Entropy production and time asymmetry in the presence of strong interactions

H. J. D. Miller* and J. Anders†
Department of Physics and Astronomy, University of Exeter, Stocker Road, Exeter EX4 4QL, England, United Kingdom
(Received 22 March 2017; published 19 June 2017)

It is known that the equilibrium properties of open classical systems that are strongly coupled to a heat bath are described by a set of thermodynamic potentials related to the system's Hamiltonian of mean force. By adapting this framework to a more general class of nonequilibrium states, we show that the equilibrium properties of the bath can be well defined, even when the system is arbitrarily far from equilibrium and correlated with the bath. These states, which retain a notion of temperature, take the form of conditional equilibrium distributions. For out-of-equilibrium processes we show that the average entropy production quantifies the extent to which the system and bath state is driven away from the conditional equilibrium distribution. In addition, we show that the stochastic entropy production satisfies a generalized Crooks relation and can be used to quantify time asymmetry of correlated nonequilibrium processes. These results naturally extend the familiar properties of entropy production in weakly coupled systems to the strong coupling regime. Experimental measurements of the entropy production at strong coupling could be pursued using optomechanics or trapped-ion systems, which allow strong coupling to be engineered.

DOI: 10.1103/PhysRevE.95.062123

I. INTRODUCTION

The central goal of stochastic thermodynamics is to provide a microscopic description of entropy production at the level of the individual trajectories traced out by the system as it is driven away from equilibrium [1–4]. Current technology now provides us with increased control over mechanically manipulated biomolecules and nanosystems, with examples including single molecule RNA unfolding experiments [5], the manipulation of light-levitated nanospheres [6], and control over trapped-ion systems [7]. As the system size is scaled down, microscopic fluctuations in entropy become appreciable and must be understood in order to optimize the thermodynamic performance of machines and devices operating at the nanoscale [8]. On a more fundamental level entropy production provides us with a quantitative description of change and irreversibility in nature, and its average increase places restrictions on allowed state transformations in accordance with the second law of thermodynamics [9,10]. More refined statements about the nature of entropy production are given by the fluctuation theorems [2,11–15], and provide universal insight into the breaking of time-reversal symmetry in a wide variety of physical systems [5,16–19].

Standard analysis of entropy production in open systems, both quantum and classical, centers on an assumption that the system *weakly* interacts with a thermal bath [4,20–22]. The benefit of this assumption is that it provides an unambiguous notion of stochastic heat, since neglecting energetic contributions from the interaction provides a clear division between the energy of the system and the bath. While the weak coupling assumption can be physically justified in macroscopic systems, the thermodynamic behavior of small-scale systems may be strongly influenced by a non-negligible interaction with their environment [23]. Thus it is of paramount importance to explore extended notions of entropy production within the strong coupling regime, which will be the subject of this paper.

The extension of thermodynamics to the strong coupling regime has been the subject of recent debate in the context of both classical [8,23–26] and quantum systems [27–33]. The central question revolves around the identification of thermodynamic potentials for the system at both the stochastic and ensemble level. An elegant solution to this problem, originally dating back to Kirkwood in 1935 [34], is to replace the isolated Hamiltonian of the system with an effective Hamiltonian that takes into account the non-negligible interaction and temperature of the environment. This allows one to define an *effective* internal energy, free energy, and entropy for the system at equilibrium [24].

Recent efforts have extended the applicability of this formalism to stochastic, nonequilibrium thermodynamics [23,25,26]. In particular, Seifert has proposed a definition of stochastic entropy production derived from a set of fluctuating thermodynamic potentials associated with the system's Hamiltonian of mean force [26]. In this paper we lend support to this approach by deriving an exact expression for the average entropy production in general nonequilibrium processes valid at arbitrary interaction strengths. Importantly it is shown that our expression converges to previously derived formulas in the limit of weak coupling [29,35]. In order to consider the thermodynamics of systems operating away from equilibrium, we introduce a class $\mathcal{D}_\beta$ of system and bath configurations in which the equilibrium properties of the bath are retained even if correlated with an arbitrary state of the system that is out of equilibrium. The entropy production is shown to increase as a result of the system and bath being driven away from configurations in $\mathcal{D}_\beta$. Furthermore, it is shown that the full statistics of stochastic entropy production obey a generalized Crooks-like fluctuation relation [11], which provides a relationship between the time asymmetry of nonequilibrium dynamics and the average entropy production.

We begin by considering an open classical system coupled to a heat bath with a time-dependent Hamiltonian

$$H(z_t;\lambda_t) = H_s(x_t;\lambda_t) + H_b(y_t) + V_{\rm int}(z_t), \tag{1}$$

*hm419@exeter.ac.uk
†janet@qipc.org

where λ_t is a time-dependent control parameter attributed to the system Hamiltonian alone, $V_{\rm int}(z_t)$ governs the interaction between system and bath, and $z_t = (x_t, y_t)$ describes a point in the collective phase space at time t, with x and y labeling the system and bath degrees of freedom, respectively. Let us first consider the equilibrium thermodynamics of the total system and assume a canonical distribution at inverse temperature β:

$$\rho^{\rm eq}(z_t;t) = \frac{e^{-\beta H(z_t;\lambda_t)}}{Z(\lambda_t)}, \tag{2}$$

where $Z(\lambda_t) = \int dz_t\ e^{-\beta H(z_t;\lambda_t)}$ is the partition function of the total system and bath. In standard thermodynamics one assumes that the interaction strength is sufficiently weak, $\beta V_{\rm int}(z_t) \ll 1$, such that the total canonical state factorizes into two uncorrelated canonical distributions for the system and bath, respectively. In this case additive thermodynamic potentials can be assigned to both system and bath via their local equilibrium distributions.

However, when $V_{\rm int}(z_t)$ is non-negligible it is not immediately clear how to assign a set of thermodynamic potentials to the system. A way to solve this problem is to introduce the Hamiltonian of mean force [23–27,32,34,36,37],

$$\tilde{H}_s(x_t;\lambda_t) := H_s(x_t;\lambda_t) - \frac{1}{\beta}\ln\langle e^{-\beta V_{\rm int}(z_t)}\rangle_b^{\rm eq}, \tag{3}$$

which acts as an effective Hamiltonian for the system that takes into account the non-negligible interaction term. Here $\langle f(z_t)\rangle_b^{\rm eq} = \int dy_t\ f(z_t)\, e^{-\beta(H_b(y_t)-F_b^{\rm eq})}$ denotes an average of arbitrary function $f(z_t)$ with respect to an isolated bath, and $F_b^{\rm eq}$ is the corresponding equilibrium free energy of the isolated bath. By averaging over the bath degrees of freedom in the canonical distribution (2), the system distribution can be expressed in an effective equilibrium state with respect to $\tilde{H}_s(x_t;\lambda_t)$:

$$\tilde{\rho}_s^{\rm eq}(x_t;t) = \frac{e^{-\beta\tilde{H}_s(x_t;\lambda_t)}}{\tilde{Z}_s(\lambda_t)}, \quad \tilde{Z}_s(\lambda_t) = \int dx_t\ e^{-\beta\tilde{H}_s(x_t;\lambda_t)}. \tag{4}$$

As was shown in [24], the partition function $\tilde{Z}_s(\lambda_t)$ can be used to obtain a set of thermodynamic potentials for the system through the standard formulas for free energy, internal energy, and entropy:

$$\begin{aligned}
\tilde{F}_s^{\rm eq}(\lambda_t) &= -\frac{1}{\beta}\ln\tilde{Z}_s(\lambda_t),\\
\tilde{U}_s^{\rm eq}(\lambda_t) &= -\partial_\beta\ln\tilde{Z}_s(\lambda_t),\\
\tilde{S}_s^{\rm eq}(\lambda_t) &= \beta\big[\tilde{U}_s^{\rm eq}(\lambda_t) - \tilde{F}_s^{\rm eq}(\lambda_t)\big].
\end{aligned} \tag{5}$$

It is well known that these thermodynamic potentials are additive with respect to the bare environment [24,27,32]. For example, the total thermodynamic entropy of the system and bath can be split into $S_{\rm tot}^{\rm eq}(\lambda_t) = \tilde{S}_s^{\rm eq}(\lambda_t) + S_b^{\rm eq}$, where $S_b^{\rm eq}$ is the entropy of the isolated canonical bath. The same additivity holds for the internal energy and free energy, implying that the presence of the interaction leaves the equilibrium properties of the bath unchanged. Instead, the influence of the interaction is attributed to the equilibrium properties of the system alone [38].

II. NONEQUILIBRIUM POTENTIALS

While the thermodynamic potentials in Eq. (5) are well defined at equilibrium, recent efforts have attempted to extend the definitions of Eq. (5) to the case where the system is no longer in an effective equilibrium state [23,25,26]. This can be achieved by first noting that the equilibrium internal energy can be expressed as $\tilde{U}_s^{\rm eq}(\lambda_t) = \langle\partial_\beta[\beta\tilde{H}_s(x_t;\lambda_t)]\rangle_s^{\rm eq}$ where $\langle..\rangle_s^{\rm eq}$ denotes an average with respect to the effective equilibrium state (4). Similarly one finds $\tilde{S}_s^{\rm eq}(\lambda_t) = -\langle\ln\tilde{\rho}_s^{\rm eq}(x_t;t)\rangle_s^{\rm eq} + \beta^2\langle\partial_\beta\tilde{H}_s(x_t;\lambda_t)\rangle_s^{\rm eq}$. These quantities can be interpreted as equilibrium averages over a set of *fluctuating* thermodynamic potentials appearing inside the brackets $\langle..\rangle_s^{\rm eq}$. We propose that the fluctuating potentials for internal energy, entropy, and free energy for states arbitrarily far from equilibrium are given, respectively, by [23,25,26]

$$\begin{aligned}
\tilde{u}_s(x_t;\lambda_t) &:= \partial_\beta[\beta\tilde{H}_s(x_t;\lambda_t)],\\
\tilde{s}_s(x_t;\lambda_t) &:= -\ln\rho_s(x_t;t) + \beta^2\partial_\beta\tilde{H}_s(x_t;\lambda_t),\\
\tilde{f}_s(x_t;\lambda_t) &:= \tilde{u}_s(x_t;\lambda_t) - \beta^{-1}\tilde{s}_s(x_t;\lambda_t).
\end{aligned} \tag{6}$$

These functions account for the temperature dependence of the mean force Hamiltonian, give the averages (5), and reduce to the standard thermodynamic potentials used in stochastic thermodynamics in the limit of weak coupling [4]. We will show that these generalized fluctuating potentials can be connected into a consistent thermodynamic framework. The average nonequilibrium internal energy will be denoted by $\tilde{U}_s(\lambda_t;t) = \langle\tilde{u}_s(x_t;\lambda_t)\rangle_s$, with $\langle..\rangle_s = \int dx_t\ \ \rho_s(x_t;t)(..)$ now an average with respect to a general nonequilibrium state of the system. Similarly the average entropy will be denoted by $\tilde{S}_s(\lambda_t;t) = \langle\tilde{s}_s(x_t;\lambda_t)\rangle_s$ and average free energy by $\tilde{F}_s(\lambda_t;t) = \langle\tilde{f}_s(x_t;\lambda_t)\rangle_s$. From Eq. (6) one sees that the nonequilibrium entropy at strong coupling involves a contribution from the Gibbs-Shannon entropy alongside a second term $\beta^2\langle\partial_\beta\tilde{H}_s(x_t;\lambda_t)\rangle_s$ that has previously been identified as an intrinsic entropy in the context of small-scale molecular motors [8].

It is not obvious that these potentials should generally be additive for a given system and bath distribution, unlike the equilibrium counterparts (5). However, let us consider a particular class $\sigma(z_t;t) \in \mathcal{D}_\beta$ of distributions defined by

$$\sigma(z_t;t) = \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t), \tag{7}$$

where we place no restriction on the system configuration and

$$\rho_b^{\rm eq}(y_t|x_t) = \frac{\rho^{\rm eq}(z_t;\lambda_t)}{\int dy_t\,\rho^{\rm eq}(z_t;\lambda_t)} \tag{8}$$

is the *equilibrium* conditional probability for bath microstate y_t given a particular microstate of the system x_t, obtained through application of Bayes's theorem. Because the system Hamiltonian cancels in the fraction in Eq. (7) the dependence on the control parameter λ_t cancels in the expression for $\rho_b^{\rm eq}(y_t|x_t)$. The class of states $\mathcal{D}_\beta$ has previously been introduced in [25] and referred to as the *stationary preparation class*, which describes a conditional equilibrium state on the bath. In this case for any microstate selected from the system the resulting conditional statistics of the bath are equivalent to that of the total canonical state (2). For this class of states

one still has a well-defined notion of temperature attributed to a thermal environment. This is manifested by a *generalized* additive relationship between the thermodynamic potentials, which we prove in Appendix A. Taking the state $\sigma(z_t;t) \in \mathcal{D}_\beta$, let us denote $U_{\rm tot}(\lambda_t;t) = \langle H(z_t;\lambda_t)\rangle$ as the internal energy of $\sigma(z_t;t)$, $S_{\rm tot}(\lambda_t;t) = -\langle \ln\, \sigma(z_t;t)\rangle$ as the Gibbs-Shannon entropy, and $F_{\rm tot}(\lambda_t;t) = U_{\rm tot}(\lambda_t;t) - \beta^{-1} S_{\rm tot}(\lambda_t;t)$ as the free energy. Then the following additive property holds:

$$\chi_{\rm tot}(\lambda_t;t) = \tilde{\chi}_s(\lambda_t;t) + \chi_b^{\rm eq}, \tag{9}$$

where $\chi \in \{F,S,U\}$. Here the thermodynamic potentials $\chi_b^{\rm eq} \in \{F_b^{\rm eq}, S_b^{\rm eq}, U_b^{\rm eq}\}$ are equivalent to those of an isolated canonical bath, and can be obtained by substituting the bath partition function Z_b into the equations given in Eq. (5), where

$$Z_b = \int dy_t\; e^{-\beta H_b(y_t)}. \tag{10}$$

The relation (9) implies that the equilibrium properties of the bath remain unchanged relative to the arbitrary state of the system, even in the presence of correlations due to strong interaction. In other words, while the bath marginal of $\sigma(z_t;t)$ is not a canonical distribution, the effect of the interaction on the bath potentials is negligible. This is physically intuitive considering that the bath is macroscopic relative to the microscopic size of the system.

Ultimately the additivity of thermodynamics potentials (5) for the class $\mathcal{D}_\beta$ will allow us to maintain a notion of temperature for states driven away from equilibrium, and will allow us to derive the second law of thermodynamics in this framework.

III. ENTROPY PRODUCTION

We will now consider a general nonequilibrium (NEQ) process operating at an arbitrarily large coupling strength and derive an exact expression for the entropy production. The NEQ process is realized over a time interval $[t_0,t]$ by varying the Hamiltonian through a parameter change λ_t with initial and final settings denoted by λ_0 and λ_t, respectively. We make two assumptions about this process.

(i) At initial time t_0 the system and bath is in a conditional equilibrium state $\sigma(z_0;t_0) \in \mathcal{D}_\beta$, with $\rho_s(x_0;t_0)$ specifying an initial arbitrary state for the system.

(ii) The total system and bath undergoes closed evolution during the time interval $[t_0,t]$ governed by Liouville's equation

$$\partial_t \rho(z_t;t) = \mathcal{L}[\rho(z_t;t)], \tag{11}$$

where $\mathcal{L}[(..)]$ is the corresponding Liouvillian resulting from the change in the Hamiltonian (1) over time. The resulting final state is specified by $\rho(z_t;t)$ with final system configuration $\rho_s(x_t;t) = \int dy_t\; \rho(z_t;t)$.

Assumption (i) is necessary in order to have a well-defined notion of both temperature and the Hamiltonian of mean force (3) prior to the NEQ process. Assumption (ii) ensures that we account for all exchanges of heat and work between the system and the bath. No restrictions are imposed on the final configuration of the system, and we denote the transformation by $\rho_s(x_0;t_0) \to \rho_s(x_t;t)$. Following the approaches taken in [23,25,26] we can use the fluctuating potentials in Eq. (6) to define the fluctuating heat dissipated from the system into the bath up to time t as

$$\tilde{Q}(z_t;t) := \tilde{u}_s(x_0;\lambda_0) - \tilde{u}_s(x_t;\lambda_t) + \int_{t_0}^{t} d\tau\; \partial_\tau \tilde{u}_s(x_\tau;\lambda_\tau), \tag{12}$$

which represents the sum of work done during the process and the decrease in internal energy of the system, in accordance with the first law of thermodynamics. Note that $\tilde{Q}(z_t;t) = \tilde{Q}(z_t[z_0];t)$ is implicitly written as a function of the initial phase space point z_0 because the evolution of point x_t depends on the deterministic evolution of the collective phase space for the system and bath, denoted by the transformation $z_0 \to z_t[z_0]$. However, the right-hand side of Eq. (12) indicates that the heat can be determined by monitoring the system degrees of freedom alone along a specific trajectory. If we take into account the full evolution of the system and bath, it is straightforward to show that the average dissipated heat is given by

$$\langle \tilde{Q}(t)\rangle = U_{\rm tot}(\lambda_t;t) - \tilde{U}_s(\lambda_t;t) - U_b^{\rm eq}, \tag{13}$$

which follows from Eq. (9) combined with initial condition (i), along with the fact that the integral in Eq. (12) is equivalent to the difference in total energy, $H(z_t;\lambda_t) - H(z_0;\lambda_0)$. This heat is nonzero because, unlike the initial state, the final state will not generally belong to the class $\mathcal{D}_\beta$ and so the additive relation (9) will not hold for the final state in general. As noted by Seifert, one can introduce a definition of fluctuating entropy production as the sum of dissipated heat and change in the fluctuating entropy of the system [26]:

$$\Sigma(z_t;t) := \tilde{s}_s(x_t;\lambda_t) - \tilde{s}_s(x_0;\lambda_0) + \beta\,\tilde{Q}(z_t;t). \tag{14}$$

For the definition (14) to be a physically relevant candidate for entropy production then it should not be negative on average, in accordance with the second law of thermodynamics. This brings us to the first main result of the paper.

IV. MAIN RESULT

Assuming the total system and bath undergoes the NEQ process specified by assumptions (i) and (ii), then the average entropy production up to time t is given by

$$\langle \Sigma(t)\rangle = D[\rho(z_t;t)||\sigma(z_t;t)], \tag{15}$$

where

$$D[\rho(z_t;t)||\sigma(z_t;t)] = \int dz_t\, \rho(z_t;t) \ln\left[\frac{\rho(z_t;t)}{\sigma(z_t;t)}\right]$$

is the Kullback-Leibler divergence between the final system and bath configuration and the corresponding conditional equilibrium state $\sigma(z_t;t) = \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t) \in \mathcal{D}_\beta$. This is the central result of the paper and the proof of Eq. (15) is provided in Appendix B. We note that this result has also been obtained independently in [39]. By Eq. (15) and the positivity of the Kullback-Leibler divergence, one has $\langle \Sigma(t)\rangle \geqslant 0$ as desired. From the definition of entropy production in Eq. (14) one obtains a form of the Clausius inequality valid for arbitrary

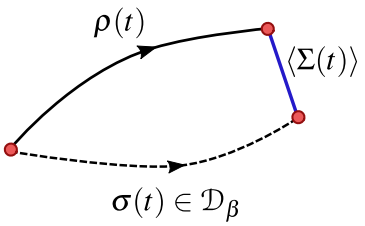

FIG. 1. Schematic representation of the equality (15). The solid line represents the actual process given by the evolving distribution $\rho(t) = \rho(z_t;t)$ whilst the dashed line represents a hypothetical quasistatic process in which the system and bath distribution stays in the conditional equilibrium state $\sigma(t) = \sigma(z_t;t) \in \mathcal{D}_\beta$. The nonnegative entropy production then quantifies the extent to which the system and bath are driven away from $\sigma(t)$, represented here as the distance of the blue line.

coupling strengths which becomes

$$\beta\langle\tilde{Q}(t)\rangle \geqslant \tilde{S}_s(\lambda_0;t_0) - \tilde{S}_s(\lambda_t;t). \tag{16}$$

Perhaps surprisingly, the Clausius inequality derived here within the strong coupling regime suggests that the change in Gibbs-Shannon entropy is generally insufficient to bound the minimum heat dissipated into the bath during a nonequilibrium process.

According to Stein's lemma [40], the divergence appearing in Eq. (15) can be interpreted as a measure of distinguishability between the final distribution and the corresponding conditional equilibrium state $\sigma(z_t;t) \in \mathcal{D}_\beta$. Thus the further the final state is driven away from the uniquely defined $\sigma(z_t;t) \in \mathcal{D}_\beta$, the greater the amount of entropy production after the process. If the dynamics governed by Eq. (11) are such that the total system and bath remains in the corresponding conditional equilibrium state in $\mathcal{D}_\beta$, the bound in Eq. (16) can be saturated at any given time t. However, in this situation the dissipated heat and entropy change are simultaneously zero; $\beta\langle\tilde{Q}(t)\rangle = \Delta\tilde{S}_s = 0$. The expression (15) can be interpreted as a generalization of a phenomenon known as *lag* encountered in closed and weakly coupled thermodynamic systems [41]. The entropy production quantifies the extent to which the configuration of the system and bath *lags* behind a hypothetical quasistatic process in which the configuration remains in the evolving conditional equilibrium state, $\sigma(z_t;t) \in \mathcal{D}_\beta$. Figure 1 illustrates this effect.

Result (15) is consistent with previously derived expressions for average entropy production when the weak coupling limit is taken. If one assumes $\beta V_{\rm int}(z_t) \ll 1$ then the Hamiltonian of mean force (3) reduces to the system Hamiltonian $H_s(x_t;\lambda_t)$ independent of temperature. As expected the heat becomes $\langle\tilde{Q}(t)\rangle \approx \langle H_b(t)\rangle - \langle H_b(t_0)\rangle$, where $\langle H_b(t)\rangle$ is the average energy of the isolated bath Hamiltonian evaluated with respect to the configuration of the bath at time t. Secondly, this also means the entropy change reduces to the change in Gibbs-Shannon entropy $\tilde{S}_s(\lambda_t;t) \approx S_s(t) = -\int dx_t \rho_s(x_t;t)\ln\ \rho_s(x_t;t)$. Finally, it can also be seen that the conditional equilibrium state $\sigma(z_t;t) \in \mathcal{D}_\beta$ reduces to a system state uncorrelated with the isolated canonical bath; $\sigma(z_t;t) \approx \rho_s(x_t;t)\rho_b^{\rm eq}(y_t)$. By comparison with Eq. (15), we obtain the same equality derived in [29,35], which is

$$\begin{aligned}\langle\Sigma(t)\rangle &\approx S_s(t) - S_s(t_0) + \beta\langle H_b(t)\rangle - \beta\langle H_b(t_0)\rangle \\ &= D\big[\rho(z_t;t)||\rho_s(x_t;t)\rho_b^{\rm eq}(y_t)\big].\end{aligned} \tag{17}$$

where $\rho_b^{\rm eq}(y_t) = e^{-\beta[H_b(y_t)-F_b^{\rm eq}]}$. It should be noted that Eq. (17) was originally derived for quantum systems in [29,35], though in the weak coupling regime the result is entirely statistical mechanical in nature and continues to hold in classical systems.

V. FLUCTUATION THEOREM

We have demonstrated that the average entropy production $\langle\Sigma(t)\rangle$ quantifies the extent to which the total system and bath is driven away from states in $\mathcal{D}_\beta$. This suggests that the fluctuations in $\Sigma(z_t;t)$ can be used to quantify time asymmetry in the dynamics of strongly coupled systems. In both weakly coupled and closed systems, fluctuation relations can be used to indicate a breaking of time-reversal symmetry by comparing the statistics of positive entropy production for a forward trajectory versus negative entropy production along the corresponding time-reversed trajectory [11,13,14,16,17,42]. We will now show that the entropy production satisfies a Crooks-like fluctuation relation. Let us again suppose that we drive a system and bath configuration $\sigma(z_0;t_0) \in \mathcal{D}_\beta$ away from $\mathcal{D}_\beta$ by varying the control parameter $\lambda_0 \to \lambda_t$, and denote the initial and final configurations of the system by $\rho_s(x_0;t_0)$ and $\rho_s(x_t;t)$, respectively. The stochastic entropy production $\Sigma(z_t;t)$ along a particular phase space trajectory fluctuates according to the sampling of the initial phase space point, and the resulting probability of occurrence can be written as follows:

$$\overrightarrow{P}(+\Sigma) = \int dz_0\, \sigma(z_0;t_0)\, \delta[\Sigma - \Sigma(z_t;t)], \tag{18}$$

where the superscript indicates that the process moves forwards in time. To compare this with the time-reversed entropy production we need to make additional assumptions. First, we require the total Hamiltonian to be time-reversal symmetric, $H(z_t;\lambda_t) = H(z_t^*;\lambda_t)$, where z_t^* indicates a conjugated phase space point in which momentum is reversed. Secondly, the initial and final configurations of the system are assumed to be time reversal symmetric; $\rho_s(x_0;t_0) = \rho_s(x_0^*;t_0)$ and $\rho_s(x_t;t) = \rho_s(x_t^*;t)$. By comparison with Eqs. (12) and (14) it is straightforward to see that these conditions imply $\Sigma(z_t;t) = -\Sigma(z_t^*;t)$. For the time-reversed process, the initial configuration is given by $\sigma(z_t^*;t) = \rho_s(x_t^*;t)\rho_b^{\rm eq}(y_t^*|x_t^*) \in \mathcal{D}_\beta$ and the control parameter is varied from $\lambda_t \to \lambda_0$. As with Eq. (18), entropy production along the reverse process has a corresponding probability of occurrence denoted by $\overleftarrow{P}(-\Sigma)$. As is proven in Appendix C, these probabilities are related by a fluctuation relation, which becomes our second main result:

$$\frac{\overrightarrow{P}(+\Sigma)}{\overleftarrow{P}(-\Sigma)} = e^{+\Sigma}, \tag{19}$$

implying that a positive entropy production along the forward trajectory is exponentially favored against its time reverse. Taking the logarithm of both sides and performing an average

over $\overrightarrow{P}(+\Sigma)$ yields an alternative expression for the average entropy production:

$$\langle \Sigma(t) \rangle = D[\overrightarrow{P}(+\Sigma)||\overleftarrow{P}(-\Sigma)]. \tag{20}$$

Following Stein's lemma again, we see that the average entropy production also quantifies the distinguishability between statistics of the forward and reverse nonequilibrium processes, respectively. By comparison with Eq. (15), if the dynamics are such that the system and bath remain in their corresponding configuration in $\mathcal{D}_\beta$ then the left-hand side of Eq. (20) reduces to zero, implying the dynamics are completely symmetric in time as expected [16,41]. This solidifies our interpretation of the entropy production (14) as a measure of time asymmetry and irreversibility generalized to the strong coupling regime.

VI. CONCLUSION

In this paper we have shown that the entropy production in a system strongly interacting with a bath demonstrates a positive increase in accordance with the second law of thermodynamics. In particular, we proved that entropy is produced when the system and bath are driven away from the conditional equilibrium distribution in $\mathcal{D}_\beta$. As we have argued, the stochastic entropy production (14) is accessible through monitoring the system's path in phase space, implying that in principle a verification of our results (15) and (19) should be accessible using standard experimental techniques [5,43]. Our results provide important modifications to Landauer's principle [35] in the presence of strong coupling, as the change in Shannon entropy is insufficient to characterize the minimum heat dissipated into the bath as the result of information erasure, as shown in the generalized Clausius inequality (16). Apparent violations of Landauer's principle resulting from correlations between the system and bath in the strong coupling regime [30,44] are naturally resolved by this modification.

Note added. After completion of this paper we became aware of similar results obtained by Strasberg and Esposito in [39].

ACKNOWLEDGMENTS

H.M. is supported by Engineering and Physical Sciences Research Council (EPSRC) through a Doctoral Training Grant. J.A. acknowledges support from EPSRC, Grant No. EP/M009165/1, and the Royal Society. We would like to thank Hamed Mohammady, Ian Ford, and Christopher Jarzynski for useful comments regarding the paper. This research was supported by the European Cooperation in Science and Technology (COST) network MP1209 "Thermodynamics in the quantum regime."

APPENDIX A: PROOF OF EQ. (9)

In this section we prove that for any conditional equilibrium distribution $\rho(z_t;t) = \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t) \in \mathcal{D}_\beta$, the nonequilibrium potentials (6) satisfy the additive property $\chi_{\rm tot}(\lambda_t;t) = \tilde{\chi}_s(\lambda_t;t) + \chi_b^{\rm eq}$. To express $\rho(z_t;t)$ in a more useful form we use the following identity [25]:

$$\begin{aligned}\rho_b^{\rm eq}(y_t|x_t) &= \frac{\rho^{\rm eq}(z_t;\lambda_t)}{\int dy_t \rho^{\rm eq}(z_t;\lambda_t)} \\ &= \frac{e^{-\beta[H_b(y_t)+V_{\rm int}(z_t)]}}{\int dy_t\, e^{-\beta[H_b(y_t)+V_{\rm int}(z_t)]}}.\end{aligned} \tag{A1}$$

We now note that the nonequilibrium internal energy is given by $\tilde{U}_s(\lambda_t;t) = \langle \partial_\beta[\beta\tilde{H}_s(x_t;\lambda_t)]\rangle_s$. To proceed we expand the fluctuating internal energy function $\tilde{u}_s(x_t;\lambda_t) = \partial_\beta[\beta\tilde{H}_s(x_t;\lambda_t)]$:

$$\begin{aligned}\tilde{u}_s(x_t;\lambda_t) &= \partial_\beta[\beta\tilde{H}_s(x_t;\lambda_t)] \\ &= H_s(x_t;\lambda_t) - \frac{\partial_\beta\langle e^{-\beta V_{\rm int}(z_t)}\rangle_b^{\rm eq}}{\langle e^{-\beta V_{\rm int}(z_t)}\rangle_b^{\rm eq}} \\ &= H_s(x_t;\lambda_t) \\ &\quad + \frac{\int dy_t\, e^{-\beta[H_b(y_t)+V_{\rm int}(z_t)]}[H_b(y_t)+V_{\rm int}(z_t)]}{\int dy_t\, e^{-\beta[H_b(y_t)+V_{\rm int}(z_t)]}} \\ &\quad + \partial_\beta\left[e^{-\beta F_b^{\rm eq}}\right] \\ &= H_s(x_t;\lambda_t) + \int dy_t\, \rho_b^{\rm eq}(y_t|x_t;\lambda_t)[H_b(y_t) \\ &\quad + V_{\rm int}(z_t)] - U_b^{\rm eq}.\end{aligned} \tag{A2}$$

Averaging both sides of Eq. (A2) with respect to $\rho_s(x_t;t)$ gives

$$\begin{aligned}\tilde{U}_s(\lambda_t;t) &= \int dz_t\, \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t)[H_s(x_t;\lambda_t) \\ &\quad + H_b(y_t) + V_{\rm int}(z_t)] - U_b^{\rm eq} \\ &= U_{\rm tot}(\lambda_t;t) - U_b^{\rm eq}.\end{aligned} \tag{A3}$$

Turning now to the entropy, we need to evaluate the Gibbs-Shannon entropy of the state $\rho(z_t;t) \in \mathcal{D}_\beta$. This can be done from the following equivalent identity:

$$\rho_b^{\rm eq}(y_t|x_t;\lambda_t) = e^{-\beta[H(z_t;\lambda_t)-\tilde{H}_s(x_t;\lambda_t)-F_b^{\rm eq}]}. \tag{A4}$$

Using this we can show the following:

$$\begin{aligned}S_{\rm tot}(\lambda_t;t) &= -\int dz_t\, \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t)d \\ &\quad \times \left[\ln \rho_s(x_t;t) + \ln \rho_b^{\rm eq}(y_t|x_t)\right] \\ &= S_s(\lambda_t;t) - \beta F_b^{\rm eq} + \beta\int dz_t\, \rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t) \\ &\quad \times [H(z_t;\lambda_t) - \tilde{H}_s(x_t;\lambda_t)] \\ &= S_s(\lambda_t;t) - \beta\left(U_{\rm tot}(\lambda_t;t) - U_b^{\rm eq}\right) + S_b^{\rm eq} \\ &\quad - \beta\int dx_t\, \rho_s(x_t;t)\tilde{H}_s(x_t;\lambda_t) \\ &= S_s(\lambda_t;t) + \beta\tilde{U}_s(\lambda_t;t) - \beta\langle\tilde{H}_s(x_t;\lambda_t)\rangle_s + S_b^{\rm eq} \\ &= \tilde{S}_s(\lambda_t;t) + S_b^{\rm eq},\end{aligned} \tag{A5}$$

where we used $U_{\rm tot}(\lambda_t;t)-U_b^{\rm eq}=\tilde{U}_s(\lambda_t;t)$ and $\beta^2\langle\partial_\beta\tilde{H}_s(x_t;\lambda_t)\rangle_s=\beta\tilde{U}_s(\lambda_t;t)-\beta\langle\tilde{H}_s(x_t;\lambda_t)\rangle_s$. Finally, the last additive relation

$$F_{\rm tot}(\lambda_t;t)=\tilde{F}_s(\lambda_t;t)+F_b^{\rm eq} \tag{A6}$$

follows trivially from Eqs. (A2) and (A5) together with the definition of fluctuating free energy, $\tilde{f}_s(x_t;\lambda_t)=\tilde{u}_s(x_t;\lambda_t)-\beta^{-1}\tilde{s}_s(x_t;\lambda_t)$. This concludes the proof of Eq. (9).

APPENDIX B: PROOF OF EQ. (15)

We begin by expressing the decrease in nonequilibrium entropy for the NEQ process specified by assumptions (i) and (ii) in the main text as follows:

$$\begin{aligned}\Delta\tilde{S}_s&=\tilde{S}_s(\lambda_0;t_0)-\tilde{S}_s(\lambda_t;t)\\&=S_{\rm tot}(\lambda_0;t_0)-S_b^{\rm eq}-\tilde{S}_s(\lambda_t;t)\\&=S_{\rm tot}(\lambda_t;t)-S_b^{\rm eq}-S_s(\lambda_t;t)-\beta^2\left\langle\partial_\beta\tilde{H}_s(x_t;\lambda_t)\right\rangle_s\\&=S_{\rm tot}(\lambda_t;t)-S_b^{\rm eq}-S_s(\lambda_t;t)-\beta\tilde{U}_s(\lambda_t;t)\\&\quad+\beta\langle\tilde{H}_s(x_t;\lambda_t)\rangle_s,\end{aligned} \tag{B1}$$

where we recall $S_s(t)=\int dx_t\rho_s(x_t;t)\ln\rho_s(x_t;t)$ represents the Gibbs-Shannon entropy of the system. In the second line we applied the additivity of the nonequilibrium entropy, according to Eq. (9). This is ensured by our choice of initial conditions given by assumption (i). In the third line we used the fact that the Gibbs-Shannon entropy is invariant under closed evolution given by Eq. (11) [17]. The remaining steps follow from the definitions of $\tilde{S}_s(\lambda_t;t)$ and $\tilde{U}_s(\lambda_t;t)$.

Now we introduce the Kullback-Leibler (KL) divergence $D[\rho(z_t;t)||\sigma(z_t;t)]$ defined in Eq. (15). Using $\sigma(z_t;t)=\rho_s(x_t;t)\rho_b^{\rm eq}(y_t|x_t)$ according to Eq. (A4), the KL divergence can be evaluated as follows:

$$\begin{aligned}&D[\rho(z_t;t)||\sigma(z_t;t)]\\&=\int dz_f\,\rho(z_t;t)\ln\left[\frac{\rho(z_t;t)}{\sigma(z_t;t)}\right]\\&=-S_{\rm tot}(\lambda_t;t)+S_s(\lambda_t;t)-\int dz_f\;\rho(z_t;t)\ln\rho_b^{\rm eq}(y_t|x_t)\\&=-S_{\rm tot}(\lambda_t;t)+S_s(\lambda_t;t)-\beta F_b^{\rm eq}\\&\quad+\beta\langle H(z_f;\lambda_f)\rangle-\beta\langle\tilde{H}_s(x_f;\lambda_f)\rangle_s\\&=-\Delta\tilde{S}_s+\beta\left[U_{\rm tot}(\lambda_t;t)-\tilde{U}_s(\lambda_t;t)-U_b^{\rm eq}\right],\end{aligned} \tag{B2}$$

where we used Eq. (B1) and $F_b^{\rm eq}=U_b^{\rm eq}-\beta^{-1}S_b^{\rm eq}$ in the final line. By using $\tilde{U}_s(\lambda_0;t_0)=U_{\rm tot}(\lambda_0;t_0)-U_b^{\rm eq}$ from Eq. (9), it is straightforward to see that the dissipated heat (13) takes the form

$$\begin{aligned}\langle\tilde{Q}(t)\rangle&=[U_{\rm tot}(\lambda_t;t)-\tilde{U}_s(\lambda_t;t)]-[U_{\rm tot}(\lambda_0;t_0)-\tilde{U}_s(\lambda_0;t_0)]\\&=U_{\rm tot}(\lambda_t;t)-\tilde{U}_s(\lambda_t;t)-U_b^{\rm eq}.\end{aligned} \tag{B3}$$

Finally, we combine Eqs. (B2) and (12) to arrive at

$$D[\rho(z_t;t)||\sigma(z_t;t)]=\beta\langle\tilde{Q}(t)\rangle-\Delta\tilde{S}_s=\langle\Sigma(t)\rangle, \tag{B4}$$

thus concluding the proof of Eq. (15).

APPENDIX C: PROOF OF EQ. (19)

To begin, first note that the fluctuating heat (13) can be expressed in terms of the difference between the fluctuating total energy and fluctuating internal energy of the system:

$$\begin{aligned}\tilde{Q}(z_t;t)&=[H(z_t;\lambda_t)-\tilde{u}_s(x_t;\lambda_t)]\\&\quad-[H(z_0;\lambda_0)-\tilde{u}_s(x_0;\lambda_0)].\end{aligned} \tag{C1}$$

Recall that the initial state for the forward process is specified by $\sigma(z_0;t_0)=\rho_s(x_0;t_0)\rho_b^{\rm eq}(y_0|x_0;\lambda_0)$, whilst for the time-reversed process the initial configuration is given by $\sigma(z_t^*;t)=\rho_s(x_t^*;t)\rho_b^{\rm eq}(y_t^*|x_t^*)\in\mathcal{D}_\beta$. Using Eq. (A4) we expand the following:

$$\begin{aligned}\ln\left[\frac{\sigma(z_0;t_0)}{\sigma(z_t^*;t)}\right]&=\ln\left[\frac{\rho_s(x_0;t_0)\rho_b^{\rm eq}(y_0|x_0)}{\rho_s(x_t^*;t)\rho_b^{\rm eq}(y_t^*|x_t^*)}\right]\\&=\ln\left[\frac{\rho_s(x_0;t_0)}{\rho_s(x_t^*;t)}\right]-\beta[H(z_0;\lambda_0)-H(z_t^*;\lambda_t)-\tilde{H}_s(x_0;\lambda_0)+\tilde{H}_s(x_t^*;\lambda_t)]\\&=\ln\left[\frac{\rho_s(x_0;t_0)}{\rho_s(x_t;t)}\right]-\beta[H(z_0;\lambda_0)-H(z_t;\lambda_t)-\tilde{H}_s(x_0;\lambda_0)+\tilde{H}_s(x_t;\lambda_t)]\\&=\tilde{s}_s(x_t;\lambda_t)-\tilde{s}_s(x_0;\lambda_0)-\beta[H(z_0;\lambda_0)-H(z_t;\lambda_t)-\tilde{H}_s(x_0;\lambda_0)+\tilde{H}_s(x_t;\lambda_t)-\beta^2\partial_\beta\tilde{H}_s(x_0;\lambda_0)+\beta^2\partial_\beta\tilde{H}_s(x_t;\lambda_t)]\\&=\tilde{s}_s(x_t;\lambda_t)-\tilde{s}_s(x_0;\lambda_0)+\beta\tilde{Q}(z_t;t),\end{aligned} \tag{C2}$$

where we used the time-reversal symmetry assumptions for $H(z_t;\lambda_t)$ and $\rho_s(x_t;t)$ and in the final line applied the definition (14). The above equality represents a detailed balanced relation that can be used to prove Eq. (19). We now evaluate the probability $\overleftarrow{P}(-\Sigma)$:

$$\begin{aligned}\overleftarrow{P}(-\Sigma)&=\int dz_t^*\;\sigma(z_t^*;t)\;\delta[\Sigma+\Sigma(z_t^*)]\\&=\int dz_0\left|\frac{\partial z_t^*}{\partial z_0}\right|^{-1}\left[\frac{\sigma(z_t^*;t)}{\sigma(z_0;t_0)}\right]\sigma(z_0;t_0)\;\delta[\Sigma-\Sigma(z_t;t)]\end{aligned} \tag{C3}$$

$$= \int dz_0\, e^{-\tilde{s}_s(x_t;\lambda_t)+\tilde{s}_s(x_0;\lambda_0)-\beta\tilde{Q}(z_t;t)}\, \sigma(z_0;t_0)\, \delta[\Sigma - \Sigma(z_t;t)]$$
$$= e^{-\Sigma} \int dz_0\, \sigma(z_0;t_0)\, \delta[\Sigma - \Sigma(z_t;t)]$$
$$= e^{-\Sigma}\, \overrightarrow{P}(+\Sigma), \tag{C4}$$

where in the second line we performed a change of variables $z_t^* \to z_0$ along with $\Sigma(z_t;t) = -\Sigma(z_t^*;t)$, in the third line we used the fact that the Jacobian is equal to unity and Eq. (C2), and in the fourth line we pulled the exponential outside the integral due to the presence of the delta function. This concludes the proof of Eq. (19).

[1] K. Sekimoto, Prog. Theor. Phys. Suppl. **130**, 17 (1998).
[2] U. Seifert, Phys. Rev. Lett. **95**, 040602 (2005).
[3] U. Seifert, Eur. Phys. J. B **64**, 423 (2008).
[4] U. Seifert, Rep. Prog. Phys **75**, 126001 (2012).
[5] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante, Nature **437**, 231 (2005).
[6] J. Millen, T. Deesuwan, P. Barker, and J. Anders, Nat. N. Tech. **9**, 425 (2014).
[7] S. An, J.-N. Zhang, M. Um, D. Lv, Y. Lu, J. Zhang, Z.-Q. Yin, H. T. Quan, and K. Kim, Nat. Phys. **11**, 193 (2014).
[8] U. Seifert, Euro. Phys. J. E **34**, 26 (2011).
[9] R. Clausius, Ann. Phys. **201**, 353 (1865).
[10] L. Landau and E. Lifshitz, Course of Theoretical Physics **5**, 230 (1958).
[11] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[12] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[13] T. Hatano and S.-I. Sasa, Phys. Rev. Lett. **86**, 3463 (2001).
[14] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, Phys. Rev. Lett. **98**, 080602 (2007).
[15] C. Jarzynski, Ann. Rev. Cond. M. Phys. **2**, 329 (2011).
[16] E. H. Feng and G. E. Crooks, Phys. Rev. Lett. **101**, 090602 (2008).
[17] J. M. R. Parrondo, C. V. den Broeck, and R. Kawai, New J. Phys. **11**, 073008 (2009).
[18] O.-P. Saira, Y. Yoon, T. Tanttu, M. Möttönen, D. V. Averin, and J. P. Pekola, Phys. Rev. Lett. **109**, 180601 (2012).
[19] A. I. Brown and D. A. Sivak, Phys. Rev. E **94**, 032137 (2016).
[20] H. Kramers, Physica **7**, 284 (1940).
[21] P. Talkner, M. Campisi, and P. Hänggi, J. Stat. Mech. (2009) P02025.
[22] S. Deffner and E. Lutz, Phys. Rev. Lett. **107**, 140404 (2011).
[23] C. Jarzynski, Phys. Rev. X **7**, 011008 (2016).
[24] M. F. Gelin and M. Thoss, Phys. Rev. E **79**, 051121 (2009).
[25] P. Talkner and P. Hänggi, Phys. Rev. E **94**, 022143 (2016).
[26] U. Seifert, Phys. Rev. Lett. **116**, 020601 (2016).
[27] C. Jarzynski, J. Stat. Mech. (2004) P09005.
[28] G.-L. Ingold, P. Hänggi, and P. Talkner, Phys. Rev. E **79**, 061105 (2008).
[29] M. Esposito, K. Lindenberg, and C. Van den Broeck, New J. Phys. **12**, 013013 (2010).
[30] S. Hilt, S. Shabbir, J. Anders, and E. Lutz, Phys. Rev. E **83**, 030102 (2011).
[31] M. Carrega, P. Solinas, M. Sassetti, and U. Weiss, Phys. Rev. Lett. **116**, 240403 (2016).
[32] T. G. Philbin and J. Anders, J. Phys. A **49**, 215303 (2016).
[33] P. Strasberg, G. Schaller, N. Lambert, and T. Brandes, New J. Phys. **18**, 33 (2016).
[34] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[35] D. Reeb and M. M. Wolf, New J. Phys. **16**, 103011 (2014).
[36] M. Campisi, P. Talkner, and P. Hänggi, Phys. Rev. Lett. **102**, 210401 (2009).
[37] S. Hilt, B. Thomas, and E. Lutz, Phys. Rev. E **84**, 031110 (2011).
[38] G. W. Ford, J. T. Lewis, and R. F. O'Connell, Phys. Rev. Lett. **55**, 2273 (1985).
[39] P. Strasberg and M. Esposito, Phys. Rev. E **95**, 062101 (2017).
[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 2012).
[41] S. Vaikuntanathan and C. Jarzynski, EPL **87**, 60005 (2009).
[42] A. Gomez-Marin, J. M. R. Parrondo, and C. Van den Broeck, EPL **82**, 50002 (2008).
[43] C. Tietz, S. Schuler, T. Speck, U. Seifert, and J. Wrachtrup, Phys. Rev. Lett. **97**, 050602 (2006).
[44] A. E. Allahverdyan and T. M. Nieuwenhuizen, Phys. Rev. Lett. **85**, 1799 (2000).