


Thermodynamics and stochastic thermodynamics of strongly coupled systems

Xiangjun Xing^{1,2,3,*} and Mingnan Ding¹

¹*Wilczek Quantum Center, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*

²*T.D. Lee Institute, Shanghai Jiao Tong University, Shanghai 200240, China*

³*Shanghai Research Center for Quantum Sciences, Shanghai 201315, China*

 (Received 4 August 2023; revised 1 February 2024; accepted 6 February 2024; published 4 March 2024)

We further develop the strong-coupling theory of thermodynamics and stochastic thermodynamics for continuous systems, constructed in the previous work [*Phys. Rev. Res.* **4**, 013015 (2022)]. A small system strongly interacting with a its environment, the dynamics of the system is assumed to be much slower than that of the bath. The system Hamiltonian is defined to be the Hamiltonian of mean force, whereas the system entropy is defined as the Gibbs-Shannon entropy. Equilibrium ensemble theories and thermodynamic theories are established for the system. Variations of three types of parameters are considered: (i) *the system parameter* λ which couples to the system and to the interaction, (ii) *the bath parameter* λ' which couples to the bath only, and (iii) the temperature $T = 1/\beta$. The work done to the system consists of three parts, proportional to $d\lambda$, $d\lambda'$, and $d\beta$ respectively. The part proportional to $d\beta$ can be understood as the work done by the bath. As long as λ' and β are not fixed, the work is not the change of total energy of the joint system. The differences between our strong-coupling equilibrium thermodynamics and the classical thermodynamics are discussed. The thermodynamic theory is promoted to the nonequilibrium level. Both the first and second laws of thermodynamics, as well as fluctuation theorems, are established for nonequilibrium processes. For processes with varying temperatures, fluctuation theorems cannot be expressed in terms of integrated work alone. Regardless of various subtleties, however, the stochastic thermodynamic theory is formulated in terms of system variables only, and $dS - \beta dQ$ is the change of total entropy. Thermodynamic quantities of the system are related to those of the joint system, and the equivalence of theories at two levels of coarse-graining is explicitly demonstrated. Finally we show that there are infinite numbers of equivalent strong-coupling theories, each determined by its definition of system Hamiltonian. Our theory is distinguished by its maximal similarity with the weak-coupling theory.

DOI: [10.1103/PhysRevE.109.034105](https://doi.org/10.1103/PhysRevE.109.034105)

I. INTRODUCTION

The traditional formulations of classical thermodynamics and stochastic thermodynamics are applicable only to systems weakly interacting with their environments. In recent years, there have been significant interests in generalizing the theories to systems strongly coupled to environments [1–16]. Broadly speaking, there are two related aspects of strong-coupling physics: the kinetic aspect and the thermodynamic aspect. The kinetic aspect addresses the effective time evolution of a system strongly interacting to its environment. Application of projection operator methods [17–22] generally yields non-Markovian nonlinear Langevin equations with colored noise and memory effects. It is only in the limit of timescale separation (TSS) that these equations become Markovian. By contrast, the thermodynamic aspect of strong-coupling physics addresses how thermodynamic functions of the system should be defined, such that thermodynamic laws and fluctuation theorems can be established without referring to the environment.

Several theoretical formalisms have been proposed [3,6,7,15] to study the thermodynamic aspect of the

strong-coupling problem. In all these theories, thermodynamic quantities are defined in terms of the *Hamiltonian of mean force* (HMF), which fully determines the equilibrium probability distribution of the system variables. The relations between these formalisms have also been systematically discussed [11,13,14]. Yet the opinions have not yet fully converged. In particular, it is not clear whether a consistent theory can be constructed without referring to the environmental variables when the interaction and/or the bath properties are time-dependent. Another related issue is how to uniquely fix the Hamiltonian of mean force using measurable quantities of the system. Finally, there are also worries about the lacking of a simple guiding principle that picks out a particular theory from an infinite number of possible theories that are consistent with thermodynamic laws. A critical review of all these issues can be found in Ref. [11].

Among all proposed strong-coupling theories of stochastic thermodynamics, probably the most influential one was developed by Seifert [3] and critically reviewed by Hanggi *et al.* [4,11,14]. This theory presumes that the interaction between the system and the bath, as well as the bath Hamiltonian, is time-independent. Work is defined as the energy change of the joint system, which can be expressed as an integral along the system's trajectory, without explicitly referring to the bath variables. The fluctuating internal energy is defined as

*xxing@sjtu.edu.cn

$\partial_\beta(\beta H_X)$, where H_X is the HMF. This theory is substantially more complex than the weak-coupling theory, because many thermodynamic quantities are modified by interactions. Furthermore, if the interaction is time-dependent, such as in some heat engine problems, the work thus defined can no longer be expressed in terms of system trajectory alone [9], and hence the theory becomes inapplicable. Finally, there has been no discussion on the cases where the temperature or other bulk properties of the bath are time-dependent.

In Ref. [15], an alternative theory of strong-coupling thermodynamics was developed under the assumption that the system dynamics is much slower than the bath dynamics. The fluctuating internal energy is identified with the HMF H_X , while the work is defined as $d\mathcal{W} \equiv (\partial_\lambda H_X)d\lambda$. Except for the subtlety that H_X generally depends on the temperature, this theory is formally identical to the weak-coupling theory and hence is much simpler than the previous theories. More importantly, this theory is applicable even if the interactions are time-dependent. It turns out that the work defined in this theory is the change of total energy *averaged over fluctuations of bath variables*. Such an average is physically justified if the dynamics of the bath is much faster than the dynamics of the system and the variation of the control parameter. For systems with fixed interaction, the equivalence between this theory and other strong-coupling theories was also established.

The purposes of the present work are to extend the theory in Ref. [15] to more general situations and to clarify several conceptual issues raised in Refs. [4,11,14]. We discuss three types of control parameters, which are the system parameter λ that couples to the system and to the interaction, the bath parameter λ' that couples only to the bath, and the temperature $T = 1/\beta$. We show that the work done to the system consists of three parts, proportional to $d\lambda$, $d\lambda'$, and $d\beta$, respectively. The latter two parts have not been studied in the previous works and arise only in strongly coupled systems. We show that the physical meaning of the part proportional to $d\beta$ is the work done by the bath to the system. Furthermore, as long as λ' or β varies with time, the work done to the system cannot be interpreted as the change of the joint system, at variance with most of the previous theories. We also prove the second law of thermodynamics and fluctuation theorems for general nonequilibrium processes where λ , λ' , β all vary with time. If the temperature is time-dependent, the fluctuation theorems can no longer be expressed in terms of work alone. Regardless of all these subtleties, however, a consistent theory of thermodynamics and stochastic thermodynamics shall be formulated without explicitly referring to the bath variables. Finally, we show that there is an infinite number of strong-coupling theories that are all equivalent to our theory. Each of these theories is completely determined by the definition of the system Hamiltonian. Our theory is distinguished by its maximal similarity with the weak-coupling theory. Overall, the present work supplies a complete theory for strong-coupling thermodynamics and stochastic thermodynamics, and also substantially extends the application domain of stochastic thermodynamics.

To make the work as pedagogical and self-contained as possible, we use Secs. II A and III A to review some relevant parts of Ref. [15]. These sections are, however, not simple repetitions of Ref. [15], since there are important notational

differences between the present work and Ref. [15]. Also the equations in these sections are heavily used in the latter parts of the work. Hence we do not recommend the readers to skip these sections.

The remainder of this work is organized as follows: In Sec. II we discuss the decomposition of Hamiltonian and entropy, with special attention paid to the nonuniqueness of HMF. In Sec. III we discuss strong-coupling equilibrium thermodynamics. In Sec. IV we discuss strong-coupling stochastic thermodynamics. In Sec. V we construct a continuum of strong-coupling theory, and demonstrate their equivalence. We also discuss the strong-coupling issue of discrete Markov systems described by master equations. In Sec. VI we use two simple examples to illustrate our theory. Finally, in Sec. VII we draw the conclusion and project future directions.

II. DECOMPOSITION OF ENERGY AND ENTROPY

As in Ref. [15], we consider a set of continuous variables \mathbf{X} , referred to as *the system*, strongly interacting with a set of continuous variables \mathbf{Y} , referred to as *the bath*. We assume TSS, which means that the dynamics of \mathbf{Y} is much faster than that of \mathbf{X} . Furthermore, there are two control parameters in the Hamiltonian (2.1). λ , the *system parameter*, controls the system and its interaction with the bath, whereas λ' , the *bath parameter*, controls the bulk property of the bath.

A. Decomposition of Hamiltonian and energy

We first briefly review the decomposition of Hamiltonian discussed in Ref. [15]. The total Hamiltonian of the joint system \mathbf{XY} is written as

$$H_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda') = H_{\mathbf{X}}^0(\mathbf{x}; \lambda) + H_{\mathbf{Y}}^0(\mathbf{y}; \lambda') + H_I^0(\mathbf{x}, \mathbf{y}; \lambda), \quad (2.1)$$

where \mathbf{x}, \mathbf{y} are the values of \mathbf{X}, \mathbf{Y} , respectively, $H_{\mathbf{X}}^0(\mathbf{x}; \lambda)$ and $H_{\mathbf{Y}}^0(\mathbf{y}; \lambda')$ are the *bare Hamiltonians* of the system and of the bath, respectively, while $H_I^0(\mathbf{x}, \mathbf{y}; \lambda)$ is the *bare interaction*. Note that three terms on the right-hand side (r.h.s.) of Eq. (2.1) are not separately measurable. Hence the decomposition (2.1) is rather arbitrary. Consider the typical situation where a small system interacts with a large bath via short-ranged potential energy,¹ the dimension of \mathbf{Y} is much larger than that of \mathbf{X} , and there is only a small fraction of fast variables significantly interacting with the slow variables. Hence both $H_{\mathbf{X}}$ and H_I^0 are small whereas $H_{\mathbf{Y}}^0(\mathbf{y}; \lambda')$ are extensive in the bath size.

Throughout this work, we use shorthand notations $\int_{\mathbf{y}} \equiv \int d^N y$ and $\int_{\mathbf{x}} \equiv \int d^N x$, and shall set the Boltzmann constant $k_B = 1$. We assume that the joint system \mathbf{XY} is further in weak interaction with a superbath. With λ, λ' remain fixed, the joint system converges to a joint equilibrium state, specified by the

¹The assumption of a short-range interaction is not as severe as it may appear. A long-range electrostatic interaction is always screened so that it becomes effectively short ranged. Gravitational interaction cannot be screened but does not play any role in the statistical physics of small systems.

joint *Gibbs-Boltzmann distribution*

$$p_{\mathbf{XY}}^{\text{EQ}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta) \equiv \frac{e^{-\beta H_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda')}}{Z_{\mathbf{XY}}(\lambda, \lambda', \beta)}, \quad (2.2)$$

where $Z_{\mathbf{XY}}(\lambda, \lambda', \beta)$ is the *joint partition function*

$$Z_{\mathbf{XY}}(\lambda, \lambda', \beta) \equiv \int_{\mathbf{xy}} e^{-\beta H_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda')}. \quad (2.3)$$

By integrating out \mathbf{y} in Eq. (2.2), we obtain the *marginal equilibrium probability density function* (pdf) of the system variables:

$$p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}; \lambda, \lambda', \beta) \equiv \int_{\mathbf{y}} \frac{e^{-\beta H_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda')}}{Z_{\mathbf{XY}}(\lambda, \lambda', \beta)}. \quad (2.4)$$

For the sake of simplicity, we often hide the parametric dependencies of pdfs on λ, λ', β below.

The *Hamiltonian of mean force* (HMF) [1,3,4,23] is defined as

$$\begin{aligned} H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta) &\equiv -T \ln \frac{\int_{\mathbf{y}} e^{-\beta H_{\mathbf{XY}}} }{\int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0}} \quad (2.5) \\ &= H_{\mathbf{X}}^0(\mathbf{x}) - T \ln \frac{\int_{\mathbf{y}} e^{-\beta(H_{\mathbf{Y}}^0 + H_{\mathbf{Y}}^1)}}{\int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0}}. \quad (2.6) \end{aligned}$$

We can rewrite Eq. (2.4) in terms of the HMF as [1,3,4,23]

$$p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) = \frac{e^{-\beta H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)}}{Z_{\mathbf{X}}(\lambda, \lambda', \beta)}, \quad (2.7)$$

where $Z_{\mathbf{X}}(\lambda, \lambda', \beta)$ satisfies

$$Z_{\mathbf{X}}(\lambda, \lambda', \beta) \equiv \int_{\mathbf{x}} e^{-\beta H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)}, \quad (2.8)$$

which is the *canonical partition function of the system*.

It is easy to verify that

$$Z_{\mathbf{X}}(\lambda, \lambda', \beta) = \frac{Z_{\mathbf{XY}}(\lambda, \lambda', \beta)}{Z_{\mathbf{Y}}^0(\lambda', \beta)}, \quad (2.9)$$

where $Z_{\mathbf{Y}}^0(\lambda', \beta)$ is the partition function of the bare bath:

$$Z_{\mathbf{Y}}^0(\lambda', \beta) \equiv \int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0(\mathbf{y}; \lambda')}. \quad (2.10)$$

If certain component of \mathbf{Y} is not coupled to \mathbf{X} , it does not appear in $H_{\mathbf{Y}}^0$. It is then easy to see from Eq. (2.6) that it does not appear in the HMF either. Hence the HMF remains finite even if the size of the bath becomes infinite. If the system consists of two subsystems \mathbf{X}_1 and \mathbf{X}_2 which are widely separated, those components of \mathbf{Y} which couple to \mathbf{X}_1 do not couple to \mathbf{X}_2 , and hence the HMF can be broken into two noninteracting components:

$$H_{\mathbf{X}_1, \mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) = H_{\mathbf{X}_1}(\mathbf{x}_1) + H_{\mathbf{X}_2}(\mathbf{x}_2). \quad (2.11)$$

For this reason, $H_{\mathbf{X}_1, \mathbf{X}_2} - H_{\mathbf{X}_1} - H_{\mathbf{X}_2}$ may be understood as the effective interaction between two subsystems $\mathbf{X}_1, \mathbf{X}_2$, as mediated by the bath. It is not difficult to construct explicit proof for these results. We think, however, that the above heuristic argument is more helpful.

The total Hamiltonian (2.1) can be decomposed into

$$H_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda') = H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta) + H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta), \quad (2.12)$$

where $H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta)$ is defined as

$$\begin{aligned} H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta) &\equiv H_{\mathbf{Y}}^0(\mathbf{y}; \lambda') + H_{\mathbf{Y}}^1(\mathbf{x}, \mathbf{y}; \lambda) \\ &+ T \ln \frac{\int_{\mathbf{y}} e^{-\beta(H_{\mathbf{Y}}^0 + H_{\mathbf{Y}}^1)}}{\int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0}}. \quad (2.13) \end{aligned}$$

We call $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ respectively the system Hamiltonian and the bath Hamiltonian, and their values the system energy and the bath energy. Such a decomposition of total Hamiltonian and energy leads to huge simplification of the strong-coupling theories of thermodynamics and stochastic thermodynamics. Note that both $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ depend on all three parameters λ, λ', β .

Let us define the *partition function* of the bath as

$$Z_{\mathbf{Y}}(\lambda', \beta) \equiv \int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta)}. \quad (2.14)$$

Using Eqs. (2.13) and (2.10), we easily see that

$$Z_{\mathbf{Y}}(\lambda', \beta) = \int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0} = Z_{\mathbf{Y}}^0(\lambda', \beta). \quad (2.15)$$

Hence, even though $H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta)$ depends on \mathbf{x} and on λ , the partition function $Z_{\mathbf{Y}}(\lambda', \beta)$ is independent of \mathbf{x} and λ anyway. The combination of Eqs. (2.15) and (2.9) leads to a decomposition of the joint partition function:

$$Z_{\mathbf{XY}}(\lambda, \lambda', \beta) = Z_{\mathbf{X}}(\lambda, \lambda', \beta) Z_{\mathbf{Y}}(\lambda', \beta). \quad (2.16)$$

Taking the derivative of Eq. (2.14) with respect to \mathbf{x} and λ , we obtain

$$\int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}} \frac{\partial H_{\mathbf{Y}}}{\partial \mathbf{x}} = \int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}} \frac{\partial H_{\mathbf{Y}}}{\partial \lambda} = 0. \quad (2.17)$$

These results will play a significant role in our theory.

B. Nonuniqueness of the Hamiltonian of mean force

While the HMF uniquely determines the equilibrium pdf of \mathbf{X} via Eqs. (2.7) and (2.8), the converse is not true. The fact that the HMF cannot be determined uniquely by observing the equilibrium distribution of the system variables has been an important part of disputes in several recent studies [4,11,13,14]. To see one origin of nonuniqueness of $H_{\mathbf{X}}$, we may change $H_{\mathbf{Y}}^0$ and $H_{\mathbf{Y}}^1$ on the r.h.s. of Eq. (2.1) simultaneously such that their sum remains fixed. According to Eq. (2.5), such a change leads to a change in HMF, but not in the total Hamiltonian.

But even the total Hamiltonian is not fully determined by the joint equilibrium distribution (2.2). We may add to it an arbitrary constant $C(\lambda, \lambda')$,

$$H_{\mathbf{XY}} \rightarrow H_{\mathbf{XY}} + C(\lambda, \lambda'), \quad (2.18)$$

which does not change the dynamic trajectories or the probability distributions of the joint system. Nonetheless, such a transformation (2.18) does change various thermodynamic potentials, such as internal energy and free energy. In other words, thermodynamics is not completely determined by dynamics alone.

As one possible way to uniquely fix the total Hamiltonian, one may impose the condition that $H_{\mathbf{XY}}$ vanishes in a

particular reference state (x_0, y_0) :

$$H_{\mathbf{XY}}(x_0, y_0; \lambda, \lambda') = 0. \quad (2.19)$$

An alternative method is to fix the joint partition function to be a constant at a particular set of parameters $\lambda_0, \lambda'_0, \beta_0$ as practiced in Ref. [13]:

$$Z_0 = Z_{\mathbf{XY}}(\lambda_0, \lambda'_0, \beta_0) = \int_{xy} e^{-\beta H_{\mathbf{XY}}(x, y; \lambda_0, \lambda'_0, \beta_0)}. \quad (2.20)$$

Different conventions may turn out to be more convenient in different situations. In the following, we always assume that one choice is made, so that the total Hamiltonian $H_{\mathbf{XY}}$ is uniquely determined by the equilibrium pdf of the joint system.

Let $\phi_{\mathbf{Y}}(\mathbf{y}; \lambda')$ be an arbitrary function of \mathbf{y} and λ' . Consider the following transformation:

$$H_{\mathbf{X}}^0 \rightarrow H_{\mathbf{X}}^0, \quad (2.21a)$$

$$H_{\mathbf{Y}}^0 \rightarrow H_{\mathbf{Y}}^0 + \phi_{\mathbf{Y}}(\mathbf{y}; \lambda'), \quad (2.21b)$$

$$H_I^0 \rightarrow H_I^0 - \phi_{\mathbf{Y}}(\mathbf{y}; \lambda'), \quad (2.21c)$$

which preserves the total Hamiltonian $H_{\mathbf{XY}}$ in Eq. (2.1). The equilibrium pdf (2.7) of system variables of course also remains invariant. According to Eqs. (2.5) and (2.13), this leads to the following transformations of $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$:

$$H_{\mathbf{X}} \rightarrow H_{\mathbf{X}} - \psi(\lambda', \beta), \quad (2.22a)$$

$$H_{\mathbf{Y}} \rightarrow H_{\mathbf{Y}} + \psi(\lambda', \beta), \quad (2.22b)$$

$$\psi(\lambda', \beta) \equiv T \ln \frac{\int_{\mathbf{y}} e^{-\beta H_{\mathbf{Y}}^0}}{\int_{\mathbf{y}} e^{-\beta(H_{\mathbf{Y}}^0 + \phi_{\mathbf{Y}}^0)}}. \quad (2.22c)$$

Note that $\psi(\lambda', \beta)$ depends λ' and β but not on λ , nor on the dynamic variables \mathbf{x}, \mathbf{y} . This shows explicitly how we may change the HMF without changing the equilibrium distribution $p_{\mathbf{X}}^{\text{EQ}}$.

We may also consider a different transformation:

$$H_{\mathbf{X}}^0 \rightarrow H_{\mathbf{X}}^0 + \phi_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda'), \quad (2.23a)$$

$$H_{\mathbf{Y}}^0 \rightarrow H_{\mathbf{Y}}^0, \quad (2.23b)$$

$$H_I^0 \rightarrow H_I^0 - \phi_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda'), \quad (2.23c)$$

which also leaves the total Hamiltonian $H_{\mathbf{XY}}$ intact. But it is easy to verify that this does not lead to any change of $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$, as defined in Eqs. (2.5) and (2.13). Finally, one may also consider the following transformation:

$$H_{\mathbf{X}}^0 \rightarrow H_{\mathbf{X}}^0 + C(\lambda'), \quad (2.24a)$$

$$H_{\mathbf{Y}}^0 \rightarrow H_{\mathbf{Y}}^0 - C(\lambda'), \quad (2.24b)$$

$$H_I^0 \rightarrow H_I^0, \quad (2.24c)$$

where $C(\lambda')$ is independent of \mathbf{x}, \mathbf{y} . But this may be understood as the combination of two transformations discussed above, with $\phi_{\mathbf{X}} = -\phi_{\mathbf{Y}} = C(\lambda')$.

To uniquely fix the HMF $H_{\mathbf{X}}$, we may impose the following condition, which is similar to Eq. (2.19):

$$H_{\mathbf{X}}(x_0; \lambda, \lambda', \beta) = 0. \quad (2.25)$$

Combining this with Eq. (2.19), we also obtain the following condition for the bath Hamiltonian:

$$H_{\mathbf{Y}}(x_0, y_0; \lambda, \lambda', \beta) = 0. \quad (2.26)$$

The fact that the HMF can be fixed only by introducing an arbitrarily chosen condition should not bother us. In general, energy-like quantities are obtained via the integration of dynamic equations and are determined only up to an additive constant.

C. Decomposition of probability distribution

As shown in Ref. [15], using the decompositions (2.12) of Hamiltonian and (2.16) of partition function, we can rewrite the joint equilibrium pdf Eq. (2.2) as

$$p_{\mathbf{XY}}^{\text{EQ}}(\mathbf{x}, \mathbf{y}) = \frac{e^{-\beta H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)} e^{-\beta H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta)}}{Z_{\mathbf{X}}(\lambda, \lambda', \beta) Z_{\mathbf{Y}}(\lambda', \beta)}. \quad (2.27)$$

Since the first term on the r.h.s. is the marginal pdf of \mathbf{x} , cf. Eq. (2.7), the second term on the r.h.s. is precisely the *conditional pdf* of \mathbf{y} given $\mathbf{X} = \mathbf{x}$:

$$p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x}) = \frac{e^{-\beta H_{\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \lambda, \lambda', \beta)}}{Z_{\mathbf{Y}}(\lambda', \beta)}, \quad (2.28)$$

such that Eq. (2.27) becomes the familiar decomposition of the joint pdf into a marginal pdf and a conditional pdf:

$$p_{\mathbf{XY}}^{\text{EQ}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x}). \quad (2.29)$$

Here the superscript EQ in $p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x})$ means equilibrium of the fast variables conditioned on the values of the slow variables.

The above decomposition can be generalized to the nonequilibrium case. The joint nonequilibrium pdf $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$ can also be decomposed as

$$p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}), \quad (2.30)$$

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbf{y}} p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}), \quad (2.31)$$

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}, \quad (2.32)$$

where $p_{\mathbf{X}}(\mathbf{x})$ and $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ are respectively the marginal pdf of \mathbf{x} and the conditional pdf of \mathbf{y} given $\mathbf{X} = \mathbf{x}$.

Since we assume that the dynamics of \mathbf{Y} is much faster than that of \mathbf{X} , it is legitimate to consider intermediate timescales $\tau_{\mathbf{Y}} \ll t \ll \tau_{\mathbf{X}}$, so that \mathbf{Y} already equilibrate conditioned on \mathbf{X} while \mathbf{X} remains out of equilibrium. The joint pdf Eq. (2.30) then reduces to

$$p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x}), \quad (2.33)$$

with $p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x})$ given by Eq. (2.28). States in the form of Eq. (2.33) are called the *stationary preparation class* by Hanggi [4].

D. Decomposition of entropy

For a nonequilibrium pdf of the joint system $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$, the total entropy is defined as the usual *Gibbs-Shannon entropy*:

$$S_{\mathbf{XY}}[p_{\mathbf{XY}}] = - \int_{\mathbf{x}, \mathbf{y}} p_{\mathbf{XY}} \ln p_{\mathbf{XY}}. \quad (2.34)$$

Throughout this work, we use $A_X[p_X]$, $A_{XY}[p_{XY}]$, etc. to denote a nonequilibrium thermodynamic quantities as functionals of the pdfs p_X , p_{XY} , whereas the equilibrium quantities will be simply denoted as A_X , A_{XY} , etc. It is known that with a fixed energy, this entropy is maximized by the microcanonical distribution, i.e., a constant pdf on the energy shell, which corresponds to the thermal equilibrium state. Furthermore, the equivalence between Gibbs-Shannon entropy and thermodynamic entropy at thermal equilibrium is well understood. The meaning of thermodynamic entropy in nonequilibrium situations is not clear in the setting of classical thermodynamics. Yet in stochastic thermodynamics, Gibbs-Shannon entropy (2.34) acquires clear physical meaning, because it can be used to construct nonequilibrium free energy, which can be proved to decrease monotonically in nonequilibrium processes.

We carry out the following decomposition of the nonequilibrium entropy (2.34):

$$S_{XY}[p_{XY}] = S_X + S_{Y|X}, \quad (2.35a)$$

$$S_X[p_X] = - \int_x p_X(x) \ln p_X(x), \quad (2.35b)$$

$$S_{Y|X}[p_{XY}] = - \int_{xy} p_{XY}(x, y) \ln p_{Y|X}(y|x), \quad (2.35c)$$

where S_X is the entropy of \mathbf{X} and $S_{Y|X}$ is the conditional entropy of \mathbf{Y} given \mathbf{X} . Note that while S_X is a functional of the marginal pdf $p_X(x)$, $S_{Y|X}$ is a functional of the joint pdf $p_{XY}(x, y)$. This decomposition is well known in information theory [24]. They are nonequilibrium generalization of decomposition of entropy discussed in Ref. [15].

There is a related concept of conditional entropy:

$$S_{Y|X=x}[p_{Y|X}] = - \int_y p_{Y|X}(y|x) \ln p_{Y|X}(y|x), \quad (2.36)$$

which is called the conditional entropy of \mathbf{Y} given $\mathbf{X} = x$. Note that $S_{Y|X}[p_{XY}]$ as defined in Eq. (2.35c) is the average of $S_{Y|X=x}[p_{Y|X}]$ over the marginal pdf $p_X(x)$:

$$S_{Y|X}[p_{XY}] = \int_x p_X(x) S_{Y|X=x}[p_{Y|X}]. \quad (2.37)$$

Another concept that plays important role in stochastic thermodynamics is the stochastic entropy [25]

$$S_{XY}[p_{XY}, \mathbf{x}, \mathbf{y}] \equiv - \ln p_{XY}(\mathbf{x}, \mathbf{y}), \quad (2.38)$$

which is understood as a functional of the pdf p_{XY} and also as a function of the microstates (\mathbf{x}, \mathbf{y}) . It admits a similar decomposition parallel to Eqs. (2.35):

$$S_{XY}[p_{XY}, \mathbf{x}, \mathbf{y}] = S_X[p_X, \mathbf{x}] + S_{Y|X=x}[p_{Y|X}, \mathbf{x}, \mathbf{y}], \quad (2.39a)$$

$$S_X[p_X, \mathbf{x}] \equiv - \ln p_X(\mathbf{x}), \quad (2.39b)$$

$$S_{Y|X=x}[p_{Y|X}, \mathbf{x}, \mathbf{y}] \equiv - \ln p_{Y|X}(\mathbf{y}|\mathbf{x}). \quad (2.39c)$$

Averaging of Eqs. (2.39) over p_{XY} yields Eqs. (2.35).

Mathematically it is also legitimate to discuss conditional pdf of slow variables given the fast variables. These distributions, however, cannot be easily measured.

III. STRONGLY COUPLED EQUILIBRIUM THERMODYNAMICS

We now use the decompositions of energy and entropy, introduced in the preceding section, to develop a consistent theory of equilibrium thermodynamics for a system strongly coupled to bath. We consider the variations of both the system parameter λ , and the bath parameter λ' , as well as the temperature $1/\beta$. As we will see, the variations of λ' and β lead to a new twist in the notion of thermodynamic work.

A. Brief review of equilibrium thermodynamics theory discussed in Ref. [15]

In this part we first briefly review the equilibrium thermodynamics of Ref. [15] that are relevant to the present work. Readers who are familiar with Ref. [15] may skip this section and go directly to Sec. III B.

We assume that the joint system (consisting of the system and the bath) is weakly coupled to a superbath. Hence the formalism of classical equilibrium statistical mechanics is applicable to the joint system. Using Eq. (2.3), various thermodynamic quantities for the joint system can be written as

$$F_{XY}(\lambda, \lambda', \beta) = -T \ln Z_{XY}(\lambda, \lambda', \beta), \quad (3.1a)$$

$$E_{XY}(\lambda, \lambda', \beta) = \int_{x,y} p_{XY}^{\text{EQ}} H_{XY}, \quad (3.1b)$$

$$S_{XY}(\lambda, \lambda', \beta) = - \int_{x,y} p_{XY}^{\text{EQ}} \ln p_{XY}^{\text{EQ}}, \quad (3.1c)$$

which satisfies the following relations:

$$E_{XY} = \frac{\partial \beta F_{XY}}{\partial \beta}, \quad (3.2a)$$

$$S_{XY} = \beta^2 \frac{\partial F_{XY}}{\partial \beta}, \quad (3.2b)$$

$$F_{XY} = E_{XY} - T S_{XY}. \quad (3.2c)$$

The partial derivatives with respect to β are taken with λ , λ' fixed.

We assume that, for fixed slow variables, the dynamics of the fast variables is ergodic. In an intermediate timescale where $\tau_Y \ll t \ll \tau_X$, the slow variables barely change, whereas the fast variables equilibrate conditioned on the slow variables $\mathbf{X} = \mathbf{x}$. The conditional equilibrium pdf of the fast variables is given by Eq. (2.28), which is Gibbs-Boltzmann pdf with respect to the bath Hamiltonian $H_Y(\mathbf{y}; \mathbf{x}, \lambda, \lambda', \beta)$. In this conditional canonical ensemble of the fast variables, \mathbf{x} serves as a fixed parameter, just like λ , λ' , and β . Using Eq. (2.14), we define the equilibrium free energy of the bath:

$$F_Y(\lambda', \beta) = -T \ln Z_Y(\lambda', \beta), \quad (3.3)$$

which depends on λ' , β but not on \mathbf{x} , λ . According to the decomposition of energy and entropy discussed in Sec. II, the internal energy and entropy of bath in the conditional

equilibrium state are

$$E_Y(\mathbf{x}) = \int_y p_{Y|X}^{\text{EQ}}(y|\mathbf{x}) H_Y(y; \mathbf{x}, \lambda, \lambda', \beta), \quad (3.4a)$$

$$S_{Y|X=x} = - \int_y p_{Y|X}^{\text{EQ}}(y|\mathbf{x}) \ln p_{Y|X}^{\text{EQ}}(y|\mathbf{x}), \quad (3.4b)$$

both of which depend on the slow variables \mathbf{x} and the system parameter λ . They can be combined to form the free energy of the bath: $F_Y(\lambda', \beta) = E_Y(\mathbf{x}) - T S_{Y|X=x}$, which is known to be independent of \mathbf{x} and λ . Note that Eq. (3.4b) is the equilibrium version of the conditional entropy Eq. (2.36).

If we observe the joint system in the very long timescale $t \gg \tau_X \gg \tau_Y$, both the slow variables and the fast variables equilibrate. The equilibrium distribution of slow variables is already shown in Eq. (2.7). This allows us to construct a reduced canonical ensemble theory for the system alone, with H_X serving as the effective Hamiltonian. The equilibrium free energy of the system is defined as

$$F_X(\lambda, \lambda', \beta) = -T \ln Z_X(\lambda, \lambda', \beta), \quad (3.5)$$

where $Z_X(\lambda, \lambda', \beta)$ is given in Eq. (2.8). According to the discussion in Sec. II, the internal energy and entropy of the equilibrium system are

$$E_X(\lambda, \lambda', \beta) = \int_x p_X^{\text{EQ}}(\mathbf{x}) H_X(\mathbf{x}; \lambda, \lambda', \beta), \quad (3.6a)$$

$$S_X(\lambda, \lambda', \beta) = - \int_x p_X^{\text{EQ}}(\mathbf{x}) \ln p_X^{\text{EQ}}(\mathbf{x}), \quad (3.6b)$$

which are related to the equilibrium free energy (3.5) via

$$F_X(\lambda, \lambda', \beta) = E_X(\lambda, \lambda', \beta) - T S_X(\lambda, \lambda', \beta). \quad (3.6c)$$

Taking the logarithm of Eq. (2.16), and using Eqs. (3.1a) and (3.3), as well as Eq. (3.5), we obtain

$$F_{XY}(\lambda, \lambda', \beta) = F_X(\lambda, \lambda', \beta) + F_Y(\lambda', \beta). \quad (3.7)$$

Hence the free energy of the joint system is the sum of the free energies of the system and of the bath.

Inserting Eqs. (2.12) and (2.29) in Eqs. (3.1b) and (3.1c), and using Eqs. (3.6) and (III.4), we find the following decomposition of internal energy and equilibrium entropy:

$$E_{XY} = E_X + \int_x p_X^{\text{EQ}}(\mathbf{x}) E_Y(\mathbf{x}), \quad (3.8a)$$

$$S_{XY} = S_X + S_{Y|X}, \quad (3.8b)$$

where $S_{Y|X}$ is the average of Eq. (3.4b) over $p_X^{\text{EQ}}(\mathbf{x})$:

$$S_{Y|X} = \int_x p_X^{\text{EQ}}(\mathbf{x}) S_{Y|X=x} = - \int_x p_X^{\text{EQ}}(\mathbf{x}) \int_y p_{Y|X}^{\text{EQ}}(y|\mathbf{x}) \ln p_{Y|X}^{\text{EQ}}(y|\mathbf{x}), \quad (3.9)$$

which is the equilibrium version of Eq. (2.35c).

B. Differential thermodynamic relations

We consider infinitesimal changes of the parameters λ, λ', β . The differential of the joint free energy (3.1a) is

$$dF_{XY} = \frac{\partial F_{XY}}{\partial T} dT + \frac{\partial F_{XY}}{\partial \lambda} d\lambda + \frac{\partial F_{XY}}{\partial \lambda'} d\lambda'. \quad (3.10)$$

Using Eqs. (3.1a), (2.3), and (3.1c), we can calculate all three partial derivatives:

$$\frac{\partial F_{XY}}{\partial T} = -S_{XY}, \quad (3.11)$$

$$\frac{\partial F_{XY}}{\partial \lambda} = \int_{x,y} p_{XY}^{\text{EQ}} \frac{\partial H_{XY}}{\partial \lambda} \equiv \langle \langle \partial_\lambda H_{XY} \rangle \rangle^{\text{EQ}}, \quad (3.12)$$

$$\frac{\partial F_{XY}}{\partial \lambda'} = \int_{x,y} p_{XY}^{\text{EQ}} \frac{\partial H_{XY}}{\partial \lambda'} \equiv \langle \langle \partial_{\lambda'} H_{XY} \rangle \rangle^{\text{EQ}}, \quad (3.13)$$

where we have introduced the notation $\langle \langle \cdot \rangle \rangle^{\text{EQ}}$ to denote the average over the joint equilibrium pdf p_{XY}^{EQ} :

$$\langle \langle \cdot \rangle \rangle^{\text{EQ}} \equiv \int_{x,y} \cdot p_{XY}^{\text{EQ}}(\mathbf{x}, y). \quad (3.14)$$

As is well known in classical statistical mechanics, $(\partial_\lambda F_{XY})d\lambda$ and $(\partial_{\lambda'} F_{XY})d\lambda'$ may be understood as the differential reversible work done by the agents who control the system parameter and the bath parameter, respectively. (Here the term *reversible* pertains because the quasistatic transitions between equilibrium states are reversible.) Below we refer to these agents as the λ agent and λ' agent, respectively. Therefore we have the following expression for the differential reversible work on the joint system:

$$\begin{aligned} dW_{XY} &= \frac{\partial F_{XY}}{\partial \lambda} d\lambda + \frac{\partial F_{XY}}{\partial \lambda'} d\lambda' \\ &= \langle \langle d_\lambda H_{XY} \rangle \rangle^{\text{EQ}} + \langle \langle d_{\lambda'} H_{XY} \rangle \rangle^{\text{EQ}}, \end{aligned} \quad (3.15)$$

where we have introduced the notation

$$d_\lambda H_{XY} \equiv (\partial_\lambda H_{XY})d\lambda, \quad (3.16a)$$

$$d_{\lambda'} H_{XY} \equiv (\partial_{\lambda'} H_{XY})d\lambda'. \quad (3.16b)$$

We can then rewrite Eq. (3.10) as

$$dF_{XY} = -S_{XY}dT + \langle \langle d_\lambda H_{XY} \rangle \rangle^{\text{EQ}} + \langle \langle d_{\lambda'} H_{XY} \rangle \rangle^{\text{EQ}}. \quad (3.17)$$

Combining Eqs. (3.17) with (3.2c), we derive the differential form for the internal energy:

$$dE_{XY} = TdS_{XY} + \langle \langle d_\lambda H_{XY} \rangle \rangle^{\text{EQ}} + \langle \langle d_{\lambda'} H_{XY} \rangle \rangle^{\text{EQ}}. \quad (3.18)$$

This is nothing but the first law of thermodynamics, where the first term on the r.h.s. is the differential heat, whereas the other two terms are the differential work. Using Eqs. (3.1c) and (3.15), we can also rewrite heat and work as

$$dQ_{XY} = TdS_{XY} = \int_{xy} H_{XY} dp_{XY}^{\text{EQ}}, \quad (3.19a)$$

$$dW_{XY} = \int_{xy} p_{XY}^{\text{EQ}} (d_\lambda H_{XY} + d_{\lambda'} H_{XY}). \quad (3.19b)$$

Hence for weakly coupled systems and for quasistatic processes, heat is the change of internal energy due to the change of pdf, whereas work is the change of internal energy due to the change of Hamiltonian. We will see that such interpretations of heat and work remain applicable also for strongly coupled small systems, both for quasistatic processes and for nonequilibrium processes.

Let us now take the differential form of the free energy of the system, defined by Eq. (3.5). We obtain

$$dF_{\mathbf{X}} = \frac{\partial F_{\mathbf{X}}}{\partial T} dT + \frac{\partial F_{\mathbf{X}}}{\partial \lambda} d\lambda + \frac{\partial F_{\mathbf{X}}}{\partial \lambda'} d\lambda', \quad (3.20)$$

which describes quasistatic transition between equilibrium states of the system. Let us first calculate the first partial derivative $\partial F_{\mathbf{X}}/\partial T$. Using Eqs. (3.5), (2.8), and (3.6b), we find

$$\begin{aligned} \frac{\partial F_{\mathbf{X}}}{\partial T} dT &= -S_{\mathbf{X}} dT + \int_{\mathbf{x}} p_{\mathbf{X}}^{\text{EQ}}(\partial_{\beta} H_{\mathbf{X}}) d\beta \\ &= -S_{\mathbf{X}} dT + \langle \partial_{\beta} H_{\mathbf{X}} \rangle^{\text{EQ}} d\beta, \end{aligned} \quad (3.21)$$

where we have introduced the notation $\langle \cdot \rangle$ to denote the average over the pdf $p_{\mathbf{X}}^{\text{EQ}}$ of slow variables:

$$\langle \cdot \rangle^{\text{EQ}} \equiv \int_{\mathbf{x}} \cdot p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}). \quad (3.22)$$

In strong contrast with Eq. (3.11), here $-\partial_T F_{\mathbf{X}}$ is not the system entropy. The extra term in the r.h.s. of Eq. (3.21) arises due to the temperature dependence of $H_{\mathbf{X}}$, and hence is a unique signature of strongly coupled systems. Its physical meaning will be discussed shortly after. For weakly coupled systems, $H_{\mathbf{X}}$ is independent of β and this extra term vanishes identically. Alternatively, if the temperature is held fixed as assumed in all previous works of strong-coupling theories [1–15], the extra term on the r.h.s. of Eq. (3.21) also vanishes.

We can also calculate the partial derivatives of $F_{\mathbf{X}}$ with respect to λ and λ' and obtain

$$\frac{\partial F_{\mathbf{X}}}{\partial \lambda} = \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}}, \quad (3.23a)$$

$$\frac{\partial F_{\mathbf{X}}}{\partial \lambda'} = \langle \partial_{\lambda'} H_{\mathbf{X}} \rangle^{\text{EQ}}. \quad (3.23b)$$

Substituting Eqs. (3.21), (3.23a), and (3.23b) back into Eq. (3.20) we find

$$dF_{\mathbf{X}} = -S_{\mathbf{X}} dT + \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda + \langle \partial_{\lambda'} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda' + \langle \partial_{\beta} H_{\mathbf{X}} \rangle^{\text{EQ}} d\beta, \quad (3.24)$$

where $d_{\beta} H_{\mathbf{X}} \equiv \partial_{\beta} H_{\mathbf{X}} d\beta$ is the differential of $H_{\mathbf{X}}$ due to variation of temperature. Combining this with Eq. (3.6c) we derive the differential of the internal energy of the system:

$$dE_{\mathbf{X}} = T dS_{\mathbf{X}} + \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda + \langle \partial_{\lambda'} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda' + \langle \partial_{\beta} H_{\mathbf{X}} \rangle^{\text{EQ}} d\beta. \quad (3.25)$$

This must be identified with the first law of thermodynamics:

$$dE_{\mathbf{X}} = \delta Q_{\mathbf{X}} + \delta W_{\mathbf{X}}. \quad (3.26)$$

For quasistatic transitions between equilibrium states, we must have $T dS_{\mathbf{X}} = \delta Q_{\mathbf{X}}$. The remaining three terms on the r.h.s. of Eq. (3.25) are then the differential *reversible* work acting on the system:

$$dW_{\mathbf{X}} = \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda + \langle \partial_{\lambda'} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda' + \langle \partial_{\beta} H_{\mathbf{X}} \rangle^{\text{EQ}} d\beta. \quad (3.27)$$

It is clear that $\langle \partial_{\lambda} H_{\mathbf{X}} \rangle$ and $\langle \partial_{\lambda'} H_{\mathbf{X}} \rangle$ are respectively the work done by the λ agent and the λ' agent. Even though λ' is not directly coupled to the system variables, it does transmit energy to the system. This is obviously achieved through the interaction between the system and the bath. The last term on the r.h.s. of Eq. (3.27), $\langle \partial_{\beta} H_{\mathbf{X}} \rangle$, then must be understood as

the work done by the bath on the system, due to a change of temperature.

Using Eqs. (3.6b) and (3.27), we can rewrite heat and the reversible work into

$$\delta Q_{\mathbf{X}} = T dS_{\mathbf{X}} = \int_{\mathbf{x}} H_{\mathbf{X}} d p_{\mathbf{X}}^{\text{EQ}}, \quad (3.28a)$$

$$\delta W_{\mathbf{X}} = \int_{\mathbf{x}} p_{\mathbf{X}}^{\text{EQ}} (d_{\lambda} H_{\mathbf{X}} + d_{\lambda'} H_{\mathbf{X}} + d_{\beta} H_{\mathbf{X}}), \quad (3.28b)$$

which again shows that heat is the change of internal energy due to the change of pdf, whereas work is the change of internal energy due to the change of Hamiltonian.

We can obtain a similar differential relation for the conditional free energy of the environment, which is defined in Eq. (3.3). Note also that there is no contribution from the variation of λ and \mathbf{x} because, as we have shown earlier, $F_{\mathbf{Y}}(\lambda', \beta)$ does not depend on these parameters. We introduce the notation

$$\langle \cdot \rangle_{\mathbf{Y}}^{\text{EQ}} \equiv \int_{\mathbf{y}} \cdot p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x}), \quad (3.29)$$

to denote the average over the conditional equilibrium pdf (2.28) of the fast variables. Comparing this with Eqs. (3.14) and (3.22), we have

$$\langle \langle \cdot \rangle \rangle^{\text{EQ}} = \langle \langle \cdot \rangle_{\mathbf{Y}}^{\text{EQ}} \rangle^{\text{EQ}}. \quad (3.30)$$

Using Eq. (2.14) we find

$$\begin{aligned} dF_{\mathbf{Y}} &= -S_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} dT + \langle \partial_{\lambda} H_{\mathbf{Y}} \rangle_{\mathbf{Y}}^{\text{EQ}} d\lambda + \langle \partial_{\beta} H_{\mathbf{Y}} \rangle_{\mathbf{Y}}^{\text{EQ}} d\beta \\ &= -S_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} dT + \langle \partial_{\lambda} H_{\mathbf{Y}} \rangle_{\mathbf{Y}}^{\text{EQ}} d\lambda - d_{\beta} H_{\mathbf{X}}. \end{aligned} \quad (3.31)$$

In the second equality of Eq. (3.31) we have used $\partial_{\beta} H_{\mathbf{X}} = -\partial_{\beta} H_{\mathbf{Y}}$ which follows from the fact that the total Hamiltonian Eq. (2.12) is independent of β . Note that we need not average $d_{\beta} H_{\mathbf{X}}$ over \mathbf{y} because $H_{\mathbf{X}}$ is independent of \mathbf{y} .

We can further average Eq. (3.31) over the equilibrium pdf $p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x})$ of the slow variables and obtain

$$dF_{\mathbf{Y}} = -S_{\mathbf{Y}|\mathbf{X}} dT + \langle \langle \partial_{\lambda} H_{\mathbf{Y}} \rangle \rangle^{\text{EQ}} d\lambda - \langle \partial_{\beta} H_{\mathbf{X}} \rangle^{\text{EQ}} d\beta, \quad (3.32)$$

where we used Eq. (3.9). Since $\langle \langle \partial_{\lambda} H_{\mathbf{Y}} \rangle \rangle$ is the work done on the bath by the λ' agent, $-\langle \partial_{\beta} H_{\mathbf{X}} \rangle$ must be understood as the work done by the system on the bath. This is of course consistent with our earlier claim that $\langle \partial_{\beta} H_{\mathbf{X}} \rangle$ is the work done by the bath on the system. Put another way, $\langle \partial_{\beta} H_{\mathbf{X}} \rangle$ is a nondissipative exchange of energy between the system and the bath as the temperature is varied.

Summing up Eqs. (3.32) and (3.24), and using Eqs. (3.7), (3.8b), and (2.12), we obtain

$$dF_{\mathbf{X}\mathbf{Y}} = -S_{\mathbf{X}\mathbf{Y}} dT + \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}} d\lambda + \langle \langle \partial_{\lambda'} H_{\mathbf{X}\mathbf{Y}} \rangle \rangle^{\text{EQ}} d\lambda'. \quad (3.33)$$

Recalling Eqs. (2.12) and (2.17), we have

$$\langle \langle \partial_{\lambda} H_{\mathbf{X}\mathbf{Y}} \rangle \rangle^{\text{EQ}} = \langle \langle \partial_{\lambda} H_{\mathbf{X}} + \partial_{\lambda} H_{\mathbf{Y}} \rangle \rangle^{\text{EQ}} = \langle \partial_{\lambda} H_{\mathbf{X}} \rangle^{\text{EQ}}. \quad (3.34)$$

Hence, Eq. (3.33) is equivalent to Eq. (3.17), as expected.

Recall that $\langle \langle \partial_{\lambda} H_{\mathbf{X}\mathbf{Y}} \rangle \rangle^{\text{EQ}}$ is the work done by the λ agent on the joint system, whereas $\langle \langle \partial_{\lambda} H_{\mathbf{X}} \rangle \rangle^{\text{EQ}}$ and $\langle \langle \partial_{\lambda} H_{\mathbf{Y}} \rangle \rangle^{\text{EQ}}$ may be understood as the work done by the λ agent on the system and on the bath. Equation (3.34) then tells us that the λ agent does no work to the bath on average, regardless of the fact

that the bath Hamiltonian H_Y depends on λ . This is of course well expected since as we already know the bath free energy F_Y is independent of λ . In fact, we will see later in this work that the λ agent does no work to the bath even if the system is out of equilibrium. In contrast, generically both $\langle d_\lambda H_X \rangle^{\text{EQ}}$ and $\langle\langle d_\lambda H_Y \rangle\rangle^{\text{EQ}}$ are nonzero, which means that the λ' agent does work both to the system and to the bath. Finally we note that the work $\langle\langle d_\lambda H_Y \rangle\rangle^{\text{EQ}}$ done by the λ' agent to the bath, which is extensive in the size of the bath, is invisible in the thermodynamic theory of the system.

C. Connections and differences with classical equilibrium thermodynamics

Here we discuss the differences between our strong-coupling theory of equilibrium thermodynamics and the classical theories of equilibrium statistical mechanics as well as equilibrium thermodynamics.

We summarize Eqs. (2.7), (2.8), (3.5), and (3.6):

$$p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) = \frac{e^{-\beta H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)}}{Z_{\mathbf{X}}(\lambda, \lambda', \beta)}, \quad (3.35a)$$

$$Z_{\mathbf{X}}(\lambda, \lambda', \beta) = \int_{\mathbf{x}} e^{-\beta H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)}, \quad (3.35b)$$

$$F_{\mathbf{X}}(\lambda, \lambda', \beta) = -T \ln Z_{\mathbf{X}}(\lambda, \lambda', \beta), \quad (3.35c)$$

$$E_{\mathbf{X}}(\lambda, \lambda', \beta) = \int_{\mathbf{x}} p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta), \quad (3.35d)$$

$$S_{\mathbf{X}}(\lambda, \lambda', \beta) = - \int_{\mathbf{x}} p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) \ln p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}). \quad (3.35e)$$

These formulas, which can be found in every textbook on statistical mechanics, constitute the core of the classical canonical ensemble theory. Using these results, other thermodynamic quantities can be computed straightforwardly. Hence, at the fundamental level, the Gibbsian formalism of equilibrium statistical mechanics, which applies to large systems, also applies to small systems strongly coupled to their environments.

However, at the level of thermodynamics, there are important differences between a strongly coupled small system and a large system. These differences are mainly due to the temperature dependence of the HMF. First, we have already seen in Eq. (3.21) that the partial derivative of the free energy with respect to the temperature is not the same as the entropy. Likewise, from Eq. (3.25) we see that the partial derivative of the internal energy with respect to the entropy is not the same as the temperature. More explicitly we have

$$\frac{\partial F_{\mathbf{X}}}{\partial T} = -S_{\mathbf{X}} + \langle \partial_T H_{\mathbf{X}} \rangle^{\text{EQ}}, \quad (3.36)$$

$$\frac{\partial E_{\mathbf{X}}}{\partial S} = T + \left(\frac{\partial S_{\mathbf{X}}}{\partial T} \right)^{-1} \langle \partial_T H_{\mathbf{X}} \rangle^{\text{EQ}}. \quad (3.37)$$

The last equation can also be written as

$$\frac{\partial E_{\mathbf{X}}}{\partial T} = T \left(\frac{\partial S_{\mathbf{X}}}{\partial T} \right) + \langle \partial_T H_{\mathbf{X}} \rangle^{\text{EQ}} = C_{\hat{\lambda}} + \langle \partial_T H_{\mathbf{X}} \rangle^{\text{EQ}}. \quad (3.38)$$

That is, the specific heat is not the same as the partial derivative of the internal energy with respect to the temperature.

Furthermore, neither $\partial_T E_{\mathbf{X}}$ nor $C_{\hat{\lambda}}$ is proportional to the variance of energy fluctuation:

$$\frac{\partial E_{\mathbf{X}}}{\partial T} \neq \beta^2 [\langle H_{\mathbf{X}}^2 \rangle^{\text{EQ}} - (\langle H_{\mathbf{X}} \rangle^{\text{EQ}})^2], \quad (3.39)$$

$$C_{\hat{\lambda}} \neq \beta^2 [\langle H_{\mathbf{X}}^2 \rangle^{\text{EQ}} - (\langle H_{\mathbf{X}} \rangle^{\text{EQ}})^2]. \quad (3.40)$$

As a consequence, neither $\partial_T E_{\mathbf{X}}$ nor $C_{\hat{\lambda}}$ is guaranteed to be positive. [The stability of the strong-coupling thermodynamics is guaranteed by the concave nature of the nonequilibrium free energy, defined in Eq. (4.9) below.] In other words, the positivity of specific heat or $\partial_T E_{\mathbf{X}}$ can no longer be deemed as a prerequisite of stability. Last but not least, the analogs of relations (3.2a) and (3.2b) do not hold in our strong-coupling thermodynamics:

$$E_{\mathbf{X}} \neq \frac{\partial \beta F_{\mathbf{X}}}{\partial \beta}, \quad S_{\mathbf{X}} \neq \beta^2 \frac{\partial F_{\mathbf{X}}}{\partial \beta}. \quad (3.41)$$

There are many other thermodynamic relations that hold in classical thermodynamics but not in our strong-coupling thermodynamics. Detailed studies of these relations are better carried out in the setting of concrete experimental systems, which are reserved for future studies. We note that in all the above partial derivatives, λ and λ' are held fixed.

There are other ways to define various thermodynamic quantities of strong-coupling thermodynamics, so that some thermodynamic relations can be restored back to the original forms in classical thermodynamics. The cost to pay is to replace Eqs. (3.35d) and (3.35e) by different expressions. Furthermore, the theory of stochastic thermodynamics becomes much more complicated. Since statistical mechanics are more fundamental than equilibrium thermodynamics, we believe it is not worth saving some thermodynamic relations at the cost of making statistical mechanics more complicated.

Other important differences also arise due to the smallness of the system being studied. In classical equilibrium thermodynamics, there are a large number of Legendre transformations which can be used to transform one thermodynamic potential into another. All these potentials, such as energy, Helmholtz free energy, Gibbs free energy, enthalpy, or grand potential, are equivalent in the sense that any of them can be used to derive thermodynamic functions and equations of states. It is well understood that the fundamental reason for this equivalence of thermodynamic potentials is the equivalence of statistical ensembles, which in turn follows from the largeness of system size and negligibility of fluctuations. In strong-coupling thermodynamics of small systems, however, fluctuations are certainly not negligible, and different statistical ensembles are not equivalent. Different thermodynamic potentials are then not related by Legendre transformation as in classical thermodynamics. Experimental setup defines the specific statistical ensemble that should be used in theoretical study.

IV. STRONGLY COUPLED STOCHASTIC THERMODYNAMICS

We now promote the strong-coupling thermodynamic theory to the nonequilibrium level. A nonequilibrium state of the joint system at the ensemble level is characterized by the joint pdf $p_{\mathbf{X}\mathbf{Y}}(\mathbf{x}, \mathbf{y})$. We consider nonequilibrium processes where

λ , λ' , β are varied according to some externally controlled protocol, which is assumed to be much slower than the dynamics of bath variables. The evolution of the joint system is described either by microscopic unitary dynamics, or a Markov dynamics, depending on whether it is coupled to a superbath. Explicit analysis using projection operator theory [22] then shows that, in the limit of TSS, and as long as we only concern timescales much longer than that of the fast dynamics, the joint pdf can be decomposition as Eq. (2.33), and the pdf of system variables $p_{\mathbf{X}}$ evolves according to a Markov process, specified either by a nonlinear Ito-Langevin equation:

$$dx^k + (L^{kj}\partial_j U - \partial_j L^{kj})dt = b^{k\alpha} dW_\alpha, \quad (4.1)$$

or by the corresponding Fokker-Planck equation [26]

$$\partial_t p_{\mathbf{X}} = \partial_i L^{ij}[\partial_j + (\partial_j U)]p_{\mathbf{X}}, \quad (4.2)$$

where $L^{ij}(\mathbf{x}; \hat{\lambda})$ is the matrix of kinetic coefficients, and $U(\mathbf{x}; \hat{\lambda})$ is the generalized potential, related to the HMF and the free energy of the system via

$$U(\mathbf{x}; \hat{\lambda}) = \beta[H_{\mathbf{X}}(\mathbf{x}; \hat{\lambda}) - F_{\mathbf{X}}(\hat{\lambda})]. \quad (4.3)$$

Here we use $\hat{\lambda}$ to denote the collection of the system parameter λ , the bath parameter λ' , and the inverse temperature β :

$$\hat{\lambda} \equiv (\lambda, \lambda', \beta), \quad (4.4)$$

and call $\hat{\lambda}$ the *parameters*. For fixed $\hat{\lambda}$, the system then converges to an equilibrium state (2.7), which can be rewritten as

$$p_{\mathbf{X}}^{\text{EQ}}(\mathbf{x}) = e^{-U(\mathbf{x}; \hat{\lambda})}. \quad (4.5)$$

It was further shown in Ref. [22] that the kinetic matrix $L^{ij}(\mathbf{x}; \hat{\lambda})$ can be expressed in terms of correlation functions of fast variables, conditioned on the value of the slow variables. The time-reversal symmetry of equilibrium states demands that the Langevin equation (4.1) satisfies the conditions of detailed balance, which are given in Eqs. (4.22). These conditions are equivalent to the *fluctuation-dissipation relation of the second kind*, as called by Kubo [27].

Our purpose is to construct a consistent theory of stochastic thermodynamics for the system, which involves the pdf $p_{\mathbf{X}}(\mathbf{x}, t)$, the HMF $H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta)$, as well as the kinetic matrix $L^{ij}(\mathbf{x}; \lambda, \lambda', \beta)$ but does not involve the bath variables. The entropy production computed in this theory equals the increase of the total entropy of the universe, assuming that the joint system can be described by a weak-coupling theory of stochastic thermodynamics. We will also see that our strong-coupling theory exhibits a few novel features, again due to the temperature dependence of the HMF.

A. Internal energy and entropy

In analogy with the notations defined in Eqs. (3.14) and (3.22), we define the following notations for averaging over nonequilibrium distributions:

$$\langle\langle \cdot \rangle\rangle \equiv \int_{\mathbf{x}, \mathbf{y}} \cdot p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}(\mathbf{y}|\mathbf{x}), \quad (4.6)$$

$$\langle \cdot \rangle \equiv \int_{\mathbf{x}} \cdot p_{\mathbf{X}}(\mathbf{x}). \quad (4.7)$$

Recall that in Sec. II we defined the system Hamiltonian as the HMF, see Eq. (2.5). The value of $H_{\mathbf{X}}$ is called the *fluctuating internal energy* of the system, using the terminology in Refs. [4,23]. The nonequilibrium internal energy is defined as the ensemble average of $H_{\mathbf{X}}$:

$$E_{\mathbf{X}}[p_{\mathbf{X}}] \equiv \int_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) H_{\mathbf{X}}(\mathbf{x}; \lambda, \lambda', \beta) = \langle H_{\mathbf{X}} \rangle. \quad (4.8)$$

Strictly speaking, the functional $E_{\mathbf{X}}[p_{\mathbf{X}}]$ depends also on the parameters λ , λ' , β . But to avoid cluttering, we shall hide these dependencies. The nonequilibrium entropy is already defined in Eq. (2.35b). The nonequilibrium free energy of the system is then defined in the standard way:

$$F_{\mathbf{X}}[p_{\mathbf{X}}] \equiv E_{\mathbf{X}}[p_{\mathbf{X}}] - TS_{\mathbf{X}}[p_{\mathbf{X}}] = \langle H_{\mathbf{X}} + T \ln p_{\mathbf{X}} \rangle, \quad (4.9)$$

which turns out to be the same as the free energy defined in several previous theories [3,4,6,11]. $F_{\mathbf{X}}[p_{\mathbf{X}}]$ is minimized by the equilibrium pdf Eq. (2.7). As we show below, in the limit of TSS, $F_{\mathbf{X}}[p_{\mathbf{X}}]$ is also invariant under coarse-graining up to an additive constant.

These nonequilibrium entropy, energy, and free energy are in fact formally identical to those in weak-coupling theory, with $H_{\mathbf{X}}$ understood as the system Hamiltonian. For an equilibrium state $p_{\mathbf{X}} = p_{\mathbf{X}}^{\text{EQ}}$, these thermodynamic variables reduce to their equilibrium counterparts, Eqs. (3.6a), (3.6b), and (3.5), respectively.

B. Nonequilibrium work and heat

Consider an infinitesimal trajectory where the system state evolves from \mathbf{x} to $\mathbf{x} + d\mathbf{x}$, while the parameters λ , λ' , β change, respectively, to $d\lambda$, $d\lambda'$, $d\beta$. Inspired by the form of equilibrium work, Eq. (3.27), we define the differential nonequilibrium work at the trajectory level as

$$d\mathcal{W}_{\mathbf{X}} \equiv d_{\lambda} H_{\mathbf{X}} + d_{\lambda'} H_{\mathbf{X}} + d_{\beta} H_{\mathbf{X}}. \quad (4.10)$$

The nonequilibrium work at the ensemble level is then the ensemble average of $d\mathcal{W}_{\mathbf{X}}$:

$$\begin{aligned} dW_{\mathbf{X}} &\equiv \int_{\mathbf{x}} p_{\mathbf{X}}(d_{\lambda} H_{\mathbf{X}} + d_{\lambda'} H_{\mathbf{X}} + d_{\beta} H_{\mathbf{X}}) \\ &= \langle d_{\lambda} H_{\mathbf{X}} \rangle + \langle d_{\lambda'} H_{\mathbf{X}} \rangle + \langle d_{\beta} H_{\mathbf{X}} \rangle. \end{aligned} \quad (4.11)$$

Similar to the reversible work (3.28b), Eq. (4.10) and (4.11) both contain three parts, which are respectively the works done by the λ agent, the λ' agent, and the bath. If the process is quasistatic, Eqs. (4.10) and (4.11) reduce to the reversible work (3.27). These definitions of work reduce to those in Ref. [15] if $d\beta = d\lambda' = 0$. Finally, using the shorthand (4.4), we may also rewrite Eqs. (4.10) and (4.11) as

$$d\mathcal{W}_{\mathbf{X}} = d_{\hat{\lambda}} H_{\mathbf{X}}, \quad (4.12)$$

$$dW_{\mathbf{X}} = \langle d_{\hat{\lambda}} H_{\mathbf{X}} \rangle. \quad (4.13)$$

The differential heat at the trajectory level and at the ensemble level are then defined as

$$d\mathcal{Q}_{\mathbf{X}} \equiv d_{\mathbf{x}} H_{\mathbf{X}}, \quad (4.14)$$

$$dQ_{\mathbf{X}} \equiv \langle\langle d\mathcal{Q}_{\mathbf{X}} \rangle\rangle, \quad (4.15)$$

where $\langle\langle \cdot \rangle\rangle$ means average over both \mathbf{X} and \mathbf{Y} . An alternative, more intuitive, expression for heat at the ensemble level is

$$dQ_{\mathbf{X}} = \int_{\mathbf{x}} H_{\mathbf{X}} d p_{\mathbf{X}}, \quad (4.16)$$

where $d p_{\mathbf{X}}$ is the change of pdf $p_{\mathbf{X}}$ during the infinitesimal process. Equivalence of Eqs. (4.15) and (4.16) can be explicitly proved using Langevin dynamics and the associated Fokker-Planck dynamics [22].

Note that Eqs. (4.11) and (4.16) are formally identical to Eqs. (3.28b) and (3.28a). Hence, as we claimed earlier, even for nonequilibrium processes, work and heat are respectively the change of internal energy of the system due to the changes of the pdf and of the Hamiltonian of the system.

The first law of thermodynamics at the trajectory level follows directly from the definitions (4.10) and (4.14):

$$dH_{\mathbf{X}} = d\mathcal{W}_{\mathbf{X}} + dQ_{\mathbf{X}}, \quad (4.17)$$

where the left-hand side (l.h.s.) is the differential of the fluctuating internal energy.

The first law at the ensemble level can be obtained either by taking the ensemble average of Eq. (4.17), or taking the differential of Eq. (4.8). Let us take the latter route. The differential is due to changes in λ , λ' , β and the evolution of $p_{\mathbf{X}}$, with \mathbf{x} behaving as a dummy variable. Hence we obtain

$$\begin{aligned} dE_{\mathbf{X}}[p_{\mathbf{X}}] &= \int_{\mathbf{x}} [(d_{\lambda} H_{\mathbf{X}} + d_{\lambda'} H_{\mathbf{X}} + d_{\beta} H_{\mathbf{X}}) p_{\mathbf{X}} + H_{\mathbf{X}} d p_{\mathbf{X}}] \\ &= \int_{\mathbf{x}} (p_{\mathbf{X}} d_{\lambda} H_{\mathbf{X}} + H_{\mathbf{X}} d p_{\mathbf{X}}) \\ &= dW_{\mathbf{X}} + dQ_{\mathbf{X}}, \end{aligned} \quad (4.18)$$

which is the first law at the ensemble level.

Let us take the differential of the nonequilibrium free energy (4.9) to obtain

$$dF_{\mathbf{X}} = dW_{\mathbf{X}} + dQ_{\mathbf{X}} - T dS_{\mathbf{X}} - S_{\mathbf{X}} dT. \quad (4.19)$$

This can be rewritten into

$$dS^{\text{tot}} = dS_{\mathbf{X}} - \beta dQ_{\mathbf{X}} = \beta(dW_{\mathbf{X}} - dF_{\mathbf{X}} - S_{\mathbf{X}} dT). \quad (4.20)$$

It was proved in Ref. [26] that, with the system dynamics described by Eqs. (4.1) and (4.2), and with entropy and heat defined by Eqs. (2.35b) and (4.16), the Clausius inequality always holds:

$$\begin{aligned} dS^{\text{tot}} &= dS_{\mathbf{X}} - \beta dQ_{\mathbf{X}} \\ &= \int_{\mathbf{x}} [\partial_k (\ln p_{\mathbf{X}} + U)] B^{kj} p [\partial_j (\ln p_{\mathbf{X}} + U)] \\ &\geq 0, \end{aligned} \quad (4.21)$$

where B^{ij} is the symmetric part of L^{ij} and is always positive definite. This is the second law of thermodynamics for a small system strongly coupled to a single heat bath.

Note that if the process is quasistatic, $p_{\mathbf{X}}$ is given by Eq. (2.7) and $d p_{\mathbf{X}}$ is solely due to changes of λ , λ' , and β . We can then explicitly calculate Eq. (4.16) and verify $dQ_{\mathbf{X}} = T dS_{\mathbf{X}}$. The change of the total entropy, as given by Eq. (4.20), then vanishes identically. Also for quasistatic processes, Eq. (4.18) reduces to Eq. (3.25), and Eq. (4.19) reduces to Eq. (3.24). All these are of course completely expected.

C. Fluctuation theorems

It is assumed that, for fixed parameters $\hat{\lambda} = (\lambda, \lambda', \beta)$, the system converges to a thermal equilibrium state which has time-reversal symmetry. It was shown in Ref. [26] that this implies the following detailed balance properties for the Langevin dynamics (4.1):

$$U(\mathbf{x}^*, \hat{\lambda}^*) = U(\mathbf{x}, \hat{\lambda}), \quad (4.22a)$$

$$\varepsilon_i L^{ij}(\mathbf{x}^*, \hat{\lambda}^*) \varepsilon_j = L^{ji}(\mathbf{x}, \hat{\lambda}), \quad (4.22b)$$

$$\int_{\mathbf{x}} e^{-U(\mathbf{x}, \hat{\lambda})} = 1, \quad (4.22c)$$

where $U(\mathbf{x}, \hat{\lambda})$ is defined in Eq. (4.3), and $\hat{\lambda}^* = (\lambda^*, (\lambda')^*, \beta)$ is the time-reversal of $\hat{\lambda}$. (Note that temperature is even in time, hence $\beta^* = \beta$.)

Using Eqs. (4.22), and assuming that both the temperature and the bath parameter λ' are fixed, various fluctuation theorems were proved for processes starting from and eventually relaxing back to equilibrium states [26]. These theorems can be formulated solely in terms of statistical properties of the total work defined along dynamic trajectories.

We now adapt the proof of fluctuation theorems in Ref. [26] to processes where all parameters $\hat{\lambda} = (\lambda, \lambda', \beta)$ vary with time. It was shown in Eq. (A27) of Appendix A of Ref. [26] that the detailed balance properties (4.22) implies the following symmetry for short-time transition probabilities of the Langevin dynamics (4.1):

$$\ln \frac{P_{\hat{\lambda}}(\mathbf{x} + d\mathbf{x} | \mathbf{x}; dt)}{P_{\hat{\lambda}^*}(\mathbf{x}^* + d\mathbf{x}^* | \mathbf{x}^*; dt)} = -d_{\mathbf{x}} U(\mathbf{x}; \hat{\lambda}), \quad (4.23)$$

where the subscripts $\hat{\lambda}$ and $\hat{\lambda}^*$ specify the dynamic processes. Hence $P_{\hat{\lambda}}(\mathbf{x} + d\mathbf{x} | \mathbf{x}; dt)$ is the probability density that the system goes from \mathbf{x} to $\mathbf{x} + d\mathbf{x}$ in the *forward process* with the fixed parameters $\hat{\lambda}$, whereas $P_{\hat{\lambda}^*}(\mathbf{x}^* + d\mathbf{x}^* | \mathbf{x}^*; dt)$ is the probability density that the system goes from $\mathbf{x}^* + d\mathbf{x}^*$ to \mathbf{x}^* in the *backward process* with the fixed parameters $\hat{\lambda}^*$.

Recalling Eqs. (4.3) and (4.14) we see that the r.h.s. of Eq. (4.23) is related to the heat via

$$-d_{\mathbf{x}} U(\mathbf{x}; \hat{\lambda}) = -\beta d_{\mathbf{x}} H_{\mathbf{X}}(\mathbf{x}; \hat{\lambda}) = -\beta dQ_{\mathbf{X}}, \quad (4.24)$$

which may be interpreted as the change of the environmental entropy during the infinitesimal process.

We now consider a *forward process* of finite duration $0 \leq t \leq \tau$, with time-dependent control parameters and temperature $\hat{\lambda}(t)$, which shall be called a *dynamic protocol*. The system starts at time $t = 0$ from an equilibrium state with initial parameters $\hat{\lambda}(0)$:

$$p_F(\mathbf{x}, 0) = e^{-U(\mathbf{x}; \hat{\lambda}(0))}. \quad (4.25)$$

The *backward process* is defined by the *backward protocol* $\tilde{\hat{\lambda}}(t) = \hat{\lambda}^*(\tau - t) = (\lambda^*(\tau - t), [\lambda'(\tau - t)]^*, \beta(\tau - t))$, and initial parameters $\hat{\lambda}^*(\tau) = (\lambda^*(\tau), [\lambda'(\tau)]^*, \beta(\tau))$. Furthermore, the initial state of the backward process is given by the equilibrium state associated with $\hat{\lambda}^*(\tau)$:

$$p_B(\mathbf{x}, 0) = e^{-U(\mathbf{x}; \hat{\lambda}^*(\tau))}. \quad (4.26)$$

Consider now a *forward trajectory* $\gamma = \mathbf{x}(t)$ in the forward process and the corresponding *backward trajectory*

$\tilde{\gamma} = \mathbf{x}^*(\tau - t)$ in the backward process. We may break both trajectories into a large number of infinitesimal segments and apply Eq. (4.23) to each pair of segments. Summing all these relations up we obtain

$$\ln \frac{P_F[\gamma|\gamma_0]}{P_B[\tilde{\gamma}|\tilde{\gamma}_0]} = - \int_{\gamma} d_x U(\mathbf{x}(t); \hat{\lambda}(t)), \quad (4.27)$$

where $\gamma_0 = \mathbf{x}(0)$ and $\tilde{\gamma}_0 = \mathbf{x}^*(\tau)$ denote respectively the initial states of the forward and backward trajectories, while $P_F[\gamma|\gamma_0]$ and $P_B[\tilde{\gamma}|\tilde{\gamma}_0]$ are respectively the conditional probability density functions of the forward (backward) trajectories in the forward (backward) processes, given their respective initial states. Further using Eqs. (4.25) and (4.26) we may form the ratio of the unconditional probabilities of trajectories:

$$\frac{P_F[\gamma]}{P_B[\tilde{\gamma}]} = \frac{P_F[\gamma|\gamma_0]}{P_B[\tilde{\gamma}|\tilde{\gamma}_0]} \frac{e^{-U(\mathbf{x}(0); \hat{\lambda}(0))}}{e^{-U(\mathbf{x}^*(\tau); \hat{\lambda}^*(\tau))}}. \quad (4.28)$$

The first ratio on the r.h.s. can be replaced using Eq. (4.27), whereas using the detailed balance (4.22a) the second ratio can be replaced by

$$e^{U(\mathbf{x}(\tau); \hat{\lambda}(\tau)) - U(\mathbf{x}(0); \hat{\lambda}(0))} = e^{\int_{\gamma} dU}, \quad (4.29)$$

where $dU = d_x U + d_{\hat{\lambda}} U$ is the complete differential of the generalized potential $U(\mathbf{x}; \hat{\lambda})$. Consequently, Eq. (4.28) can be rewritten as

$$\ln \frac{P_F[\gamma]}{P_B[\tilde{\gamma}]} = - \int_{\gamma} d_x U + \int_{\gamma} dU. \quad (4.30)$$

While the first term on the r.h.s. is the change of environmental entropy along the trajectory [cf. Eq. (4.24)], the second term is the change of the stochastic entropy along the forward trajectory of the forward process, with the understanding that the system is in equilibrium both in the initial time and in the final time of the process. Hence Eq. (4.30) may be understood as the entropy production $\Sigma_F[\gamma]$ along the forward trajectory of the forward process. Since the forward and backward trajectories and processes are related to each other by time reversal, Eq. (4.30) may also be understood as negative the entropy production $\Sigma_B[\tilde{\gamma}]$ along the backward trajectory of the backward process. Hence we arrive at

$$\Sigma_F[\gamma] = -\Sigma_B[\tilde{\gamma}] = \int_{\gamma} d_{\hat{\lambda}} U = \int_{\gamma} (d_{\lambda} U + d_{\lambda'} U + d_{\beta} U). \quad (4.31)$$

Further repeating the proof in Sec. IV E of Ref. [26], we may prove the following fluctuation theorems for the pdfs of entropy production in the forward and backward processes:

$$p_F(\sigma) = e^{\sigma} p_B(-\sigma), \quad (4.32a)$$

$$\langle e^{-\sigma} \rangle_F = 1. \quad (4.32b)$$

To make contact between the entropy production and the nonequilibrium work previously defined, we use Eqs. (4.3) and (4.10) to rewrite Eq. (4.31) into

$$\begin{aligned} \Sigma_F[\gamma] &= \int_{\gamma} \beta d\mathcal{W}_{\mathbf{X}} + \int_{\gamma} H_{\mathbf{X}} d\beta \\ &+ \beta(0)F_{\mathbf{X}}(\hat{\lambda}(0)) - \beta(\tau)F_{\mathbf{X}}(\hat{\lambda}(\tau)). \end{aligned} \quad (4.33)$$

If the temperature is fixed along the process, then Eq. (4.33) reduces to

$$\Sigma_F[\gamma] = \beta \mathcal{W}_{\mathbf{X}}[\gamma] - \beta \Delta F_{\mathbf{X}}, \quad (4.34)$$

and Eqs. (IV.32) reduce to the familiar Crooks fluctuation theorem and Jarzynski equality:

$$p_F(\mathcal{W}_{\mathbf{X}}) = e^{\beta(\mathcal{W}_{\mathbf{X}} - \Delta F_{\mathbf{X}})} p_B(-\mathcal{W}_{\mathbf{X}}), \quad (4.35a)$$

$$\langle e^{-\beta \mathcal{W}_{\mathbf{X}}} \rangle_F = e^{-\beta \Delta F_{\mathbf{X}}}, \quad (4.35b)$$

where $\langle \cdot \rangle_F$ means average over trajectories in the forward process. If, however, the temperature is varied during the process, the entropy production as given by Eq. (4.33) cannot be expressed in terms of work alone. The fluctuation theorems (IV.32) then are genuinely different from the Crooks fluctuation theorem and Jarzynski equality.

D. Coarse-graining and physical meanings of work and heat

For the moment we assume that the joint system \mathbf{XY} is weakly coupled with a superbath whose dynamics are even faster than that of \mathbf{Y} . The statistical physics of the joint system is then described by the weakly coupled stochastic thermodynamics, with $H_{\mathbf{XY}}$ being the Hamiltonian. Below we refer to this theory as the *fine-grained theory*, whereas the strong-coupling stochastic thermodynamics for the slow variables will be referred to as the *coarse-grained theory*. With the aid of TSS, we discuss the connection between these two theories of stochastic thermodynamics and thereby clarify the physical meanings of work and heat in the strong-coupling theory.

The thermodynamic quantities of the fine-grained theory are defined as

$$E_{\mathbf{XY}}[p_{\mathbf{XY}}] \equiv \int_{\mathbf{XY}} p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) H_{\mathbf{XY}}(\lambda, \lambda'), \quad (4.36a)$$

$$S_{\mathbf{XY}}[p_{\mathbf{XY}}] \equiv - \int_{\mathbf{XY}} p_{\mathbf{XY}} \ln p_{\mathbf{XY}}, \quad (4.36b)$$

$$\begin{aligned} F_{\mathbf{XY}}[p_{\mathbf{XY}}] &\equiv E_{\mathbf{XY}}[p_{\mathbf{XY}}] - T S_{\mathbf{XY}}[p_{\mathbf{XY}}] \\ &= \int_{\mathbf{x}} p_{\mathbf{XY}} (H_{\mathbf{XY}} + T \ln p_{\mathbf{XY}}). \end{aligned} \quad (4.36c)$$

Using Eqs. (2.12) and (2.33), they may be rewritten as

$$E_{\mathbf{XY}}[p_{\mathbf{XY}}] = E_{\mathbf{X}}[p_{\mathbf{X}}] + \int_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) E_{\mathbf{Y}}(\mathbf{x}), \quad (4.37a)$$

$$S_{\mathbf{XY}}[p_{\mathbf{XY}}] = S_{\mathbf{X}}[p_{\mathbf{X}}] + S_{\mathbf{Y}|\mathbf{X}}[p_{\mathbf{X}} p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}], \quad (4.37b)$$

$$F_{\mathbf{XY}}[p_{\mathbf{XY}}] = F_{\mathbf{X}}[p_{\mathbf{X}}] + F_{\mathbf{Y}}(\lambda', \beta), \quad (4.37c)$$

where $E_{\mathbf{Y}}(\mathbf{x})$ is defined in Eq. (3.4a), $S_{\mathbf{Y}|\mathbf{X}}$ is the conditional entropy of \mathbf{Y} given \mathbf{X} :

$$S_{\mathbf{Y}|\mathbf{X}}[p_{\mathbf{X}} p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}] = - \int_{\mathbf{xy}} p_{\mathbf{X}} p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}} \ln p_{\mathbf{Y}|\mathbf{X}}^{\text{EQ}}, \quad (4.38)$$

whereas $F_{\mathbf{Y}}(\lambda', \beta)$ is the bath free energy defined in Eq. (3.3).

Equation (4.37c) says that the nonequilibrium free energy $F_{\mathbf{XY}}[p_{\mathbf{XY}}]$ of the fine-grained theory differs from that of the coarse-grained theory, $F_{\mathbf{X}}[p_{\mathbf{X}}]$, only by an additive constant $F_{\mathbf{Y}}(\lambda', \beta)$ that is independent of $p_{\mathbf{X}}$. In other words, the nonequilibrium free energy is invariant under coarse-graining in the limit of TSS, up to an additive constant. This partially

explains why the nonequilibrium free energy plays such an important role in stochastic thermodynamics.

The work and heat of the fine-grained theory are defined using the standard theory of stochastic thermodynamics. At the trajectory level, we have

$$d\mathcal{W}_{\text{XY}} \equiv d_\lambda H_{\text{XY}} + d_{\lambda'} H_{\text{XY}}, \quad (4.39)$$

$$dQ_{\text{XY}} \equiv d_{xy} H_{\text{XY}}. \quad (4.40)$$

Note that there is no term $d_\beta H_{\text{XY}}$ in the work, because H_{XY} is independent of β . At the ensemble level, we have

$$dW_{\text{XY}} \equiv \int_{xy} p_{\text{XY}}(d_\lambda H_{\text{XY}} + d_{\lambda'} H_{\text{XY}}), \quad (4.41)$$

$$dQ_{\text{XY}} \equiv \int_{xy} H_{\text{XY}} d p_{\text{XY}}. \quad (4.42)$$

Evidently, the two terms in the works $d\mathcal{W}_{\text{XY}}$ and dW_{XY} are due to the λ agent and the λ' agent, respectively. As a well-known feature of weak-coupling stochastic thermodynamics, the work in the fine-grained theory is the change of the total energy of the universe, assuming that there is no parameters other than λ, λ' varying over time.

We now establish the connection between the work defined in the fine-grained theory and that in the coarse-grained theory. First using Eq. (2.12), we have

$$d_\lambda H_{\text{XY}} = d_\lambda H_{\text{X}} + d_\lambda H_{\text{Y}}. \quad (4.43)$$

The l.h.s. is the work done to the joint system by the λ agent at the trajectory level, whereas the two terms on the r.h.s. are respectively the work done to the system and to the bath by the λ agent. Note that $d_\lambda H_{\text{Y}}$ is generically nonzero. Yet if we average $d_\lambda H_{\text{Y}}$ over the conditional equilibrium pdf (2.28) of the fast variables, we obtain identically zero due to Eq. (2.17). That is, the average work done by the λ agent to the bath vanishes, even though the work along a particular trajectory may not vanish.

Now, consider a time interval Δt satisfying $\tau_{\text{X}} \gg \Delta t \gg \tau_{\text{Y}}$. Time-average of physical quantities within this interval should be equivalent to the ensemble average over the conditional equilibrium pdf (2.28), at least for a typical trajectory. This means that, for a typical trajectory, $d_\lambda H_{\text{Y}}$ is a rapidly oscillating function with the typical timescale set by τ_{Y} , such that its time-average becomes vanishingly small if $\Delta t/\tau_{\text{Y}} \gg 1$. Hence, in the limit of TSS, and for typical trajectories, the λ agent does no work on the bath.

The story is different for the work done by the λ' agent, which can also be decomposed into two parts:

$$d_{\lambda'} H_{\text{XY}} = d_{\lambda'} H_{\text{X}} + d_{\lambda'} H_{\text{Y}}. \quad (4.44)$$

Here, neither term on the r.h.s. vanishes, even if averaged over the fast variables. Hence the λ' agent generically does work both to the system and to the bath.

Using Eqs. (4.43) and (4.44), together with the fact that $H_{\text{XY}} = H_{\text{X}} + H_{\text{Y}}$ is independent of β , we may rewrite Eq. (4.39) as

$$\begin{aligned} d\mathcal{W}_{\text{XY}} &= d_\lambda H_{\text{X}} + d_{\lambda'} H_{\text{X}} + d_\beta H_{\text{X}} + d_\lambda H_{\text{Y}} + d_{\lambda'} H_{\text{Y}} + d_\beta H_{\text{Y}} \\ &= d\mathcal{W}_{\text{X}} + d\mathcal{W}_{\text{Y}}, \end{aligned} \quad (4.45)$$

where $d\mathcal{W}_{\text{X}}$ is the work in the coarse-grained theory as given by Eq. (4.10), and $d\mathcal{W}_{\text{Y}}$ is the work on the bath at the trajectory level:

$$d\mathcal{W}_{\text{Y}} = d_\lambda H_{\text{Y}} + d_{\lambda'} H_{\text{Y}} + d_\beta H_{\text{Y}}. \quad (4.46)$$

Among all three terms on the r.h.s., the first term $d_\lambda H_{\text{Y}}$ is a rapidly oscillating function and averages to zero, as we have already shown above, whereas $d_{\lambda'} H_{\text{Y}}$ is generically non-vanishing, and is extensive in the bath size. The last term $d_\beta H_{\text{Y}} = -d_\beta H_{\text{X}}$ represents the work done by the system due to variations of the temperature. We emphasize that $d\mathcal{W}_{\text{Y}}$ is generically nonzero. Hence the work $d\mathcal{W}_{\text{X}}$ in the coarse-grained theory is not the same as the work $d\mathcal{W}_{\text{XY}}$ in the fine-grained theory.

In most previous theories on strong-coupling stochastic thermodynamics, it is always assumed that the bare interaction Hamiltonian H_I^0 remains fixed, and the work is always defined as the change of the total energy. Since H_{Y}^0 is independent of λ , it can be seen easily from Eqs. (2.6) and (2.1) that if H_I^0 is independent of λ ,

$$d_\lambda H_{\text{X}} = d_\lambda H_{\text{X}}^0 = d_\lambda H_{\text{XY}}. \quad (4.47)$$

Hence if both λ and β are also fixed, we have

$$d\mathcal{W}_{\text{X}} = d_\lambda H_{\text{X}} = d_\lambda H_{\text{XY}}, \quad (4.48)$$

which means that the work defined in our theory (the l.h.s.) agrees with the work defined in the previous theories (the r.h.s.). However, if the bare interaction Hamiltonian H_I^0 depends on λ , $d_\lambda H_{\text{XY}} = d_\lambda H_{\text{X}}^0(\mathbf{x}; \lambda) + d_\lambda H_I^0(\mathbf{x}, \mathbf{y}; \lambda)$ generically involves both \mathbf{x} and \mathbf{y} . Definition of work as $d_\lambda H_{\text{XY}}$ then fails to yield a strong-coupling theory involving only system variables. For this reason, the previous strong-coupling theories cannot be used to study small systems with variable coupling to their environments.

We can similarly establish the connection between heat in the fine-grained theory and that in the coarse-grained theory. We only do it at the ensemble level. To further simplify the matter, we also assume that the bath parameter and temperature remain fixed, whereas the system parameter λ is varied. First we take the differential of Eq. (2.33) and obtain

$$d p_{\text{XY}}(\mathbf{x}, \mathbf{y}) = d p_{\text{X}} p_{\text{Y}|\text{X}}^{\text{EQ}} + p_{\text{X}} d p_{\text{Y}|\text{X}}^{\text{EQ}}, \quad (4.49)$$

where $d p_{\text{Y}|\text{X}}^{\text{EQ}}$ is due to the change of λ , while $d p_{\text{X}}(\mathbf{x})$ arises due to the time evolution of p_{X} . Let us use this and Eq. (2.12) to rewrite Eq. (4.42) into

$$\begin{aligned} dQ_{\text{XY}} &= \int_{xy} (H_{\text{X}} + H_{\text{Y}})(d p_{\text{X}} p_{\text{Y}|\text{X}}^{\text{EQ}} + p_{\text{X}} d p_{\text{Y}|\text{X}}^{\text{EQ}}) \\ &= \int_{\text{X}} H_{\text{X}} d p_{\text{X}} \int_{\text{Y}} p_{\text{Y}|\text{X}}^{\text{EQ}} + \int_{\text{X}} H_{\text{X}} p_{\text{X}} \int_{\text{Y}} d p_{\text{Y}|\text{X}}^{\text{EQ}} \\ &\quad + d \int_{xy} H_{\text{Y}} p_{\text{XY}} - \int_{\text{X}} p_{\text{X}} \int_{\text{Y}} (d_\lambda H_{\text{Y}}) p_{\text{Y}|\text{X}}^{\text{EQ}}. \end{aligned} \quad (4.50)$$

On the r.h.s., the first term is just the heat dQ_{X} , defined in Eq. (4.16), since the Y integral of $p_{\text{Y}|\text{X}}^{\text{EQ}}$ yields unity, whereas the second term vanishes due to the normalization condition of $p_{\text{Y}|\text{X}}^{\text{EQ}}$, the third term is the differential of the internal energy $\langle H_{\text{Y}} \rangle$ of the bath, and the final term vanishes identically due

to Eq. (2.17). Summarizing, we have

$$dQ_{XY} = dQ_X + d\langle H_Y \rangle. \quad (4.51)$$

Let us now revisit the conditional entropy of the bath, Eq. (4.38). Using Eq. (2.28), we can rewrite it as

$$S_{Y|X} = \int_{xy} p_{XY} (\ln Z_Y + \beta H_Y) = \ln Z_Y + \beta \langle H_Y \rangle. \quad (4.52)$$

Recalling that Z_Y is independent of λ , we can take the differential of the preceding equation and obtain

$$dS_{Y|X} = \beta d\langle H_Y \rangle. \quad (4.53)$$

Combining this with Eq. (4.51), we obtain

$$dQ_{XY} - dQ_X = d\langle H_Y \rangle = T dS_{Y|X}. \quad (4.54)$$

The physical meaning of this result is very clear: The heat absorbed by the joint system dQ_{XY} consists of two parts. The first part dQ_X is the heat absorbed by the system, whereas the second part $d\langle H_Y \rangle$ is the change of the internal energy of the bath. Because the bath remains in conditional thermal equilibrium, the second part also equals to the change of the conditional entropy of the bath, multiplied by the temperature.

If the joint system is thermally closed, i.e., if we take the bath to be the rest of the universe, $dQ_{XY} = 0$, whence

$$-dQ_X = T dS_{Y|X} = d\langle H_Y \rangle. \quad (4.55)$$

Hence $-\beta dQ_X$ can be understood as the change of the conditional entropy of the environment, while dQ_X can be understood as the change of the internal energy of the environment.

It is known that the combination $dS_{XY} - \beta dQ_{XY}$ may be interpreted as the entropy production, i.e., the change in the total entropy of the universe. We may use Eqs. (2.33) and (2.28) in Eqs. (4.36b) and (4.42) to obtain

$$dS^{\text{tot}} = dS_{XY} - \beta dQ_{XY} = dS_X - \beta dQ_X. \quad (4.56)$$

Hence $dS_X - \beta dQ_X$, which we proved in Eq. (4.21) to be positive definite, is indeed the entropy production.

Alternatively, we may take Y to be the rest of the universe and use Eq. (4.55). We obtain

$$dS_X - \beta dQ_{XY} = dS_X + dS_{Y|X} = dS_{XY} = dS^{\text{tot}}. \quad (4.57)$$

In all the above discussions, the conditional equilibrium nature of the bath, and the independence of Z_Y on λ play essential roles. Both properties follow from TSS and the judicious decomposition of Hamiltonian, Eq. (2.12).

V. ALTERNATIVE THEORIES

The entire theory constructed in this work follows from the particular decomposition of Hamiltonian (2.12). It is, however, possible to start from a different decomposition of Hamiltonian and derive a consistent strong-coupling theory of thermodynamics and stochastic thermodynamics. Such an alternative theory no longer has many of the nice features discussed in this work. For example, the system entropy is no longer the Gibbs-Shannon entropy, and the equilibrium pdf of the system variables is no longer the Gibbs-Boltzmann distribution.

We consider an alternative strong-coupling theory where the system Hamiltonian is chosen to be

$$\tilde{H}(\mathbf{x}) \equiv H(\mathbf{x}) + \chi(\mathbf{x}), \quad (5.1)$$

where $\chi(\mathbf{x})$ is an unspecified function. (From now on, we drop all subscripts \mathbf{X} , since there is no danger of confusion.) The nonequilibrium thermodynamic quantities in this new theory are denoted as \tilde{E} , \tilde{S} , \tilde{F} , $d\tilde{W}$, $d\tilde{Q}$, etc., where H , E , S , F , dW , dQ will be used to denote thermodynamic quantities in our theory. We only consider thermodynamic quantities at the ensemble level since they also uniquely determine thermodynamic quantities at the trajectory level.

To construct the alternative strong-coupling theory, which shall be called *the tilde theory*, we only need to express \tilde{E} , \tilde{S} , \tilde{F} , $d\tilde{W}$, $d\tilde{Q}$, etc. in terms of E , S , F , dW , dQ , and $\chi(\mathbf{x})$. We do so by imposing the following three consistency conditions,² which guarantee that the tilde theory is equivalent to our theory from the thermodynamic and stochastic-thermodynamic point of view:

(i) The tilde theory yields the same nonequilibrium free energy as our theory:

$$F[p] = E[p] - TS[p] = \tilde{F}[p] = \tilde{E}[p] - T\tilde{S}[p], \quad (5.2)$$

where $F[p]$ is given by Eq. (4.9). This guarantees that two theories predict the same equilibrium state.

(ii) The tilde theory satisfies

$$dE - dW - dQ = d\tilde{E} - d\tilde{W} - d\tilde{Q}. \quad (5.3)$$

This guarantees that the first laws of thermodynamics in two theories are equivalent.

(iii) Two theories yield the same entropy production:

$$dS^{\text{tot}} = dS - \beta dQ = d\tilde{S} - \beta d\tilde{Q}. \quad (5.4)$$

This guarantees that the second laws in the two theories are equivalent.

The internal energy $\tilde{E}[p]$ is just the ensemble average of the system Hamiltonian $\tilde{H}(\mathbf{x})$. Using Eq. (5.1) we have

$$\tilde{E}[p] = E[p] + \int_{\mathbf{x}} \chi(\mathbf{x}) p(\mathbf{x}). \quad (5.5)$$

Then condition (i) demands that $\tilde{S}[p]$ is given by

$$\tilde{S}[p] = S[p] + \beta \int_{\mathbf{x}} \chi(\mathbf{x}) p(\mathbf{x}). \quad (5.6)$$

Further assume that the work $d\tilde{W}$ is given by

$$d\tilde{W} = dW + d\phi(\mathbf{x}), \quad (5.7)$$

where $d\phi(\mathbf{x})$ is a (generally inexact) differential form. Then condition (ii) demands that the tilde heat is given by

$$d\tilde{Q} = dQ + d \int_{\mathbf{x}} \chi(\mathbf{x}) p(\mathbf{x}) - d\phi(\mathbf{x}). \quad (5.8)$$

²It is important to note that our consistency conditions are fundamentally different from those imposed by Seifert [3], which are Eqs. (V.20).

Finally we impose condition (iii). Using Eqs. (5.6) and (5.8) in Eq. (5.4), we determine $d\tilde{\phi}(\mathbf{x})$ in terms of $\chi(\mathbf{x})$:

$$d\tilde{\phi}(\mathbf{x}) = -\beta^{-1}d\beta \int_{\mathbf{x}} \chi(\mathbf{x})p(\mathbf{x}). \quad (5.9)$$

Substituting this back into Eqs. (5.7) and (5.8), we obtain

$$d\tilde{W} = dW - \beta^{-1}d\beta \int_{\mathbf{x}} \chi(\mathbf{x})p(\mathbf{x}), \quad (5.10)$$

$$d\tilde{Q} = dQ + \beta^{-1}d\left(\beta \int_{\mathbf{x}} \chi(\mathbf{x})p(\mathbf{x})\right). \quad (5.11)$$

Equations (5.1)–(5.6) and (5.10), (5.11) determine all thermodynamic quantities in the tilde theory in terms of an arbitrary function $\chi(\mathbf{x})$. In general, these quantities are much more complicated than those in our theory.

The choice made by Seifert [3], which has been followed by many others, is

$$\chi(\mathbf{x}) = \beta\partial_{\beta}H(\mathbf{x}; \lambda, \lambda', \beta), \quad (5.12)$$

so that the system Hamiltonian, i.e., the fluctuating internal energy, becomes

$$\tilde{H}(\mathbf{x}; \lambda, \lambda', \beta) = \partial_{\beta}(\beta H), \quad (5.13)$$

which is the same as our definition only if the HMF is independent of temperature. Various thermodynamic quantities are

$$\tilde{E}[p] = \int_{\mathbf{x}} p\partial_{\beta}(\beta H), \quad (5.14)$$

$$\tilde{S}[p] = \int_{\mathbf{x}} p(-\ln p + \beta^2\partial_{\beta}H), \quad (5.15)$$

$$d\tilde{W} = \int_{\mathbf{x}} p(d_{\lambda}H + d_{\lambda'}H), \quad (5.16)$$

$$d\tilde{Q} = \int_{\mathbf{x}} \tilde{H}dp + 2 \int_{\mathbf{x}} pd_{\beta}H + \beta \int_{\mathbf{x}} p\partial_{\beta}(d_{\lambda}H + d_{\lambda'}H + d_{\beta}H). \quad (5.17)$$

Comparing Eqs. (5.16) with (4.11), we see that the work in this tilde theory does not contain $d_{\beta}H$. Other thermodynamic quantities in this theory, especially the heat, are much more complicated than those in our theory. Note that in Seifert's theory and many of the following works, it is always assumed that H_I^0 [see Eqs. (4.47) and (4.48), as well as the discussion around them], and that λ', β do not vary. Hence the tilde theory we constructed here is an extension of Seifert's theory. For the particular case where β, λ' are fixed, Eqs. (5.13)–(5.17) reduce to the third column of Table I in Ref. [15]. It is evident that among all these theories, our theory with $\chi = 0, \tilde{H} = H$ is the simplest and exhibits the maximal similarity with the weak-coupling theory of stochastic thermodynamics.

Work and heat at the trajectory level in this alternative theory can also be easily determined:

$$d\tilde{\mathcal{W}} = d_{\lambda}H + d_{\lambda'}H, \quad (5.18)$$

$$d\tilde{\mathcal{Q}} = d_{\mathbf{x}}\tilde{H} + 2(d_{\beta}H) + \beta[\partial_{\beta}(d_{\lambda}H + d_{\lambda'}H + d_{\beta}H)], \quad (5.19)$$

whose ensemble averages yield Eqs. (5.16) and (5.17).

In the equilibrium state, the internal energy in the tilde theory is

$$\tilde{E} = \int \frac{1}{Z} e^{-\beta H} \partial_{\beta}(\beta H) = \frac{\partial}{\partial \beta}(\beta F), \quad (5.20a)$$

where $F = -T \ln \int e^{-\beta H}$ is the equilibrium free energy. It then follows that the tilde equilibrium entropy becomes

$$\tilde{S} = \beta^2 \frac{\partial}{\partial \beta} F. \quad (5.20b)$$

Equations (5.20) were called ‘‘thermodynamic consistency conditions’’ by Seifert. Note, however, that they imply neither the first nor the second law of thermodynamics.

We note that a similar issue also arises in the stochastic thermodynamics of discrete systems described by master equations. Generally the interaction between a discrete system and its environment is strong. It is well known that the *energy function* ϵ_k appearing in the equilibrium Gibbs-Boltzmann distribution $e^{\beta(F-\epsilon_k)}$, which is the analog of HMF in our problem, is actually a constrained free energy of the mesoscopic state k . It is generally a function of temperature and can be decomposed into an energy part and an entropy part:

$$\epsilon_k = e_k - T s_k, \quad (5.21a)$$

$$e_k \equiv \frac{\partial \beta \epsilon_k}{\partial \beta}, \quad (5.21b)$$

$$s_k \equiv \beta^2 \frac{\partial \epsilon_k}{\partial \beta}. \quad (5.21c)$$

The Gibbs-Shannon entropy $-p_k \ln p_k$ is strictly speaking not the real thermodynamic entropy. Nonetheless, in the standard theory of stochastic thermodynamics, we choose to define the system energy as ϵ_k and the system entropy as $-p_k \ln p_k$. (For a lucid discussion on this topic, we refer the readers to Sec. 3.3 of Ref. [28].) This is consistent with our definitions of energy and entropy for strongly coupled systems. As an alternative choice which share the same spirit as Seifert's strong-coupling theory, we may define the system energy as e_k and system entropy as

$$S = \sum p_k (-\ln p_k + s_k). \quad (5.22)$$

The resulting theory of stochastic thermodynamics would be much more complicated and much different from the standard theory. Hence our strong-coupling theory of stochastic thermodynamics for continuous systems is more naturally connected to the standard theory of stochastic thermodynamics for discrete systems.

VI. TWO EXAMPLES

We illustrate our strong-coupling theory using two simple examples.

A. A small piston

We consider an ideal gas consisting of N noninteracting point particles confined in a cylinder with a frictionless piston. The volume of the gas is Ax , where A is the cross section and x is the length, which varies as the piston moves. The gas is

immersed in a liquid with temperature $T = 1/\beta$ and pressure P . We further assume the walls of the cylinder conduct heat well so that it remains in equilibrium with the liquid outside. (This assumption may be rather unrealistic. But remember that we are discussing this example only to illustrate our theory.) Assuming the overdamped limit for the piston, the only slow variable is the length x , whose equilibrium pdf is

$$p^{\text{EQ}}(x) = \frac{1}{Z} e^{-\beta H(x;T,P)} = \frac{1}{Z} e^{-N \ln(N\xi^3/Axe) - \beta PAx}, \quad (6.1)$$

where $\xi = \xi(T) = h/\sqrt{2\pi mT}$ is the de Broglie thermal wavelength. This corresponds to the HMF

$$H(x;T,P) = NT \ln \frac{N\xi^3}{Axe} + PAx, \quad (6.2)$$

which can be understood as the Helmholtz free energy of the gas and the liquid, up to an additive constant. Alternatively, $H(x;T,P)$ may be also understood as the fluctuating Gibbs energy of the gas inside the cylinder. Note that $H(x;T,P)$ explicitly depends on temperature and pressure. The partition function Z is then determined by the condition of normalization:

$$Z = \int dx e^{-\beta H(x;T,P)}. \quad (6.3)$$

We take the view point that the system consists of solely the piston, and the bath consists of both the gas inside the cylinder and the liquid outside. Note that an isolated piston has vanishing potential energy, hence both terms on the r.h.s. of Eq. (6.2) are due to the interaction between the piston and the environment. It then follows that there is no system parameter in this example and A, P, N , and ξ should be all understood as the bath parameters. Remarkably, even such a simple example of stochastic thermodynamics demands a strong-coupling theory! Also note that as long as we figure out the HMF of x , there is no need to know the detailed Hamiltonian of the gas and the liquid.

We assume that β, P are tunable, but A, ξ are fixed. Using Eqs. (4.10) and (4.14), heat and work at the trajectory level are given by

$$dQ = d_x H = -NT \frac{dx}{x} + PA dx, \quad (6.4)$$

$$\begin{aligned} dW &= d_\beta H + d_P H \\ &= N \ln \left(\frac{N\xi^3}{Axe^{5/2}} \right) dT + Ax dP. \end{aligned} \quad (6.5)$$

It is interesting to compare our theory with Seifert's theory for this particular example. Using Eq. (5.13) we see that the fluctuating internal energy is

$$\tilde{H} = PAx, \quad (6.6)$$

which a linear function of the length of the cylinder x . The heat in Seifert's theory is very complicated, as one can work out using Eq. (5.17).

While the entropy in our theory is defined as the Gibbs-Shannon entropy of the piston, the entropy in Seifert's theory can also be worked out using Eq. (5.15):

$$\tilde{S} = - \int_x p(x) \ln p(x) + \int_x p(x) N \ln \frac{Axe^{5/2}}{N\xi^3}. \quad (6.7)$$

While the first term on the r.h.s. is the Gibbs-Shannon entropy of the piston, the second term turns out to be the thermodynamic entropy of the gas inside the cylinder. Hence for this particular model, the difference between our theory and Seifert's theory lies in whether we treat the gas confined in the cylinder as part of the system or as part of the bath. We must emphasize, of course, that Seifert's theory cannot be used to describe processes with variable temperature and pressure.

B. Modified harmonic-oscillator bath model

In this section, we illustrate our theory using a toy model of Brownian motion, with coordinate and momentum $\{x, p\}$, coupled to a large number of harmonic oscillators, with canonical variables $\{q_i, p_i\}$. This model is a generalization of the harmonic-oscillator model studied by Zwanzig [29,30] as well as by Caldeira and Leggett [31,32]. The total Hamiltonian is given by

$$\begin{aligned} H^{\text{tot}} &= \left(\frac{p^2}{2m} + V(x) \right) - \lambda(t) \sum_j \gamma_j q_j x \\ &+ \sum_j^N \left(\frac{1}{2} p_j^2 + \frac{\omega_j^2}{2} q_j^2 \right), \end{aligned} \quad (6.8)$$

where the three parts on the r.h.s. are respectively H_X^0, H_p^0, H_Y^0 in Eq. (2.1). Hence, in this model, $\lambda(t)$ and $\{\gamma_j\}$ are system parameters, whereas $\{\omega_j\}$ are bath parameters.

The timescale of bath dynamics is characterized by all frequencies $\{\omega_j\}$. We assume that all inverse frequencies $1/\omega_j$ are sufficiently large so that all oscillators always remain in conditional equilibrium given the instantaneous values of $\lambda(t)$ and x, p . This of course also implies that the variation of the control parameter $\lambda(t)$ needs to be much slower than the dynamics of the oscillators. In the limit of TSS, the effective dynamics of the Brownian particle obeys a nonlinear Langevin equation with white noise, where the time-dependent friction coefficient is given by

$$\gamma(t) = \beta \lambda(t)^2 \int_0^{+\infty} dt \sum_j \frac{\gamma_j^2}{\omega_j^2} \cos \omega_j t. \quad (6.9)$$

A detailed derivation of Eq. (6.9) is given in Ref. [22].

Note that, unlike in previous works [29–34], we have not explicitly introduced the counter-terms in the Hamiltonian. Note also that we have introduced a control parameter $\lambda(t)$ in front of the coupling terms, which can be varied systematically. We can follow the strategy discussed in Sec. II A and decompose the total Hamiltonian in the form of Eq. (2.12), with

$$H_X = \frac{p^2}{2m} + V(x) - \lambda(t)^2 \left(\sum_j \frac{\gamma_j^2}{2\omega_j^2} \right) x^2, \quad (6.10)$$

$$H_Y = \sum_j^N \left[\frac{1}{2} p_j^2 + \frac{\omega_j^2}{2} \left(q_j - \frac{\lambda(t)\gamma_j}{\omega_j^2} x \right)^2 \right]. \quad (6.11)$$

Note that both H_X and H_Y depend explicitly on $\lambda(t)$ and $\{\gamma_j\}$. The bath partition function and free energy, however, are independent of λ and $\{\gamma_j\}$, as demonstrated in Eq. (2.15).

Assuming that the external potential $V(x)$ is fixed, the heat and work at the trajectory level are respectively

$$d\mathcal{Q} = \frac{p}{m} dp + \left[V'(x) - 2\lambda^2 \left(\sum_j \frac{\gamma_j^2}{2\omega_j^2} \right) x \right] dx, \quad (6.12)$$

$$d\mathcal{W} = 2\lambda \left(\sum_j \frac{\gamma_j^2}{2\omega_j^2} \right) x^2 d\lambda. \quad (6.13)$$

Consider now a process where we tune the control parameter $\lambda(t)$ systematically. For a given trajectory of the joint system, the external agent does work both to the system and to the bath, as given by Eq. (4.43). The total work done to the system and the bath in general cannot be expressed in terms of the system variables alone. As we have argued in Sec. IV D, however, since the dynamics of the oscillators is very fast, along a typical trajectory, the integrated work done to the bath by the external agent is vanishingly small. Hence the integrated work $\mathcal{W}_{\mathbf{X}} = \int_{\gamma} d\lambda H_{\mathbf{X}}$ is the change of total energy, up to a negligibly small and rapidly fluctuating error. Neglecting this error, the stochastic thermodynamics of this process is fully described by our theory which involves only system variables. This conclusion is both highly nontrivial and conceptually satisfactory.

The problem studied above is special in the sense that the HMF is independent of temperature so that the bath does no work to the system as the temperature is varied. We can easily modify the model such that the HMF is temperature dependent. Consider for example the following total Hamiltonian:

$$H^{\text{tot}} = \left(\frac{p^2}{2m} + V(x) \right) + \frac{\lambda(t)x^2}{2} \sum_j c_j^2 q_j^4 + \sum_j^N \left(\frac{1}{2} p_j^2 + \frac{\omega_j^2}{2} q_j^2 \right), \quad (6.14)$$

which involves nonlinear couplings between x and q_j . The HMF can be easily computed using Eq. (2.6):

$$H_{\mathbf{X}} = \frac{p^2}{2m} + V(x) - T \sum_j \ln \int \frac{dy}{\sqrt{\pi}} e^{-\frac{y^2}{2} - \frac{\lambda c_j x^2 y^4}{\beta \omega_j^2}}. \quad (6.15)$$

It is evident that $H_{\mathbf{X}}$ depends on both β and λ .

VII. CONCLUSION

In this work, we have demonstrated how to extend the standard theories of thermodynamics and stochastic thermodynamics for continuous systems to the strong-coupling regime, where system parameters, bath parameters, and temperature may be systematically varied. There is no need to explicitly refer to the environmental degree of freedom, as long as we define the system Hamiltonian as the Hamiltonian of mean force (HMF), and the system entropy as the usual Gibbs-Shannon entropy. Differences with the classical theories, however, do arise due to the temperature dependence of HMF, both in the equilibrium theory and in the nonequilibrium theory. The most important new feature is that the environment does work on the system when the temperature is varied. We have also constructed an infinite number of equivalent strong-coupling theories, each characterized by its definition of system Hamiltonian. Among all these theories, our theory is distinguished by its maximal similarity to the weak-coupling theory, and by its natural connection with the standard theory of stochastic thermodynamics of discrete systems.

To appreciate the descriptive power of our theory, consider a colloid immersed in a nematic liquid crystal and confined by an optical trap. The system parameter λ would control the optical trap or the surface anchoring strength of nematogens, whereas the bath parameter λ' would be the magnetic field acting on the nematogens but not on the colloid. As another example, consider a heat engine working between two heat baths. The system parameter λ may control the coupling between the engine and the baths, which vanishes during adiabatic processes, but are nonvanishing during isothermal processes. These problems can be naturally addressed using our theory, but not using the other strongly coupling theories of stochastic thermodynamics.

ACKNOWLEDGMENTS

The authors acknowledge support from NSFC Grant #12375035 as well as Shanghai Municipal Science and Technology Major Project (Grant No. 2019SHZDZX01). M.D. thanks Chenxing Post Doctoral Incentive Project for financial support.

-
- [1] C. Jarzynski, NonEQ work theorem for a system strongly coupled to a thermal environment, *J. Stat. Mech.* (2004) P09005.
 - [2] M. F. Gelin and M. Thoss, Thermodynamics of a sub-ensemble of a canonical ensemble, *Phys. Rev. E* **79**, 051121 (2009).
 - [3] U. Seifert, First and second law of thermodynamics at strong coupling, *Phys. Rev. Lett.* **116**, 020601 (2016).
 - [4] P. Talkner and P. Hänggi, Open system trajectories specify fluctuating work but not heat, *Phys. Rev. E* **94**, 022143 (2016).
 - [5] P. Strasberg, G. Schaller, N. Lambert, and T. Brandes, Non equilibrium thermodynamics in the strong coupling and non-Markovian regime based on a reaction coordinate mapping, *New J. Phys.* **18**, 073007 (2016).
 - [6] P. Strasberg and M. Esposito, Stochastic thermodynamics in the strong coupling regime: An unambiguous approach based on coarse graining, *Phys. Rev. E* **95**, 062101 (2017).
 - [7] C. Jarzynski, Stochastic and macroscopic thermodynamics of strongly coupled systems, *Phys. Rev. X* **7**, 011008 (2017).
 - [8] H. J. D. Miller and J. Anders, Entropy production and time asymmetry in the presence of strong interactions, *Phys. Rev. E* **95**, 062123 (2017).
 - [9] E. Aurell, On work and heat in time-dependent strong coupling, *Entropy* **19**, 595 (2017).
 - [10] E. Aurell, Unified picture of strong-coupling stochastic thermodynamics and time reversals, *Phys. Rev. E* **97**, 042112 (2018).

- [11] P. Talkner and P. Hänggi, Colloquium: Statistical mechanics and thermodynamics at strong coupling: Quantum and classical, *Rev. Mod. Phys.* **92**, 041002 (2020).
- [12] R. de Miguel and J. Miguel Rubí, Strong coupling and nonextensive thermodynamics, *Entropy* **22**, 975 (2020).
- [13] P. Strasberg and M. Esposito, Measurability of non-equilibrium thermodynamics in terms of the Hamiltonian of mean force, *Phys. Rev. E* **101**, 050101(R) (2020).
- [14] P. Talkner and P. Hänggi, Comment on “measurability of nonequilibrium thermodynamics in terms of the Hamiltonian of mean force,” *Phys. Rev. E* **102**, 066101 (2020).
- [15] M. Ding, Z. Tu, and X. Xing, Strong coupling thermodynamics and stochastic thermodynamics from the unifying perspective of time-scale separation, *Phys. Rev. Res.* **4**, 013015 (2022).
- [16] R. de Miguel and J. Miguel Rubí, Statistical mechanics at strong coupling: A bridge between Landsberg’s energy levels and Hill’s nanothermodynamics, *Nanomaterials* **10**, 2471 (2020).
- [17] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, Oxford, England, 2001).
- [18] S. Chaturvedi and F. Shibata, Time-convolutionless projection operator formalism for elimination of fast variables. Applications to Brownian motion, *Z. Phys. B* **35**, 297 (1979).
- [19] H. Grabert, *Projection Operator Techniques in Nonequilibrium Statistical Mechanics* (Springer, 2006), Vol. 95.
- [20] H. P. Breuer, B. Kappler, and F. Petruccione, The time-convolutionless projection operator technique in the quantum theory of dissipation and decoherence, *Ann. Phys. (NY)* **291**, 36 (2001).
- [21] S. Bo and A. Celani, Multiple-scale stochastic processes: Decimation, averaging and beyond, *Phys. Rep.* **670**, 1 (2017).
- [22] M. Ding and X. Xing, Multi-scale projection operator method and coarse-graining of covariant Fokker-Planck theory, *Phys. Rev. Res.* **5**, 013193 (2023).
- [23] M. Campisi, P. Talkner, and P. Hänggi, Fluctuation theorem for arbitrary open quantum systems, *Phys. Rev. Lett.* **102**, 210401 (2009).
- [24] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).
- [25] U. Seifert, Entropy production along a stochastic trajectory and an integral fluctuation theorem, *Phys. Rev. Lett.* **95**, 040602 (2005).
- [26] M. Ding and X. Xing, Covariant nonequilibrium thermodynamics from Ito-Langevin dynamics, *Phys. Rev. Res.* **4**, 033247 (2022).
- [27] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics* (Springer Science & Business Media, 2012), Vol. 31.
- [28] L. Peliti and S. Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, 2021).
- [29] R. Zwanzig, Nonlinear generalized Langevin equations, *J. Stat. Phys.* **9**, 215 (1973).
- [30] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, Oxford, 2001).
- [31] A. O. Caldeira and A. J. Leggett, Path integral approach to quantum Brownian motion, *Physica A (Amsterdam, Neth.)* **121**, 587 (1983).
- [32] A. O. Caldeira and A. J. Leggett, Quantum tunnelling in a dissipative system, *Ann. Phys. (NY)* **149**, 374 (1983).
- [33] C. Jarzynski, Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale, *Annu. Rev. Condens. Matter Phys.* **2**, 329 (2011).
- [34] U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines, *Rep. Prog. Phys.* **75**, 126001 (2012).