large C ($C_{tot}=C+dN/2$ becomes very large for very large C).

VI. CONCLUSIONS

We devised a finite bath fluctuation theorem that gives information about the probability of work on systems that have been thermalized with a finite heat bath. This corresponds to physical situations which are situated between the two ideal cases of absent bath (microcanonical ensemble) and infinite bath (canonical ensemble). The finite bath fluctuation theorem interpolates between microcanonical and canonical fluctuation theorems. It thus generalizes these theorems and reveals a common underlying mathematical structure.

The validity of the finite bath statistics is illustrated by means of numerical simulations of a 2D gas of hard disks in a box with perfectly reflecting walls, see Fig. 1, and the validity of the finite bath fluctuation theorem is confirmed both analytically and numerically, cf. Fig. 2, for our system.

Similarity and differences between the finite bath fluctuation theorem and the canonical and microcanonical fluctuation theorems have been discussed, as well as its interrelation with the second law of thermodynamics. In contrast with the canonical fluctuation theorem, two temperatures, instead of one, appear in the finite bath fluctuation theorem. The physical meaning of these two temperatures has been clarified by considering a rethermalization process.

As shown in Sec. II, the finite bath statistics in Eq. (6) is a special instance of the general statistical formula according to which the bath density of states determines the shape of the system pdf. Based on quasiadiabatic perturbation theory of chaotic systems, Jarzynski [31] found that a slow particle coupled to a small bath with fast chaotic degrees of freedom thermalizes and reaches a stationary pdf whose shape is dictated by the density of states of the bath. Our simulations provide an example that such behavior of the system pdf occurs even if there is no time-scale separation between system and bath. In any case, thermalization of the subsystem toward a pdf of the form in Eq. (6) is expected only if the total system is ergodic.

An important assumption underlying our main finding is that we used a specific heat that is energy independent: whether a finite bath fluctuation theorem exists also in the case of more realistic energy dependent specific heats remains an open challenge.

ACKNOWLEDGMENTS

Financial support by the DFG via the collaborative research center SFB-486, project A10, via the project no. 1517/26–2, the German Excellence Initiative via the *Nanosystems Initiative Munich* (NIM) and the Volkswagen Foundation (project I/80424) is gratefully acknowledged.

APPENDIX A: SPECIFIC HEAT OF A BATH OF n HARD SPHERES

Although straightforward, the calculation of the microcanonical specific heat of a gas of hard spheres is not discussed in statistical mechanics textbooks. We present this calculation below.

The Hamiltonian of a gas of n d-dimensional hard spheres of radius a reads

$$H_B(\{\vec{p}_i\},\{\vec{q}_i\}) = \sum_{i=1}^{n}\frac{\vec{p}_i^{\,2}}{2m} + \sum_{i<j} V(|\vec{q}_i - \vec{q}_j|), \tag{A1}$$

where $\vec{p}_i$, $\vec{q}_i$ are the d-dimensional momentum and position vectors of the i^{th} sphere, and

$$V(x) = \begin{cases} 0 & x \geq a \\ +\infty & x < a \end{cases} \tag{A2}$$

is the hard-core interaction potential. The phase space volume Φ_B with energy below E_B becomes

$$\Phi_B(E_B) = \int \prod_{i=1}^{n} d\vec{q}_i \int \prod_{i=1}^{n} d\vec{p}_i \times \theta\left(E_B - \sum_{i=1}^{n} \frac{\vec{p}_i^{\,2}}{2m} - \sum_{i<j} V(|\vec{q}_i - \vec{q}_j|) \right), \tag{A3}$$

where each integral in $d\vec{q}_i$ is restricted to the region $\mathcal{V}$, of volume V, of the box. For values of $|\vec{q}_i - \vec{q}_j|$ smaller than a, the integrand vanishes, thus reducing the spatial integration domain to the region $\mathcal{M} \subset \mathcal{V}^n$ where $|\vec{q}_i - \vec{q}_j| > a$, for each couple i, j. In this region the interaction term is zero and one obtains

$$\Phi_B(E_B) = V'^n \int \prod_{i=1}^{n} d\vec{p}_i \theta\left(E_B - \sum_{i=1}^{n} \frac{\vec{p}_i^{\,2}}{2m} \right), \tag{A4}$$

where $V'^n = \int_{\mathcal{M}} \prod_{i=1}^{n} d\vec{q}_i$, is independent of E_B. We shall refer to V' as to the reduced volume. The integration over the momenta then yields [32]

$$\Phi_B(E_B) = A_{dn} (2m)^{dn/2} V'^n E_B^{dn/2} \tag{A5}$$

where $A_N := \pi^{N/2} / \Gamma(N/2+1)$. By differentiating $\Phi_B(E_B)$ with respect to E_B, one finally obtains the density of states of the gas of hard spheres

$$\Omega_B(E_B) = A_{dn} (dn/2) (2m)^{dn/2} V'^N E_B^{dn/2-1}. \tag{A6}$$

The only difference with the density of states of an ideal gas is that the actual volume V is replaced by the reduced volume V'. The temperature $T_B(E_B) = \Phi_B(E_B)/\Omega_B(E_B)$, is given by the same formula as for the ideal gas, i.e., $T_B(E_B) = 2E_B/(dn)$ and so is the specific heat, i.e., $C(E_B) = dn/2$. For simplicity, in Eq. (A1) we neglected the spheres rotational degrees of freedom. These however would add to the total specific heat an energy independent contribution.

APPENDIX B: EXISTENCE AND (NON)UNIQUENESS OF SOLUTIONS OF Eq. (5)

We prove that, given U and λ, it is always possible to find a T such that Eq. (5) is satisfied. For this purpose we define the function

$$I_\lambda(U,T) := \int_0^{CT+U} de \Omega_\lambda(e)(e-U)(CT-e+U)^{C-1} \tag{B1}$$

which is continuous with respect to both U and T. The symbol $\Omega_\lambda(e)$ denotes the density of states of the Hamiltonian $H(\mathbf{z},\lambda)$. Equation (5) can be equivalently expressed as:

$$I_\lambda(U,T) = 0. \tag{B2}$$

For $T=0$ it is

$$I_\lambda(U,0) = \int_0^{U} de \Omega_\lambda(e)(e-U)(U-e)^{C-1} \tag{B3}$$

Since $\Omega_\lambda(e) \geq 0$, and $e - U \leq 0$ in the integration domain, we have

$$I_\lambda(U,0) \leq 0. \tag{B4}$$

On the other hand for $T \gg U/C$, we find

$$I_\lambda(U,T) \simeq \int_0^{CT} de \Omega_\lambda(e)(e-U)(CT-e)^{C-1} \tag{B5}$$

where we neglected the terms U as compared to CT. By making the change of variable $x = CT - e$, and neglecting again the term U as compared to CT, we obtain:

$$I_\lambda(U,T) \simeq \int_0^{CT} dx \Omega_\lambda(CT-x)(CT-x)x^{C-1}. \tag{B6}$$

All three terms forming the integrand are non-negative, hence

$$I_\lambda(U, T \gg U/C) \geq 0. \tag{B7}$$

Thus $I_\lambda(U,T)$ is nonpositive for $T=0$ and non-negative for very large T. This implies, that there must be at least one *non-negative* value of T, for which $I_\lambda(U,T)=0$. Uniqueness, however is not guaranteed.

In a similar way it is also possible to prove that

$$I_\lambda(0,T) \geq 0, \quad I_\lambda(U \gg CT, T) \leq 0 \tag{B8}$$

showing that one can also fix T and find a U such that $I_\lambda(U,T)=0$. Also in this case only existence is guaranteed but not uniqueness.

Examples for which two or more different energies correspond to the same temperature were reported in [33,34] for microcanonical ($C=0$) gases with interparticle interaction of the Lennard-Jones type. These systems undergo a microcanonical phase transition whose signature is the appearance of oscillations in the function $T(U)$, which, therefore, is not invertible (i.e, $U(T)$ is multivalued). These oscillations are expected to appear also if these Lennard-Jones type systems are thermalized by means of a finite bath with specific heat $C>0$. Based on the observation that no oscillation appear in the canonical treatment [34], one expects that the amplitude of these oscillations decreases with increasing C.

[1] D. J. Evans, E. G. D. Cohen, and G. P. Morriss, Phys. Rev. Lett. **71**, 2401 (1993).
[2] G. Gallavotti and E. G. D. Cohen, Phys. Rev. Lett. **74**, 2694 (1995).
[3] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[4] C. Jarzynski, C. R. Phys. **8**, 495 (2007).
[5] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[6] P. Talkner and P. Hänggi, J. Phys. A: Math. Theor. **40**, F569 (2007).
[7] P. Talkner, E. Lutz, and P. Hänggi, Phys. Rev. E **75**, 050102(R) (2007).
[8] M. Campisi, P. Talkner, and P. Hänggi, Phys. Rev. Lett. **102**, 210401 (2009).
[9] C. Maes, in *Poincaré Séminaire 2003*, edited by J. Balibard, B. Duplantier, and V. Rivasseau (Birkhauser, Basel, 2004), p. 145.
[10] U. M. B. Marconi, A. Puglisi, L. Rondoni, and A. Vulpiani, Phys. Rep. **461**, 111 (2008).
[11] C. Jarzynski, Eur. Phys. J. B **64**, 331 (2008).
[12] B. Cleuren, C. Van den Broeck, and R. Kawai, Phys. Rev. Lett. **96**, 050601 (2006).
[13] P. Talkner, P. Hänggi, and M. Morillo, Phys. Rev. E **77**, 051131 (2008).
[14] C. J. Thompson, *Classical Equilibrium Statical Mechanics* (Oxford University Press, Oxford, 1988).
[15] H. B. Prosper, Am. J. Phys. **61**, 54 (1993).
[16] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics* (Dover, New York, 1949).
[17] Sometimes this pdf is expressed in the system energy space, E, rather than in its phase space $\mathbf{z}$. For example in Ref. [15], the energy pdf of the system is given as: $P(E)=\text{const}(E_{tot}-E)^{3n/2-1}E^{3m/2-1}$. The term $E^{3m/2-1}$ comes from the system's density of states, where the system is assumed to be itself an ideal gas of m particles. Eq. (2) is more general in that the system is not assumed to be ideal.
[18] In this work we adopt the convention of measuring temperature in units of energy. Thus k_B, the Boltzmann constant is equal to 1 and the specific heat is dimensionless.
[19] M. P. Almeida, Physica A **300**, 424 (2001).
[20] M. Campisi, Phys. Lett. A **366**, 335 (2007).
[21] This freedom is known as *duality* [20,22,24] and also occurs for other types of statistical ensembles (e.g., the Gaussian ensemble) [22].
[22] M. Campisi, Physica A **385**, 501 (2007).
[23] G. Gallavotti, *Statistical Mechanics. A Short Treatise* (Springer Verlag, Berlin, 1995).
[24] M. Campisi and G. B. Bagci, Phys. Lett. A **362**, 11 (2007).
[25] In taking the canonical limit we use the T parameterization.
[26] A. B. Adib, J. Stat. Phys. **117**, 581 (2004).
[27] M. Campisi, Phys. Rev. E **78**, 012102 (2008).
[28] M. Campisi, Phys. Rev. E **78**, 051123 (2008).
[29] M. Campisi, Stud. Hist. Philos. Mod. Phys. **39**, 181 (2008).
[30] A. E. Allahverdyan and T. M. Nieuwenhuizen, Phys. Rev. E **71**, 046107 (2005).
[31] C. Jarzynski, Phys. Rev. Lett. **74**, 2937 (1995).
[32] K. Huang, *Statistical Mechanics* (John Wiley & Sons, Singapore, 1963).
[33] S. Hilbert and J. Dunkel, Phys. Rev. E **74**, 011120 (2006).
[34] J. Dunkel and S. Hilbert, Physica A **370**, 390 (2006).