Aurell, Erik

Unified picture of strong-coupling stochastic thermodynamics and time reversals

Published in:
Physical Review E

DOI:
10.1103/PhysRevE.97.042112

Published: 09/04/2018

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Aurell, E. (2018). Unified picture of strong-coupling stochastic thermodynamics and time reversals. Physical Review E, 97(4), 1-10. [042112]. https://doi.org/10.1103/PhysRevE.97.042112

Unified picture of strong-coupling stochastic thermodynamics and time reversals

Erik Aurell[*]

KTH–Royal Institute of Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden
and Departments of Computer Science and Applied Physics, Aalto University, Espoo, FIN-00076 Aalto, Finland

(Received 5 January 2018; published 9 April 2018)

Strong-coupling statistical thermodynamics is formulated as the Hamiltonian dynamics of an observed system interacting with another unobserved system (a bath). It is shown that the entropy production functional of stochastic thermodynamics, defined as the log ratio of forward and backward system path probabilities, is in a one-to-one relation with the log ratios of the joint initial conditions of the system and the bath. A version of strong-coupling statistical thermodynamics where the system-bath interaction vanishes at the beginning and at the end of a process is, as is also weak-coupling stochastic thermodynamics, related to the bath initially in equilibrium by itself. The heat is then the change of bath energy over the process, and it is discussed when this heat is a functional of the system history alone. The version of strong-coupling statistical thermodynamics introduced by Seifert and Jarzynski is related to the bath initially in conditional equilibrium with respect to the system. This leads to heat as another functional of the system history which needs to be determined by thermodynamic integration. The log ratio of forward and backward system path probabilities in a stochastic process is finally related to log ratios of the initial conditions of a combined system and bath. It is shown that the entropy production formulas of stochastic processes under a general class of time reversals are given by the differences of bath energies in a larger underlying Hamiltonian system. The paper highlights the centrality of time reversal in stochastic thermodynamics, also in the case of strong coupling.

DOI: 10.1103/PhysRevE.97.042112

I. INTRODUCTION

Stochastic thermodynamics describes mesoscopic systems which can be controlled individually while also interacting with a surrounding uncontrolled environment, here, for brevity, called a bath. Work done on such systems is, as in classical macroscopic thermodynamics, the total change in energy of the system and the bath during a process. In a general setting this could depend on the bath, but for conservative dynamics where only the system Hamiltonian H_S depends explicitly on time, work defined this way reduces to $\int \partial_t H_S dt$, a functional of the system history only [1–4].

In its standard formulation, stochastic thermodynamics assumes that the energy stored in the coupling between the system and the bath is negligible compared to the system energy. The internal energy change can then be taken to be the change of system energy only, and as work is then a quantity determined by the system history only. Heat can similarly be taken to be the change of bath energy. By itself this is not measurable on the system, but it can be deduced from the system history in many standard models in nonequilibrium physics, in particular, for master equations (for discrete states) and for Langevin equations (for continuous states). Work, heat, and change in internal energy then obey a trajectory-wise first law where all three quantities are measurable functionals of the system history. The theoretical and fundamental interest in stochastic thermodynamics stems to a considerable extent from work and heat also satisfying exact equalities collectively known as fluctuation relations [1,5].

"Strong coupling" refers to the setting where the variations of the energy stored in the coupling between the system and the bath are comparable to or greater than the variations in system energy. It is not obvious if such a change should be counted with the change of bath energy as heat, or if it should be counted with the change of system energy as an internal energy change, or if its variation should somehow be split between the two. In the related quantum problem, internal energy has in fact in different publications been assumed to include none, half, and all of the system-bath interaction energy (for a recent critical discussion, see Ref. [6]). It is therefore not obvious that there is a meaningful trajectory-wise first law in strong-coupling statistical thermodynamics, nor if there are meaningful strong-coupling fluctuation relations. The issue was first raised in Ref. [7] and answered for Jarzynski equality (JE) soon after in Ref. [8], where this fluctuation relation was restated as

$$\langle e^{-\beta\delta W}\rangle_{\mathrm{eq}} = e^{-\beta\Delta\tilde{F}_S}. \tag{1}$$

In the above, $\beta = \frac{1}{k_B T}$ is the inverse temperature, δW is the work, and the average is over realizations starting from a joint equilibrium of the system and the bath. The left-hand side is hence the same as in standard stochastic thermodynamics and measurable on the system alone. The quantity $\tilde{F}_S$ on the right-hand side is, on the other hand, a free energy at mean force [8–14]. This depends on the equilibrium state of the system and the bath together. It is not measurable on the system alone, but has to be deduced by thermodynamic integration, i.e., by following changes in $\tilde{F}_S$ as the temperature or other

[*] eaurell@kth.se

parameters are varied. Importantly, the right-hand side of (1), however, does not depend on the protocol for changing the system energy H_S while the left-hand side does. This shows that there is a meaningful strong-coupling JE, and also that fluctuating strong-coupling work is a meaningful quantity.

Other strong-coupling fluctuation relations have been slower to obtain. In fact, up to the recent proposals in Refs. [15,16], there was no strong-coupling definition of total entropy change in the combined system and bath that would satisfy the integral fluctuation theorem (IFT),

$$\langle e^{-\Delta S_{\mathrm{TOT}}} \rangle = 1. \tag{2}$$

Heat would be related to such a quantity as $\delta Q = \beta(\Delta S_{\mathrm{TOT}} - \Delta S_S)$, where a general definition of ΔS_S, the entropy change of the system, has also been lacking. The proposal of Ref. [15], to be discussed below, was criticized in Ref. [17], where the authors reached the conclusion that open system trajectories only specify work and not heat. Following upon Refs. [15,16], two important steps were later taken in Ref. [18] where the proposal was derived by a time-scale separation argument (coarse graining), and in Ref. [19], where it was related to a time reversal.

The first goal of this paper is to restate the issue of strong-coupling thermodynamics as one of time reversals in a combined system and bath. It will emerge that the entropy production functional of stochastic thermodynamics is equal to the log ratio of probabilities of the initial states in the larger system. Although quite simple, this result was, to the author's knowledge, first explicitly stated quite recently [19].

Entropy production functionals as log ratios lead to fluctuation relations as "tautologies" [20,21]. The second goal of this paper is hence to show that different initial probabilities and different time reversals of a system and a bath lead to different entropy production functionals which all satisfy fluctuation relations. This also gives a different perspective on entropy production and time reversals in stochastic differential systems, whenever these can be seen as the effective dynamics of a system also interacting with a bath.

The paper is organized as follows. In Sec. II, I relate ratios of forward and backward path probabilities of the system to initial probability distributions of the combined system and bath in the forward and backward process. In Sec. III, I discuss three different examples. In the first, the system-bath interaction is assumed to vanish at the beginning and the end of the process, and the bath is initially in equilibrium, while the system state can be arbitrary. This gives an additional term in the work, as recently discussed at length in Ref. [22], but heat is simply the change of bath energy, the same as in weak-coupling statistical thermodynamics. The second example is a reformulation of the approach of Refs. [15,16,18,19] where the bath is initially in conditional equilibrium with respect to the system. The last example is finally, as in the discussion around (1) above, of the case when the system and the bath are assumed initially jointly in equilibrium. This leads to formulas for heat which at first glance look unfamiliar, but which can be reduced to the case of conditional equilibria. In Sec. IV, I consider entropy production and time general reversals in stochastic dynamics when that dynamics results from an interaction with a bath, and show that related entropy production functions equal the differences of bath energies in units of $k_B T$. In Sec. V, I discuss and compare the results. Three Appendixes contain technical details or material which is either standard or already presented elsewhere.

II. FORWARD-BACKWARD PATH PROBABILITIES AND BATHS

I will assume that the system and the bath together are one big closed conservative system. The total Hamiltonian is hence

$$H_{\mathrm{TOT}}(x,y) = H_S(x) + H_I(x,y) + H_B(y), \tag{3}$$

where the three parts refer to the system, the interaction, and the bath, respectively. The phase space of the system is parametrized by x (coordinates and momenta of the system) and the phase space of the bath is parametrized by y (coordinates and momenta of the bath). I will assume either that only H_S depends explicitly on time, or that only H_S and H_I depend explicitly on time. The initial state of the system and the bath is $\rho_i(x^i,y^i)$, where the subscript indicates "initial." Special classes will be considered later.

Let us assume that the system has D degrees of freedom and the bath N degrees of freedom. Observing the system at $n = \frac{N}{D}$ time points $t_1,t_2,\dots,t_n$ should generically give the same information as observing the bath at the initial time t_i. We may therefore postulate an equivalence between the probability distribution ρ_i over initial conditions of the total system, and the joint probability distribution $P^F(x_0,x_1,\dots,x_n)$ of the coordinates and momenta of the system at time points $t_i = t_0,t_1,t_2,\dots,t_n = t_f$. By the law of conservation of probability, this equivalence is

$$P^F(x_0,x_1,\dots,x_n)\prod_{k=0}^{n} dx_k = \rho_i(x^i,y^i)dx^i dy^i. \tag{4}$$

The shift from x^i,y^i to $x_0,x_1,\dots,x_n$ is a change of variables. Equation (4) can therefore also be written

$$P^F(x_0,x_1,\dots,x_n) = \rho_i(x^i,y^i)\left|\frac{\partial\left(\{x_k\}_{k=0}^n\right)}{\partial(x^i,y^i)}\right|^{-1}, \tag{5}$$

where the second term is a Jacobian of the transformation.

Let us now consider a time-reversed process parametrized by a reversed time $t^* = t_f - t$. This process starts at $t_i^* = 0$ $(t = t_f)$ and runs until $t_f^* = t_f - t_i$ $(t = t_i)$. The general concept of time reversal in stochastic thermodynamics was discussed in great detail by Chétrite and Gawędzki in Ref. [23]. I will assume that time reversal is implemented by a functional $\mathcal{I}$ such that the time-reversed coordinates (x^*_{t*},y^*_{t*}) are $\mathcal{I}(x_t,y_t)$ and the time-reversed Hamiltonian $H^*_{t^*}$ is $\mathcal{I}H_t$. Time-reversed dynamics is thus Hamiltonian dynamics in the coordinates (t^*,x^*_{t*},y^*_{t*}) with the Hamiltonian $H^*_{t^*}(x^*_{t*},y^*_{t*})$. The time-reversal functional is assumed to have the following general properties:

Involution: $\mathcal{I}$ is an involution on (x,y,H), i.e., $(\mathcal{I})^2 = \mathbf{1}$.

Separability: $\mathcal{I}$ acts separately on the system, i.e., $[\mathcal{I}(x,y)]_{\mathrm{system}} = \mathcal{I}x$.

Volume preservation: $\mathcal{I}$ separately preserves system phase space volume and bath phase space volume.

A main example which satisfies all the above is standard time inversion of all the generalized coordinates as $q^*_{t*} = q_t$ and the generalized momenta as $p^*_{t*} = -p_t$, and the Hamiltonian, when there is no magnetic field, as $H^*_{t*} = H_t$. When there is a nonzero magnetic field, time reversal can be done by changing the sign of the magnetic field [24], but other time reversals are also possible [25].

Let the initial density of the time-reversed process be ρ^*_i. Then the backward path probability satisfies

$$P^B(x^*_0,\dots,x^*_n)\prod_{k=0}^{n} dx^*_k = \rho^*_i[(x^*)^i,(y^*)^i]d(x^*)^i d(y^*)^i. \quad (6)$$

Combining (4) and (6), one has

$$\frac{P^F(x_0,\dots,x_n)}{P^B(x^*_0,\dots,x^*_n)} = \frac{\rho_i(x^i,y^i)}{\rho^*_i[(x^*)^i,(y^*)^i]}\left|\frac{\frac{\partial\left(\{x^*_k\}_{k=0}^n\right)}{\partial[(x^*)^i,(y^*)^i]}}{\frac{\partial\left(\{x_k\}_{k=0}^n\right)}{\partial(x^i,y^i)}}\right|. \quad (7)$$

The ratio of Jacobians in (7) can be combined with $\frac{\partial(\{x^*_k\})}{\partial(\{x_k\})}$ and $\frac{\partial[(x^*)^i,(y^*)^i]}{\partial(x^f,y^f)}$, both of which have an absolute value of one. The absolute ratio of Jacobians in (7) therefore has the same value as $|\frac{\partial(x^f,y^f)}{(x^i,y^i)}|^{-1}$, which is one, because Hamiltonian dynamics preserves the total phase space volume.

Instead of (7), we therefore have much more simply

$$\frac{P^F(x_0,x_1,\dots,x_n)}{P^B(x^*_0,x^*_1,\dots,x^*_n)} = \frac{\rho_i(x^i,y^i)}{\rho^*_i[(x^*)^i,(y^*)^i]}. \quad (8)$$

Equation (8) says that for classical conservative systems, path probabilities are only consequences of uncertainties in the initial conditions, and the ratios of path probabilities are given by the ratios of probabilities of the initial conditions.

III. SCENARIOS FOR STRONG-COUPLING HEAT

In this section, I will give self-contained descriptions of three scenarios. The scenarios differ only in what is assumed for the initial states $\rho_i(x^i,y^i)$ and $\rho_i[(x^*)^i,(y^*)^i]$. The descriptions end with a summary of what strong-coupling heat has to be in each scenario to satisfy the integrated fluctuation theorem (2).

A. Factorized equilibria with time-dependent system-bath coupling

In standard stochastic thermodynamics the interaction between the system and the bath is weak and the bath is initially in equilibrium by itself. The smallest deviation from this scenario that allows one to treat also strong coupling is to assume that the interaction is time dependent, and vanishing at the beginning and the end of the process. As then both H_S and H_I depend explicitly on time, the work is

$$\Delta H_{\text{TOT}} = \delta W^{(J)} + \delta W_{if} = \int \partial_t H_S dt + \int \partial_t H_I dt. \quad (9)$$

The first term in the above is, as in (1), the Jarzynski work, while the second term was introduced in Ref. [22]. It is a functional of the system history only for some models of the bath and the system-bath interaction. In particular, it is, however, so for the Zwanzig model (Caldeira-Leggett model), which leads to Kramers-Langevin system dynamics [26–28]. A summary with some extensions is given in Appendix A.

The factorized initial conditions, where the bath is in equilibrium, are

$$\rho_i(x^i,y^i) = \rho^i_S(x^i)\rho^{\text{eq}}_B(y^i), \quad (10)$$

where the system state $\rho^i_S(x^i)$ can be anything and

$$\rho^{\text{eq}}_B(y) = e^{-\beta(H_B(y)-F_B)}. \quad (11)$$

There is no dependence on the interaction Hamiltonian in (11) since that has been assumed to vanish at the beginning of the process.

The initial conditions of the backwards process are analogously

$$\rho_i[(x^*)^i,(y^*)^i] = \rho^{i,*}_S[(x^*)^i]\rho^{\text{eq}}_B[(y^*)^i], \quad (12)$$

which give an entropy production

$$\begin{aligned}\Delta S^{(\text{fact.eq.})}_{\text{TOT}} &= \log\frac{P^F(x_0,x_1,\dots,x_n)}{P^B(x^*_0,x^*_1,\dots,x^*_n)}\\ &= \log\rho^i_S - \log\rho^{i,*}_S\\ &\quad + \log\rho^{\text{eq}}_B(y^i) - \log\rho^{\text{eq}}_B[(y^*)^i]. \quad (13)\end{aligned}$$

In Sec. IV and Appendix C, I consider a class of examples where the comparison is made between $\log\rho^{\text{eq}}_B[(y^*)^i]$ and $\log\rho^{\text{eq}}_B(y^i)$ and where $(y^*)^i$ is determined from the whole system path. In a general setting, $(y^*)^i$ will hence not be a simple transformation of y^f only. Assuming here that the equilibrium state of the bath is time-reversal invariant, that is, $\rho^{\text{eq}}_B[(y^*)^i] = \rho^{\text{eq}}_B(y^f)$, which holds for the "canonical" time reversal of Sec. IV, the difference in the last line in (13) is $\beta\Delta H_B$, the change in bath energy in units of k_BT.

If further the initial state of the time-reversed system ($\rho^{i,*}_S$) is identical to the final state of the system going forwards (ρ^f_S), one recognizes in (13) from standard stochastic thermodynamics the stochastic entropy $-\Delta\log\rho$, the negative log change in probability density from an initial position at the initial time to a final position [5]. It is simple to then rewrite (13) as

$$\Delta S^{(\text{fact.eq.})}_{\text{TOT}} = -\Delta\log P + \beta(\delta W^{(J)} + \delta W_{if}) - \beta\Delta H_S. \quad (14)$$

The heat functional in this scenario is thus

$$\delta Q^{(\text{fact.eq.})} = \delta W^{(J)} + \delta W_{if} - \Delta H_S = \Delta H_B. \quad (15)$$

Since the interaction energy has been assumed to vanish at the boundaries, heat is only the change in bath energy during the process, the same as in standard (weak-coupling) stochastic thermodynamics. If $\delta Q^{(\text{fact.eq.})}$ in (15) is a functional measurable on the system alone however, depends on the second term δW_{if} (see Appendix A).

B. Conditional equilibria with time-reversal symmetric states

Next, I turn to the approach of Refs. [15,16,19]. Only H_S now depends explicitly on time, and the work functional is, as in (1), only the Jarzynski work,

$$\Delta H_{\text{TOT}} = \delta W^{(J)} = \int \partial_t H_S dt. \quad (16)$$

The bath is assumed to be initially in equilibrium conditional of the system,

$$\rho_i(x^i,y^i) = \rho^i_S(x^i)\sigma(y^i|x^i), \quad (17)$$

where $\rho_S^i(x^i)$ can be anything and

$$\sigma(y^i|x^i) = \frac{e^{-\beta(H_I(x^i,y^i)+H_B(y^i))}}{\int dy' e^{-\beta(H_I(x^i,y')+H_B(y'))}}. \qquad (18)$$

The initial conditions of the time-reversed process are also such that the bath is in equilibrium conditional to the system, and adopting analogous assumptions to the above (also stated in Ref. [19]), I will assume that the conditional distribution of the bath is time-reversal symmetric. This means

$$\rho_i[(x^*)^i,(y^*)^i] = \rho_S^{i,*}[(x^*)^i]\sigma(y^f|x^f), \qquad (19)$$

with the same conditional probability as in (18). The total entropy change is then

$$\begin{aligned}\Delta S_{TOT}^{\text{(cond.eq.)}} &= \log\frac{P^F(x_0,x_1,\dots,x_n)}{P^B(x_0^*,x_1^*,\dots,x_n^*)}\\ &= \log\rho_S^i - \log\rho_S^{i,*}\\ &\quad + \log\sigma(y^i|x^i) - \log\sigma(y^f|x^f). \qquad (20)\end{aligned}$$

In the same setting as in the previous section, where the initial state of the system going backwards ($\rho_S^{i,*}$) is the same as final state of the system going forwards (ρ_S^f), it was shown in Ref. [15] that (20) can be rewritten as

$$\Delta S_{\text{TOT}}^{\text{(cond.eq.)}} = \Delta\tilde{s}_S + \beta\delta W^{(J)} - \beta\Delta\tilde{u}_S, \qquad (21)$$

where $\tilde{u}_S$ is an energylike function, $\tilde{f}_S$ is the constant in a Gibbs-Boltzmann-like distribution $P^{\text{(cond.eq.)}} = e^{\beta(\tilde{f}_S-\tilde{u}_S)}$, and $\tilde{s}_S = -\log P^{\text{(cond.eq.)}}$ is the corresponding entropylike quantity. For completeness, this derivation is repeated in Appendix B.

The heat functional in this scenario is thus

$$\delta Q^{\text{(cond.eq.)}} = \delta W^{(J)} - \Delta\tilde{u}_S. \qquad (22)$$

As $\tilde{F}_S$ in (1), the quantities $\tilde{f}_S$, $\tilde{u}_S$, and $\tilde{s}_S$ depend on the bath. A parameter variation, i.e., thermodynamic integration, is needed to determine an arbitrary constant in $\tilde{u}_S$ and $\tilde{f}_S$ which would otherwise render (21) and (22) indeterminate.

The explicit form of $\tilde{u}_S$, rederived in Appendix B and stated in (B6), is $H_S - \partial_\beta \log\langle e^{-\beta H_I}\rangle_B$, where $\langle\cdots\rangle_B$ indicates an average with respect to the Gibbs state $e^{\beta(F_B-H_B)}$. The change $\Delta\tilde{u}_S$ hence includes the change in system energy ΔH_S and the change in average both of the bath and interaction energy with respect to a conditional bath Gibbs state $e^{\beta(F'_B-H_B-H_I)}$ (H_I and F'_B depend on the system coordinate). The heat in (22) includes the corresponding fluctuating quantities. It is quite interesting that the proposal in Ref. [15] hence does not reduce to any of the simpler earlier suggestions that counted in the heat some definite fractions of respectively ΔH_B and ΔH_I.

C. Joint equilibrium of the system and the bath

The last scenario adheres closely to the the equilibrium strong-coupling theory and several earlier contributions [8–14]. Of the three terms in (3), again only the system Hamiltonian H_S depends explicitly on time and the work is given by (16). The assumption is now that the bath and the system are jointly in equilibrium at the beginning of the process,

$$\rho_i(x^i,y^i) = \rho_{\text{TOT}}^{\text{eq}}(x^i,y^i) = \frac{1}{Z_{\text{TOT}}^i}e^{-\beta H_{\text{TOT}}^i}. \qquad (23)$$

The initial conditions of the backwards process are analogously taken to be

$$\rho_i[(x^*)^i,(y^*)^i] = \frac{1}{Z_{\text{TOT}}^f}e^{-\beta H_{\text{TOT}}^f}. \qquad (24)$$

From (8) we then have

$$\begin{aligned}\Delta S_{\text{TOT}}^{\text{(tot.eq.)}} &= \log\frac{P^F(x_0,x_1,\dots,x_n)}{P^B(x_0^*,x_1^*,\dots,x_n^*)}\\ &= \beta\delta W^{(J)} + \log Z_{\text{TOT}}^f - \log Z_{\text{TOT}}^i. \qquad (25)\end{aligned}$$

The Jarzynski work is a functional of the system history and gives, for this scenario, all the coordinate dependence. The statistical properties of $\Delta S_{\text{TOT}}^{\text{(tot.eq.)}}$ and $\delta W^{(J)}$ are therefore in this scenario the same.

The last two terms (constants) in (25) can be referred to the total free energy with respect to that of the bath alone,

$$\tilde{F}_S = \frac{1}{\beta}\log\frac{Z_B}{Z_{\text{TOT}}} = F_{\text{TOT}} - F_B, \qquad (26)$$

and are thus the change of a free energy at mean force, as already used in (1) above,

$$\log Z_{\text{TOT}}^f - \log Z_{\text{TOT}}^i = -\beta\Delta\tilde{F}_S. \qquad (27)$$

The free energy at mean force can be written

$$\tilde{F}_S = \tilde{U}_S - \frac{1}{\beta}\tilde{S}_S, \qquad (28)$$

where the internal energy (or potential) at mean force is

$$\tilde{U}_S = \partial_\beta(\beta\tilde{F}_S) = U_{\text{TOT}} - U_B, \qquad (29)$$

and the corresponding entropy is

$$\tilde{S}_S = \beta(\tilde{U}_S - \tilde{F}_S) = S_{\text{TOT}} - S_B. \qquad (30)$$

With these conventions, (25) can be rewritten,

$$\Delta S_{\text{TOT}}^{\text{(tot.eq.)}} = \Delta\tilde{S}_S + \beta\delta W^{(J)} - \beta\Delta\tilde{U}_S, \qquad (31)$$

and the heat functional is

$$\delta Q^{\text{(tot.eq.)}} = \delta W^{(J)} - \Delta\tilde{U}_S. \qquad (32)$$

To compare (32) to (22) we must recognize that the time reversals are qualitatively different. The heat in (22) was derived under the assumption that the initial state of the system in the backward process is the same as the final state of the forward process. This is not the same as in (32) where the initial state of the system in the backward process is the marginal of a total equilibrium state, while the final state of the forward process is generally something else. To compare, we must instead go back to the total entropy productions in (20) and (25) and identify the initial system states of the forward and backward states in (20) as $= e^{\beta(\mathcal{F}_S-\mathcal{H}_S)}$, where $\mathcal{H}_S$ and $\mathcal{F}_S$ are the potential and free energy at mean force of Onsager and

Kirkwood [8–10,13,14]. With this, (20) reduces to (25)

$$\begin{aligned}\Delta S_{\mathrm{TOT}}^{(\text{cond.eq.-red.})} &= \beta\Delta\mathcal{H}_S - \beta\Delta\mathcal{F}_S \\ &= \beta\Delta H_{\mathrm{TOT}} - \beta\Delta H_S + \Delta\log\langle e^{-\beta H_I}\rangle_B, \quad (33)\end{aligned}$$

where in the second line I have used (B3) from Appendix B.

IV. TIME REVERSALS IN STOCHASTIC DYNAMICS

In this section the focus is not on strong coupling. The interaction will hence be taken weak, or assumed to depend on time as in Sec. III A. The focus is instead on using the general result in Sec. II to give a different perspective on time reversals in stochastic dynamics [23]. To lighten the presentation, technical details are given in Appendix C.

It is well known that a Kramers-Langevin equation $\dot{x} = \frac{p}{m}$ and $\dot{p} = -\partial_x V(x,t) - \gamma\frac{p}{m} + \sqrt{2k_B T\gamma}\xi$ can be derived from the total Hamiltonian dynamics of a system interacting linearly with a bath of harmonic oscillators which are initially in thermal equilibrium [26,29,30]. *Complete time reversal* refers to standard time inversion of all the coordinates and momenta, of both the bath and the system. On the level of the system this is a process conditioned by the future, that at the final time the bath will be in equilibrium, and is therefore not a Markov process. It follows immediately from (8) that the entropy production in such a time reversal is zero because the right-hand side of (8) can also be written $\frac{\rho_f(x^f,y^f)}{\rho_i^*[(x^*)^i,(y^*)^i]}$ (preservation of phase space volume) and this ratio is one (time reversal preserves phase space volume). This is logical because when the motion of both the system and the bath is reversed, they will evolve back to their initial state, and no information will be lost.

The closest to complete time reversal defined on the level of the system is *natural time reversal* [23]. This is standard time reversal on the system and transforming the dynamics to $\frac{dx^*}{dt^*} = \frac{p^*}{m}$ and $\frac{dp^*}{dt^*} = -\partial_x V(x,t_f - t) + \gamma\frac{p^*}{m} + \sqrt{2k_B T\gamma}\xi^*$, where ξ^* is a noise with the same statistical properties as ξ. The antifriction ($\gamma\frac{p^*}{m}$) shows that this equation does not originate from the system interacting with a bath initially in thermal equilibrium. In the other direction it was shown in Ref. [23] that the entropy production associated to natural time reversal is $(t_f - t_i)\gamma/m$; natural time reversal is therefore different from complete time reversal. For completeness, a sketch of a derivation of this fact is given Appendix C.

We turn now instead to *canonical time reversal* [23], where the backward process also obeys a Kramers-Langevin with positive friction, $\frac{dx^*}{dt^*} = \frac{p^*}{m}$ and $\frac{dp^*}{dt^*} = -\partial_x V(x,t_f - t) - \gamma\frac{p^*}{m} + \sqrt{2k_B T\gamma}\xi'$, and ξ' again is a noise with the same statistical properties as ξ. It is convenient to consider a wider class of *general time reversals*, introduced in Ref. [23] by splitting the drift field (time derivative of the system coordinate). We split the system potential in two parts that transform differently, $V_t = V_t^+ + V_t^-$ and the time-reversed total Hamiltonian $H_{t^*}^*$ will contain the piece $V_{t^*}^{*,+} - V_{t^*}^{*,-} = V_t - 2V_t^-$. Canonical time reversal is the case when $V^- = 0$. The system equation under such general time reversal is $\frac{dx^*}{dt^*} = \frac{p^*}{m}$ and $\frac{dp^*}{dt^*} = -\partial_x V + 2\partial_x V^- - \gamma\frac{p^*}{m} + \sqrt{2k_B T\gamma}\xi''$. Introducing the notation of Ref. [23] that $u_+ = -\gamma p/m - \partial_x V^-$ is the part of the drift field that transforms as a vector and $u_- = -\partial_x V^+$ is the part that transforms as a pseudovector, and identifying $D = k_B T\gamma$ as the diffusion coefficient, one has

$$\Delta S_{\mathrm{TOT}} = \log\frac{P^F}{P^B} = -\log\Delta P + \int(\dot{p} - u_-)\frac{1}{D}u_+ dt, \tag{34}$$

which is a main result of Ref. [23], adapted to this situation. Using explicit expressions for the dynamics of the continuum of harmonic oscillators that make up the bath, it is, on the other hand, straightforward to show that

$$H_B^*[(y^*)^i] - H_B(y^i) = \int(\dot{p} - u_-)\frac{1}{\gamma}u_+ dt, \tag{35}$$

with the same definitions of u_- and u_+ as above. Detailed derivations of (34) and (35) are given in Appendix C. The entropy production formula under general time reversal is thus, in fact, the energy difference in a microscopic bath model in units of $k_B T$. For canonical time reversal, (35) simplifies to

$$\text{Eq. (35) (canonical reversal)} = \int -\frac{p}{m}dp - \partial_x V dx, \tag{36}$$

where the right-hand side equals the work ($\delta W = \Delta H_{\mathrm{TOT}}$) minus the total change of system energy (ΔH_S). For this time reversal, $H_B^*[(y^*)^i] - H_B(y^i)$ hence equals $H_B(y^f) - H_B(y^i)$, the change in bath energy in the forward process, and $e^{-\beta H_B^*[(y^*)^i]} = e^{-\beta H_B(y^f)}$.

The above examples extend naturally to when the system-bath coupling is nonlinear in the system. As already found in Ref. [26], this leads to a friction term that is nonlinear in the system coordinate and a noise term which satisfies an Einstein relation with the friction term. More recently, perturbative solutions have been found when the bath is weakly anharmonic [31,32]. Although these contributions establish a form of fluctuation-dissipation theorems, they can also be interpreted as showing that naive versions of fluctuation-dissipation theorems do not hold. Hence, at least some general diffusion processes where the noise terms do not satisfy an Einstein relation with the friction term also have representations in terms of explicit baths.

Time reversals in overdamped stochastic systems, where the diffusion tensor D can depend on the coordinate effected by the noise ($dx = \cdots + \sqrt{2D}dW$, $D = k_B T/\gamma$), can be embedded in the underdamped case discussed above ($dx = \frac{p}{m}dt$, $dp = \cdots - \gamma dx + \sqrt{2k_B T\gamma}dW$). When temperature is constant, the overdamped limit gives no new contributions to the entropy production [33]. Entropy production under a general time reversal of an overdamped stochastic with a possibly space-dependent friction coefficient γ can hence also be related to an energy change in a bath, as above.

V. DISCUSSION AND CONCLUSIONS

Entropy production is related to irreversibility and how a system transforms under time reversal. As such, it has long been a fundamental topic in statistical physics [34], where the two central results that hold near to equilibrium are the fluctuation-dissipation theorem and Onsager's relations.

When terms in the dynamical equations of macroscopic quantities can be classified as reversible (conservative) and irreversible (dissipative), results generalizing Onsager's reciprocal relations have been obtained far from equilibrium [35,36]. Connections between time reversal and entropy production have also more recently been made in macroscopic fluctuation theory [37].

This paper has addressed time reversal and entropy production in the context of stochastic thermodynamics where the system is assumed small (mesoscopic or microscopic), the system history is observable, and fluctuations of the quantities which would be constant in the thermodynamic limit are important. I have considered the dynamics of a system interacting with an unobserved environment (bath) in a Hamiltonian framework with an arbitrarily strong system-bath coupling. I have shown that the log ratios of forward and backward path probabilities of a system are related to the log ratios of the initial state of the total system (system and bath) in a forward and backward process.

Depending on what is assumed for the initial state of the bath, one gets different entropy productions for the system. This is not surprising because different initial states of the bath correspond to different levels of control, and time reversal then leads to a different loss of information. Here, I compare (21) and (14). In both cases the initial state of the system can formally be anything. In practice, it is, however, reasonable to assume in the first case either that the system and the bath are jointly in equilibrium (discussed above in Sec. III C), or that the system has been fixed for some time in position x^i so that the bath will have had time to relax to conditional equilibrium $\sigma(y^i|x^i)$. I hence assume that this is the scenario for both the forward and backward process. Using the explicit expression of $\tilde{u}_S$ from (B6), we then have

$$\begin{aligned}\delta Q^{(\text{fact.eq.})} &- \delta Q^{(\text{cond.eq.})} \\ &= \delta W_{if} - \langle H_I + H_B \rangle_{x^f} + \langle H_I + H_B \rangle_{x^i}. \end{aligned} \tag{37}$$

The difference in heat is hence in one part the extra work δW_{if} needed to change the system-bath interaction, and in the other part the change in the expected value of the bath energy and interaction energy, conditioned on the system state. For factorized equilibrium this second term vanishes while for conditional bath equilibrium it is counted in the change of internal energy. The two different forms of strong-coupling heat are hence mutually compatible. The critique of Ref. [17] that strong-coupling heat is not a uniquely defined concept can therefore partly be reformulated as saying that different versions correspond to different physical situations.

Finally, in this work I have shown that the entropy production functional of stochastic thermodynamics applied to diffusive systems defined as the log ratio of path probabilities can be interpreted as the change of bath energy in an underlying, more detailed, microscopic model. This is a different connection between the mathematical and physical notions of entropy production, and a further strong argument in favor of the physical soundness of stochastic thermodynamics.

ACKNOWLEDGMENTS

I thank Raphaël Chétrite for important clarifying remarks on time reversals in stochastic dynamics and Janet Anders, Peter Hänggi, and Udo Seifert for discussions and valuable comments. This research was supported by the Academy of Finland through its Center of Excellence COIN, Grant No. 251170.

APPENDIX A: WORK WITH TIME-DEPENDENT SYSTEM-BATH INTERACTION

This Appendix summarizes the discussion in Ref. [22] of time-dependent strong coupling, with some extensions.

I will now write the system as $x = (Q,P)$ and the bath as $y = (q,p)$, and I will assume that the system and the bath only interact through their generalized coordinates

$$H_{\text{TOT}} = H_S(Q,P,t) + H_I(Q,q,t) + H_B(q,p), \tag{A1}$$

where the explicit time dependencies have been indicated. The equation of motion of the system is

$$\dot{Q} = \partial_P H_S(Q,P,t) \quad \left(\text{typically} = \frac{P}{M}\right) \tag{A2}$$

and

$$\dot{P} = -\partial_Q H_S(Q,P,t) - \partial_Q H_I(Q,q,t). \tag{A3}$$

The second term, which depends on bath coordinate q, is a force acting on the system, conventionally said to be from the bath on the system. For the Zwanzig (Caldeira-Leggett) model, the bath is a collection of harmonic oscillators and the interaction term is

$$H_I = -QqC_q(t) + \frac{1}{2m_q\omega_q^2}Q^2C_q^2(t) \quad (\text{Zwanzig}). \tag{A4}$$

In the above, $C_q(t)$ is the time-dependent interaction coefficient between the system and bath oscillator q, m_q and ω_q are the mass and angular frequency of that oscillator, and the last term (which does not depend on q) is the Caldeira-Leggett counterterm. The force from the bath on the system is then

$$-\partial_Q H_I = qC_q(t) - \frac{1}{m_q\omega_q^2}QC_q^2(t) \quad (\text{Zwanzig}). \tag{A5}$$

It is well known that for an Ohmic bath with all C_q constant, this tends to the sum of the friction force and the random force in a Kramers-Langevin equation. In Ref. [22] the situation was considered where for all interaction coefficients $C_q \propto \sqrt{\eta(t)}$, where $\eta(t)$ is a time-dependent friction coefficient. In that setting, the force from the bath on the system is

$$-\partial_Q H_I \approx -\eta\dot{Q} - \frac{\dot{\eta}}{2\eta}Q + \sqrt{\frac{2\eta}{\beta}}\xi \quad (\text{from Ref. [22]}), \tag{A6}$$

where ξ is standard white noise.

From the structure of the interaction term it is now easy to determine the second contribution to the work for the Caldeira-Leggett model. Namely,

$$\partial_t H_I = \left(-\frac{\dot{C}_q(t)}{C_q(t)}Q\right)(-\partial_Q H_I), \tag{A7}$$

which, when $C_q \propto \sqrt{\eta(t)}$, leads to

$$\begin{aligned}\delta W_{if} &= \int \partial_t H_I dt\\ &= \int \left(-\frac{\dot\eta}{2\eta}Q\right)\left(-\eta\dot Q - \frac{\dot\eta}{2\eta}Q + \sqrt{\frac{2\eta}{\beta}}\xi\right)dt\\ &\quad \text{(from Ref. [22]).}\end{aligned} \tag{A8}$$

Summarizing, the effective motion of the system in the Caldeira-Leggett model with time-dependent friction is

$$\dot Q = \frac{P}{M}, \quad \dot P = -\partial_Q V + F_S, \tag{A9}$$

where the generalized Sekimoto force F_S is

$$F_S = -\eta\dot Q - \frac{\dot\eta}{2\eta}Q + \sqrt{\frac{2\eta}{\beta}}\xi. \tag{A10}$$

The change of internal energy is for this model

$$\begin{aligned}\Delta U = \Delta H_S &= \int (\partial_t H_S + \dot P\partial_P H_S + \dot Q\partial_Q H_S)dt\\ &= \delta W^{(J)} + \int \frac{P}{M}F_S dt,\end{aligned} \tag{A11}$$

and the work δW_{if} from (A8) is

$$\delta W_{if} = \int \left(-\frac{\dot\eta}{2\eta}Q\right)F_S dt. \tag{A12}$$

Finally, the heat is

$$\delta Q = \delta H_B = \delta W - \Delta U = \int F_S\left(-\frac{\dot\eta}{2\eta}Q - \frac{P}{M}\right)dt. \tag{A13}$$

Work, heat, and internal energy change are hence for this model in equal measure functionals of the system history only.

The above approach can be generalized to interactions of the type

$$H_I(Q,q,t) = A(Q)B(q)C(t), \tag{A14}$$

where $A(Q)$ is a known function of the system, and $C(t)$ is a known function of time. The bath will then exert a force on the system as

$$-\partial_Q H_I(Q,q,t) = -\partial_Q A[B(q)C(t)]. \tag{A15}$$

When the acceleration of the system can be measured, this force is a system observable since

$$-\partial_Q H_I(Q,q,t) = \dot P + \partial_Q H_S(Q,P,t). \tag{A16}$$

On the other hand,

$$\begin{aligned}\partial_t H_I(Q,q,t) &= [A(Q)B(q)]\partial_t C\\ &= -\frac{\partial_t \log C}{\partial_Q \log A}[-\partial_Q H_I(Q,q,t)].\end{aligned} \tag{A17}$$

The second contribution to the work is then a functional of system history as

$$\begin{aligned}&\int \partial_t H_I(Q,q,t)dt\\ &= \int \left(-\frac{\partial_t \log C}{\partial_Q \log A}\right)[dP + \partial_Q H_S(Q,P,t)dt].\end{aligned} \tag{A18}$$

APPENDIX B: STRONG-COUPLING SYSTEM ENTROPY, INTERNAL ENERGY, AND FREE ENERGY

This Appendix contains the details of the transition from (20) to (21) in Sec. III B above. We repeat the starting point as

$$\Delta S_{\mathrm{TOT}} = \Delta(-\log\rho_S) - \Delta[\log\sigma(y|x)]. \tag{B1}$$

Using the assumption stated in Ref. [19] below Eq. (17), the two parts of the last term in (B1) can be written

$$\begin{aligned}\log\sigma(y^i|x^i) &= -\beta H^i_{\mathrm{TOT}} + \beta H^i_S - \log\langle e^{-\beta H_I}\rangle^i_B,\\ \log\sigma(y^f|x^f) &= -\beta H^f_{\mathrm{TOT}} + \beta H^f_S - \log\langle e^{-\beta H_I}\rangle^f_B,\end{aligned}$$

where we have introduced the notation of Ref. [15],

$$\langle\cdots\rangle_B = e^{\beta F_B}\int dy' e^{-\beta H_B(y')}(\cdots). \tag{B2}$$

We thus have a contribution to (B1) as

$$-\Delta\log\sigma(y|x) = \beta\Delta H_{\mathrm{TOT}} - \beta\Delta H_S + \Delta\log\langle e^{-\beta H_I}\rangle_B \tag{B3}$$

The contributions of the free energy of the bath (F_B) cancel and do not contribute to (B3).

The difference ΔH_{TOT} in (B3) is the work δW. Under the assumption that only H_S depends explicitly on time, δW is the Jarzynski work $\delta W^{(J)}$. The second difference ΔH_S in (B3) is the change of the system internal energy as usually defined, for many models of system-bath interactions that can also be taken as a functional of the system history only.

The logarithmic terms in (B3) can, on the other hand, be rewritten,

$$\log\langle e^{-\beta H_I}\rangle_B = \beta^2\partial_\beta\left(-\frac{1}{\beta}\log\langle e^{-\beta H_I}\rangle_B\right) + \beta\partial_\beta\log\langle e^{-\beta H_I}\rangle_B. \tag{B4}$$

The first term can be included in a strong-coupling system entropy,

$$\tilde s_S = -\log\rho_S + \beta^2\partial_\beta\left(-\frac{1}{\beta}\log\langle e^{-\beta H_I}\rangle_B\right), \tag{B5}$$

while the second can be combined with the bare change of the system internal energy as

$$\begin{aligned}\tilde u_S &= H_S - \partial_\beta\log\langle e^{-\beta H_I}\rangle_B\\ &= \partial_\beta\left(\beta\left(H_S - \frac{1}{\beta}\log\langle e^{-\beta H_I}\rangle_B\right)\right).\end{aligned} \tag{B6}$$

With these definitions we hence have (21), which we copy also here as

$$\Delta S_{\mathrm{TOT}} = \Delta\tilde s_S + \beta\delta W - \beta\Delta\tilde u_S. \tag{B7}$$

The definitions of $\tilde s_S$ and u_S can be related to a strong-coupling system free energy,

$$\tilde f_S = \tilde u_S - \frac{1}{\beta}\tilde s_S = H_S + \frac{1}{\beta}\log\rho_S - \frac{1}{\beta}\log\langle e^{-\beta H_I}\rangle_B, \tag{B8}$$

through the standard thermodynamic relations (Legendre transforms)

$$\tilde{u}_S = \partial_\beta(\beta\tilde{f}_S) = \tilde{f}_S + \beta\partial_\beta\tilde{f}_S, \tag{B9}$$

$$\tilde{s}_S = \beta(\tilde{u}_S - \tilde{f}_S) = \beta\partial_\beta(\beta\tilde{f}_S) - \beta\tilde{f}_S. \tag{B10}$$

APPENDIX C: DETAILS ON TIME REVERSALS IN STOCHASTIC DYNAMICS

This Appendix provides technical details for Sec. IV in the main text. The Kramers-Langevin equation $\dot{x} = \frac{p}{m}$ and $\dot{p} = -\partial_x V(x,t) - \gamma\frac{p}{m} + \sqrt{2k_BT\gamma}\xi$ is to be interpreted in the Stratonovich convention [23]. Over a small time interval t to $t' = t + \epsilon$ this means

$$x' - x = \epsilon\frac{\overline{p}}{m}, \tag{C1}$$

$$p' - p = -\epsilon\gamma\frac{\overline{p}}{m} - \epsilon\partial_x V(x,t) + \sqrt{2k_BT\gamma}\Delta\Xi, \tag{C2}$$

where $\overline{p} = \frac{p+p'}{2}$ and $\Delta\Xi$ is a centered normal variable of variance ϵ. Terms higher than ϵ have been suppressed. It follows that the probability distribution of p' conditioned on p is

$$P(p'|p) = \frac{1}{(4\pi k_BT\gamma\epsilon)^{\frac{d}{2}}}\exp\left(-\frac{\left[p' - p + \epsilon\gamma\frac{\overline{p}}{m} + \epsilon\partial_x V(x,t)\right]^2}{4k_BT\gamma\epsilon}\right)\left(1 + \epsilon\frac{\gamma d}{2m}\right), \tag{C3}$$

where d is the dimension of space, and the last term arises from the Jacobian when transforming from $\Delta\Xi$ to p'. *Natural time reversal* of the Kramers-Langevin equation means $\frac{dx^*}{dt^*} = \frac{p^*}{m}$ and $\frac{dp^*}{dt^*} = -\partial_{x^*}V(x^*,t^*) + \gamma\frac{p^*}{m} + \sqrt{2k_BT\gamma}\xi^*$, where $x^*_{t^*} = x_t$, $p^*_{t^*} = -p_t$, and ξ^* is a noise with the same characteristics as ξ. The probability distribution of $(p^*)'$ conditional on p^* over a short time t^* to $(t^*)' = t^* + \epsilon$ is thus

$$P((p^*)'|p^*) = \frac{1}{(4\pi k_BT\gamma\epsilon)^{\frac{d}{2}}}\exp\left(-\frac{\left[(p^*)' - p^* - \epsilon\gamma\frac{\overline{p^*}}{m} + \epsilon\partial_{x^*}V(x^*,t^*)\right]^2}{4k_BT\gamma\epsilon}\right)\left(1 - \epsilon\frac{\gamma d}{2m}\right). \tag{C4}$$

Inserting $(p^*)' = -p$ and $(p^*) = -p'$ one can form the ratio

$$\frac{P(p'|p)}{P[(p^*)'|p^*]} = \left(1 + \epsilon\frac{\gamma d}{m}\right) + O(\epsilon^2), \tag{C5}$$

which leads to an entropy production in the environment, over the whole process, as

$$\delta S^{\text{natural}}_{\text{env}} = \log\frac{P^F}{P^B} = (t_f - t_i)\frac{\gamma d}{m}. \tag{C6}$$

The *general time reversal* of the Kramers-Langevin equation as discussed in the main text means $\frac{dx^*}{dt^*} = \frac{p^*}{m}$ and $\frac{dp^*}{dt^*} = -\partial_{x^*}V + 2\partial_{x^*}V^- - \gamma\frac{p^*}{m} + \sqrt{2k_BT\gamma}\xi''$, where ξ'' is, as above, a noise with the same characteristics as ξ. In this case the ratio of the two propagators over a short time interval is

$$\frac{P(p'|p)}{P[(p^*)'|p^*]} = \exp\left(\frac{1}{k_BT\gamma}(p' - p + \partial_x V^+)\right.$$
$$\left.\times(-\gamma p/m - \partial_x V^-)\right). \tag{C7}$$

Introducing the notation of Ref. [23], that $u_+ = -\gamma p/m - \partial_x V^-$ is the part of the drift field that transforms as a vector and $u_- = -\partial_x V^+$ is the part that transforms as a pseudovector, and identifying $D = k_BT\gamma$ as the diffusion coefficient, one has

$$\delta S^{\text{general}}_{\text{env}} = \int(\dot{p} - u_-)\frac{1}{D}u_+dt, \tag{C8}$$

which is the formula quoted as (34) in the main text. For *canonical time reversal*, the special case of the above when $V^- = 0$, a more detailed discussion along the same lines as above can be found in Ref. [33]. Mathematically rigorous derivations of (C6) and (C8), as well as other time reversals of diffusion processes, can be found in Ref. [23].

1. Microscopic model

I will now show that (C8) can also be derived as the change of bath energy in an explicit model of a bath as harmonic oscillators initially in thermal equilibrium. The oscillators are labeled by their frequencies ω, have mass m_ω and density of states $f(\omega)$, and interact with the system with coupling strength C_ω. An Ohmic spectrum that satisfies

$$\frac{f(\omega)C^2(\omega)}{m_\omega} = \frac{2}{\pi}\gamma\omega^2 \tag{C9}$$

leads to Kramers-Langevin dynamics for the system with friction coefficient γ [26,29,30].

It is convenient to introduce the following terms for mappings:

$\mathcal{I}$ is as before the mapping $(x,y,H) \to (x^*,y^*,H^*)$. On the system $\mathcal{I}$ acts as, in general time reversal above; on the level of the bath, the action of $\mathcal{I}$ is to be determined.

$\mathcal{T}$ is the forward evolution of the system and the bath from time t_i and initial conditions (x^i,y^i) to time t_f and final conditions (x^f,y^f) under Hamiltonian H.

$\mathcal{T}^*$ is the time-reversed evolution of the system and the bath from time $t^*_i = 0$ and initial conditions $[(x^*)^i,(y^*)^i]$ to time $t^*_f = t_f - t_i$ and final conditions $[(x^*)^f,(y^*)^f]$ under Hamiltonian H^*.

$\mathcal{F}$ is the determination of (x^i,y^i), the initial conditions in the forward process, in terms of $\{x_k\}_{k=0}^n$, the forward trajectory of the system. Note that $x^i = x_0$, i.e., this mapping is trivial on the system.

$\mathcal{F}^*$ is the determination of $(y^*)^i$, the initial conditions in the time-reversed process, in terms of $\{x^*_k\}_{k=0}^n$, the time-reversed trajectory of the system. Also here, $(x^*)^i = x^*_0$.

All mappings are assumed to be smooth and invertible as needed. We can then define

$$\mathcal{I}(x^f,y^f) = \mathcal{F}^*\mathcal{I}\mathcal{F}^{-1}\mathcal{T}^{-1}(x^f,y^f), \tag{C10}$$

$$\mathcal{I}[(x^*)^f,(y^*)^f] = \mathcal{F}\mathcal{I}\mathcal{F}^{*-1}\mathcal{T}^{*-1}[(x^*)^f,(y^*)^f]. \tag{C11}$$

In other words, the above says that the time-reversed final conditions of the bath, in either process, are what they have to be as initial conditions so that the whole trajectory of the system is time reversed. With these (formal) definitions, $\mathcal{I}$ is an involution, as illustrated by the following diagram,

$$\begin{tikzcd}
((x^*)^f,(y^*)^f) \arrow[dd, "\mathcal{I}"] & & & ((x^*)^i,(y^*)^i) \arrow[lll, "\mathcal{T}^*"'] \\
& \{x_k\}_{k=0}^n \arrow[dl, "\mathcal{F}"'] \arrow[r, leftrightarrow, "\mathcal{I}"] & \{x_k^*\}_{k=0}^n \arrow[ur, "\mathcal{F}^*"] & \\
(x^i,y^i) \arrow[rrr, "\mathcal{T}"] & & & (x^f,y^f) \arrow[uu, "\mathcal{I}"']
\end{tikzcd}$$

2. Phase space volume

To show that $\mathcal{I}$ preserves phase space volume, we have to consider the Jacobians corresponding to (C10) and (C11). To avoid undercounting in the continuously sampled limit, take the forward system path $\{x_k\}_{k=0}^n$ to be specified by initial system coordinates and momenta $x_0 = (X^i,P^i)$ and $2n$ momenta increments $x_k = (\Delta P_{2k-1},\Delta P_{2k})$, and similarly for the time-reversed path.

The initial conditions of the bath are only reflected in the noise term of the Kramers-Langevin equation, that is,

$$\mathcal{F}^{-1} : \sqrt{2k_BT\gamma}\xi = \int_0^\infty f(\omega)C(\omega)\left[q_\omega \cos\omega t + \frac{p_\omega}{m_\omega\omega}\sin\omega t\right]d\omega, \tag{C12}$$

and similarly for the backward process,

$$\mathcal{F}^{*-1} : \sqrt{2k_BT\gamma}\xi'' = \int_0^\infty f(\omega)C(\omega)\left[q_\omega^* \cos\omega t^* + \frac{p_\omega^*}{m_\omega\omega}\sin\omega t^*\right]d\omega. \tag{C13}$$

Equation (C12) determines how the momentum increments $(\Delta P_k, k > 0)$ depend on the initial conditions of the bath (q_ω,p_ω), and analogously for the time-reversed path. The initial conditions of the paths can be solved for by an inverse Fourier transform,

$$\mathcal{F} : \begin{cases} q_\omega = \frac{1}{\pi}\frac{1}{f(\omega)C(\omega)}\int(\dot{p} + \partial_x V + \gamma p/m)\cos\omega t dt, \\ p_\omega = \frac{1}{\pi}\frac{m_\omega\omega}{f(\omega)C(\omega)}\int(\dot{p} + \partial_x V + \gamma p/m)\sin\omega t dt, \end{cases} \tag{C14}$$

and similarly,

$$\mathcal{F}^* : \begin{cases} q_\omega^* = \frac{1}{\pi}\frac{1}{f(\omega)C(\omega)}\int(\dot{p} + \partial_x V - 2\partial_x V^- - \gamma p/m)\cos\omega t^* dt^*, \\ p_\omega^* = \frac{1}{\pi}\frac{m_\omega\omega}{f(\omega)C(\omega)}\int(\dot{p} + \partial_x V - 2\partial_x V^- - \gamma p/m)\sin\omega t^* dt^*. \end{cases} \tag{C15}$$

Equation (C10) defines the determinant of the Jacobian of $\mathcal{I}$ as

$$\left|\frac{\partial\mathcal{I}(x^f,y^f)}{\partial(x^f,y^f)}\right| = \left|\frac{\partial[(x^*)^i,(y^*)^i]}{\partial\{x_k^*\}_{k=0}^n}\right|\left|\frac{\partial\{x_k^*\}_{k=0}^n}{\partial\{x_k\}_{k=0}^n}\right|\left|\frac{\partial\{x_k\}_{k=0}^n}{\partial(x^i,y^i)}\right|\left|\frac{\partial(x^i,y^i)}{\partial(x^f,y^f)}\right| = \left|\frac{\partial\{x_k^*\}_{k=1}^n}{\partial(y^*)^i}\right|^{-1}\left|\frac{\partial\{x_k\}_{k=1}^n}{\partial y^i}\right|. \tag{C16}$$

In the above, it has been used that $|\frac{\partial\{x_k^*\}_{k=0}^n}{\partial\{x_k\}_{k=0}^n}|$ is one because $\mathcal{I}$ preserves the system volume, that $\frac{\partial(x^i,y^i)}{\partial(x^f,y^f)}$ is one because the full dynamics is conservative, and that $\mathcal{F}$ acts trivially on the system. The whole expression is finally one because by (C12) and (C13) the two Jacobians $\frac{\partial\{x_k^*\}_{k=1}^n}{\partial(y^*)^i}$ and $\frac{\partial\{x_k\}_{k=1}^n}{\partial y^i}$ are the same.

3. Change of bath energy

Finally, we consider the energy change of the bath between the starting positions of the backward and forward process,

$$\Delta H_B = \int_0^\infty f(\omega)\left(\frac{1}{2m_\omega}((p_\omega^*)^2 - (p_\omega)^2) + \frac{1}{2}m_\omega\omega^2((q_\omega^*)^2 - (q_\omega)^2)\right)d\omega, \tag{C17}$$

where the contributions from a given ω are

$$\begin{aligned} \frac{(p_\omega^*)^2 - p_\omega^2}{2m_\omega} + \frac{1}{2}m_\omega\omega^2\big((q_\omega^*)^2 - q_\omega^2\big) = \left(\frac{m_\omega\omega}{\pi f(\omega)C(\omega)}\right)^2 \iint \cos\omega(t-t')[-2(\dot{p} + \partial_x V^+) \\ \times(\gamma p/m + \partial_x V^-)' - 2(\dot{p} + \partial_x V^+)'(\gamma p/m + \partial_x V^-)]dt\,dt'. \end{aligned} \tag{C18}$$

In the above, primed quantities refer to time t' and unprimed to time t. Using (C9), the notation in (C8), (C17), and $\int \cos\omega(t-t')d\omega = 2\pi\delta(t-t')$, this leads to

$$\Delta H_B = \int (\dot{p} - u_-)\frac{1}{\gamma}u_+ dt, \tag{C19}$$

which is Eq. (35) in the main text.

[1] C. Jarzynski, Annu. Rev. Condens. Matter Phys. **2**, 329 (2011).

[2] K. Sekimoto, *Stochastic Energetics*, Lecture Notes in Physics Vol. 799 (Springer, Berlin, 2010).

[3] E. Sevick, R. Prabhakar, S. R. Williams, and D. J. Searles, Annu. Rev. Phys. Chem. **59**, 603 (2008).

[4] M. Esposito, U. Harbola, and S. Mukamel, Rev. Mod. Phys. **81**, 1665 (2009).

[5] U. Seifert, Rep. Prog. Phys. **75**, 1 (2012).

[6] M. Esposito, M. A. Ochoa, and M. Galperin, Phys. Rev. B **92**, 235440 (2015).

[7] E. G. D. Cohen and D. Mauzerall, J. Stat. Mech. (2004) P07006.

[8] C. Jarzynski, J. Stat. Mech. (2004) P09005.

[9] L. Onsager, Chem. Rev. **13**, 73 (1933).

[10] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).

[11] G. W. Ford, J. T. Lewis, and R. F. O'Connell, J. Stat. Phys. **53**, 439 (1988).

[12] P. Hänggi, G.-L. Ingold, and P. Talkner, New J. Phys. **10**, 115008 (2008).

[13] M. F. Gelin and M. Thoss, Phys. Rev. E **79**, 051121 (2009).

[14] M. Campisi, P. Hänggi, and P. Talkner, Rev. Mod. Phys. **83**, 771 (2011).

[15] U. Seifert, Phys. Rev. Lett. **116**, 020601 (2016).

[16] C. Jarzynski, Phys. Rev. X **7**, 011008 (2017).

[17] P. Talkner and P. Hänggi, Phys. Rev. E **94**, 022143 (2016).

[18] P. Strasberg and M. Esposito, Phys. Rev. E **95**, 062101 (2017).

[19] H. J. D. Miller and J. Anders, Phys. Rev. E **95**, 062123 (2017).

[20] C. Maes, J. Stat. Phys. **95**, 367 (1999).

[21] K. Gawędzki, arXiv:1308.1518.

[22] E. Aurell, Entropy **19**, 595 (2017).

[23] R. Chetrite and K. Gawędzki, Commun. Math. Phys. **282**, 469 (2008).

[24] R. Kubo, J. Phys. Soc. Jpn. **12**, 570 (1957).

[25] S. Bonella, A. Coretti, L. Rondoni, and G. Ciccotti, Phys. Rev. E **96**, 012160 (2017).

[26] R. Zwanzig, J. Stat. Phys. **9**, 215 (1973).

[27] A. Caldeira and A. Leggett, Physica A **121**, 587 (1983).

[28] H. Grabert, P. Schramm, and G.-L. Ingold, Phys. Rep. **168**, 115 (1988).

[29] W. Bez, Z. Phys. B: Condens. Matter **39**, 319 (1980).

[30] P. Hänggi, in *Stochastic Dynamics*, edited by L. Schimansky-Geier and T. Pöschel, Lecture Notes in Physics Vol. 484 (Springer, Berlin, 1997), p. 15.

[31] C. Bhadra and D. Banerjee, J. Stat. Mech. (2016) 043404.

[32] M. Krüger and C. Maes, J. Phys.: Condens. Matter **29**, 064004 (2017).

[33] A. Celani, S. Bo, R. Eichhorn, and E. Aurell, Phys. Rev. Lett. **109**, 260603 (2012).

[34] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed. (Elsevier, Amsterdam, 2007).

[35] C. Maes and K. Netočný, J. Stat. Phys. **110**, 269 (2003).

[36] M. Pavelka, V. Klika, and M. Grmela, Phys. Rev. E **90**, 062131 (2014).

[37] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, Rev. Mod. Phys. **87**, 593 (2015).