

Exchange Fluctuation Theorem for correlated quantum systems

Sania Jevtic¹, Terry Rudolph², and David Jennings²

¹*Mathematical Sciences, Room 501, John Crank Building,
Brunel University, Uxbridge UB8 3PH, United Kingdom*

²*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

(Dated: December 5, 2014)

We extend the Exchange Fluctuation Theorem for energy exchange between thermal quantum systems beyond the assumption of molecular chaos, and describe the non-equilibrium exchange dynamics of correlated quantum states. The relation quantifies how the tendency for systems to equilibrate is modified in high-correlation environments. Our results elucidate the role of measurement disturbance for such scenarios. We show a simple application by finding a semi-classical maximum work theorem in the presence of correlations.

I. INTRODUCTION

Fluctuation theorems describe non-equilibrium 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]. However, in addition to focussing on the extraction of mechanical work from a single system, an equally fundamental topic is the thermodynamic tendency of multipartite systems to equilibrate. The canonical example of this is heat exchange between two thermal systems at different temperatures and leads us instead to fluctuation theorems for heat that provide a quantitative description of the fluctuations in energy exchange between two hot bodies.

The thermodynamic arrow [5, 6] 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, $\rho_A \otimes \rho_B \rightarrow U \rho_A \otimes \rho_B U^\dagger$, we define heat flow into A as $Q_A = \text{tr}[H_A \rho'_A] - \text{tr}[H_A \rho_A]$ 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

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

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 [4], which states that for two systems A and B , initially at temperatures T_A and T_B , the probability $P(q)$ of a *sharp* exchange of

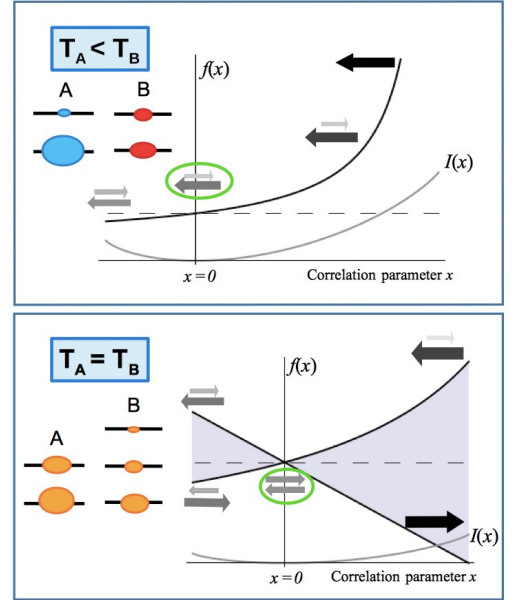


FIG. 1: (Color online). **Why correlations matter:** Two systems A and B are at temperatures T_A and T_B initially. The generalised XFT, equation (17), is plotted as a function $f(x)$ of a correlation parameter x and the mutual information $I(x)$ (light grey curve) is also included. As $I(x)$ increases, 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 grey 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 grey region) and we see, even though $T_A = T_B$, correlations can induce heat flow on average, thereby violating the principle of detailed balance. This example is discussed in section V.

energy q from B to A obeys the relation

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

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. A simple application of Jensen's inequality to (2) leads to an averaged inequality $\langle q \rangle (1/T_A - 1/T_B) \geq 0$. This seems to suggest that (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. In general, a sharp (rank-1 projective) energy measurement will produce non-classical disturbances in the quantum state of a system, and even though the expression $\langle q \rangle (1/T_A - 1/T_B) \geq 0$ still provides a thermodynamic directionality, it is no longer identical to (1), which does not involve any measurement disturbance.

A. The assumption of molecular chaos and the role of correlations.

The scope of the XFT 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 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, equation (1), and the Jarzynski-Wójcik Fluctuation Theorem, equation (2), are therefore limited in application, and 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 there should exist *no directional constraint* whatsoever, and for such situations no equality such as equation (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 pure state thermality is of importance and turns out to be the *typical* scenario with respect to the Haar measure. Specifically, for a randomly chosen multipartite state $|\Psi\rangle$ of a system with fixed energy, the reduced state for a small subsystem is exponentially likely to be Gibbsian [12], 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 [13], often through their parallel formulations as resource theories [14]. 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 are allowed to evolve under non-equilibrium dynamics 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.

The statistics of energy exchange (and even particle exchange) has been investigated previously in a number of noteworthy publications [20–23]. However, 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 this broader framework.

This paper is structured in the following way. In section 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 section III we analyse the heat flow in a parallel manner to the original XFT [4], however without assuming molecular chaos. We find that dropping this assumption enforces the use of a sharp mutual information measure, quantifying 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 allowing for quantum fluctuations stemming from intrinsic randomness in pure quantum states. In section IV we present a derivation for a non-equilibrium equality in the presence of correlations. The approach is from a more abstract setting using probability theory and random variables. The upshot of this is the acquirement of the ‘‘efficacy factor’’ that gives

a notion of the strength of correlations and their effect on the directionality of thermodynamics processes. Section V brings all these ideas together through an exactly solvable toy model: a qubit-qudit system with an energy conserving interaction. In section VIC 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 VID 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. The thermodynamic scenario

In defining heat exchange in [4], 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: $\rho_{AB} \rightarrow \mathcal{M}_1[\rho_{AB}] \rightarrow U \circ \mathcal{M}_1[\rho_{AB}] \rightarrow \mathcal{M}_2 \circ U \circ \mathcal{M}_1[\rho_{AB}]$.

As mentioned, the central assumptions of the XFT are time-reversal symmetry of the underlying dynamics and the initial thermality of the individual subsystems. Specifically, we assume that $\Theta^\dagger H_A \Theta = H_A$ and $\Theta^\dagger H_B \Theta = H_B$, where Θ is the anti-unitary time-reversal operator. In what follows we use joint energy eigenstates $|\phi, \chi\rangle$, 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 $\{|\phi\rangle\}$ 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 $\Theta(|\phi\rangle\langle\phi| \otimes |\chi\rangle\langle\chi|)$ is always in the basis set $\{|\phi\rangle\langle\phi| \otimes |\chi\rangle\langle\chi|\}$ for any ϕ and χ . The thermal marginal

states of ρ_{AB} are then given by

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

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

where Z_A and Z_B are the usual partition functions for A and B , and β_A, β_B are their 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 what follows we 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

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

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 $\text{tr}[\rho_{AB}(H_A + H_B)] = \text{tr}[U\rho_{AB}U^\dagger(H_A + H_B)]$. 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 in the XFT, 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 $\rho_{\text{therm}} = Z^{-1} \sum_k e^{-\beta E_k} |E_k\rangle\langle E_k|$.

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

$$\Delta\epsilon = \langle\phi', \chi'|(H_A + H_B)|\phi', \chi'\rangle - \langle\phi, \chi|(H_A + H_B)|\phi, \chi\rangle \quad (6)$$

and with an energy transfer q into A defined as

$$q = \langle\phi'|H_A|\phi'\rangle - \langle\phi|H_A|\phi\rangle. \quad (7)$$

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

III. AN EXCHANGE FLUCTUATION THEOREM FOR CORRELATED QUANTUM STATES

We can now formulate a generalized XFT that drops assumption (III) of molecular chaos (see section 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 equation (5), has probability

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

where $U \equiv e^{-iHt}$ and $H \equiv H_A + H_B + H_{\text{int}}$ 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, $\Theta^\dagger H \Theta = H$, and the anti-unitarity of Θ it follows that $U = \Theta^\dagger U^\dagger \Theta$, and so we have that

$$|\langle\phi', \chi'|U|\phi, \chi\rangle|^2 = |(\Theta|\phi, \chi\rangle, U\Theta|\phi', \chi'\rangle)|^2. \quad (9)$$

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

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

For a sharp energy measurement, we simply have $\mathcal{M}_1 = \{M_\phi \otimes N_\chi\}$ where $\{M_\phi = |\phi\rangle\langle\phi|\}$ and $\{N_\chi = |\chi\rangle\langle\chi|\}$ are the rank-1 projectors in the energy eigenbases¹, and when ρ_{AB} is a correlated quantum state having thermal marginals as in equations (3) and (4), we find that $\mathcal{M}_1$ maps ρ_{AB} into a classically correlated state, diagonal in the energy eigenbasis. Moreover, it is readily seen that $\text{tr}_{B(A)}[\mathcal{M}_1[\rho_{AB}]] = \rho_{A(B)}$, and so the state $\mathcal{M}_1[\rho_{AB}]$ has the same thermal marginals as ρ_{AB} ².

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

$$p(|\phi, \chi\rangle) = e^{-\beta_A E_\phi - \beta_B E_\chi - \log(Z_A Z_B) + \mathcal{I}[\rho_{AB}; M_\phi, N_\chi]}, \quad (11)$$

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

$$p(|\phi, \chi\rangle) = p(|\phi', \chi'\rangle_*) e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)}, \quad (12)$$

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

$$\Delta\mathcal{I}(\gamma) = \mathcal{I}[\rho_{AB}; M_{\phi'_*}, N_{\chi'_*}] - \mathcal{I}[\rho_{AB}; M_\phi, N_\chi] \quad (13)$$

is the appropriate correlation measure, dependent only on the initial state ρ_{AB} and the initial measurement $\mathcal{M}_1$. A derivation of (12) is provided in the appendix. Note that with this definition of $\Delta\mathcal{I}(\gamma)$, the assumption of molecular chaos gives $\Delta\mathcal{I}(\gamma) = 0$.

Combining (12) with (9) we find that

$$\frac{\text{Prob}[\gamma]}{\text{Prob}[\gamma^*]} = e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)}, \quad (14)$$

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

$$\gamma^* \equiv (\rho_{AB}; |\phi', \chi'\rangle_* \xrightarrow{U} |\phi, \chi\rangle_*). \quad (15)$$

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 (14) 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 $\Gamma(q, \Delta\epsilon)^* = \Gamma(-q, -\Delta\epsilon)$. The probability of the former is given by

$$\text{Prob}[\Gamma(q, \Delta\epsilon)] = \sum_{\gamma \in \Gamma(q, \Delta\epsilon)} e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}(\gamma)} \text{Prob}[\gamma^*]. \quad (16)$$

¹ The function $\mathcal{I}$ may be related to the classical relative entropy of the joint measurement outcomes through the relation $I_c(\mathcal{M} : \mathcal{N}) = \sum_{i,j} \text{tr}[M_i \otimes N_j \rho_{AB}] \mathcal{I}(\rho_{AB}; M_i, N_j)$.

² 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 $I[\sigma_{AB}; A : B] = S[\sigma_A] + S[\sigma_B] - S[\sigma_{AB}]$.

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

$$e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_l} \leq \frac{\text{Prob}[\Gamma(q, \Delta\epsilon)]}{\text{Prob}[\Gamma(-q, -\Delta\epsilon)]} \leq e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_u}. \quad (17)$$

The XFT in equation (17) is a generalization of Jarzynski and Wójcik in equation (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. As can be seen, 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 (14) and sum over $\gamma \in \Gamma$, to obtain the non-equilibrium equality for an initially correlated state

$$\langle e^{-\Delta\beta q - \beta_B \Delta\epsilon + \Delta\mathcal{I}} \rangle = 1 \quad (18)$$

and then using Jensen's inequality we have that

$$\Delta\beta \langle q \rangle + \beta_B \langle \Delta\epsilon \rangle - \langle \Delta\mathcal{I} \rangle \geq 0. \quad (19)$$

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 $\langle \Delta\epsilon \rangle = 0$ we have $\Delta\beta \langle q \rangle - \langle \Delta\mathcal{I} \rangle \geq 0$, which reduces to (1) the Clausius relation $Q(1/T_A - 1/T_B) \geq 0$ 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 $Q(1/T_A - 1/T_B) \geq \langle \Delta\mathcal{I} \rangle$ 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 equation (17) 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. A NON-EQUILIBRIUM EQUALITY IN THE PRESENCE OF CORRELATIONS

In the previous section we derived a non-equilibrium equality, equation (18), for initially correlated systems using the concrete idea of histories and time reversal. Here, we follow the compact approach of [11] to isolate the ‘‘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 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

$$p_\mu = \text{tr}[M_\mu \otimes N_\mu \rho M_\mu^\dagger \otimes N_\mu^\dagger]. \quad (20)$$

The initial state ρ is again assumed to have thermal marginals from which we define the uncorrelated probability distribution

$$p_\mu^0 = \text{tr}[M_\mu \rho_A] \text{tr}[N_\mu \rho_B] = \frac{e^{-\beta_A \phi}}{Z_A} \frac{e^{-\beta_B \chi}}{Z_B}. \quad (21)$$

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 labelled by $\nu = (\phi', \chi')$ and probability

$$p_{\nu|\mu} = \text{tr}[M_\nu \otimes N_\nu \rho'_\mu M_\nu^\dagger \otimes N_\nu^\dagger] = |\langle\phi', \chi'|U|\phi, \chi\rangle|^2. \quad (22)$$

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

$$p_{\mu\nu} = \langle\phi, \chi|\rho|\phi, \chi\rangle |\langle\phi', \chi'|U|\phi, \chi\rangle|^2 \quad (23)$$

is simply $\text{Prob}[\gamma]$ in equation (8).

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

$$P_X(x) = \sum_{\mu\nu} \delta(x - X_{\mu\nu}) p_{\mu\nu} \quad (24)$$

where $X_{\mu\nu}$ is a discrete random variable distributed according to $p_{\mu\nu}$. Define the function $F_{\tilde{X}}(-x) = P_X(x)e^{-x}$ that is analogous to the time-reversed probability in the XFT. It may be shown that $F_{\tilde{X}}$ is a probability density function for the random variable $\tilde{X} := -X$ if $p_{\mu\nu}e^{-X_{\mu\nu}}$ is a probability distribution (see the Appendix for details).

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

$$X_{\mu\nu} = \ln p_\mu - \ln f_\nu + \Delta\mathcal{I}_{\mu\nu}. \quad (25)$$

where the correlation

$$\Delta\mathcal{I}_{\mu\nu} = \mathcal{I}_\nu - \mathcal{I}_\mu = \ln\left(\frac{p_\mu}{p_\mu^0}\right) - \ln\left(\frac{f_\nu}{f_\nu^0}\right)$$

and $f_\nu = \text{tr}[M_\nu \otimes N_\nu \rho M_\nu^\dagger \otimes N_\nu^\dagger]$, $f_\nu^0 = \frac{e^{-\beta_A \phi'}}{Z_A} \frac{e^{-\beta_B \chi'}}{Z_B}$ are the probabilities of the final measurement on the correlated ρ and product $\rho_A \otimes \rho_B$ states.

By taking an average of both sides of $F_{\tilde{X}} = P_X e^{-x}$, this choice of variables gives us the thermodynamic relation

$$\langle e^{-\Delta\beta q - \beta_B \Delta\epsilon} \rangle_p = \langle e^{-\Delta\mathcal{I}} \rangle_f. \quad (26)$$

The subscripts p and f indicate that the averages are to be taken with respect to the probability distributions $p_{\mu\nu}$ (equation (23)) and

$$f_{\mu\nu} = \langle \phi', \chi' | \rho | \phi', \chi' \rangle \langle \phi, \chi | U^\dagger | \phi', \chi' \rangle^2. \quad (27)$$

The sharp heat into A is $q = \phi' - \phi$, the inverse temperature difference $\Delta\beta = \beta_A - \beta_B$, and the global energy change $\Delta\epsilon = \phi' + \chi' - \phi - \chi$. This formulation separates out the “correlation factor” $\eta := \langle e^{-\Delta\mathcal{I}} \rangle_f$ that quantifies the deviation from the assumption of molecular chaos. In comparing equations (26) and (18) we notice that the above analysis implies that “taking $\Delta\mathcal{I}$ to the other side” in (18) results in now having to take the average with respect to $f_{\mu\nu}$, the time-reversed probability distribution $p_{\mu\nu}$. A further difference between equations (18) and (26) is that applying Jensen’s inequality here gives

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

This matches equation (19) if and only if the random variable $\Delta\mathcal{I} = \text{const}$.

V. AN EXACTLY SOLVABLE TOY MODEL

We now compare three “lenses” through which to view heat exchange between correlated systems: the XFT from equation (17), its averaged form in equation (19), and the exponentiated average in (26). For convenience we list them below in order

$$e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_l} \leq \frac{\text{Prob}[\Gamma(q, \Delta\epsilon)]}{\text{Prob}[\Gamma(-q, -\Delta\epsilon)]} \leq e^{\Delta\beta q + \beta_B \Delta\epsilon - \Delta\mathcal{I}_u}, \quad (28a)$$

$$\Delta\beta \langle q \rangle + \beta_B \langle \Delta\epsilon \rangle - \langle \Delta\mathcal{I} \rangle \geq 0, \quad (28b)$$

$$\langle e^{-\Delta\beta q - \beta_B \Delta\epsilon} \rangle_p = \langle e^{-\Delta\mathcal{I}} \rangle_f =: \eta. \quad (28c)$$

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

$$H_i = \sum_{n=0}^{d_i-1} n |n\rangle \langle n| \quad (29)$$

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

The energy-conserving interaction Hamiltonian that commutes with the free part $H_A + H_B$ is

$$H_{\text{int}} = \sum_{j=1}^{d-1} \omega_j (|0, j\rangle \langle 1, j-1| + |1, j-1\rangle \langle 0, j|) \quad (30)$$

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 $j = 1, \dots, d-1$. Coherent evolution under this Hamiltonian is restricted to the energy-degenerate, two-dimensional subspaces spanned by $\{|0, j\rangle, |1, j-1\rangle\}$.

In deriving the thermodynamic relations, an initial projective measurement $\mathcal{M}_1$ is made on the bipartite state ρ , this kills off any coherence in the free energy eigenbasis. As such we simply take as our initial state a classically correlated density matrix

$$\rho_{AB} \equiv \rho_{AB}(t=0) = \sum_{m=0}^1 \sum_{n=0}^{d-1} \lambda_{mn} |mn\rangle \langle mn| \quad (31)$$

where $\sum_{mn} \lambda_{mn} = 1$ and $\lambda_{mn} \geq 0$ 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 $\text{tr}_{\setminus i}[\rho_{AB}] = \frac{1}{Z_i} e^{-\beta_i H_i}$, 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 $H = H_A + H_B + H_{\text{int}}$ and then a final measurement is made in the free energy basis.

In this toy example the three thermodynamic relations become

$$\min \left\{ \frac{\lambda_{0,j}}{\lambda_{1,j-1}} \right\} \leq \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]} \leq \max \left\{ \frac{\lambda_{0,j}}{\lambda_{1,j-1}} \right\} \quad (32a)$$

$$\Delta\beta \langle q \rangle \geq \sum_j (\lambda_{0j} - \lambda_{1,j-1}) \left(\Delta\beta + \ln \frac{\lambda_{1,j-1}}{\lambda_{0,j}} \right) \sin^2 \omega_j t \quad (32b)$$

$$\eta = 1 + \sum_{j=1}^{d-1} (\lambda_{0j}(e^{-\Delta\beta} - 1) + \lambda_{1,j-1}(e^{\Delta\beta} - 1)) \sin^2 \omega_j t \quad (32c)$$

We refer the reader to the Appendix for details.

To analyse these results, we concentrate on the smallest d that gives non-trivial 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 $\text{tr}_A[\tau_{AB}] = 0$ and $\text{tr}_B[\tau_{AB}] = 0$ to ensure that ρ_{AB} has thermal marginals. Furthermore, we must have that $\text{tr}[\tau_{AB}] = 0$ and $\rho_A \otimes \rho_B + \tau_{AB} \geq 0$ to ensure that ρ_{AB} is a genuine quantum state. Let τ_{AB} be a diagonal matrix, initially it has $2d$ parameters, but there are three constraints therefore we reduce to $2d - 3$ independent parameters, and in fact we always look for the smallest number of 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 and analyse the systems at the time where $\omega t = \pi/2$.

A. A two-qubit system: $d = 2$

The matrix $\tau_{AB} = \zeta \text{diag}(x, -x, -x, x)$ 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

$$-e^{-(a+b)} \leq x \leq \min\{e^{-a}, e^{-b}\}. \quad (33)$$

Let's take B to be hotter than A at the start, then $\Delta\beta > 0$ 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. Transitions occur only in between the energy degenerate states $|01\rangle$ and $|10\rangle$ and we have $\lambda_{01} = \zeta(e^{-b} - x)$ and $\lambda_{10} = \zeta(e^{-a} - x)$.

The three thermodynamics relations become

$$\begin{aligned} \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]} &= \frac{e^{-b} - x}{e^{-a} - x}, \\ \Delta\beta\langle q \rangle &\geq \zeta(e^{-b} - e^{-a}) \left(\Delta\beta + \ln \frac{e^{-a} - x}{e^{-b} - x} \right), \\ \eta &= 1 + \zeta((e^{-b} - x)(e^{-\Delta\beta} - 1) + (e^{-a} - x)(e^{\Delta\beta} - 1)). \end{aligned}$$

We analyse these relations in different extremal settings.

Equal temperature $\Delta\beta = 0$: When the qubits are at equal temperatures, the likelihoods of the forward (heat from B to A : $q > 0$) to backward (heat from A to B : $q < 0$) transitions are equal $\text{Prob}[\Gamma(q=1)] = \text{Prob}[\Gamma(q=-1)]$. Thus *detailed balance* is preserved no matter the size of initial correlations.

Maximum value of $x = e^{-a}$ with $\Delta\beta > 0$: When x takes this value, the λ_{10} eigenvalue 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 is the forward one: $|01\rangle \rightarrow |10\rangle$. This is a *deterministic* transfer of heat from B to A and is possible due to the correlations, and regardless of the initial temperatures as long as $T_B > T_A$. This behaviour is reflected in the

thermodynamic relations. Let

$$R := \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]}$$

then in this case, the ratio diverges $R \rightarrow \infty$ as expected since the backward transition (the denominator) does not occur $\text{Prob}[\Gamma(q=-1)] = 0$. In this limit the Clausius relation trivialises $\Delta\beta \geq -\infty$. In contrast, the correlation factor remains finite $\eta = 1 - \zeta e^{-b}(1 - e^{-\Delta\beta})^2 < 1$. As this is equal to $\langle e^{-\Delta\beta q} \rangle$, and by assumption $\Delta\beta > 0$, this 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 .

Minimum value $x = -e^{-(a+b)}$ with $\Delta\beta > 0$: We just saw that it is possible for correlations to switch off the backwards transition by setting $\lambda_{10} = 0$, however, we can never completely suppress the forward transition. At the minimum value of x 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 ratio $R \approx e^{\Delta\beta}$, this is a reflection on the fact that at low temperatures the reduced states are more pure and therefore cannot be highly correlated. Otherwise $R < e^{\Delta\beta}$, since $e^{-a} < e^{-b}$ by assumption, so that for maximal x , the ratio of forward to backward transfer can be suppressed compared to the uncorrelated case. However there is no amount of correlation that makes $R < 1$ which would mean that negative heat flow is always more likely to occur.

The Clausius relation becomes

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

so we see that with such correlations, we are guaranteed that a finite amount of averaged heat will transfer from hot to cold (unless $a = b$), but the probability for doing so shrinks compared to the uncorrelated case.

Finally the correlation factor

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

is 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 behaviour as captured by our three thermodynamic relations. The XFT is depicted in the top of figure 1. We now consider a qubit-qutrit system in which correlations lead to even more non-classical features.

B. A qubit-qutrit system: $d = 3$

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 some $x, y \in \mathbb{R}$. Positivity of the matrix $\rho_{AB} = \rho_A \otimes \rho_B + \tau_{AB}$ leads to

$$-1 \leq x \leq e^{-a} \quad (34)$$

$$-e^{-b} \leq y \leq e^{-(a+b)} \quad (35)$$

$$-e^{-(a+2b)} \leq x+y \leq e^{-2b} \quad (36)$$

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

This time transitions occur within two subspaces: $\{|01\rangle, |10\rangle\}$ and $\{|02\rangle, |11\rangle\}$. We have

$$\lambda_{01} = \zeta(e^{-b} + y) \quad (37)$$

$$\lambda_{10} = \zeta(e^{-a} - x) \quad (38)$$

$$\lambda_{02} = \zeta(e^{-2b} - (x+y)) \quad (39)$$

$$\lambda_{11} = \zeta(e^{-(a+b)} - y) \quad (40)$$

In this case the three thermodynamic relations are

$$\min \left\{ \frac{e^{-b} + y}{e^{-a} - x}, \frac{e^{-2b} - (x+y)}{e^{-(a+b)} - y} \right\} \leq \frac{\text{Prob}[\Gamma(q=1)]}{\text{Prob}[\Gamma(q=-1)]} \geq \max \left\{ \frac{e^{-b} + y}{e^{-a} - x}, \frac{e^{-2b} - (x+y)}{e^{-(a+b)} - y} \right\} \quad (41)$$

$$\Delta\beta\langle q \rangle \geq \Delta\beta\zeta((1+e^{-b})\delta + y) + \zeta \left((\delta + x + y) \ln \frac{e^{-a} - x}{e^{-b} + y} + (e^{-b}\delta - x) \ln \frac{e^{-(a+b)} - y}{e^{-2b} - (x+y)} \right) \quad (42)$$

$$\eta = 1 + \zeta((e^{-b}(1+e^{-b}) - x)(e^{-\Delta\beta} - 1) + (e^{-a}(1+e^{-b}) - (x+y))(e^{\Delta\beta} - 1)) \quad (43)$$

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

Equal temperature $\Delta\beta = 0$: The situation now is entirely different to the two qubit case: at equal temperatures the ratio of probabilities R in the XFT is not equal to unity hence correlations distort detailed balance (for finite a)! Depicted in the bottom of figure 1. There are two choices that make the upper bound diverge for

$$\frac{e^{-a}(1 - e^{-a}) - y}{e^{-2a} - y} \leq R < \infty, \quad x = e^{-2a} \quad (44)$$

$$\frac{e^{-a}(1 + e^{-a})}{e^{-a} - x} \leq R < \infty, \quad y = e^{-2a} \quad (45)$$

but it is not possible to simultaneously set $x = y = e^{-2a}$ so we cannot attain $R \rightarrow \infty$.

Consider the top relation. We tighten the range of R when the lower bound is maximised, that is, fixing $y = e^{-a}(e^{-a} - 1)$. Doing this we find that the lower bound is greater than unity if $T_A = T_B > (\ln 2)^{-1}$. So by setting the temperature high enough, we are more likely to observe heat flow into system A. At a lower temperature, this is not guaranteed as the lower bound can go to zero in which case $R \in [0, \infty]$ and the bounds are not tight at finite temperatures. Regarding the bottom inequality for R , we may maximise the lower bound

by setting $x = 0$ and find $R \in [1, \infty]$ meaning that that qutrit is more likely to heat the qubit. We see that correlations make heat flow asymmetric even when A and B are at the same temperature. Observing a bias in the direction of heat flow when $\Delta\beta = 0$ therefore provides a way of revealing the difference between system sizes.

For $\Delta\beta = 0$ the Clausius inequality is not illuminating, however the correlation factor

$$\langle e^{-\Delta\mathcal{I}} \rangle = 1$$

is non-trivial because $\Delta\mathcal{I} \neq 0$ in general for the qubit-qutrit case when $\Delta\beta = 0$ (the expression for $\Delta\mathcal{I}$ is given in the Appendix, and it is equal to zero when $\Delta\beta = 0$ for the qubit-qubit system). We can interpret this expression as a ‘‘correlation fluctuation theorem’’ for systems at equal temperature.

Unequal temperature $\Delta\beta \neq 0$: There are choices of x, y that can completely switch off one transition $|0, j\rangle \leftrightarrow |1, j-1\rangle$ thus reducing the qubit-qutrit system to an effective qubit-qubit one. For instance $y = e^{-(a+b)}$ and $x = e^{-b}(e^{-b} - e^{-a})$ sets $\lambda_{11} = \lambda_{02} = 0$ and only the $|01\rangle \leftrightarrow |10\rangle$ transition is allowed, and we still have the freedom to make it deterministic so that $|01\rangle \rightarrow |10\rangle$ only, if additionally $e^{-b}(e^{-b} - e^{-a}) = e^{-a}$ is satisfied. Simi-

larly choosing $y = -e^{-b}$ and $x = e^{-a}$ switches of the $|01\rangle \leftrightarrow |10\rangle$ and, if $e^{-a} - e^{-b} = e^{-2b}$ is satisfied, then $|02\rangle \leftrightarrow |11\rangle$ is deterministic. These settings recover the qubit-qubit case where we can get deterministic heat flow from hot to cold by picking the correlations correctly.

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/macrosopic regime [25], 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 [10], in which the extractable work is given by $W \leq kTI(X : Y)$, 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 [26].

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

That we have obtained an *inequality* in equation (17) 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 $-W \leq -\Delta F$. However it is easy to see that equation (17) 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 equation (26), this deviation is parameterised 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 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 $\sigma_X \rightarrow \sigma_{XE} = |\psi\rangle_{XE}\langle\psi|$, 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 $\mathcal{M}_2 = \{M_{\phi'}^{(2)} \otimes N_{\chi'}^{(2)}\}$ we then require a generalization of (9). Using the time-reversal invariance of the unitary interaction we have $\text{tr}[M_{\phi'}^{(2)} \otimes N_{\chi'}^{(2)} U \sigma_{AB} U^\dagger] = \text{tr}[(\Theta \sigma_{AB} \Theta^\dagger) U \Theta M_{\phi'}^{(2)} \otimes N_{\chi'}^{(2)} \Theta^\dagger U^\dagger]$, 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 . (A similar requirement arises in the derivation of the non-equilibrium equality in section IV.) 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.

D. Application to a semi-classical maximum work theorem

The above results, and in particular (19), find simple application in a semi-classical maximum work theorem scenario [15] 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 (19) leads to

$$dW_C \leq -dU_A + T_B dS_A - T_B \langle d\mathcal{I} \rangle \quad (46)$$

where $T_A dS_A := \langle dq \rangle$ corresponds 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. We also make an identification of S_A as the thermodynamic entropy, although again this requires more care if the system is taken to finish out of equilibrium. 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 [16–19].

E. 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 results 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 semi-classical 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 takes the form of a mutual information. A similar result has already been obtained for the case of the work Fluctuation Theorem [24] in which one allows feedback control. There the relevant mutual information is between the controller and the primary system. Furthermore the impact of this mutual information within the scenario has already been experimentally verified [26], and suggests that the generalized XFT obtained here should also be realisable in a similar manner with existing technologies.

Acknowledgments

D.J. is supported by the Royal Commission for the Exhibition of 1851. S.J. is supported by EPSRC grant EP/K022512/1. 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.

-
- [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] C. Jarzynski and D. K. Wójcik, Phys. Rev. Lett. **92**, 230602 (2004).
 - [5] H. Price, *Time's Arrow and Archimedes' Point*, Oxford University Press (1996).
 - [6] H. D. Zeh, *The Physical Basis of The Direction of Time*, Springer 4th Ed. (2001).
 - [7] L. Boltzmann, *Lectures on Gas Theory*, Dover, (2011).
 - [8] M. H. Partovi, Phys. Rev. E **77**, 021110, (2008).
 - [9] D. Jennings and T. Rudolph, Phys. Rev. E, **81**, 061130, (2010).
 - [10] W. H. Zurek, in *Proceedings Frontiers of Nonequilibrium Quantum Statistical Mechanics*, edited by G. T. Moore and M. O. Scully (Plenum, 1986), pp. 145150.
 - [11] T. Albash, D. A. Lidar, M. Marvian, and P. Zanardi, Phys. Rev. E **88**, 032146 (2013).
 - [12] S. Popescu and A. Short and A. Winter, Nature Physics **2** 754 (2006).
 - [13] F. G. S. L. Brandão and M. Plenio, Nature Physics **4**, 873, (2008).
 - [14] F. G. S. L. Brandão and M. Horodecki and J. Oppenheim, and J. M. Renes and R. W. Spekkens, arXiv:1111.3882 (2011).
 - [15] H. Callen, *Thermodynamics and an Introduction to Thermostatistics*, Wiley 2nd ed., (1985).
 - [16] L. Szilard, *Z. f. Physik* **53** 840-856 (1929)
 - [17] C. H. Bennett, Int. J. Theor. Phys. **21** 905-940 (1982)
 - [18] T. Sagawa and M. Ueda, Phys. Rev. Lett. **100**, 080403 (2008).
 - [19] S. Jevtic and D. Jennings and T. Rudolph, Phys. Rev. Lett. **108**, 110403 (2012)
 - [20] S. Deffner and E. Lutz, Phys. Rev. Lett. **107**, 140404 (2011)
 - [21] J. M. R. Parrondo, and C. Van den Broeck and R. Kawai, N. J. Phys. **11** 073008 (2009)
 - [22] M. Campisi, and P. Talkner and P. Hänggi, Phys. Rev.

- Lett. **105** 140601 (2010)
- [23] J. M. Horowitz, Phys. Rev. E **85** 031110 (2012)
- [24] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2012)
- [25] B. Groisman, and S. Popescu and A. Winter, Phys. Rev. A , **72**, 032317 (2005)
- [26] S. Toyabe, and T. Sagawa and M. Ueda, and E. Muneyuki and M. Sano, Nature Physics **6**, 988 (2010)
- [27] Nicolas Brunner, Noah Linden, Sandu Popescu and Paul Skrzypczyk, Phys. Rev. E , **85**, 051111 (2012)

Appendix A: Derivation of equation (12)

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

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

To show (12) we consider the sharp energy measurement $\mathcal{M}_1 = \{M_\phi \otimes N_\chi\}$ where $\{M_\phi = |\phi\rangle\langle\phi|\}$ and $\{N_\chi = |\chi\rangle\langle\chi|\}$ are the rank-1 projectors in the 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 $p(|\phi, \chi\rangle) = \langle\phi, \chi|\rho_{AB}|\phi, \chi\rangle = \text{tr}[M_\phi \otimes N_\chi \rho_{AB}]$. However, from the definition of $\mathcal{I}$ we have that

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

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

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

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

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

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 the abstract fluctuation theorem

Here we fill in the details leading up to equation (26). Using the discretised expression for $P(x)$ we have

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

In the first to second we have used a property of delta 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 $\tilde{f}_{\mu\nu} := p_{\mu\nu}e^{-X_{\mu\nu}}$ and $\tilde{X}_{\mu\nu} := -X_{\mu\nu}$. The third line is a probability density function for the new random variable $\tilde{X}_{\mu\nu}$ if $\tilde{f}_{\mu\nu}$ is a probability distribution.

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

$$X_{\mu\nu} = \ln p_\mu - \ln f_\nu + \Delta\mathcal{I}_{\mu\nu}. \quad (\text{B1})$$

where the correlation

$$\Delta\mathcal{I}_{\mu\nu} = \mathcal{I}_\nu - \mathcal{I}_\mu = \ln\left(\frac{p_\mu}{p_\mu^0}\right) - \ln\left(\frac{f_\nu}{f_\nu^0}\right)$$

and $f_\nu = \text{tr}[M_\nu \otimes N_\nu \rho M_\nu^\dagger \otimes N_\nu^\dagger]$, $f_\nu^0 = \frac{e^{-\beta_A \phi'}}{Z_A} \frac{e^{-\beta_B \chi'}}{Z_B}$ are the probabilities of the final measurement on the correlated ρ and product $\rho_A \otimes \rho_B$ states. With this $X_{\mu\nu}$ we have simply $X_{\mu\nu} = \ln p_\mu^0 - \ln f_\nu^0$ and using the expressions for these uncorrelated probability distributions we obtain

$$X_{\mu\nu} = \beta_A q_{\mu\nu}^A + \beta_B q_{\mu\nu}^B, \quad (\text{B2})$$

in terms of the sharp heat into A and B

$$q_{\mu\nu}^A = \phi' - \phi, \quad q_{\mu\nu}^B = \chi' - \chi. \quad (\text{B3})$$

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 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 with this we average the left hand side of the non-equilibrium equality $\langle e^{-x} \rangle = \langle e^{-\Delta\beta q - \beta_B \Delta\epsilon} \rangle_p$, where the subscript p indicates that the average is with respect to the $p_{\mu\nu}$ probability distribution given in equation (23).

To calculate the correlation factor $\eta := \langle F_{\tilde{X}}(-x) \rangle$ let us formally write $\tilde{f}_{\mu\nu} = f_\nu \tilde{f}_{\mu|\nu}$. We have by definition $\tilde{f}_{\mu\nu} = p_{\mu\nu}e^{-X_{\mu\nu}}$. Since $p_{\mu\nu} = p_\mu p_{\nu|\mu}$ and $e^{-X_{\mu\nu}} = \frac{f_\nu}{p_\mu} e^{-\Delta\mathcal{I}_{\mu\nu}}$ we can deduce

$$\tilde{f}_{\mu|\nu} := p_{\nu|\mu} e^{-\Delta\mathcal{I}_{\mu\nu}}. \quad (\text{B4})$$

Therefore

$$F_{\tilde{X}}(-x) = \sum_{\mu\nu} \delta(x + \tilde{X}_{\mu\nu}) f_\nu p_{\nu|\mu} e^{-\Delta\mathcal{I}_{\mu\nu}}. \quad (\text{B5})$$

Is $f_\nu p_{\nu|\mu}$ a valid probability distribution? The f_ν part is fine since it is the probability of projecting the state ρ onto $M_\nu \otimes N_\nu$. The conditional $p_{\nu|\mu}$ is given in equation (22), note that it is equivalent to its time-reversed expression $|\langle \phi, \chi | U^\dagger | \phi', \chi' \rangle|^2 =: f_{\mu|\nu}$, c.f. equation (9). This is the probability that the state starts in $|\phi'\rangle \otimes |\chi'\rangle$, evolves under $U^\dagger$ and is projected onto $|\phi\rangle \otimes |\chi\rangle$, and it satisfies $\sum_\mu f_{\mu|\nu} = 1$, therefore $\sum_{\mu\nu} f_\nu f_{\mu|\nu} = 1$ and we do indeed have a valid probability $f_{\mu\nu} := f_\nu f_{\mu|\nu}$ (note carefully the difference between tildes and no tildes).

Appendix C: Details for the toy example

Consider first the generalised exchange fluctuation theorem in equation (17)

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

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

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

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

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

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

$$\mathcal{I}(\rho_{AB}; m, n) = \ln \left[\frac{\langle mn | \rho_{AB} | mn \rangle}{\langle m | \rho_A | m \rangle \langle n | \rho_B | n \rangle} \right] \quad (\text{C3})$$

$$= \beta_A m + \beta_B n + \ln \frac{\lambda_{mn}}{Z_A Z_B} \quad (\text{C4})$$

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

$$\Delta\mathcal{I}(\gamma[q = 1|j]) = \mathcal{I}(\rho_{AB}; 1, j-1) - \mathcal{I}(\rho_{AB}; 0, j) \quad (\text{C5})$$

$$= \Delta\beta + \ln \frac{\lambda_{1,j-1}}{\lambda_{0,j}} \quad (\text{C6})$$

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 $\Delta\mathcal{I}_u = \Delta\beta + \max_j \left\{ \ln \frac{\lambda_{1,j-1}}{\lambda_{0,j}} \right\}_{j=1}^{d-1}$ and $\Delta\mathcal{I}_l = \Delta\beta + \min_j \left\{ \ln \frac{\lambda_{1,j-1}}{\lambda_{0,j}} \right\}_{j=1}^{d-1}$.

Substituting these expressions into equation (17) we obtain

$$\min \left\{ \frac{\lambda_{0,j}}{\lambda_{1,j-1}} \right\} \leq \frac{\text{Prob}[\Gamma(q = 1)]}{\text{Prob}[\Gamma(q = -1)]} \leq \max \left\{ \frac{\lambda_{0,j}}{\lambda_{1,j-1}} \right\} \quad (\text{C7})$$

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

function is

$$\langle \Delta \mathcal{I} \rangle = \sum_j (\text{Prob}[q = \pm 1|j] \Delta \mathcal{I}(\gamma[q = \pm 1|j])) \quad (\text{C8})$$

$$= \sum_j (\lambda_{0j} - \lambda_{1,j-1}) (\Delta \beta + \ln \frac{\lambda_{1,j-1}}{\lambda_{0,j}}) \sin^2 \omega_j t \quad (\text{C9})$$

Let us now turn our attention to the correlation factor

$$\eta = \langle e^{-\Delta \mathcal{I}} \rangle_f$$

from equation (26). The initial and final measurement

labels are $\mu = (0, j)$ and $\nu = (1, j-1)$, we have $\Delta \mathcal{I}_{\mu\nu} = \Delta \beta + \ln \frac{\lambda_{1,j-1}}{\lambda_{0j}}$, and the probability $f_{\mu\nu} = f_\nu f_{\mu|\nu} = \sum_j \lambda_{1,j-1} \sin^2 \omega_j t$ since $f_{\mu|\nu} = |\langle 0, j | U^\dagger | 1, j-1 \rangle|^2$ and $f_\nu = \langle 1, j-1 | \rho_{AB} | 1, j-1 \rangle$. Including also the remaining transitions $\mu = (1, j-1)$ and $\nu = (0, j)$, we obtain

$$\begin{aligned} \langle e^{-\Delta \mathcal{I}} \rangle_f &= \sum_{j=1}^{d-1} (\lambda_{1,j-1} e^{-\Delta \beta - \ln \frac{\lambda_{1,j-1}}{\lambda_{0j}}} \\ &\quad + \lambda_{0j} e^{\Delta \beta - \ln \frac{\lambda_{0j}}{\lambda_{1,j-1}}}) \sin^2 \omega_j t \end{aligned} \quad (\text{C10})$$

and this may be simplified to give the η in the main text.