Exchange fluctuation theorem for correlated quantum systems

Sania Jevtic
Institut für Theoretische Physik, Appelstr. 2, Hannover, D-30167, Germany

Terry Rudolph and David Jennings
Controlled Quantum Dynamics Theory Group, Level 12, EEE, Imperial College London, London SW7 2AZ, United Kingdom

Yuji Hirono, Shojun Nakayama, and Mio Murao
Department of Physics, The University of Tokyo, Hongo 7-3-1 Bunkyo-ku Tokyo 113-0033, Japan
(Received 17 May 2012; revised manuscript received 3 July 2015; published 6 October 2015)

We extend the exchange fluctuation theorem for energy exchange between thermal quantum systems beyond the assumption of molecular chaos, and describe the nonequilibrium exchange dynamics of correlated quantum states. The relation quantifies how the tendency for systems to equilibrate is modified in high-correlation environments. In addition, a more abstract approach leads us to a "correlation fluctuation theorem". Our results elucidate the role of measurement disturbance for such scenarios. We show a simple application by finding a semiclassical maximum work theorem in the presence of correlations. We also present a toy example of qubit-qudit heat exchange, and find that non-classical behaviour such as deterministic energy transfer and anomalous heat flow are reflected in our exchange fluctuation theorem.

DOI: 10.1103/PhysRevE.92.042113 PACS number(s): 05.30.−d, 03.65.Ud, 03.67.Mn, 05.70.Ln

I. INTRODUCTION

Fluctuation theorems describe nonequilibrium transformations of a thermodynamic system and constitute a refinement of the second law of thermodynamics, the most well-known incarnations being the work fluctuation theorems due to Jarzynski and Crooks [1–3]. These fluctuation theorems focus on the extraction of mechanical work from a single system in contact with a heat bath. However, an equally fundamental topic is the tendency of thermal systems to equilibrate with one another. The canonical example of this is heat exchange between two thermal systems at different temperatures and leads us instead to "exchange fluctuation theorems" that provide a quantitative description of the fluctuations in energy exchange between two hot bodies.

The thermodynamic arrow [4,5] is one particular manifestation of the second law of thermodynamics, and in its canonical form states that, on average, heat will flow from a hotter body to a colder one. Specifically, given two thermal states ρ_A and ρ_B at temperatures T_A and T_B with respect to Hamiltonians H_A and H_B, and an energy-conserving unitary evolution of the joint state, $\rho_A \otimes \rho_B \to U\rho_A \otimes \rho_B U^\dagger$, we define heat flow into A as $Q_A = \mathrm{tr}[H_A\rho'_A] - \mathrm{tr}[H_A\rho_A]$, where ρ'_A is the final reduced state for A, and we assume that the free Hamiltonians for A and B do not change. The fact that Gibbsian states minimize the free energy yields the Clausius inequality

$$Q_A\left(\frac{1}{T_A} - \frac{1}{T_B}\right) \geqslant 0, \tag{1}$$

and so if $T_A < T_B$ we have that the Q_A is strictly positive, and *on average* energy is transferred from the hotter body to the colder one. This is the standard thermodynamic arrow for heat flow.

However, a sharper expression of the directionality for heat flow exists in the recent exchange fluctuation theorem (XFT) due to Jarzynski and Wójcik [6], which states that for two systems A and B, initially at temperatures T_A and T_B, the probability $P(q)$ of observing an exchange of heat q from B to A during their interaction obeys the relation

$$\frac{P(q)}{P(-q)} = \exp[\Delta\beta q], \tag{2}$$

where $\Delta\beta = (kT_A)^{-1} - (kT_B)^{-1}$. This relation quantifies the relative likelihood of a fixed exchange process and its time-reversed twin, and shows that heat flow from a colder to a hotter object is exponentially suppressed. Averaging over all q and applying Jensen's inequality leads to an averaged inequality $\langle q\rangle(1/T_A - 1/T_B) \geqslant 0$. This seems to suggest that Eq. (1) automatically follows from (2); however, it must be emphasized that while Q_A equals $\langle q\rangle$ for *classical* states, this need not be true for more general quantum mechanical states. To obtain q, rank-1 projective energy measurements must be performed individually on the interacting systems; this produces nonclassical disturbances in their quantum states. Even though the expression $\langle q\rangle(1/T_A - 1/T_B) \geqslant 0$ still provides a thermodynamic directionality, it is no longer identical to (1) outside of the classical setting, which does not involve any measurement disturbance.

A. Assumption of molecular chaos and the role of correlations

The scope of the XFT in [6] is extremely broad, being valid for arbitrary unitary interactions between A and B that conserve energy, with the resultant form relying on two key assumptions of (I) initial Gibbsian states, and (II) the assumed time-reversal invariance of the underlying dynamics.

However the strict *directionality* of the thermodynamic heat-flow relies on a tacit third assumption (III) that the systems involved are initially uncorrelated—namely Boltzmann's assumption of "molecular chaos" [7].

The molecular chaos assumption is required both in classical and quantum mechanics and, irrespective of any inherent quantum randomness, plays the central role in thermodynamic directionality of heat flow. Indeed it has been shown explicitly that if you drop the assumption of molecular chaos then you

weaken the thermodynamic arrow (as has been shown in [8,9] and references therein).

As such, both the Clausius relation, Eq. (1), and the Jarzynski-Wójcik fluctuation theorem, Eq. (2), are limited in application since they will fail to hold within the domain of high-correlation environments. Indeed, with the extremal case of a globally pure, multipartite quantum state with thermal subsystems [where assumptions (I) and (II) are satisfied] there should exist *no directional constraint* whatsoever, and for such situations no equality such as Eq. (2) should hold. The reason for this is that pure quantum states are states of maximal knowledge and may be reversibly interconverted through the appropriate unitary transformations. Such local thermality in globally pure states turns out to be the *typical* scenario with respect to the Haar measure. Specifically, for a randomly chosen multipartite state $|\Psi\rangle$ of a multipartite system with fixed total energy, the reduced state for a small subsystem is exponentially likely to be Gibbsian [10], with the thermality arising due to quantum entanglement.

Quantum correlations can be far stronger than their classical counterparts, and in addition to the Gibbsian typicality in pure states, entanglement theory has other deep connections with thermodynamics [11], often through their parallel formulations as resource theories [12]. Beyond the foundational interest of studying the dissolution of the thermodynamic arrow due to strong correlations, there is also rapid experimental progress in the precise manipulation of small quantum systems designed to function as engines at nanoscales, and as such, it is also of practical importance to determine the fundamental limitations and behavior of heat exchange in such quantum systems.

The purpose of this paper is to remove the third fundamental assumption (III) of molecular chaos, and extend the existing XFT into high-correlation environments, in which initially correlated quantum systems satisfying assumption (I) are allowed to evolve under nonequilibrium dynamics that satisfy assumption (II) and exchange heat. In doing so we identify the appropriate thermodynamic measure for the effect of correlations on sharp energy exchanges and describe how it may contribute in work-extraction primitives, such as the maximum work theorem scenario.

Our results complement the body of work concerned with extending fluctuation theorems to quantum systems. Some advances in this area include a fluctuation theorem (FT) for entropy production [13] in an open quantum system far from equilibrium [14], and an FT for a forced harmonic oscillator coupled to thermal reservoir, using quantum stochastic thermodynamics [15]. In [16] an exchange fluctuation theorem for energy and particle exchange is derived when multiple measurements are made during the interaction. In [17] FTs for generalized thermodynamic observables are found for the broad setting of quantum dynamics described by completely positive trace preserving maps. Yet, the important difference between these prior approaches and the work herein is the absence of initial correlations. As far as we are aware, assumption (III) of molecular chaos has always been made, and it is not clear that previous approaches may be easily extended to the broader framework that we consider.

This paper is structured in the following way. In Sec. II we present an overview of the components required for deriving the XFT, and highlight some conceptual points that relate to thermality due to quantum fluctuations. In Sec. III we analyze the heat flow in a parallel manner to the original XFT [6], however without assuming molecular chaos. We find that dropping this assumption enforces the use of a "sharp" mutual information measure that quantifies the correlations between the two subsystems and how these correlations impact the thermodynamic heat flow. This in turn provides us with a generalized form of the XFT, valid in a correlated quantum environment and allows for quantum fluctuations stemming from intrinsic randomness in pure quantum states. In Sec. IV we present an independent derivation for nonequilibrium heat exchange in the presence of correlations using a more abstract setting. The upshot of this is the acquirement of the "correlation factor" that gives a notion of the strength of correlations and their effect on the directionality of thermodynamics processes. In Sec. V we illustrate our thermodynamic results by applying them to a qubit-qudit system. In Sec. VI C we draw attention to the strong classicality of the XFT, due to the demand of projective measurements, and highlight the technical obstacles to an extension involving more gentle, generalized POVMs. Section VI D provides a simple application of our results to a maximum work theorem, and illustrates the potential work value that correlations between two thermal systems can possess.

II. OVERVIEW OF THE GENERALIZED SETTING

In this section we give an overview of the basic setting employed, including the form of our time-reversal assumptions. Our central goal will be to establish an exchange fluctuation theorem for a bipartite quantum system, whose reduced states ρ_A and ρ_B are initially thermal, but where we drop the assumption of the initial factorization of the joint state ρ_{AB} and allow genuine quantum mechanical coherence and entanglement to either evolve or be initially present.

A. Thermodynamic scenario

In defining heat exchange in [6], the isolated bipartite system is assumed to undergo a three step process. An initial energy measurement $\mathcal{M}_1$ is first performed on the two subsystems which are then allowed to subsequently interact and evolve under a unitary U, until a final energy measurement $\mathcal{M}_2$ is performed. Thus the bipartite quantum state ρ_{AB} undergoes the following sequence of quantum operations: $\rho_{AB} \to \mathcal{M}_1[\rho_{AB}] \to U \circ \mathcal{M}_1[\rho_{AB}] \to \mathcal{M}_2 \circ U \circ \mathcal{M}_1[\rho_{AB}]$.

As mentioned, the central assumptions of the XFT are (I) the initial thermality of the individual subsystems and (II) time-reversal symmetry of the underlying dynamics. Let H_A and H_B be the subsystem Hamiltonians. Then, a more specific statement of (II) is that we assume $\Theta^\dagger H_A \Theta = H_A$ and $\Theta^\dagger H_B \Theta = H_B$, where Θ is the antiunitary time-reversal operator. In what follows we use joint energy eigenstates $|\phi,\chi\rangle := |\phi\rangle \otimes |\chi\rangle$, such that $H_A|\phi,\chi\rangle = E_\phi|\phi,\chi\rangle$ and $H_B|\phi,\chi\rangle = E_\chi|\phi,\chi\rangle$, for which we deduce that $H_A\Theta|\phi,\chi\rangle = E_\phi\Theta|\phi,\chi\rangle$ (with a similar expression for B). We take the local energy eigenstates $\{|\phi\rangle\}$ and $\{|\chi\rangle\}$ to be complete orthonormal bases for A and B so that $H_A = \sum_\phi E_\phi|\phi\rangle\langle\phi|$ and $H_B = \sum_\chi E_\chi|\chi\rangle\langle\chi|$, and since Θ is a symmetry of classical states we assume that $\Theta(|\phi\rangle \otimes |\chi\rangle)$ is always in the basis set $\{|\phi\rangle \otimes |\chi\rangle\}$ for any ϕ and χ. The

thermal marginal states of ρ_{AB}, as required by assumption (I), are then given by

$$\rho_A = Z_A^{-1} e^{-\beta_A H_A} = Z_A^{-1} \sum_{\phi} e^{-\beta_A E_\phi} |\phi\rangle\langle\phi|, \tag{3}$$

$$\rho_B = Z_B^{-1} e^{-\beta_B H_B} = Z_B^{-1} \sum_{\chi} e^{-\beta_B E_\chi} |\chi\rangle\langle\chi|, \tag{4}$$

where Z_A and Z_B are the usual partition functions for A and B, and β_A, β_B are the inverse temperatures.

The measurements $\mathcal{M}_1$ and $\mathcal{M}_2$ used to determine the energies of the subsystems can in general be POVM measurements; however, in the rest of this paper we follow the formulation of [6] and restrict both the initial and final measurement to be rank-1 projective measurements onto the local energy eigenbases of A and B, namely $\{|\phi\rangle\}$ and $\{|\chi\rangle\}$, and only at the end discuss the challenges of extending beyond such sharp measurements.

Given this setting, it is useful to introduce the notion of a *history* for the composite quantum system AB. A history is denoted

$$\gamma = \left(\rho_{AB}; |\phi\rangle \otimes |\chi\rangle \xrightarrow{U} |\phi'\rangle \otimes |\chi'\rangle\right), \tag{5}$$

where ρ_{AB} is the initial quantum state that is first projected into the energy eigenstate $|\phi\rangle \otimes |\chi\rangle$ under $\mathcal{M}_1$ and then evolves unitarily to $U(|\phi\rangle \otimes |\chi\rangle)$, which is then measured and projected into the energy eigenstate $|\phi'\rangle \otimes |\chi'\rangle$ under $\mathcal{M}_2$.

We denote by Γ the full set of all histories $\{\gamma\}$ comprised of first beginning in the state ρ_{AB}, measuring out some energy eigenstate, evolving under some U, and then measuring out some final energy eigenstate.

The thermodynamic condition of energy conservation we use is simply that $\mathrm{tr}[\rho_{AB}(H_A + H_B)] = \mathrm{tr}[U\rho_{AB}U^\dagger(H_A + H_B)]$. We note that U involves the interaction Hamiltonian that is assumed to be smoothly switched on and off, but the energies of the subsystems are always measured with respect to their appropriate free Hamiltonian.

B. Intrinsic fluctuations due to quantum coherence

While the exchange fluctuation theorem is a refinement on the thermodynamic arrow for heat flow, in that it deals with subensembles of postselected outcomes with sharp energy transfers q, not all of the uncertainty is attributable to the statistical mixture of different energy states. Pure quantum mechanical states allow the possibility of intrinsic quantum fluctuations, and so while we might step beyond classical statistical fluctuations by focusing on individual pure state outcomes as is done in the original treatment of XFT in [6], we might also allow the possibility of quantum coherence evolving under the unitary dynamics and generating new indeterminacy. For example, with respect to the average statistics of energy measurements $\{|E_k\rangle\langle E_k|\}$, the pure quantum state $|\psi\rangle \propto \sum_k \sqrt{e^{-\beta E_k}}|E_k\rangle$ is indistinguishable from a thermal mixed state $\rho_{\mathrm{therm}} = Z^{-1}\sum_k e^{-\beta E_k}|E_k\rangle\langle E_k|$.

Nevertheless, the energy measurement of such superpositions $|\psi\rangle$ will display quantum fluctuations, some of which may increase the total energy of AB, some decrease it, but on average no net energy should be gained from the fluctuations. It is then simple to allow histories with positive energy fluctuations that increase the total energy of AB, or negative fluctuations that decrease the energy of AB, and also histories with no fluctuations at all. As such, a useful and physically intuitive division of the set of histories is into sets of histories with similar energy transformations. In particular we write $\Gamma = \cup_{q,\Delta\epsilon}\Gamma(q,\Delta\epsilon)$, where $\Gamma(q,\Delta\epsilon)$ is the set of γ [of the form (5)] with fluctuations of the total energy of AB

$$\Delta\epsilon = \langle\phi',\chi'|(H_A + H_B)|\phi',\chi'\rangle - \langle\phi,\chi|(H_A + H_B)|\phi,\chi\rangle. \tag{6}$$

This change in energy may be interpreted as work that is extracted or performed on the system. The energy transfer q into A is defined as

$$q = \langle\phi'|H_A|\phi'\rangle - \langle\phi|H_A|\phi\rangle. \tag{7}$$

In the next section we make use of this setting and derive the generalized XFT theorem.

III. EXCHANGE FLUCTUATION THEOREM FOR CORRELATED QUANTUM STATES

We can now formulate a generalized XFT that drops assumption (III) of molecular chaos, discussed in Sec. I A, and allow a general bipartite quantum state ρ_{AB} with thermal marginals. Given this initial state ρ_{AB}, the occurrence of a single history $\gamma \in \Gamma(q,\Delta\epsilon)$ in Eq. (5) has probability

$$\mathrm{Prob}[\gamma] = \langle\phi,\chi|\rho_{AB}|\phi,\chi\rangle|(|\phi',\chi'\rangle, U|\phi,\chi\rangle)|^2, \tag{8}$$

where $U \equiv e^{-iHt}$ and $H \equiv H_A + H_B + H_{\mathrm{int}}$ is the total Hamiltonian, including the interaction H_{int} between A and B, which is switched on at the initial time, and for clarity we write the Hilbert space inner product as $(\cdot,\cdot)$.

From time-reversal invariance, we require $\Theta^\dagger H\Theta = H$ as well as $\Theta^\dagger H_A\Theta = H_A$ and $\Theta^\dagger H_B\Theta = H_B$. From the antiunitarity of Θ it follows that $U = \Theta^\dagger U^\dagger\Theta$, and so we have that

$$|(|\phi',\chi'\rangle, U|\phi,\chi\rangle)|^2 = |(\Theta|\phi,\chi\rangle, U\Theta|\phi',\chi'\rangle)|^2. \tag{9}$$

Thus time-reversal symmetry alone implies the probability to go from the initial state $|\phi,\chi\rangle$ to $|\phi',\chi'\rangle$ is always equal to the probability to go from $\Theta|\phi',\chi'\rangle$ to $\Theta|\phi,\chi\rangle$ under the *same* unitary interaction. In what follows, we use a star to denote time-reversed objects, for example $|\phi',\chi'\rangle_* := \Theta|\phi',\chi'\rangle$.

To quantify correlations in the quantum state as they relate to the XFT we define, for any joint local POVMs $\{M_i\}$ on A and $\{N_j\}$ on B, the quantity $\mathcal{I}(\rho_{AB}; M_i, N_j)$ via the expression

$$\mathcal{I}(\rho_{AB}; M_i, N_j) := \ln\left(\frac{\mathrm{tr}[M_i \otimes N_j\rho_{AB}]}{\mathrm{tr}[M_i\rho_A]\mathrm{tr}[N_j\rho_B]}\right). \tag{10}$$

Note that when the state factorizes $\rho_{AB} = \rho_A \otimes \rho_B$ we have $\mathcal{I}(\rho_{AB}; M_i, N_j) = 0$.

For the sharp energy measurement, we simply have $\mathcal{M}_1 = \mathcal{M}_2 = \{M_\phi \otimes N_\chi\}$, where $\{M_\phi = |\phi\rangle\langle\phi|\}$ and $\{N_\chi = |\chi\rangle\langle\chi|\}$

are the rank-1 projectors in the energy eigenbases,[1] and when ρ_{AB} is a correlated quantum state having thermal marginals as in Eqs. (3) and (4), we find that performing $\mathcal{M}_1$ and forgetting the outcome maps ρ_{AB} into a classically correlated state χ_{AB} that is diagonal in the energy eigenbasis. Moreover, it is readily seen that the pre- and postmeasured states (under the nonselective measurement $\mathcal{M}_1$) have the same thermal marginals $\mathrm{tr}_{B(A)}[\rho_{AB}] = \mathrm{tr}_{B(A)}[\chi_{AB}] = \rho_{A(B)}$.

For this particular initial measurement, the probability $p(|\phi,\chi\rangle)$ of projecting ρ_{AB} (or χ_{AB}) into the state $|\phi,\chi\rangle$ under $\mathcal{M}_1$ can be written as

$$p(|\phi,\chi\rangle) := \langle\phi,\chi|\rho_{AB}|\phi,\chi\rangle \tag{11}$$

$$= e^{-\beta_A E_\phi - \beta_B E_\chi - \log(Z_A Z_B) + \mathcal{I}[\rho_{AB};M_\phi,N_\chi]}, \tag{12}$$

while a comparison with the probability of obtaining the $|\phi',\chi'\rangle_*$ outcome implies that

$$p(|\phi,\chi\rangle) = p(|\phi',\chi'\rangle_*) e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)}, \tag{13}$$

where $\Delta\beta = \beta_A - \beta_B$, and crucially

$$\Delta\mathcal{I}(\gamma) = \mathcal{I}[\rho_{AB}; M_{\phi'_*}, N_{\chi'_*}] - \mathcal{I}[\rho_{AB}; M_\phi, N_\chi] \tag{14}$$

is the appropriate correlation measure, dependent only on the initial state ρ_{AB} and the initial measurement $\mathcal{M}_1$. A derivation of (13) is provided in the Appendix A. When we assume molecular chaos (III), both terms in $\Delta\mathcal{I}(\gamma)$ are individually equal to zero; hence $\Delta\mathcal{I}(\gamma) = 0$.

Combining (13) with (9) and (8) we find that

$$\frac{\mathrm{Prob}[\gamma]}{\mathrm{Prob}[\gamma^*]} = e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)}, \tag{15}$$

where γ^* is the time-reversed twin of γ given by

$$\gamma^* \equiv (\rho_{AB}; |\phi',\chi'\rangle_* \xrightarrow{U} |\phi,\chi\rangle_*). \tag{16}$$

In particular, the history γ involves a quantity of energy q being transferred into A and a net increase of total energy $\Delta\epsilon$, while γ^* involves the opposite changes, and so $\Gamma(q,\Delta\epsilon)^* = \Gamma(-q,-\Delta\epsilon)$. We also note that (15) is independent of the specific form of the dynamics (beyond time-reversal invariance), and depends solely on the properties of the initial quantum state ρ_{AB}.

One can now compare the ratio of probabilities of the set $\Gamma(q,\Delta\epsilon)$ and its time-reversed twin set $\Gamma(q,\Delta\epsilon)^* = \Gamma(-q,-\Delta\epsilon)$ since, in an experiment, one can only measure q but not the specific history γ. The probability of the former is given by

$$\mathrm{Prob}[\Gamma(q,\Delta\epsilon)] = \sum_{\gamma\in\Gamma(q,\Delta\epsilon)} e^{\Delta\beta q + \beta_B\Delta\epsilon - \Delta\mathcal{I}(\gamma)} \mathrm{Prob}[\gamma^*]. \tag{17}$$

While the term $e^{\Delta\beta q + \beta_B\Delta\epsilon}$ may be factored out of the sum, the correlation term cannot as it will generally vary over the set $\Gamma(q,\Delta\epsilon)$. Instead we necessarily obtain bounds for the ratio of the probabilities. To fix the lower and upper bounds, we respectively define $\Delta\mathcal{I}_l = \max_{\gamma\in\Gamma(q,\Delta\epsilon)}[\Delta\mathcal{I}(\gamma)]$ and $\Delta\mathcal{I}_u = \min_{\gamma\in\Gamma(q,\Delta\epsilon)}[\Delta\mathcal{I}(\gamma)]$, and immediately deduce that

$$e^{\Delta\beta q + \beta_B\Delta\epsilon - \Delta\mathcal{I}_l} \leqslant \frac{\mathrm{Prob}[\Gamma(q,\Delta\epsilon)]}{\mathrm{Prob}[\Gamma(-q,-\Delta\epsilon)]} \leqslant e^{\Delta\beta q + \beta_B\Delta\epsilon - \Delta\mathcal{I}_u}. \tag{18}$$

The XFT in Eq. (18) is a generalization of Jarzynski and Wójczik in Eq. (2) and it is a constraint on the relative likelihood of a forward transition to a backward transition given an initially correlated quantum state with thermal subsystems. However, the initial sharp energy measurement $\mathcal{M}_1$ means that the generalized XFT is not sensitive to any correlations beyond those of classical correlated states. Nevertheless, by moving away from the assumption of molecular chaos we obtain $\Delta\mathcal{I} \neq 0$ and one gradually weakens the constraint on the thermodynamic arrow, as expected. Moreover, it is not possible to tighten these bounds without making additional assumptions as to the particular form of the dynamics.

Beyond the relative likelihood of the forward and reverse processes, one can take (15) and sum over all $\gamma \in \Gamma$, to obtain the nonequilibrium equality for an initially correlated state

$$\langle e^{-\Delta\beta q - \beta_B\Delta\epsilon + \Delta\mathcal{I}}\rangle = 1 \tag{19}$$

and then using Jensen's inequality we have that

$$\Delta\beta\langle q\rangle + \beta_B\langle\Delta\epsilon\rangle - \langle\Delta\mathcal{I}\rangle \geqslant 0. \tag{20}$$

Here, $\langle\Delta\mathcal{I}\rangle$ represents the difference in the classical mutual information of measurement outcomes between the initial and final states.

Given the assumption that average total energy is conserved $\langle\Delta\epsilon\rangle = 0$ we have $\Delta\beta\langle q\rangle - \langle\Delta\mathcal{I}\rangle \geqslant 0$, which reduces to (1) the Clausius relation $Q(1/T_A - 1/T_B) \geqslant 0$ for $Q = \langle q\rangle$ and the assumption of molecular chaos. More importantly, it displays the energetic value of correlations in providing a modified lower bound of $Q(1/T_A - 1/T_B) \geqslant \langle\Delta\mathcal{I}\rangle$ with the function $\mathcal{I}(\rho_{AB}; M_i, N_j)$ as the appropriate sharp-outcome measure for the initial bipartite quantum state. This must be compared with averaged results obtained previously [8,9] in which $\langle\Delta\mathcal{I}\rangle$ is replaced with the change in the quantum mutual information of the state ρ_{AB}. The origin of the difference is that the XFT demands sharp energies at the initial and final stages, as opposed to bluntly looking at expectation values of energy for pure quantum states.

While it is natural to impose energy conservation, either at the level of commuting Hamiltonians or expectation values, the XFT given by Eq. (18) makes predictions for a particular type of state with local temperatures and global correlations, and provides only the relative likelihood of seeing one forward thermodynamic process compared to its reverse—not whether it occurs at all. As such, the specific interaction Hamiltonian that is used only serves to predict the absolute likelihood of these different processes.

IV. NONEQUILIBRIUM EQUALITY IN THE PRESENCE OF CORRELATIONS

In the previous section we derived XFTs for initially correlated systems using the concrete idea of histories and

[1]The function $\mathcal{I}$ may be related to the classical relative entropy of the joint measurement outcomes through the relation $I_c(\mathcal{M}:\mathcal{N}) = \sum_{i,j}\mathrm{tr}[M_i\otimes N_j\rho_{AB}]\mathcal{I}(\rho_{AB};M_i,N_j)$. In addition, we have that $I_c(\mathcal{M}:\mathcal{N}) = I[\mathcal{M}_1[\rho_{AB}];A:B]$, where $I[\sigma_{AB};A:B]$ is the quantum mutual information of the bipartite state σ_{AB}, defined as $I[\sigma_{AB};A:B] = S[\sigma_A] + S[\sigma_B] - S[\sigma_{AB}]$.

time reversal. Here, we follow the compact approach of [17] to present an alternative formulation of the XFT and isolate a "correlation factor" that quantifies the deviation from molecular chaos.

We adopt the same prepare-evolve-measure setting as before. The initial bipartite quantum state ρ_{AB} is projected onto the energy basis $\mathcal{M}_1 = \{M_\phi \otimes N_\chi\} = \{|\phi\rangle\langle\phi| \otimes |\chi\rangle\langle\chi|\}$. For simplicity we use $\mu = (\phi,\chi)$ to label the outcome of $\mathcal{M}_1$ that prepares the state

$$\rho_\mu = \frac{1}{p_\mu} M_\mu \otimes N_\mu \rho M_\mu^\dagger \otimes N_\mu^\dagger = |\phi,\chi\rangle\langle\phi,\chi|$$

with probability

$$p_\mu = \mathrm{tr}[M_\mu \otimes N_\mu \rho_{AB}]. \tag{21}$$

The initial state ρ_{AB} is again assumed to have thermal marginals from which we define the uncorrelated probability distribution

$$p_\mu^0 = \mathrm{tr}[M_\mu \rho_A]\mathrm{tr}[N_\mu \rho_B] = \frac{e^{-\beta_A E_\phi}}{Z_A}\frac{e^{-\beta_B E_\chi}}{Z_B}. \tag{22}$$

The prepared state ρ_μ evolves under the unitary U to $\rho'_\mu = U\rho_\mu U^\dagger$ and after the interaction the final energy measurement projects this state onto $\rho'_{\nu|\mu} = \frac{1}{p_{\nu|\mu}} M_\nu \otimes N_\nu \rho_\mu M_\nu \otimes N_\nu = |\phi',\chi'\rangle\langle\phi',\chi'|$ with the outcome labeled by $\nu = (\phi',\chi')$ and probability

$$p_{\nu|\mu} = \mathrm{tr}[M_\nu \otimes N_\nu \rho'_\mu] = |\langle\phi',\chi'|U|\phi,\chi\rangle|^2. \tag{23}$$

The total probability to obtain outcome ν is $p_\nu = \sum_\mu p_\mu p_{\nu|\mu} := \sum_\mu p_{\mu\nu}$, where

$$p_{\mu\nu} = \langle\phi,\chi|\rho|\phi,\chi\rangle|\langle\phi',\chi'|U|\phi,\chi\rangle|^2 \tag{24}$$

is simply Prob[γ] in Eq. (8).

We convert $p_{\mu\nu}$ into a probability density function on $\mathbb{R}$ for a continuous random variable x by writing

$$\sum_{\mu\nu} \delta(x - X_{\mu\nu}) p_{\mu\nu} =: P_X(x), \tag{25}$$

where $X_{\mu\nu}$ is a discrete random variable distributed according to $p_{\mu\nu}$. Define the function $F_{\widetilde{X}}(-x) = P_X(x)e^{-x}$; this is analogous to time reversing, with $\widetilde{X} := -X$.

We now choose the random variable $X_{\mu\nu}$ to be given by

$$X_{\mu\nu} = \ln p_\mu - \ln f_\nu + \Delta\mathcal{I}_{\mu\nu}, \tag{26}$$

where f_ν, f_ν^0 refer to time-reversed distributions

$$f_\nu := \mathrm{tr}[M_{\nu_*} \otimes N_{\nu_*} \rho_{AB}], \tag{27}$$

$$f_\nu^0 := \mathrm{tr}[M_{\nu_*}\rho_A]\mathrm{tr}[N_{\nu_*}\rho_B] = \frac{e^{-\beta_A E'_\phi}}{Z_A}\frac{e^{-\beta_B E'_\chi}}{Z_B}, \tag{28}$$

$$M_{\nu_*} = |\phi'\rangle\langle\phi'|_*, \quad N_{\nu_*} = |\chi'\rangle\langle\chi'|_* \tag{29}$$

and $\Delta\mathcal{I}_{\mu\nu}$ is equivalent to Eq. (14) for the history γ corresponding to (μ,ν).

Integrating both sides of $F_{\widetilde{X}}(-x) = P_X(x)e^{-x}$ over all x, we arrive at the thermodynamic relation

$$\langle e^{-\Delta\beta q - \beta_B \Delta\epsilon}\rangle = \langle e^{-\Delta\mathcal{I}}\rangle_*, \tag{30}$$

where the star on the right-hand side indicates that the average is to be taken with respect to the time-reversed probability distribution. Note that we could have reached this equality by rearranging and subsequently integrating Eq. (15). Appendix B provides the details for this calculation.

This formulation separates out the "correlation factor" $\eta := \langle e^{-\Delta\mathcal{I}}\rangle_*$ that quantifies the deviation from the assumption of molecular chaos. Indeed, when the systems begin in a product state, $\mathcal{I} = 0$ and both Eqs. (30) and (19) reduce to $\langle e^{-\Delta\beta q - \beta_B\Delta\epsilon}\rangle = 1$.

Applying Jensen's inequality to (30) gives

$$\Delta\beta\langle q\rangle + \beta_B\langle\Delta\epsilon\rangle + \ln\eta \geqslant 0.$$

This matches Eq. (20) if and only if the random variable $\Delta\mathcal{I}$ is a constant.

V. EXACTLY SOLVABLE TOY MODEL

We now compare the three "lenses" through which to view heat exchange between correlated systems: the XFT from Eq. (18), its averaged form in Eq. (20), and the exponentiated average in (30). For convenience we list them below in order:

$$e^{\Delta\beta q + \beta_B\Delta\epsilon - \Delta\mathcal{I}_l} \leqslant \frac{\mathrm{Prob}[\Gamma(q,\Delta\epsilon)]}{\mathrm{Prob}[\Gamma(-q,-\Delta\epsilon)]} \leqslant e^{\Delta\beta q + \beta_B\Delta\epsilon - \Delta\mathcal{I}_u}, \tag{31a}$$

$$\Delta\beta\langle q\rangle + \beta_B\langle\Delta\epsilon\rangle - \langle\Delta\mathcal{I}\rangle \geqslant 0, \tag{31b}$$

$$\langle e^{-\Delta\beta q - \beta_B\Delta\epsilon}\rangle = \langle e^{-\Delta\mathcal{I}}\rangle_* =: \eta. \tag{31c}$$

To interpret these relations, we consider a setting that admits a complete solution. Significant differences arise between them even for a low-dimensional scenario, for which we have the usual caveat that distributions can become quite broad, and so any expectation values must correspond to multiple runs on the systems in the i.i.d. limit.

We work with a joint Hilbert space $\mathcal{H}_{d_A} \otimes \mathcal{H}_{d_B}$ describing two subsystems of dimension $d_A = 2$ and $d_B = d$, with d unspecified for now. The free Hamiltonian of system $i \in \{A,B\}$ is

$$H_i = \sum_{n=0}^{d_i - 1} n|n\rangle\langle n|, \tag{32}$$

where we have set all energy separations to be unity and the ground state is zero.

The evolution is chosen to be *sharply* energy conserving so that $\Delta\epsilon = 0$ for all histories. The interaction Hamiltonian H_{int} that achieves this commutes with the free part $H_A + H_B$ and in its most general form it is

$$H_{\mathrm{int}} = \sum_{j=1}^{d-1} \omega_j(|0,j\rangle\langle 1,j-1| + |1,j-1\rangle\langle 0,j|). \tag{33}$$

The eigenvectors of H_{int} are $|\pm_j\rangle = \frac{1}{\sqrt{2}}(|1,j-1\rangle \pm |0,j\rangle)$ with eigenvalues $\pm\omega_j$, for $j = 1,\ldots,d-1$. Coherent evolution under this Hamiltonian is restricted to the energy-degenerate, two-dimensional subspaces spanned by $\{|0,j\rangle,|1,j-1\rangle\}$. The subspaces can be thought of as virtual qubits with virtual temperatures assigned to them as in [18].

In deriving the thermodynamic relations, an initial projective measurement $\mathcal{M}_1$ is made on the bipartite state ρ_{AB}; this

kills off any coherence in the free energy eigenbasis. As such we may simply take as our initial state a classically correlated density matrix

$$\rho_{AB} \equiv \rho_{AB}(t=0) = \sum_{m=0}^{1}\sum_{n=0}^{d-1} \lambda_{mn}|mn\rangle\langle mn|, \tag{34}$$

where $\sum_{mn}\lambda_{mn}=1$ and $\lambda_{mn}\geqslant 0$ for all m,n. A further constraint on the λ_{mn} comes from the requirement that the subsystems $i=\{A,B\}$ must be thermal states $\mathrm{tr}_{\setminus i}[\rho_{AB}] = \frac{1}{Z_i}e^{-\beta_i H_i}$; the $\setminus i$ notation means the complement of i. The bipartite state evolves to $\rho_{AB}(t)=U(t)\rho_{AB}U(t)^\dagger$, where $U(t)=e^{-iHt}$, with $H=H_A+H_B+H_{\mathrm{int}}$ and then a final measurement is made in the free energy basis.

In this toy example the three thermodynamic relations become

$$\min\left\{\frac{\lambda_{0,j}}{\lambda_{1,j-1}}\right\} \leqslant \frac{\mathrm{Prob}[\Gamma(q=1)]}{\mathrm{Prob}[\Gamma(q=-1)]} \leqslant \max\left\{\frac{\lambda_{0,j}}{\lambda_{1,j-1}}\right\}, \tag{35a}$$

$$\Delta\beta\langle q\rangle \geqslant \sum_j(\lambda_{0j}-\lambda_{1,j-1})\left(\Delta\beta+\ln\frac{\lambda_{1,j-1}}{\lambda_{0,j}}\right)\sin^2\omega_j t, \tag{35b}$$

$$\eta = 1+\sum_{j=1}^{d-1}[\lambda_{0j}(e^{-\Delta\beta}-1)+\lambda_{1,j-1}(e^{\Delta\beta}-1)]\sin^2\omega_j t. \tag{35c}$$

We refer the reader to Appendix C for details.

To analyze these results, we concentrate on the smallest d that gives nontrivial results. Note that the condition that ρ_{AB} is physical has not been imposed yet. The initial density matrix is required to have thermal marginals. A convenient way of describing the correlated bipartite state ρ_{AB} on $\mathcal{H}_{d_A}\otimes\mathcal{H}_{d_B}$ is by writing it as $\rho_{AB}=\rho_A\otimes\rho_B+\tau_{AB}$, where the operator τ_{AB} must obey $\mathrm{tr}_A[\tau_{AB}]=0$ and $\mathrm{tr}_B[\tau_{AB}]=0$ to ensure that ρ_{AB} has thermal marginals. Furthermore, we must have that $\mathrm{tr}[\tau_{AB}]=0$ and $\rho_A\otimes\rho_B+\tau_{AB}\geqslant 0$ to ensure that ρ_{AB} is a genuine quantum state. Since ρ_{AB},ρ_A,ρ_B are diagonal, then so is τ_{AB}. In this case, initially τ_{AB} has $2d$ parameters, but the three trace constraints reduce this to $2d-3$ independent parameters. For less cumbersome notation, define $\zeta := (Z_A Z_B)^{-1}$, $a:=\beta_A$, and $b:=\beta_B$, so that $\Delta\beta=a-b$. Also in the following we set $\omega_j=\omega$ for all j in H_{int} and analyze the systems at the time where $\omega t=\pi/2$, which corresponds to a complete transfer of energy q between the coupled microscopic energy levels.

We will see that positivity of ρ_{AB} gives us a range of compatible spectra. This range contains the product state $\rho_{AB}=\rho_A\otimes\rho_B$ as a special case, but in general ρ_{AB} will be classically correlated, and the size of these correlations may be measured by the quantum mutual information $I[\rho_{AB}] = S(\rho_A)+S(\rho_B)-S(\rho_{AB})$, where $S(\rho)=-\mathrm{tr}(\rho\log\rho)$ is the von Neumann entropy of ρ.

First we consider a two-qubit system $d=2$. The matrix $\tau_{AB}=\zeta\,\mathrm{diag}(x,-x,-x,x)$ satisfies the trace conditions on τ_{AB} for some $x\in\mathbb{R}$. Positivity of the matrix $\rho_{AB}=\rho_A\otimes\rho_B+\tau_{AB}$ leads to the constraint

$$-e^{-(a+b)} \leqslant x \leqslant \min\{e^{-a},e^{-b}\}, \tag{36}$$

and the state ρ_{AB} is correlated whenever $x\neq 0$. Let's take B to be hotter than or equal to A at the start; then $\Delta\beta=a-b\leqslant 0$ and $\min\{e^{-a},e^{-b}\}=e^{-a}$. Notice that high temperatures (i.e., small a,b) widen the range of x because larger temperatures are synonymous with reduced states being more mixed and this permits greater correlations. For this τ_{AB} we have

$$\begin{aligned}\rho_{AB} &= \mathrm{diag}(\lambda_{00},\lambda_{01},\lambda_{10},\lambda_{11})\\ &= \zeta\,\mathrm{diag}(1+x,e^{-b}-x,e^{-a}-x,e^{-(a+b)}+x),\end{aligned}$$

where the operators are diagonal in the $\{|00\rangle,|01\rangle,|10\rangle,|11\rangle\}$ basis. Given a,b, the mutual information is a function of x only and is shown in the top of Fig. 1. With our choice of

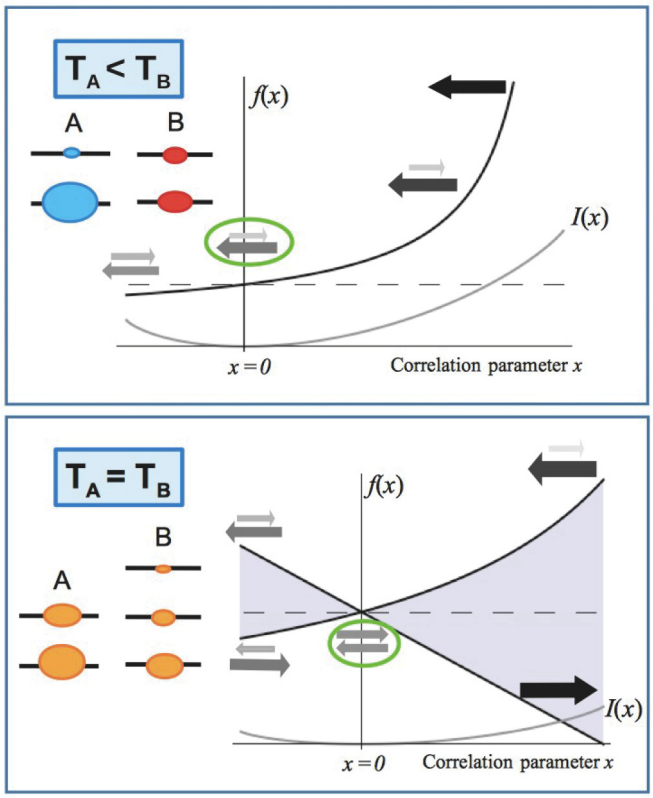

FIG. 1. (Color online) Why correlations matter: two systems A and B are at temperatures T_A and T_B initially. The generalized XFT, Eq. (18), is plotted as a function $f(x)$ of a correlation parameter x and the mutual information $I(x)$ (light gray curve) is also included. As $I(x)$ increases, the XFT deviates from its value in the "molecular chaos" assumption where $x=0$ (dashed line). The likelihood of heat flow direction is indicated by the gray arrows for different x. A circle round the arrows indicates the uncorrelated regime $x=0$. For some x, backward heat flow can be completely suppressed. *Top figure:* a two-qubit system where B is hotter than A; $f(x)$ is a single black curve. *Bottom figure:* a qubit-qutrit system at equal temperature. Unlike two qubits, the XFT is only defined by a range (light gray region) and we see, even though $T_A=T_B$, correlations can induce heat flow on average, thereby violating the principle of detailed balance.

energy-conserving Hamiltonian, transitions occur only in the energy-degenerate subspace spanned by $|01\rangle$ and $|10\rangle$.

The three thermodynamics relations become

$$f(x) := \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]} = \frac{e^{-b}-x}{e^{-a}-x},$$

$$\Delta\beta\langle q\rangle \geqslant \zeta(e^{-b}-e^{-a})\left(\Delta\beta + \ln\frac{e^{-a}-x}{e^{-b}-x}\right),$$

$$\eta = 1+\zeta[(e^{-b}-x)(e^{-\Delta\beta}-1)+(e^{-a}-x)(e^{\Delta\beta}-1)],$$

where we have defined $f(x)$ as the XFT and is displayed in the top of Fig. 1. The convention B hotter than A means that $q>0$ ($q<0$) when heat flows from B to A (A to B) and this is called forward (backward) transition. We analyze this thermodynamic setting in different extremal cases.

When the qubits are at equal temperatures $\Delta\beta = 0$, the likelihoods of the sharp forward to backward transitions are equal $\text{Prob}[\Gamma(q=1)] = \text{Prob}[\Gamma(q=-1)]$. Thus *detailed balance* is preserved no matter and no heat flows on average. This remains true regardless of the size of the initial correlations, which are determined by $x \in [-e^{2a}, e^{-a}]$.

A. Deterministic transitions

Within the permissible set of classically correlated states of AB, it is possible to find an initial state that exhibits *deterministic* heat transfer from hot B to cold A, and the XFT $f(x)$ diverges. System A has a virtual temperature of zero [18].

When x achieves its upper bound $x = e^{-a}$, the λ_{10} eigenvalue of ρ_{AB} is set to zero. Since the interaction is between the $|01\rangle$ and $|10\rangle$ states only, switching off λ_{10} means that the only transition allowed during the transfer time $\omega t = \pi/2$ is the forward one: $|01\rangle \to |10\rangle$. This is a deterministic transfer of heat from B to A and is made possible due to correlations, and happens regardless of the initial temperatures as long as $T_B > T_A$.

In this case, the XFT diverges $f(x = e^{-a}) \to \infty$ as expected since the backward transition does not occur $\text{Prob}[\Gamma(q=-1)] = 0$. The Clausius relation $\Delta\beta\langle q\rangle \geqslant -\infty$ tells us nothing about the direction of heat flow. In contrast, the correlation factor remains finite $\eta = 1-\zeta e^{-b}(1-e^{-\Delta\beta})^2 < 1$. Since it is equal to $\langle e^{-\Delta\beta q}\rangle$, and by assumption we have $\Delta\beta > 0$, it reflects the fact that the transition probability distribution is skewed so that the $q>0$ transition is far more likely than the backward one. Unlike the first two thermodynamic relations, the correlation factor is sensitive to the temperatures of the two qubits even in this extremal case. As long as $x = e^{-a}$, then $\lambda_{01} = 0$; however, $\lambda_{10} = \zeta(e^{-b}-e^{-a})$ remains temperature dependent and varies between 0 for $a = b$ and $\frac{1}{4}$ for $b \ll a \ll 1$, the limit where A and B are very hot but A at significantly cooler than B. Since η is linear in λ_{ij} it varies with the choice of a and b through $\lambda_{10} = \zeta(e^{-b}-e^{-a})$. We find the smallest value η can attain is $\frac{3}{4}$ for the most skewed distribution permitted $\{\lambda_{00},\lambda_{01},\lambda_{10},\lambda_{11}\} = \{\frac{1}{2},\frac{1}{4},0,\frac{1}{4}\}$, which delivers the biggest amount of heat $q=1$ per bipartite system ρ.

B. Anomalous heat flow

Aside from selecting deterministic heat transfer from hot to cold, a different set of correlations can enhance negative, or anomalous, heat flow from cold to hot.

At the minimum value of $x = -e^{-(a+b)}$ we have $\lambda_{01} = \zeta e^{-b}(1+e^{-a})$ and $\lambda_{10} = \zeta e^{-a}(1+e^{-b})$ giving

$$R = e^{\Delta\beta}\left(\frac{1+e^{-a}}{1+e^{-b}}\right).$$

When both temperatures are low so that $e^{-a}, e^{-b} \ll 1$ then we approach the uncorrelated ($x=0$) value for the ratio $f(x) \approx e^{\Delta\beta}$. This reflects the fact that at low temperatures the reduced states are purer and therefore cannot be highly correlated; thus the XFT tends to its molecular chaos form. In general, $f(x = -e^{-(a+b)}) < e^{\Delta\beta}$, since $e^{-a} < e^{-b}$ by assumption, so that, for maximal x, the XFT $f(x = -e^{-(a+b)})$ of forward to backward transfer is suppressed compared to the uncorrelated case $f(x=0) = e^{\Delta\beta}$ and "negative" heat transfer is enhanced. Note that for all allowed values of x in Eq. (36), we always have $f(x) > 1$; hence the likelihood of negative heat flow never surpasses that of normal heat flow from hot to cold. This directionality can, however, be lost in a high correlation environment, see for instance [8], where negative heat flow can be achieved on average when A and B begin in a pure entangled state.

The Clausius relation

$$\Delta\beta\langle q\rangle \geqslant \zeta(e^{-b}-e^{-a})\ln\left(\frac{1+e^{-b}}{1+e^{-a}}\right) > 0$$

suggests that a strictly positive amount of heat transfers from hot to cold. The correlation factor

$$\eta = 1+\zeta(e^{-a}-e^{-b})^2$$

is strictly greater than unity because the forward process is reduced and we are relatively more likely to observe a sharp amount of heat $q=-1$ being transferred, even though on average $\langle q\rangle > 0$.

Even in the elementary system of two qubits, we observe rich heat-exchange behavior as captured by the three thermodynamic relations. The XFT $f(x)$ and the correlation factor η provide the sharpest descriptions of the qubit-qubit thermodynamics. Let us now consider a qubit-qutrit system in which correlations lead to even stronger nonclassical features.

C. Distortion of detailed balance

In the preceding discussion we observed that the ratio of forward to backward heat flow is balanced when A and B are qubits at equal temperature. In this section, we will find that, remarkably, correlations distort this detailed balance when we consider a slightly more general system. In the following, A is a qubit and B is a qutrit.

It is easily checked that the matrix $\tau_{AB} = \zeta\,\text{diag}[x, y, -(x+y), -x, -y, x+y]$ satisfies the trace conditions on τ_{AB} for $x, y \in \mathbb{R}$. Positivity of the matrix $\rho_{AB} = \rho_A \otimes \rho_B + \tau_{AB}$ leads to

$$-1 \leqslant x \leqslant e^{-a}, \tag{37}$$

$$-e^{-b} \leqslant y \leqslant e^{-(a+b)}, \tag{38}$$

$$-e^{-(a+2b)} \leqslant x+y \leqslant e^{-2b}. \tag{39}$$

As for the $d = 2$ case, higher temperatures, corresponding to lower values of a and b, allow a greater variation of initial correlations.

This time transitions occur within the two energy-degenerate subspaces spanned by $\{|01\rangle,|10\rangle\}$ and $\{|02\rangle,|11\rangle\}$. The initial occupancies of these states are given by the eigenvalues of ρ_{AB}:

$$\lambda_{01} = \zeta(e^{-b} + y), \tag{40}$$

$$\lambda_{10} = \zeta(e^{-a} - x), \tag{41}$$

$$\lambda_{02} = \zeta[e^{-2b} - (x + y)], \tag{42}$$

$$\lambda_{11} = \zeta(e^{-(a+b)} - y). \tag{43}$$

The three thermodynamic relations are

$$\begin{aligned} \min &\left\{\frac{e^{-b}+y}{e^{-a}-x}, \frac{e^{-2b}-(x+y)}{e^{-(a+b)}-y}\right\} \\ &\leqslant f(x,y) := \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]} \\ &\leqslant \max\left\{\frac{e^{-b}+y}{e^{-a}-x}, \frac{e^{-2b}-(x+y)}{e^{-(a+b)}-y}\right\}, \end{aligned} \tag{44}$$

$$\begin{aligned} \Delta\beta\langle q\rangle \geqslant\ & \Delta\beta\zeta[(1+e^{-b})\delta + y] \\ &+\zeta\left((\delta + x + y)\ln\frac{e^{-a}-x}{e^{-b}+y}\right. \\ &\left.+(e^{-b}\delta - x)\ln\frac{e^{-(a+b)}-y}{e^{-2b}-(x+y)}\right), \end{aligned} \tag{45}$$

$$\begin{aligned} \eta = 1 &+ \zeta\{[e^{-b}(1+e^{-b}) - x](e^{-\Delta\beta} - 1) \\ &+[e^{-a}(1+e^{-b}) - (x+y)](e^{\Delta\beta} - 1)\}, \end{aligned} \tag{46}$$

and $\delta := e^{-b} - e^{-a}$.

The situation now is entirely different to the two qubit case. At equal temperatures the XFT $f(x,y) \neq 1$, in general; hence correlations can distort detailed balance. The XFT $f(x,y=0)$ for $\Delta\beta = 0$ is depicted in the bottom of Fig. 1 [the XFT is labeled simply $f(x)$]. Let $w = e^{-a} = e^{-b}$, and we consider two extremal choices (a) $x = -y = w$ and (b) $x = 0, y = w^2$. These choices effectively reduce the qutrit-qubit system down to a qubit-qubit one. However, there is a greater freedom arising from the qubit-qutrit correlations and this enables the distributions of the energy levels to be more skewed thereby admitting an asymmetric heat flow between the qubit A and qutrit B.

In (a), $\lambda_{01} = \lambda_{10} = 0$; hence no energy exchange occurs in the $\{|01\rangle,|10\rangle\}$ subspace. All of the dynamics is restricted to $\{|11\rangle,|02\rangle\}$, and since $\lambda_{11} = w^2 + w > \lambda_{02} = w^2$ the $|11\rangle \to |02\rangle$ transition is more likely to occur and this corresponds to heat flowing from A to B ($q < 0$). The XFT reflects this as it is upper bounded by 1:

$$f(x = w, y = -w) \in \left[\frac{2w-1}{w+1}, 1\right].$$

The lower bound must be positive; this restricts $w \geqslant 1/2$ or $T_A = T_B \geqslant (\ln 2)^{-1}$. Hence this situation can only occur when A, B are hot enough.

In (b), there is now only heat exchange in the $\{|01\rangle,|10\rangle\}$ subspace with $|01\rangle \to |10\rangle$ being more likely, that is, heat into A. This is captured by the XFT since

$$f(x = 0, y = w^2) \in [1, 1 + w]$$

and so there is a greater probability of observing an amount of heat $q = 1$ into the qubit A from the qutrit B. Hence correlations can skew heat flow even when the systems are at equal temperatures.

For $\Delta\beta = 0$ the Clausius inequality is not useful, and the correlation factor $\langle e^{-\Delta\mathcal{I}}\rangle = 1$ reduces to the value in the molecular chaos setting, even though the correlation measure $\Delta\mathcal{I} \neq 0$ for the dynamical trajectories. For instance, in case (a), for transitions in the $\{|11\rangle,|02\rangle\}$ subspace at $\Delta\beta = 0$, we find $\Delta\mathcal{I} = \ln(\lambda_{11}/\lambda_{02}) = \ln(1 + 1/w)$ (see Appendix C). This is quite different to the qubit-qubit case above where $\Delta\mathcal{I} = 0$. Nevertheless, for the qubit-qutrit system, the fluctuations in $\Delta\mathcal{I}$ are distributed in such a way that the "correlation fluctuation theorem" remains $\eta = 1$.

VI. DISCUSSION OF PHYSICAL ASSUMPTIONS

Some of the assumptions and terms we have introduced require clarification and comparison with current literature on the topic of XFTs.

A. Macroscopic significance of the function $\mathcal{I}$

The quantity $\mathcal{I}$ might at first glance seem to be merely a mathematical measure of correlation without any operational significance; however, this is not the case and it is simply a sharp version of the mutual information $I = \langle\mathcal{I}\rangle$, which in turn arises in extremely natural macroscopic and operational situations. For example, it is known to have the operational meaning as the work required to decorrelate a system in the asymptotic or macroscopic regime [19], while in other thermodynamic contexts it is identified as the correct measure of correlations in thermodynamic processes of bipartite quantum thermal systems [8,9] for averaged measurement outcomes. Finally, the role of mutual information in thermodynamics also arises in the context of Maxwell Demon scenarios [20], in which the extractable work is given by $W \leqslant kTI(X:Y)$, where Y is the measurement statistics of the demon and X is the actual microstate of the physical system. This energetic value of correlations can be cast into the form of a fluctuation theorem that amounts to a work extraction version of the results presented here, and recently has been experimentally verified in the context of feedback control of microscale thermodynamic systems [21].

B. But shouldn't fluctuation theorems be equalities, not inequalities?

That we have obtained an *inequality* in Eq. (18) to describe the high-correlation scenario might seem as a step in the wrong direction, given that fluctuation theorems give *equalities* that generalize the more traditional inequalities such as the Clausius relation $-W \leqslant -\Delta F$. However, it is easy to see that Eq. (18) is indeed a generalization of the traditional Jarzynski-Wójcik XFT. At the simplest level, it transitions to the traditional equality for zero initial correlations and

energy conserving dynamics—as it should. The breaking of the equality means the ratio of the forward and backward probabilities $R = \frac{\text{Prob}[\Gamma(+q)]}{\text{Prob}[\Gamma(-q)]}$ is now only located within a fixed, finite interval of size $\Delta R = e^{\Delta\beta q}(e^{-\Delta\mathcal{I}_u} - e^{-\Delta\mathcal{I}_l})$, governed by the correlative structure in the initial quantum state. This is again to be expected, since in the absence of specifying finer details of the interaction dynamics we cannot *a priori* tell whether a particular interaction is sensitive to the correlations. Put another way, some interactions are better at activating the correlations than others, and as we increase the correlations we widen this finite interval. Equivalently, in the exponentiated XFT *equality* in Eq. (30), this deviation is parametrized by the correlation factor η.

The increase of ΔR is exactly the distortion of the usual thermodynamic arrow; however, it is important to note a distinction between the fluctuation theorem setting and the setting based on traditional expectation values. As already mentioned, when we measure heat flow via $Q = \text{tr}[H_A(U\rho U^\dagger - \rho)]$ we are not introducing any local measurement disturbance into the initial system. Any entanglement present initially can influence the subsequent interactions and so can provide dramatic distortions of thermodynamic directionality. Indeed, for the most extreme case of a *pure* multipartite state with local thermal states *no* restriction exists beyond energy conservation and any such transformation can be done deterministically, including a maximal flow of heat from the colder to the hotter system (see [8] and [9] for details).

Recall that any mixed state σ_X admits a purification $\sigma_X \to \sigma_{XE} = |\psi\rangle_{XE}\langle\psi|$, which is unique up to arbitrary unitaries on the purifying environment E. If one adopts this perspective, one has that for *any* fixed thermal states ρ_A and ρ_B, the issue of how large ΔR is amounts to asking how much of the purifying correlations is present in the state ρ_{AB} for the composite system AB. Such states ρ_{AB} range between the product state (molecular chaos) $\rho_A \otimes \rho_B$, and the situation where $\rho_{AB} = |\Psi_{AB}\rangle\langle\Psi_{AB}|$, and B is a purification of A.

C. Going beyond sharp energy measurements

As mentioned, the sharp energy measurements $\mathcal{M}_1$ used are quite destructive of coherence, and so one might wonder whether an XFT can be obtained for more gentle POVMs. In other words, can we perform the time-reverse pairing trick using mixed quantum states?

Given a preparation of some σ_{AB} by the initial measurement $\mathcal{M}_1$, we wish to do the pairing trick with the state σ_{AB} and a time-reversed twin. If we drop the assumption that $\mathcal{M}_2$ is a sharp energy measurement, but leave it unspecified as $\mathcal{M}_2 = \{M^{(2)}_{\phi'} \otimes N^{(2)}_{\chi'}\}$ we then require a generalization of (9). Using the time-reversal invariance of the unitary interaction we have $\text{tr}[M^{(2)}_{\phi'} \otimes N^{(2)}_{\chi'} U\sigma_{AB}U^\dagger] = \text{tr}[(\Theta\sigma_{AB}\Theta^\dagger)U\Theta M^{(2)}_{\phi'} \otimes N^{(2)}_{\chi'}\Theta^\dagger U^\dagger]$, and from this we see that, for the pairing trick to work, the POVM elements of $\mathcal{M}_2$ must *themselves be valid quantum states* of the same form prepared by $\mathcal{M}_1$ and the set of elements should be closed under the time-reversal operator Θ. This on its own is a highly restrictive condition, and explains why forming a theorem for more general POVMs than the projective case is difficult.

Our overarching goal is to characterize intrinsically quantum mechanical effects in thermodynamics. Therefore, we must consider new methods for arriving at thermodynamical relations that fully account for "quantumness." One suggestion is presented in [22] which defines a random variable for energy that avoids the need to measure the initial system. Such an approach is beneficial since otherwise coherence and entanglement are destroyed by the initial measurement, yet it has been shown that both these properties have physical consequences in thermodynamic processes [23–25]. Another useful tool in quantum thermodynamics is the "single-shot" approach to thermodynamics [26], which originated from entanglement theory, and has recently been applied to fluctuations theorems [27] and follows the original measure-evolve-measure setting.

D. Application to a semiclassical maximum work theorem

The above results, and in particular (20), find simple application in a semiclassical maximum work theorem scenario [28] in which a quantity of ordered energy is extracted from a primary quantum subsystem A. The primary system is free to dump entropy in the form of heat into a heat sink B, with fast relaxation times, and exchange mechanical work with a third (classical) adiabatic system C.

On the assumption of conservation of energy for the composite system ABC and the adiabaticity of C the averaged relation (20) leads to

$$dW_C \leqslant -dU_A + dQ - T_B\langle d\mathcal{I}\rangle, \tag{47}$$

where $dQ := \langle dq\rangle$ corresponds to heat flowing into A, and we assume for simplicity that no net work is done on A. This does make the identification of $\Delta\epsilon$ with mechanical work, which can be debated as more or less sensible given that in extreme quantum regimes this can have broad distributions. In the case in which the system starts or finishes in equilibrium then we may identify the heat flow dQ with the thermodynamic entropy in the usual way. We do not expand on these points here, but at the simplest level the main point of this application is to illustrate the contribution that the initial correlations between the primary subsystem and the reversible heat sink provide to the usual maximum work theorem, and in the process illustrate the well-known energetic value of correlations [25,29–31].

VII SUMMARY AND OUTLOOK

We have extended the Jarzynski-Wójcik exchange fluctuation theorem to the situation where we drop the assumption of molecular chaos, and allow correlations to exist in the composite state. These correlations result in a modification of the XFT relation and can enhance the probability of heat flowing in the backward direction. We have applied our results to deriving a semiclassical maximum work theorem for correlated systems. Our work highlights the difficulty of obtaining further results for situations without initial and final measurements of energy. Our result show a deviation of the traditional XFT due to correlations present, and depends on a term that takes the form of a sharp mutual information. A similar result has already been obtained for the case of the work fluctuation theorem [32], in which one allows feedback

control. There the relevant mutual information is between the controller and the primary system. Furthermore, such mutual information has already been shown to be experimentally relevant to microscale and nanoscale thermodynamics [21].

ACKNOWLEDGMENTS

D.J. is supported by the Royal Commission for the Exhibition of 1851. S.J. is supported by EPSRC Grant No. EP/K022512/1, by the ERC grants QFTCMPS, and SIQS by the cluster of excellence EXC 201 Quantum Engineering and Space-Time Research. T.R. is supported by the UK Engineering and Physical Sciences Research Council. Y.H. is supported by the Japan Society for the Promotion of Science for Young Scientists. S.N. and M.M. are supported by Project for Developing Innovation Systems of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

APPENDIX A: DERIVATION OF EQ. (13)

For any joint local POVMs $\{M_i\}$ on A and $\{N_j\}$ on B, we have defined the quantity $\mathcal{I}(\rho_{AB}; M_i, N_j)$ via the expression

$$\mathcal{I}(\rho_{AB}; M_i, N_j) := \ln\left(\frac{\mathrm{tr}[M_i \otimes N_j \rho_{AB}]}{(\mathrm{tr}[M_i \rho_A]\mathrm{tr}[N_j \rho_B])}\right). \quad \text{(A1)}$$

To show (13) we consider the sharp energy measurement $\mathcal{M}_1 = \{M_\phi \otimes N_\chi\}$, where $\{M_\phi = |\phi\rangle\langle\phi|\}$ and $\{\mathcal{N}_\chi = |\chi\rangle\langle\chi|\}$ are the rank-1 projectors in the local energy eigenbases.

For this particular initial measurement, the probability $p(|\phi,\chi\rangle)$ of projecting into the state $|\phi,\chi\rangle$ under $\mathcal{M}_1$ is simply given by $p(|\phi,\chi\rangle) = \langle\phi,\chi|\rho_{AB}|\phi,\chi\rangle = \mathrm{tr}[M_\phi \otimes N_\chi \rho_{AB}]$. However, from the definition of $\mathcal{I}$ we have that

$$\mathrm{tr}[M_\phi \otimes N_\chi \rho_{AB}] = e^{\mathcal{I}(\rho_{AB}; M_\phi, N_\chi)} \mathrm{tr}[M_\phi \rho_A]\mathrm{tr}[N_\chi \rho_B].$$

By assumption the state ρ_{AB} has thermal marginals and so we have that

$$\mathrm{tr}[M_\phi \rho_A]\mathrm{tr}[N_\chi \rho_B] = e^{-\beta_A E_\phi - \beta_B E_\chi}/(Z_A Z_B).$$

Substitution of these terms into $p(|\phi,\chi\rangle)$ gives

$$p(|\phi,\chi\rangle) = e^{-\beta_A E_\phi - \beta_B E_\chi - \log(Z_A Z_B) + \mathcal{I}[\rho_{AB}; M_\phi, N_\chi]},$$

while the probability of obtaining $|\phi',\chi'\rangle_*$ in the same measurement on ρ_{AB} is given by

$$p(|\phi',\chi'\rangle_*) = e^{-\beta_A E_{\phi'} - \beta_B E_{\chi'} - \log(Z_A Z_B) + \mathcal{I}[\rho_{AB}; M_{\phi'_*}, N_{\chi'_*}]},$$

because time-reversal symmetry means that $E_{\phi'} = E_{\phi'_*}$ and $E_{\chi'} = E_{\chi'_*}$. Note that the $\mathcal{I}$ term makes $p(|\phi',\chi'\rangle_*)$ independent of $p(|\phi',\chi'\rangle)$. Taking the ratio of these two probabilities leads us to the desired result

$$p(|\phi,\chi\rangle) = p(|\phi',\chi'\rangle_*) e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)}, \quad \text{(A2)}$$

where $\Delta\beta = \beta_A - \beta_B$, and

$$\Delta\mathcal{I}(\gamma) = \mathcal{I}[\rho_{AB}; M_{\phi'_*}, N_{\chi'_*}] - \mathcal{I}[\rho_{AB}; M_\phi, N_\chi],$$

as claimed.

APPENDIX B: DERIVATION OF EQ. (30)

Here we fill in the details leading up to Eq. (30). Using the discretized expression for $P(x)$, we have

$$\begin{aligned} P_X(x)e^{-x} &= \sum_{\mu\nu} \delta(x - X_{\mu\nu}) p_{\mu\nu} e^{-x} \\ &= \sum_{\mu\nu} \delta(x - X_{\mu\nu}) p_{\mu\nu} e^{-X_{\mu\nu}} \\ &\equiv \sum_{\mu\nu} \delta(x - (-\widetilde{X}_{\mu\nu}))\widetilde{f}_{\mu\nu} \\ &:= F_{\widetilde{X}}(-x). \end{aligned}$$

In the first to second we have used a property of δ functions, $g(x)\delta(x - x_0) = g(x_0)\delta(x - x_0)$, for some function $g(x)$, and in the second to third line we have defined $\widetilde{f}_{\mu\nu} := p_{\mu\nu} e^{-X_{\mu\nu}}$ and $\widetilde{X}_{\mu\nu} := -X_{\mu\nu}$. The third line is a probability density function for the new random variable $\widetilde{X}_{\mu\nu}$ if $\widetilde{f}_{\mu\nu}$ is a probability distribution.

With the choices made for $X_{\mu\nu}, f_\nu, f^0_\nu$ in Eqs. (26)–(28) we have simply $X_{\mu\nu} = \ln p^0_\mu - \ln f^0_\nu$ and using the expressions for these uncorrelated probability distributions we obtain

$$X_{\mu\nu} = \beta_A q^A_{\mu\nu} + \beta_B q^B_{\mu\nu}, \quad \text{(B1)}$$

in terms of the sharp heat into A and B, $q^A_{\mu\nu} = E'_\phi - E_\phi$ and $q^B_{\mu\nu} = E'_\chi - E_\chi$. Since $\delta(x - X_{\mu\nu})e^{-X_{\mu\nu}} = \delta(x - X_{\mu\nu})e^{-x}$, we are allowed to drop the μ,ν labels and convert $X_{\mu\nu}$ into the continuous random variable $x = \beta_A q^A + \beta_B q^B$. Define $q := q^A$, $\Delta\epsilon := q^B + q$, and $\Delta\beta = \beta_A - \beta_B$, then

$$\begin{aligned} \int P_X(x)e^{-x}dx &= \sum_{\mu,\nu} \int \delta(x - X_{\mu\nu}) p_{\mu\nu} e^{-x} dx \\ &\equiv \sum_\gamma \mathrm{Prob}[\gamma] e^{-\Delta\beta q - \beta_B \Delta\epsilon} \\ &= \langle e^{-\Delta\beta q - \beta_B \Delta\epsilon}\rangle. \end{aligned}$$

In the first to second line we have used the fact that $X_{\mu\nu}$ only picks out the values of x that are allowed by the histories γ, for which $p_{\mu\nu} = \mathrm{Prob}[\gamma]$.

Now consider $F_{\widetilde{X}}(-x)$. Let us formally write $\widetilde{f}_{\mu\nu} = f_\nu \widetilde{f}_{\mu|\nu}$. We have by definition $\widetilde{f}_{\mu\nu} = p_{\mu\nu} e^{-X_{\mu\nu}}$. Since $p_{\mu\nu} = p_\mu p_{\nu|\mu}$ and $e^{-X_{\mu\nu}} = \frac{f_\nu}{p_\mu} e^{-\Delta I_{\mu\nu}}$ we can deduce

$$\widetilde{f}_{\mu|\nu} := p_{\nu|\mu} e^{-\Delta\mathcal{I}_{\mu\nu}}. \quad \text{(B2)}$$

Therefore

$$F_{\widetilde{X}}(-x) = \sum_{\mu\nu} \delta(x + \widetilde{X}_{\mu\nu}) f_\nu p_{\nu|\mu} e^{-\Delta\mathcal{I}_{\mu\nu}}. \quad \text{(B3)}$$

Is $f_\nu p_{\nu|\mu}$ a valid probability distribution? The f_ν part is the probability of projecting the state ρ_{AB} onto $M_{\nu_*} \otimes N_{\nu_*} = |\phi',\chi'\rangle\langle\phi',\chi'|_*$. To be consistent, the conditional $f_{\mu|\nu} := |(|\phi,\chi\rangle_*, U|\phi',\chi'\rangle_*|^2$, but, from time-reversal symmetry in Eq. (9), we have $f_{\mu|\nu} = p_{\nu|\mu}$. Clearly, $\sum_\mu f_{\mu|\nu} = 1$ and $\sum_{\mu\nu} f_\nu f_{\mu|\nu} = 1$; hence $f_\nu p_{\nu|\mu} =: f_{\mu\nu}$ is indeed a valid probability (note carefully the difference between tildes and no tildes); in fact it is equal to $\mathrm{Prob}[\gamma^*]$.

Finally we may evaluate the correlation factor

$$\eta := \int F_{\widetilde{X}}(-x)dx = \sum_{\mu\nu}\int \delta(x+\widetilde{X}_{\mu\nu}) f_{\mu\nu} e^{-\Delta\mathcal{I}_{\mu\nu}} dx$$
$$= \sum_{\mu\nu} f_{\mu\nu} e^{-\Delta\mathcal{I}_{\mu\nu}} \equiv \sum_{\gamma^*} \mathrm{Prob}[\gamma^*] e^{-\Delta\mathcal{I}(\gamma^*)} = \langle e^{-\Delta\mathcal{I}}\rangle_*.$$

We have used the fact that $\widetilde{X}_{\mu\nu} = -X_{\mu\nu}$ is the time-reversed version of $X_{\mu\nu}$, this gets us from the second to third line, and the asterisk on the bottom line in $\langle e^{-\Delta\mathcal{I}}\rangle_*$ indicates that this average is taken with respect to the time-reversed probability distribution.

APPENDIX C: DETAILS FOR THE TOY EXAMPLE

Consider first the generalized exchange fluctuation theorem in Eq. (18),

$$e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_l} \leqslant \frac{\mathrm{Prob}[\Gamma(q,\Delta\epsilon)]}{\mathrm{Prob}[\Gamma(-q,-\Delta\epsilon)]} \leqslant e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_u}. \tag{C1}$$

We focus on the XFT for positive heat $q = 1$ flows into A; in this example these are the histories

$$\gamma[q=1|j]: \ (m,n) = (0,j) \to (m',n') = (1,j-1),$$

for $j = 1,\ldots,d-1$ and we have chosen the time-reversed state to be the spin-flipped one. For these transitions,

$$\Delta\epsilon = \langle 1,j-1|H_A + H_B|1,j-1\rangle - \langle 0,j|H_A+H_B|0,j\rangle = 0. \tag{C2}$$

Note that the $\Delta\epsilon = 0$ even for the reverse transition $\gamma[q=-1|j]$, and these are the only histories permitted by the interaction.

The probabilities for these transitions are

$$\mathrm{Prob}[\gamma[q=1|j]] = \langle 0,j|\rho_{AB}|0,j\rangle |\langle 1,j-1|U|0,j\rangle|^2$$
$$= \lambda_{0,j}\sin^2\omega_j t,$$
$$\mathrm{Prob}[\gamma[q=-1|j]] = \langle 1,j-1|\rho_{AB}|1,j-1\rangle|$$
$$\times \ \langle 0,j|U^\dagger|1,j-1\rangle|^2$$
$$= \lambda_{1,j-1}\sin^2\omega_j t.$$

The correlation function for projective energy measurement $M_m \otimes N_n = |m\rangle\langle m| \otimes |n\rangle\langle n|$ is

$$\mathcal{I}(\rho_{AB};m,n) = \ln\left[\frac{\langle mn|\rho_{AB}|mn\rangle}{\langle m|\rho_A|m\rangle\langle n|\rho_B|n\rangle}\right]$$
$$= \beta_A m + \beta_B n + \ln\frac{\lambda_{mn}}{Z_A Z_B}.$$

The sharp heat into A is $q = \langle m'|H_A|m'\rangle - \langle m|H_A|m\rangle = m' - m = 0, \pm 1$ since $m = 0,1$. The change in the correlation function is

$$\Delta\mathcal{I}(\gamma[q=1|j]) = \mathcal{I}(\rho_{AB};1,j-1) - \mathcal{I}(\rho_{AB};0,j)$$
$$= \Delta\beta + \ln\frac{\lambda_{1,j-1}}{\lambda_{0,j}}.$$

Later we will also make use of

$$\Delta\mathcal{I}(\gamma[q=-1|j]) = -\Delta\beta + \ln\frac{\lambda_{0,j}}{\lambda_{1,j-1}} = -\Delta\mathcal{I}(\gamma[q=1|j]).$$

The upper u and lower l bounds on $\Delta\mathcal{I}$ are given by $\Delta\mathcal{I}_u = \Delta\beta + \max_j\{\ln\frac{\lambda_{1,j-1}}{\lambda_{0,j}}\}_{j=1}^{d-1}$ and $\Delta\mathcal{I}_l = \Delta\beta + \min_j\{\ln\frac{\lambda_{1,j-1}}{\lambda_{0,j}}\}_{j=1}^{d-1}$.

Substituting these expressions into Eq. (18) we obtain

$$\min\left\{\frac{\lambda_{0,j}}{\lambda_{1,j-1}}\right\} \leqslant \frac{\mathrm{Prob}[\Gamma(q=1)]}{\mathrm{Prob}[\Gamma(q=-1)]} \leqslant \max\left\{\frac{\lambda_{0,j}}{\lambda_{1,j-1}}\right\}. \tag{C3}$$

The second thermodynamic inequality is simply $\Delta\beta\langle q\rangle \geqslant \langle\Delta\mathcal{I}\rangle$ because $\Delta\epsilon = 0$ for this energy-conserving interaction. The average difference of the correlation function is

$$\langle\Delta\mathcal{I}\rangle = \sum_j \mathrm{Prob}[\gamma[q=\pm1|j]]\Delta\mathcal{I}(\gamma[q=\pm1|j])$$
$$= \sum_j(\lambda_{0j}-\lambda_{1,j-1})\left(\Delta\beta + \ln\frac{\lambda_{1,j-1}}{\lambda_{0,j}}\right)\sin^2\omega_j t.$$

Finally we turn our attention to the correlation factor

$$\eta = \langle e^{-\Delta\mathcal{I}}\rangle_* = \sum_{\mu\nu} f_{\mu\nu} e^{-\Delta\mathcal{I}_{\mu\nu}}$$

from Eq. (30). The terms appearing in the sum are defined in Sec. IV and Appendix B. The initial and final measurement outcome labels are $\mu = (0,j)$ and $\nu = (1,j-1)$; we have $\Delta\mathcal{I}_{\mu\nu} = \Delta\beta + \ln\frac{\lambda_{1,j-1}}{\lambda_{0j}}$ with probability $f_{\mu\nu} = f_\nu f_{\mu|\nu}$, where $f_{\mu|\nu} = |\langle 0,j|U^\dagger|1,j-1\rangle|^2 = \sin^2\omega_j t$ and $f_\nu = \langle 1,j-1|\rho_{AB}|1,j-1\rangle = \lambda_{1,j-1}$. Including also the reverse transitions $\mu = (1,j-1)$ and $\nu = (0,j)$, and not forgetting the contributions from the trivial transitions in the one-dimensional energy subspaces $\{|00\rangle\}$ and $\{|1,d-1\rangle\}$ for which $\Delta\mathcal{I}_{\mu\nu} = 0$, we obtain

$$\eta = \lambda_{00} + \lambda_{1,d-1} + \sum_{j=1}^{d-1}(\lambda_{1,j-1}e^{-\Delta\beta - \ln\frac{\lambda_{1,j-1}}{\lambda_{0j}}}$$
$$+ \lambda_{0j}e^{\Delta\beta - \ln\frac{\lambda_{0j}}{\lambda_{1,j-1}}})\sin^2\omega_j t$$
$$= \lambda_{00} + \lambda_{1,d-1} + \sum_{j=1}^{d-1}(\lambda_{0,j}e^{-\Delta\beta} + \lambda_{1,j-1}e^{-\Delta\beta})\sin^2\omega_j t.$$

Finally, using $\sum_{j=0}^{d-1}(\lambda_{0,j}+\lambda_{1,j}) = 1$, this form of η may be rewritten to give one in Eq. (35c).

[1] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[2] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997).
[3] G. Crooks, J. Stat. Phys. **90**, 1481 (1998).
[4] H. Price, *Time's Arrow and Archimedes' Point* (Oxford University Press, Oxford, 1996).
[5] H. D. Zeh, *The Physical Basis of The Direction of Time*, 4th ed. (Springer, New York, 2001).
[6] C. Jarzynski and D. K. Wójcik, Phys. Rev. Lett. **92**, 230602 (2004).
[7] L. Boltzmann, *Lectures on Gas Theory* (Dover, New York, 2011).
[8] M. H. Partovi, Phys. Rev. E **77**, 021110 (2008).
[9] D. Jennings and T. Rudolph, Phys. Rev. E **81**, 061130 (2010).
[10] S. Popescu, A. Short, and A. Winter, Nat. Phys. **2**, 754 (2006).
[11] F. G. S. L. Brandão and M. Plenio, Nat. Phys. **4**, 873 (2008).
[12] F. G. S. L. Brandão, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens, Phys. Rev. Lett. **111**, 250404 (2013).
[13] J. M. R. Parrondo, C. Van den Broeck, and R. Kawai, New J. Phys. **11**, 073008 (2009).
[14] S. Deffner and E. Lutz, Phys. Rev. Lett. **107**, 140404 (2011).
[15] J. M. Horowitz, Phys. Rev. E **85**, 031110 (2012).
[16] M. Campisi, P. Talkner, and P. Hänggi, Phys. Rev. Lett. **105**, 140601 (2010).
[17] T. Albash, D. A. Lidar, M. Marvian, and P. Zanardi, Phys. Rev. E **88**, 032146 (2013).
[18] N. Brunner, N. Linden, S. Popescu, and P. Skrzypczyk, Phys. Rev. E **85**, 051117 (2012).
[19] B. Groisman, S. Popescu, and A. Winter, Phys. Rev. A **72**, 032317 (2005).
[20] W. H. Zurek, in *Proceedings Frontiers of Nonequilibrium Quantum Statistical Mechanics*, edited by G. T. Moore and M. O. Scully (Plenum, New York, 1986), pp. 145–150.
[21] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, Nat. Phys. **6**, 988 (2010).
[22] A. E. Allahverdyan, Phys. Rev. E **90**, 032137 (2014).
[23] M. Lostaglio, D. Jennings, and T. Rudolph, Nat. Commun. **6**, 6383 (2015).
[24] M. Lostaglio, K. Korzekwa, D. Jennings, and T. Rudolph, Phys. Rev. X **5**, 021001 (2015).
[25] S. Jevtic, D. Jennings, and T. Rudolph, Phys. Rev. Lett. **108**, 110403 (2012).
[26] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, New J. Phys. **13**, 053015 (2011).
[27] N. Y. Halpern, A. J. P. Garner, O. C. O. Dahlsten, and V. Vedral, New J. Phys. **17**, 095003 (2015).
[28] H. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, New York, 1985).
[29] L. Szilard, Z. Phys. **53**, 840 (1929).
[30] C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).
[31] T. Sagawa and M. Ueda, Phys. Rev. Lett. **100**, 080403 (2008).
[32] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).