Trajectory Class Fluctuation Theorem

Gregory Wimsatt,* Alexander B. Boyd,† and James P. Crutchfield‡
Complexity Sciences Center and Physics and Astronomy Department, University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: April 27, 2024)

The Trajectory Class Fluctuation Theorem (TCFT) substantially strengthens the Second Law of Thermodynamics—that, in point of fact, can be a rather weak bound on resource fluxes. Practically, it improves empirical estimates of free energies, a task known to be statistically challenging, and has connected the microscopic dynamics with the mesoscopic information processing in experimentally-implemented Josephson-junction information engines. The development here justifies that empirical analysis, explicating its mathematical foundations.

The TCFT reveals the thermodynamics induced by macroscopic system transformations for each measurable subset of system trajectories. In this, it directly combats the statistical challenge of extremely rare events that dominate thermodynamic calculations. And, it reveals new forms of free energy—forms that can be solved analytically and practically estimated. For engineered systems, it provides a toolkit for diagnosing the thermodynamics responsible for system functionality. Conceptually, the TCFT unifies a host of previously-established fluctuation theorems, interpolating from Crooks' Detailed Fluctuation Theorem (single trajectories) to Jarzynski's Equality (full trajectory ensembles).

Keywords: integral fluctuation theorem, detailed fluctuation theorem, free energy

I. INTRODUCTION

The century-old study of thermodynamic fluctuations was rejuvenated with the discovery of fluctuation theorems in the very late twentieth century [1–5]. Jarzynski's Equality [3] and Crooks' Detailed Fluctuation Theorem [6], in particular, have been used to infer and relate thermodynamic properties of small systems driven by transformations very far from equilibrium. Fluctuation theorems revealed that stochastic deviations from equilibrium in small-scale systems obey specific functional forms. That is, fluctuations are lawful.

In fact, the equilibrium Second Law and its nonequilibrium generalization can be derived from the stronger equalities provided by fluctuation theorems [5]. In addition, equilibrium and nonequilibrium free energy changes are readily obtained from those same stronger equalities [3, 7]. And, this allows estimating free energies given sufficient sampling of a thermodynamic process. This has been carried out successfully for RNA and DNA configurational free energies [8–10] and quantum harmonic oscillators [11]. However, obtaining free energies is generally quite challenging statistically due to the existence of rare events that dominate the exponential average work [12, 13].

Many of these results rely on averaging thermodynamic quantities over trajectory ensembles. We recently introduced the *trajectory class fluctuation theorem* (TCFT) that focuses instead on subsets of trajectories—trajectory *classes* [14]. While a restricted form of the TCFT had been noted previously [13], we present a theorem that applies to arbitrary measurable subsets of trajectories. And, we derive a suite of results that lift prior limitations. Namely, by considering information about one or more trajectory classes, we markedly strengthen statements of the Second Law [14]. Practically, too, by using trajectory classes with high probability, we overcome limitations in estimating free-energy differences due to finite sampling.

The TCFT introduces a new level of flexibility central to extracting free energy differences in a wide variety of empirical settings. The *detailed fluctuation theorem* (DFT) [6], the basis from which these results are derived, relies on comparing state trajectory probabilities evaluated from a *forward experiment* and and a *reverse experiment* to evaluate the entropy production in the forward experiment. However, the DFT's predictive capacity is severely hampered by the fact that the state trajectories are typically so numerous that their individual probabilities are extremely small, if not zero. It is then virtually impossible to sample sufficient data to reliably estimate those probabilities. Moreover, it is rare in an experiment to have complete information about a system trajectory.

To meet these challenges, Ref. [14]'s TCFT provides a practical computational advantage by estimating entropy from a much smaller space of trajectory classes—classes that can be tailored to specific experimental data and constraints. The following further expands on the TCFT's experimental relevance by generalizing to the case in which the reverse experiment does not necessarily start in a special distribution—i.e., the distribution conjugate to

* gwwimsatt@ucdavis.edu
† abboyd@ucdavis.edu
‡ chaos@ucdavis.edu; Corresponding author

the ending distribution of the forward experiment. This generalization requires introducing a new thermodynamic quantity known as the *entropy difference*, which can be interpreted as entropy production in special cases, but has important thermodynamic consequences regardless. To illustrate, we apply the TCFT to metastable processes—processes where the system begins and ends in metastable distributions—and show how to derive metastable free energies with an appropriately initialized set of experiments [15].

Theoretically, the TCFT unites a wide variety of prior fluctuation theorems. These include theorems that range systematically from Crooks' DFT [4] to Jarzynski's *integral fluctuation theorem* (IFT) [3]. That noted, the TCFT itself can be derived from broader theorems still [16, 17]; see App. A. The TCFT's strength then is in its balance of specificity and generality.

Developing the TCFT proceeds as follows. Section II presents fluctuation theorem building blocks, culminating in Crooks' (DFT) and the basic IFTs. Section III uses the DFT to introduce and prove the TCFT. Section IV shows how it strengthens the Second Law in light of process data. However, the Second Law and its strengthened forms are much more useful when the free-energy difference is known so that bounds on work can be established. And so, Sec. V shows how to use the TCFT to solve for free-energy differences beyond just the equilibrium free-energy difference. Section VI demonstrates how the TCFT overcomes the tyranny of rare events when estimating free energies from data. Section VII highlights related results and surveys how the TCFT encapsulates them. Finally, we briefly discuss several subtle aspects of applying the TCFT in Sec. VIII, considering the application to a recent nanoscale flux qubit experiment [14]. Section IX concludes.

II. BACKGROUND

A. Model, Probability Densities, and Time Reversal

Consider a system interacting with both a control device and a thermal environment over a time interval $[0, \tau]$ from time 0 to time τ. The device enacts a control protocol $\overrightarrow{\lambda}$ over the time interval to influence the system. Specifically, at any time t, $\overrightarrow{\lambda}(t)$ specifies the function from system state to system energy, called the *energy landscape*, at time t. The protocol therefore both affects how the system evolves and requires energy, which we call *work*, to be exchanged between the system and the control device. The thermal environment has inverse temperature β so that energy, denoted *heat*, flows between the system and environment. No other interactions exist. Altogether, this induces a stochastic dynamic in the system as it evolves from time 0 to time τ.

We model time as either continuous or discrete. In general, the resultant set T of times is some subset of the interval $[0, \tau]$ that includes 0 and τ.

We model the system states as either its microstates or as some coarse-graining of its microstates. Denote a particular system *state* as z and a particular system state *trajectory* as $\overrightarrow{z}$, with $\overrightarrow{z}(t)$ the realized system state at time $t \in T$. We let $\mathcal{Z}$ denote the set of all possible states and $\overrightarrow{\mathcal{Z}}$ the set of all possible trajectories.

The energy of the system in state z at time t is denoted $E_t(z)$. The energy change of the system over an entire trajectory $\overrightarrow{z}$ is then:

$$\Delta E(\overrightarrow{z}) = E_\tau(\overrightarrow{z}(\tau)) - E_0(\overrightarrow{z}(0)) \ .$$

We designate positive work or heat to mean that energy flowed into the system from the control device or thermal environment, respectively.

We require that the net work $W(\overrightarrow{z})$ is a function of trajectory. This can be achieved if T and $\mathcal{Z}$ are sufficiently refined. (For example, $T = [0, \tau]$ and $\mathcal{Z}$ is the system's set of microstates.) Then, by conservation of energy, heat $Q(\overrightarrow{z})$ must also be a function of trajectory and we have the First Law for each trajectory:

$$\Delta E(\overrightarrow{z}) = W(\overrightarrow{z}) + Q(\overrightarrow{z}) \ .$$

We describe probabilities of states with state probability densities, which we refer to as *distributions*. These are functions of $\mathcal{Z}$ that when integrated over a region of state space give the corresponding probability of occupying that region. As an important example, the system's equilibrium distribution π_t for energy landscape E_t is the corresponding Boltzmann distribution:

$$\pi_t(z) = e^{-\beta(E_t(z) - F_t^{\text{eq}})} \ ,$$

where F_t^{eq} is the system's equilibrium free energy at time t and β is the inverse temperature as denoted in statistical mechanics. Note that if $\mathcal{Z}$ is discrete, then a state probability density evaluated at a state is in fact the corresponding probability of that state. That is, integration over discrete spaces can simply be taken to be summation.

Similarly, we describe probabilities of trajectories with trajectory probability densities. These densities are functions of $\overrightarrow{z}$ that, when integrated over a region of trajectory space, give the corresponding probability of a trajectory occupying that region.

If the state space $\mathcal{Z}$ is a Euclidean space of dimension d, as is typical for spaces of microstates, integration over $\mathcal{Z}$

can of course be done with a d-dimensional Riemannian integral. However, the trajectory space $\overrightarrow{\mathcal{Z}}$ is much too large to be Euclidean when the set of times T is continuous. This then requires a more powerful notion of integration. A solution can be found via measure theory but we treat the subject only briefly here. (A sequel provides the details [18].) A measure is a function on the regions of a space that returns an amount of some "quantity", such as probability, in a region. We choose a particular measure on trajectory space and refer to it as a *base measure*. A probability density is then a function that, when Lebesgue-integrated via the base measure over a region of space, yields the corresponding probability of that region. Defining appropriate base measures on a continuous-time trajectory space is rather technical, though, and so we leave the discussion to the sequel.

We specify the dynamic induced by a protocol $\overrightarrow{\lambda}$ via a set of trajectory probability densities, one for each possible initial state z. This gives the probability density of evolving the system trajectory $\overrightarrow{z}$ conditioned on starting in a state z:

$$\Pr_{\overrightarrow{\lambda}}(\overrightarrow{Z} = \overrightarrow{z} | Z_0 = z) \ ,$$

where $\overrightarrow{Z}$ and Z_0 are random variables for the trajectory and the initial state, respectively. Call such a set of trajectory probability densities a *state-conditioned process*.

For system state z, the *time-reverse state* $z^\dagger$, or simply *reverse state*, is the same state with all components odd under time reversal flipped in sign. (Recall momentum or spin.) For state trajectory $\overrightarrow{z}$, $\overrightarrow{z}^\dagger$ is the *reverse state trajectory*: $(\overrightarrow{z}^\dagger)(t) = (\overrightarrow{z}(\tau - t))^\dagger$ for $0 \leq t \leq \tau$. If κ is a distribution over system states, then $\kappa^\dagger$ is the reverse distribution, defined by $(\kappa^\dagger)(z) = \kappa(z^\dagger)$. Note that time reversal of a state, trajectory, or distribution is an involution, meaning that time reversal acted twice on any such object returns the original object.

For a given protocol $\overrightarrow{\lambda}$, consider the corresponding *time-reverse protocol* $\overrightarrow{\lambda}^\dagger$. $\overrightarrow{\lambda}$ dictates a set of forces and fields that are applied to the system as a function of time. Enacting $\overrightarrow{\lambda}^\dagger$ then requires applying these same influences but in the reverse order as well as flipping the sign of time-odd influences, such as magnetic fields. Time reversing is therefore also involutional on protocols.

For simplicity when working with time-reversal, we require:

- $\tau - t$ to be in T for each t in T and
- $z^\dagger$ to be in $\mathcal{Z}$ for each z in $\mathcal{Z}$.

These basic symmetry requirements are satisfied in typical models in statistical mechanics and nonequilibrium thermodynamics.

B. Forward and Reverse: Experiments and Processes

The main objects of study are a system's *forward* and *reverse processes*, which result from forward and reverse experiments. The *forward experiment* consists of an initial distribution ρ and the forward control protocol $\overrightarrow{\lambda}$, which evolves the distribution over the time interval $(0, \tau)$. Similarly, for some state distribution σ, the *reverse experiment* applies the reverse control protocol $\overrightarrow{\lambda}^\dagger$ to the initial distribution $\sigma^\dagger$. We refer to ρ and σ as *privileged distributions*, emphasizing that different choices for these distributions result in different predictions given by the TCFTs and other fluctuation theorems.

Through the control protocol $\overrightarrow{\lambda}$, the forward experiment produces the *forward state-conditioned process* p, where we denote the probability density of a trajectory $\overrightarrow{z}$ conditioned on the initial state z as:

$$p(\overrightarrow{z}|z) \equiv \Pr_{\overrightarrow{\lambda}}(\overrightarrow{Z} = \overrightarrow{z} | Z_0 = z) \ ,$$

Similarly, the *reverse state-conditioned process* r' is obtained under the reverse protocol $\overrightarrow{\lambda}^\dagger$:

$$r'(\overrightarrow{z}|z) \equiv \Pr_{\overrightarrow{\lambda}^\dagger}(\overrightarrow{Z} = \overrightarrow{z} | Z_0 = z) \ .$$

We suppose throughout that microscopic reversibility holds for the system. This means that when the system evolves along trajectory $\overrightarrow{z}$, the environment's net entropy change is:

$$\begin{aligned} \Delta S_{\text{env}}(\overrightarrow{z}) &= -\beta Q(\overrightarrow{z}) \\ &= \ln \frac{p(\overrightarrow{z}|\overrightarrow{z}(0))}{r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger)} \ . \end{aligned} \tag{1}$$

Recall that microscopic reversibility can be derived, for example, under Markov [6], Hamiltonian [12, 19], or Langevin [20] assumptions.

The *forward process* is a trajectory probability density P specified by the initial distribution ρ and the forward state-conditioned process. Under P, the probability density of a trajectory $\overrightarrow{z}$ is:

$$P(\overrightarrow{z}) \equiv \rho(\overrightarrow{z}(0)) p(\overrightarrow{z}|\overrightarrow{z}(0)) \ .$$

For any time t, we marginalize P to find the evolved state distribution ρ_t. For each region A in system state space, let $P_t(A)$ be the probability of the system's state being in A at time t, and let C be the set of trajectories that occupy region A at time t. Then ρ_t is the state probability

density that satisfies:

$$P_t(A) = \int_A dz\, \rho_t(z) = \int_C d\overrightarrow{z} P(\overrightarrow{z}) \ .$$

Analogously, the *reverse process* is a trajectory probability density R' determined by the initial distribution $\sigma^\dagger$ and the reverse state-conditioned process. Under R', the probability density of a trajectory $\overrightarrow{z}$ is:

$$R'(\overrightarrow{z}) \equiv (\sigma^\dagger)(\overrightarrow{z}(0)) r'(\overrightarrow{z}|\overrightarrow{z}(0)) \ . \tag{2}$$

To simplify the following, we use an alternate representation for the reverse process—the *formal reverse representation* R—another trajectory probability density. For each trajectory $\overrightarrow{z}$, we define:

$$R(\overrightarrow{z}) \equiv R'(\overrightarrow{z}^\dagger) \ . \tag{3}$$

To keep the two representations distinct, the original representation R' of the reverse process is called the *physical reverse representation*.

C. Detailed Fluctuation Theorem

Applying the principle of microscopic reversibility to forward and reverse processes leads directly to a *detailed fluctuation theorem* (DFT). First, define the *system state entropy*, or *system state surprisal*, of a given state z for a given distribution κ [20] as:

$$s_{\text{sys}}(z;\kappa) = -\ln \kappa(z) \ .$$

Second, define the *system entropy difference* for a trajectory $\overrightarrow{z}$ in terms of the two privileged distributions ρ and σ:

$$\begin{aligned} \Delta s_{\text{sys}}(\overrightarrow{z}) &\equiv s_{\text{sys}}(\overrightarrow{z}(\tau);\sigma) - s_{\text{sys}}(\overrightarrow{z}(0);\rho) \\ &= \ln \frac{\rho(\overrightarrow{z}(0))}{\sigma(\overrightarrow{z}(\tau))} \ . \end{aligned} \tag{4}$$

This is similar, but more general than the *change in system entropy* [20] of the forward experiment. The latter is the difference in surprisal of the system in the forward experiment:

$$\ln \frac{\rho(\overrightarrow{z}(0))}{\rho_\tau(\overrightarrow{z}(\tau))} \ .$$

The system entropy difference is the change in system entropy if the reverse experiment is initialized in the time reversal of the final distribution of the forward experiment, meaning $\sigma = \rho_\tau$.

We designate the *entropy difference* as the difference in system state entropy and the change in environmental entropy:

$$\Sigma(\overrightarrow{z}) = \Delta s_{\text{sys}}(\overrightarrow{z}) + \Delta S_{\text{env}}(\overrightarrow{z}) \ . \tag{5}$$

Again, this is similar, but more general than another familiar quantity. When we choose $\sigma = \rho_\tau$, then $\Sigma(\overrightarrow{z})$ gives the entropy change of the system plus that of the environment for $\overrightarrow{z}$, which is known as the *entropy production* of the forward experiment.

Together in Eq. (5), Eqs. (1) and (4) yield an expression for the entropy difference:

$$\Sigma(\overrightarrow{z}) = \ln \frac{\rho(\overrightarrow{z}(0)) p(\overrightarrow{z}|\overrightarrow{z}(0))}{\sigma(\overrightarrow{z}(\tau)) r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger)} \ . \tag{6}$$

The numerator is $P(\overrightarrow{z})$. And, by Eqs. (2) and (3):

$$\begin{aligned} R(\overrightarrow{z}) &= (\sigma^\dagger)((\overrightarrow{z}^\dagger)(0)) r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}^\dagger)(0)) \\ &= (\sigma^\dagger)((\overrightarrow{z}(\tau))^\dagger) r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger) \\ &= \sigma(\overrightarrow{z}(\tau)) r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger) \ . \end{aligned}$$

These two observations translate Eq. (6) into a DFT:

$$\Sigma(\overrightarrow{z}) = \ln \frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} \ . \tag{7}$$

This is a fluctuation theorem obtained previously in a Langevin setting [20] and a generalization of the Crooks fluctuation theorem [6] to the case of arbitrary privileged distributions.

Continuing in this way, we introduce a constraint on the forward and reverse processes:

$$R(\overrightarrow{z}) = 0 \text{ when } P(\overrightarrow{z}) = 0 \ . \tag{8}$$

The motivation is that the forward process needs to "cover" all the trajectories that are significant to the reverse process. The failure of Condition (8) generally introduces subprobabilistic measures and densities for the reverse process that complicate the development and, in any case, may not be experimentally accessible. When considering the TCFT applied to a trajectory class C, described shortly in Sec. III, such complications are avoided so long as Condition (8) holds for all $\overrightarrow{z} \in C$. For simplicity of discussion, we assume it holds for all $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$.

Assuming that the heat $Q(\overrightarrow{z})$ is finite for all $\overrightarrow{z}$, then microscopic reversibility Eq. (1) guarantees $r(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger) = 0$ wherever $p(\overrightarrow{z}|\overrightarrow{z}(0)) = 0$. Then Condition (8) is met so long as $\sigma(\overrightarrow{z}(\tau)) = 0$ for any $\overrightarrow{z}$

where $\rho(\overrightarrow{z}(0)) = 0$. The most straightforward way to ensure this is to let ρ have at least a small amount of probability density on all system states. Note that if E_0 is everywhere finite, then π_0 has full support.

D. Work and Free Energy

The following shows that, for any given trajectory, the entropy difference decomposes into the requisite work in the forward experiment minus a *difference in nonequilibrium free energy*. Note that the latter is more general than the change in nonequilibrium free energy of the forward experiment. This realization yields important versions of the fluctuation theorems. In particular, it allows extracting the change of free energy—an important privileged-distribution-dependent but protocol-independent quantity—from the work.

To see this, first define the *state free energy* for a distribution κ and system energy function E:

$$f(z; \kappa, E) = E(z) + \beta^{-1} \ln \kappa(z) \ .$$

An important example occurs when κ is the equilibrium distribution for E. In that case, the state free energy is constant over all z and is the equilibrium free energy.

Second, define the *trajectory free-energy difference* for the forward and reverse processes as:

$$\begin{aligned} \Delta f(\overrightarrow{z}) &\equiv f(\overrightarrow{z}(\tau); \sigma, E_\tau) - f(\overrightarrow{z}(0); \rho, E_0) \\ &= \Delta E(\overrightarrow{z}) + \beta^{-1} \ln \frac{\sigma(\overrightarrow{z}(\tau))}{\rho(\overrightarrow{z}(0))} \ . \end{aligned}$$

Again, the latter echoes a familiar thermodynamic quantity: the *change in nonequilibrium free energy* ΔF^{neq} for the forward experiment [15]. The free-energy difference reduces to the nonequilibrium free energy change when $\sigma = \rho_\tau$.

Using the first law—$\Delta E(\overrightarrow{z}) = W(\overrightarrow{z}) + Q(\overrightarrow{z})$—rewrite the entropy difference in terms of the work and free-energy difference:

$$\begin{aligned} \Sigma(\overrightarrow{z}) &= -\ln \frac{\sigma(\overrightarrow{z}(\tau))}{\rho(\overrightarrow{z}(0))} - \beta Q(\overrightarrow{z}) \\ &= \beta\big(\Delta E(\overrightarrow{z}) - \Delta f(\overrightarrow{z}) - Q(\overrightarrow{z})\big) \\ &= \beta\big(W(\overrightarrow{z}) - \Delta f(\overrightarrow{z})\big) \ , \end{aligned} \tag{9}$$

And so, the entropy difference is the work in the forward experiment minus the difference in nonequilibrium free energy.

E. Ensemble Fluctuation Theorems and the Second Law

From Eq. (7)'s DFT, it is easy to derive two general fluctuation theorems. First, there is the nominal ensemble fluctuation theorem:

$$\begin{aligned} \langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &= -\int d\overrightarrow{z} P(\overrightarrow{z}) \ln \frac{R(\overrightarrow{z})}{P(\overrightarrow{z})} \\ &= \mathrm{D}_{\mathrm{KL}}\left[P \,||\, R\right]_{\overrightarrow{\mathcal{Z}}} \ . \end{aligned} \tag{10}$$

Here, $\langle \cdot \rangle_{\overrightarrow{\mathcal{Z}}}$ denotes an ensemble average over all trajectories $\overrightarrow{\mathcal{Z}}$. And, $\mathrm{D}_{\mathrm{KL}}\left[P \,||\, R\right]_{\overrightarrow{\mathcal{Z}}}$ is the Kullback-Leibler divergence between the forward and reverse process taking all trajectories $\overrightarrow{\mathcal{Z}}$ as argument.

The divergence tracks the mismatch between the distributions. Generally, it is nonnegative and vanishes only when the distributions are equal over all events. In the present case, the ensemble average entropy difference is zero only when $P(\overrightarrow{z}) = R(\overrightarrow{z})$ for all $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$.

The divergence's nonnegativity is tantamount to a generalized Second Law of thermodynamics—one that bounds average entropy differences:

$$\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} \geq 0 \ . \tag{11}$$

This includes the familiar bound on entropy production, but different choices of the privileged distributions lead to new bounds on thermodynamic quantities.

Applying Eq. (9), the average free energy change bounds the work done in the forward experiment:

$$\begin{aligned} \langle W \rangle_{\overrightarrow{\mathcal{Z}}} &\geq \langle \Delta f \rangle_{\overrightarrow{\mathcal{Z}}} \\ &= \int d\overrightarrow{z} P(\overrightarrow{z}) \Delta f(\overrightarrow{z}) \\ &= \int d\overrightarrow{z} P(\overrightarrow{z}) f(\overrightarrow{z}(\tau); \sigma, E_\tau) \\ &\quad - \int d\overrightarrow{z} P(\overrightarrow{z}) f(\overrightarrow{z}(0); \rho, E_0) \\ &= \int dz \rho_\tau(z) f(z; \sigma, E_\tau) - \int dz \rho(z) f(z; \rho, E_0). \end{aligned} \tag{12}$$

The average work is determined by the forward process exclusively and therefore must not have any actual dependence on the second privileged distribution σ, despite the latter's appearance in the first term on the RHS. And yet, the bound must hold for whichever distribution is chosen for σ.

This begs the question, for which σ is Eq. (12) tightest?

The answer is the final-time distribution ρ_τ:

$$\begin{aligned}\int dz\rho_\tau(z)f(z;\sigma,E_\tau) &= \int dz\rho_\tau(z)[E_\tau(z)+\beta^{-1}\ln\sigma(z)] \\ &= \int dz\rho_\tau(z)[E_\tau(z)+\beta^{-1}\ln\rho_\tau(z) \\ &\quad -\beta^{-1}\ln\rho_\tau(z)+\beta^{-1}\ln\sigma(z)] \\ &= \int dz\rho_\tau(z)f(z;\rho_\tau,E_\tau) \\ &\quad -\beta^{-1}\mathrm{D}_{\mathrm{KL}}\left[\rho_\tau \,||\, \sigma\right] \\ &\leq \int dz\rho_\tau(z)f(z;\rho_\tau,E_\tau)\ ,\end{aligned}$$

where:

$$\mathrm{D}_{\mathrm{KL}}\left[\rho_\tau \,||\, \sigma\right] = \int dz\rho_\tau(z)\ln\frac{\rho_\tau(z)}{\sigma(z)}$$

is nonnegative. Therefore, the free-energy difference is generally less than the change in nonequilibrium free energy, which gives the strongest bound on work production:

$$\langle W\rangle_{\overrightarrow{\mathcal{Z}}} \geq \langle\Delta F^{\mathrm{neq}}\rangle_{\overrightarrow{\mathcal{Z}}} \geq \langle\Delta f\rangle_{\overrightarrow{\mathcal{Z}}}\ . \tag{13}$$

Even though the nonequilibrium free-energy change for the forward process provides the tightest bound on work when it is used in Eq. (12), there are other useful alternatives. This flexibility is helpful as it may be difficult to determine the precise final-time distribution ρ_τ. Or, we may be more interested in the system after it relaxes to its equilibrium state π_τ determined by the final-time energy function E_τ:

$$\begin{aligned}\int dz\rho_\tau(z)f(z;\pi_\tau,E_\tau) &= \int dz\rho_\tau(z)(E_\tau(z) \\ &\quad +\beta^{-1}\ln[e^{\beta(E_\tau(z)-F_\tau^{\mathrm{eq}})}]) \\ &= F_\tau^{\mathrm{eq}}\ .\end{aligned}$$

If we start the system in equilibrium $\rho = \pi_0$, a similar calculation shows that, from Eq. (12):

$$\langle W\rangle_{\overrightarrow{\mathcal{Z}}} \geq \Delta F^{\mathrm{eq}}\ ,$$

with:

$$\Delta F^{\mathrm{eq}} = F_\tau^{\mathrm{eq}} - F_0^{\mathrm{eq}}\ .$$

Thus, this gives the equilibrium Second Law which applies to systems that start in equilibrium, but end in arbitrary distributions close to or far from equilibrium. And so, it also applies without using any information about the final distribution of the forward experiment ρ_τ.

This all said, the assumption of equilibrium is rather restrictive. It is often possible to set more informed bounds on work invested by using incomplete information about the initial distribution ρ_0 and final distribution ρ_τ. Such distributions are metastable, but provide a convenient method of improving work production estimates. Section V B discusses this shortly.

Finally, from the DFT we can obtain the exponential ensemble fluctuation theorem:

$$\begin{aligned}\left\langle e^{-\Sigma}\right\rangle_{\overrightarrow{\mathcal{Z}}} &= \int d\overrightarrow{z}\,P(\overrightarrow{z})\frac{R(\overrightarrow{z})}{P(\overrightarrow{z})} \\ &= \int d\overrightarrow{z}\,R(\overrightarrow{z}) \\ &= 1\ .\end{aligned} \tag{14}$$

Again assuming equilibrium privileged distributions, the equilibrium free-energy difference can be extracted from the average:

$$\begin{aligned}\left\langle e^{-\Sigma}\right\rangle_{\overrightarrow{\mathcal{Z}}} &= \left\langle e^{-\beta(W-\Delta F^{\mathrm{eq}})}\right\rangle_{\overrightarrow{\mathcal{Z}}} \\ &= \left\langle e^{-\beta W}\right\rangle_{\overrightarrow{\mathcal{Z}}} e^{\beta\Delta F^{\mathrm{eq}}}\ ,\end{aligned}$$

giving Jarzynski's Equality:

$$\left\langle e^{-\beta W}\right\rangle_{\overrightarrow{\mathcal{Z}}} = e^{-\beta\Delta F^{\mathrm{eq}}}\ . \tag{15}$$

Remarkably, this allows extracting equilibrium free-energy differences from work statistics of highly nonequilibrium processes. However, the exponential free-energy difference Δf cannot generally be extracted from $\left\langle e^{-\beta(W-\Delta f)}\right\rangle_{\overrightarrow{\mathcal{Z}}}$ for any choice of ρ besides π_0. Additionally, estimating even the free-energy difference from experiment using Eq. (15) can lead to sampling issues due to rare but resource-dominant events. We use the TCFT in Secs. V and VI to confront these two problems.

III. TRAJECTORY CLASS FLUCTUATION THEOREM

The preceding results on nonequilibrium thermodynamic processes are statements concerning either individual trajectories or ensemble averages—that is, averages over all trajectories. As we will see, though, a markedly broader picture emerges when considering averages that lie between. Specifically, the following treats arbitrary subsets of trajectories, called *trajectory classes*, as the main players in analyzing fluctuations. The resulting *trajectory class fluctuation theorem* (TCFT) reveals relationships involving probabilities of trajectory classes, averages conditioned on trajectory classes, and thermodynamic quantities of interest. And, the results include strengthened versions of the Second Law, solving for general free energies from works, and statistically efficient methods for finding those

free energies from data. Additionally, the TCFT provides a general form that subsumes many fluctuation theorems.

The section begins by introducing trajectory classes and relevant relationships and probabilities involving them. Then it establishes the TCFT from these building blocks. It ends with a discussion of the TCFT's scope, suggesting a way to treat zero-probability classes and providing an example use.

A. Trajectory Classes

Every trajectory class is a subset of $\overrightarrow{\mathcal{Z}}$ for which the forward and reverse processes assign probability. Formally, the set of all trajectory classes $\mathcal{C}$ must constitute a σ-algebra over the trajectories $\overrightarrow{\mathcal{Z}}$. However, to stay with the physics, the following does not focus on measure-theoretic details. (The sequel focuses on the latter.) Instead, we first discuss several example classes and thereafter interpret any subset of $\overrightarrow{\mathcal{Z}}$ of practical interest to be part of the assumed σ-algebra and, therefore, to be a trajectory class.

Clearly, the nature of $\mathcal{Z}$ and T, such as whether these sets are discrete or continuous, must determine the form of the trajectories and therefore determine the form of the trajectory classes. However, many intuitive types of trajectory classes exist quite broadly, as illustrated by the following list of Examples:

1. All trajectories that at a given time $t \in T$ are in a specified finite volume of state space,
2. All trajectories that at a given time $t \in T$ are in a particular state,
3. All trajectories that have an entropy difference in a specified finite range of values,
4. All trajectories that have a particular value of the entropy difference,
5. All trajectories: $\overrightarrow{\mathcal{Z}}$, and
6. The singleton $\{\overrightarrow{z}\}$ for any trajectory $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$.

The singleton trajectory classes of Example (6) provides one instance where the model of the system determines whether a trajectory class exists. For a finite number of times T, we can assume the singleton trajectory classes exist. However, for technical reasons, such trajectory classes often fail to exist for continuous-time processes. See Sec. VII for more examples as they apply to known results.

B. Trajectory Class Quantities

For each trajectory class C, we denote the forward and reverse process probabilities as $P(C)$ and $R(C)$, respectively. They are given by:

$$P(C) = \int_C d\overrightarrow{z}\, P(\overrightarrow{z}) \quad \text{and} \quad R(C) = \int_C d\overrightarrow{z}\, R(\overrightarrow{z}) \ .$$

To derive the TCFT for a trajectory class C, we require that $P(C)$ be nonzero. However, classes like Examples (2), (4), and (6) above will often have zero probability. Section III E discusses the use of the TCFT in such cases. Until then, we will assume that $P(C) \neq 0$.

The forward and reverse class-conditioned trajectory probability densities are, for $\overrightarrow{z} \in C$:

$$P(\overrightarrow{z}|C) \equiv \frac{P(\overrightarrow{z})}{P(C)} \quad \text{and} \quad R(\overrightarrow{z}|C) \equiv \frac{R(\overrightarrow{z})}{R(C)} \ ,$$

respectively. The class-conditioned densities vanish for $\overrightarrow{z} \notin C$. When $R(C) = 0$, we allow $R(\overrightarrow{z}|C)$ to be any probability.

We also make frequent use of class-conditioned expectation values:

$$\langle f \rangle_C \equiv \int d\overrightarrow{z}\, P(\overrightarrow{z}|C) f(\overrightarrow{z}) \ ,$$

for arbitrary functions f of $\overrightarrow{\mathcal{Z}}$.

The reverse of a trajectory class C is defined as:

$$C^\dagger = \{\overrightarrow{z}^\dagger | \overrightarrow{z} \in C\} \ .$$

Then the physical reverse probability $R'(C^\dagger)$ of obtaining trajectory class $C^\dagger$ during the reverse experiment is equal to the formal reverse probability $R(C)$:

$$R(C) = R'(C^\dagger) \ .$$

With the above equality, we can then obtain an empirical estimate of $R(C)$ from reverse experiment data.

We then define two important quantities for any class. The *class reverse surprisal* measures how much more surprising an occurrence of class C is in the reverse process than in the forward process:

$$\Theta_C \equiv \ln \frac{P(C)}{R(C)} \ .$$

While Θ_C does not have explicit dependence on any specific trajectory $\overrightarrow{z}$, the *class irreversibility* ψ_C does:

$$\psi_C(\overrightarrow{z}) \equiv \ln \frac{P(\overrightarrow{z}|C)}{R(\overrightarrow{z}|C)} \ .$$

C. Fluctuation Theorem

We now introduce two fluctuation theorems that arise from the preceding setup. Given their close relation, together they constitute the TCFT.

The class reverse surprisal and class irreversibility form a key decomposition of the entropy difference for $\overrightarrow{z} \in C$:

$$\begin{aligned}\Sigma(\overrightarrow{z}) &= \ln \frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} \\ &= \ln \frac{P(C)P(\overrightarrow{z}|C)}{R(C)R(\overrightarrow{z}|C)} \\ &= \Theta_C + \psi_C(\overrightarrow{z}) \ . \end{aligned} \tag{16}$$

Equation (16) can fail in some cases. For, example if $R(C) = 0$ then it fails when the trajectory is outside of the class, which allows nonzero trajectory probability $R(\overrightarrow{z}) \neq 0$. But a trajectory $\overrightarrow{z} \in C$ for which Eq. (16) fails must occur with zero probability in the forward process. So, any instance of Σ can be substituted with $\Theta_C + \psi_C$ in any class-conditioned average. To derive the TCFT, we will only use Σ in class-conditioned averages and so we will treat Eq. (16) as valid in all cases.

The class irreversibility ψ_C takes two important forms. The first when averaged directly; the second when averaging its exponential. When class averaging directly, we obtain:

$$\begin{aligned}\Psi_C &\equiv \langle \psi_C \rangle_C \\ &= \mathrm{D_{KL}}\left[P \,||\, R\right]_C \ , \end{aligned} \tag{17}$$

where:

$$\mathrm{D_{KL}}\left[P \,||\, R\right]_C \equiv \int d\overrightarrow{z} P(\overrightarrow{z}|C) \ln \frac{P(\overrightarrow{z}|C)}{R(\overrightarrow{z}|C)} \ , \tag{18}$$

is the class-conditioned divergence between P and R. It is a nonnegative quantity, being a Kullback-Leibler divergence, that measures how closely the reverse process emulates the forward process when conditioned on the class C.

Directly class averaging the entropy difference of Eq. (16) gives the following.

Theorem 1. *Nominal Class Fluctuation Theorem (NCFT): For any trajectory class C where $P(C) \neq 0$:*

$$\begin{aligned}\langle \Sigma \rangle_C &= \Theta_C + \Psi_C \\ &= \ln \frac{P(C)}{R(C)} + \mathrm{D_{KL}}\left[P \,||\, R\right]_C \ . \end{aligned} \tag{19}$$

When Sec. IV considers refinements of the Second Law, this equality proves its worth in describing the average entropy difference while conveniently isolating the precise trajectory information into the nonnegative class irreversibility.

Turning now to exponential class averages, we have a fruitful identity:

$$\begin{aligned}\left\langle e^{-\psi_C} \right\rangle_C &= \int d\overrightarrow{z} P(\overrightarrow{z}|C) \frac{R(\overrightarrow{z}|C)}{P(\overrightarrow{z}|C)} \\ &= \int d\overrightarrow{z} R(\overrightarrow{z}|C) \\ &= 1 \ . \end{aligned}$$

And, Eq. (16) gives:

$$\begin{aligned}\left\langle e^{-\Sigma} \right\rangle_C &= \left\langle e^{-\Theta_C - \psi_C} \right\rangle_C \\ &= e^{-\Theta_C} \left\langle e^{-\psi_C} \right\rangle_C \ . \end{aligned}$$

Combining these yields the following.

Theorem 2. *Exponential Class Fluctuation Theorem (ECFT):*

$$\begin{aligned}\langle e^{-\Sigma} \rangle_C &= e^{-\Theta_C} \\ &= \frac{R(C)}{P(C)} \ . \end{aligned} \tag{20}$$

The equality's significance lies in relating the entropy difference to the rather simple class reverse surprisal without any possibly-detailed specification of trajectory probabilities beyond the class probabilities. We use it, shortly, to develop straightforward equalities about entropy difference, work, free energy changes, and forward and reverse class probabilities.

D. Scope: Nonzero Probability Classes

The TCFT spans a large collection of fluctuation theorems. By choosing C to be all possible trajectories $\overrightarrow{\mathcal{Z}}$, we obtain the ensemble fluctuation theorems. That is:

$$\begin{aligned}\Theta_{\overrightarrow{\mathcal{Z}}} &= \ln \frac{P(\overrightarrow{\mathcal{Z}})}{R(\overrightarrow{\mathcal{Z}})} \\ &= 0, \end{aligned}$$

and:

$$\Psi_{\overrightarrow{\mathcal{Z}}} = \mathrm{D_{KL}}\left[P \,||\, R\right]_{\overrightarrow{\mathcal{Z}}} \ .$$

So that:

$$\begin{aligned}\langle\Sigma\rangle_{\overrightarrow{\mathcal{Z}}} &= \mathrm{D_{KL}}\left[P \mid\mid R\right]_{\overrightarrow{\mathcal{Z}}} \text{ and} \\ \left\langle e^{-\Sigma}\right\rangle_{\overrightarrow{\mathcal{Z}}} &= e^{-0} \\ &= 1 \ .\end{aligned}$$

These recover the ensemble fluctuation theorems of Eqs. (10) and (14).

Choosing any proper subset C of $\overrightarrow{\mathcal{Z}}$ identifies a new set of FTs—the TCFT applied to the more refined class C. We will consider several types of classes in Secs. V and VI. Also see Sec. VII for examples from the literature.

So consider the opposite extreme, letting $C = \{\overrightarrow{z}\}$ consist of a single trajectory $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$. We require $P(\{\overrightarrow{z}\}) > 0$ in order to apply the TCFT, but note that $P(\{\overrightarrow{z}\})$, the probability of obtaining the particular trajectory $\overrightarrow{z}$ in the forward process, is typically zero for continuous-state or continuous-time processes. Keep in mind that $P(\{\overrightarrow{z}\})$ is distinct from the probability density $P(\overrightarrow{z})$. Then:

$$\Theta_{\{\overrightarrow{z}\}} = \ln \frac{P(\{\overrightarrow{z}\})}{R(\{\overrightarrow{z}\})}$$

and:

$$\begin{aligned}\Psi_{\{\overrightarrow{z}\}} &= \mathrm{D_{KL}}\left[P \mid\mid R\right]_{\{\overrightarrow{z}\}} \\ &= 0 \ .\end{aligned}$$

And, so:

$$\langle\Sigma\rangle_{\{\overrightarrow{z}\}} = \ln \frac{P(\{\overrightarrow{z}\})}{R(\{\overrightarrow{z}\})} \ .$$

Integrating $P(\overrightarrow{z})$ over $\{\overrightarrow{z}\}$ yields $P(\{\overrightarrow{z}\})$, so $P(\{\overrightarrow{z}\}) = P(\overrightarrow{z})d\overrightarrow{z}$. Similarly, $R(\{\overrightarrow{z}\}) = R(\overrightarrow{z})d\overrightarrow{z}$, meaning:

$$\langle\Sigma\rangle_{\{\overrightarrow{z}\}} = \ln \frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} \ .$$

And:

$$\begin{aligned}\langle\Sigma\rangle_{\{\overrightarrow{z}\}} &= \int_{\{\overrightarrow{z}\}} d\overrightarrow{z}' P(\overrightarrow{z}'|\{\overrightarrow{z}\})\Sigma(\overrightarrow{z}') \\ &= \Sigma(\overrightarrow{z}) \ .\end{aligned}$$

From the above equalities, we recover the DFT as expressed in Eq. (7).

E. Scope: Zero Probability Classes

For a sufficiently large trajectory space $\overrightarrow{\mathcal{Z}}$, such as with continuous-time processes, the vast majority of singleton classes $\{\overrightarrow{z}\}$ must have zero probability. This is because $\overrightarrow{\mathcal{Z}}$ is then uncountable and only countably many disjoint events can have nonzero probability. In general, many classes of interest, like those whose trajectories with a specific work value, will have zero probability and so will not be directly subject to the TCFT.

Fortunately, we can still apply the TCFT less directly to a class C such that $P(C) = 0$. One method has practical appeal. Consider a second trajectory class $C' \supset C$ that is nearly identical to C except for extensions in some dimensions of trajectory space such that $P(C') > 0$. For example, if C is all trajectories that pass through a specific state z at a particular time, one might let C' be all trajectories that pass through a small but nonzero-probability neighborhood of states surrounding z at that time. Then, one considers the TCFT applied to the broadened class C' in place of the original class C. Since it is necessary that experimentally-sampled classes have nonzero probability in any case, this approach is attractive. The remaining art is to choose and use an appropriate alternative class C' for the class of interest C.

Carrying this further, a second possibility suggests itself. One that is more satisfying theoretically and yields results involving probability densities. Consider a limiting procedure that applies the TCFT to smaller and smaller classes containing a class of interest. To give one such scheme, consider a trajectory class C and a sequence of classes $C_1, C_2, \ldots$ such that:

- $C_1 \supseteq C_2 \supseteq \ldots$,
- $\bigcap_{n\in\mathbb{N}} C_n = C$, and
- $P(C_n) > 0$ for all n .

Then consider Eqs. (19) and (20) for each C_n and limiting behavior as $n \to \infty$ to evaluate entropy differences for classes with zero probability.

As an example, consider the class C of trajectories with work value $\widetilde{W}$ generated by a process:

$$C = \{\overrightarrow{z} | W(\overrightarrow{z}) = \widetilde{W}\} \ .$$

If the work distribution's values are continuous for the process, then each particular work value has zero probability of occurring. However, consider a class C' that allows a range of works:

$$C' = \{\overrightarrow{z} | \widetilde{W} - \epsilon < W(\overrightarrow{z}) < \widetilde{W} + \epsilon\} \ ,$$

for some $\epsilon > 0$. Such a class generically has nonzero probability and can be used in place of C.

Considering Thm. 2's ECFT with equilibrium privileged

distributions, we have:

$$\left\langle e^{-\beta(W-\Delta F^{\text{eq}})} \right\rangle_{C'} = \frac{R(C')}{P(C')} \ .$$

As ϵ decreases, the work distribution for C' necessarily narrows, so that $e^{-\beta(W-\Delta F^{\text{eq}})}$ approaches $e^{-\beta(\widetilde{W}-\Delta F^{\text{eq}})}$. And, if the work distributions for the forward and reverse processes are continuous functions of work value, then $R(C')$ and $P(C')$ will eventually shrink at the same, constant rate. In this case, $R(C')/P(C')$ converges to a ratio of work densities at $\widetilde{W}$.

This procedure recovers Crooks' work fluctuation theorem [4]:

$$e^{-\beta(W-\Delta F^{\text{eq}})} = \frac{R(W)}{P(W)} \ , \tag{21}$$

where $P(W)$ and $R(W)$ are the probability densities of obtaining work W in the forward and reverse processes, respectively. Note that for a trajectory $\overrightarrow{z}$ with work W, the work under the reverse protocol $\overleftarrow{\lambda}^{\dagger}$ and reverse trajectory $\overrightarrow{z}^{\dagger}$ is $-W$. So $R(W) = R'(-W)$.

In fact, the TCFT introduced here can be strengthened to directly address classes of zero probability without the need for approximations or limiting schemes. This strengthening is done with the measure-theoretical notion of conditional expectation. However, an exposition requires a more thorough treatment of measure theory and so we treat it in the sequel.

IV. STRENGTHENING THE SECOND LAW

Having established the TCFT and outlined how it subsumes existing fluctuation theorems, the following turns attention to bounds on the entropy difference that are similar to but stronger than the Second Law. For these, we need only to determine, experimentally or computationally, the probabilities of trajectories in the forward and reverse processes. First, we find a Second Law for individual trajectory classes. Second, this yields a fluctuation theorem involving multiple trajectory classes that together partition all trajectories. This fluctuation theorem then produces the *Trajectory Partition Second Law* that sets a strictly stronger bound on the ensemble-average entropy difference than the traditional Second Law.

A. Trajectory Class Second Law

Discarding the class-average class irreversibility Ψ_C in Eq. (19)—a nonnegative quantity—gives the *trajectory class second law* (TCSL):

$$\begin{aligned} \langle \Sigma \rangle_C &\geq \Theta_C \\ &= \ln \frac{P(C)}{R(C)} \ . \end{aligned} \tag{22}$$

Thus, the class reverse surprisal Θ_C—a quantity that only depends on $P(C)$ and $R(C)$— bounds the class averaged entropy difference $\langle \Sigma \rangle_C$.

This is similar to Eq. (11)'s Second Law that bounds the ensemble-averaged entropy difference to be nonnegative. However, Eq. (22) is more precise since it uses more information—the class probabilities in the forward and reverse processes. Section IV C elaborates on this advantage over the ensemble Second Law.

Equation (19) says that the average entropy difference $\langle \Sigma \rangle_C$ is close to the class reverse surprisal Θ_C when the class-average class irreversibility Ψ_C is small. Appendix B shows that having such a small class irreversibility is equivalent to the class C having a narrow entropy-difference distribution.

B. Trajectory Partition Fluctuation Theorem

Now partition all trajectories $\overrightarrow{\mathcal{Z}}$ into trajectory classes forming a collection Q. Then averaging Eq. (19) over all classes in Q gives an equality obtained in Ref. [21] that is generalized to arbitrary privileged distributions:

$$\begin{aligned} \langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &= \sum_{C \in Q} P(C) \langle \Sigma \rangle_C \\ &= \sum_{C \in Q} P(C) (\Theta_C + \Psi_C) \\ &= \langle \Theta \rangle_Q + \mathbf{\Psi}_Q \ , \end{aligned} \tag{23}$$

where the *partition averaged class reverse surprisal* is:

$$\begin{aligned} \langle \Theta \rangle_Q &\equiv \sum_{C \in Q} P(C) \Theta_C \\ &= \mathrm{D_{KL}} \left[P \ || \ R \right]_Q \ . \end{aligned}$$

This is a Kullback-Leibler divergence over classes in the partition:

$$\mathrm{D_{KL}} \left[P \ || \ R \right]_Q \equiv \sum_{C \in Q} P(C) \ln \frac{P(C)}{R(C)} \ . \tag{24}$$

And, where the *partition-class averaged class irreversibility* is:

$$\begin{aligned}\mathbf{\Psi}_Q &\equiv \sum_{C \in Q} P(C) \Psi_C \\ &= \sum_{C \in Q} P(C) \, \mathrm{D}_{\mathrm{KL}} \left[P \mid\mid R \right]_C \ ,\end{aligned}$$

a weighted sum of divergences. So, the ensemble average entropy difference decomposes into the mismatch between forward and reverse class probabilities plus the mismatch between specific forward and reverse trajectory probabilities in a class, averaged over all classes.

C. Trajectory Partition Second Law

Since divergences are nonnegative, Eq. (23) leads directly to the *trajectory partition second law* (TPSL) by discarding the partition-class averaged class irreversibility $\mathbf{\Psi}_Q$:

$$\begin{aligned}\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &\geq \langle \Theta \rangle_Q \\ &= \mathrm{D}_{\mathrm{KL}} \left[P \mid\mid R \right]_Q \ ,\end{aligned} \tag{25}$$

where, notably, the RHS leaves out detailed trajectory information, relying only on class probabilities. One can use this expression to bound the entropy difference of a system from a wide array of limited observations of the system. This includes coarse graining time [22, 23] and system state space [15, 22, 23], as well as many other possibilities [24–26].

The information that is left out in going from Eq. (23) to Eq. (25) is the class irreversibility averaged over all trajectories in a class and then averaged over all classes in the partition. So, in accordance with Sec. IV A, if the classes are chosen to have narrow entropy-difference distributions, the class reverse surprisals will tightly bound the ensemble average entropy difference. Specifically, Eq. (25) is a tight-bound.

Contrast this with Eq. (11)'s ensemble Second Law. Since it only states that the average entropy difference $\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}}$ is nonnegative, the trajectory partition second law always provides a nonnegative improvement over the Second Law in estimating ensemble-average entropy differences.

To emphasize, Eq. (25)'s TPSL can be made arbitrarily tight by considering finer and finer partitions Q whose classes have narrower and narrower entropy-difference distributions, independent of the process and how poorly Eq. (11) bounds the average entropy difference.

That said, partitioning trajectories into finer classes complicates solving for and relating class probabilities. Ideally, there is middle ground with a relatively simplified partition composed of classes that carry sufficient information about the trajectory probabilities to tightly-bound the average entropy difference. Reference [14] provides a compelling example of the experimental usefulness of this result, as it uses naturally defined classes to obtain strong work estimates for trajectories in experimentally implemented bit erasure in a flux qubit.

V. FREE ENERGIES VIA CONSTANT-DIFFERENCE CLASSES

Paralleling the bounds on work production derived from the Second Law, Eq. (9) converts the bounds of Sec. IV into statements involving works and trajectory free-energy differences. Thus, determining a bound on the average work required over an arbitrary trajectory class requires obtaining the trajectory free-energy difference. The following shows how to find these free-energy differences given access to the work statistics for the forward experiment and trajectory class statistics for both the forward and reverse experiment. Our only requirement is that the trajectories in the class all have the same free-energy difference. The resulting trajectory class free energy differences provide average work bounds for a wide variety of processes. We then consider an important type of nonequilibrium process—a metastable process—for an example application.

A. Constant Free-Energy Differences

For any pair of forward and reverse state-conditioned processes defined by a forward protocol $\overrightarrow{\lambda}$, a choice of equilibrium privileged distributions $\rho = \pi_0$ and $\sigma = \pi_\tau$ ensures a constant free-energy difference $\Delta f(\overrightarrow{z})$ over all trajectories $\overrightarrow{z}$: the equilibrium free energy change ΔF^{eq}. For nonequilibrium privileged distributions, $\rho \neq \pi_0$ or $\sigma \neq \pi_\tau$, the free-energy difference varies over trajectories. And, this precludes extraction of the free-energy difference from the exponential average entropy difference as was done to obtain Eq. (15) for $\Delta f = \Delta F^{\mathrm{eq}}$. By focusing on trajectory classes of constant free-energy difference, though, we can actually extract these free-energy differences from class averages.

Suppose every trajectory $\overrightarrow{z}$ in class C has the same free energy difference $\Delta f_C = \Delta f(\overrightarrow{z})$. Then, we can extract Δf_C from the class average of the exponential entropy difference:

$$\begin{aligned}\left\langle e^{-\Sigma} \right\rangle_C &= \left\langle e^{-\beta(W - \Delta f)} \right\rangle_C \\ &= e^{\beta \Delta f_C} \left\langle e^{-\beta W} \right\rangle_C \ .\end{aligned}$$

Then, the ECFT Eq. (20) gives:

$$\Delta f_C = -\beta^{-1} \ln \langle e^{-\beta W} \rangle_C + \beta^{-1} \ln \frac{R(C)}{P(C)} \ . \quad (26)$$

This equality relates free-energy differences to statistics on works and class probabilities. Jarzynski's Equality Eq. (15) is the special case where $C = \overrightarrow{\mathcal{Z}}$, $\rho = \pi_0$, and $\sigma = \pi_\tau$.

B. Metastable Process Work Bounds

Having established Eq. (26), we now demonstrate its application to bounding the ensemble average work of a particular type of process that we call a *metastable process*.

As a motivating example, consider an information-storing biomolecule whose configurational space is too complex to fully model but which has a coarser description of state that is robust to thermal noise, like the various functional configurations of a protein or RNA molecule. With the results here, one can obtain refined bounds for the work production in altering the occupancy of these coarsened states along with free energies associated with these states, without knowing the exact details of the underlying Hamiltonian. A similar analysis to ours was done to obtain the change in free energy of RNA through stretching [8, 10]. However, our procedure comes with the added possibility of treating distributions over multiple possible coarsened states.

1. Metastable Processes

Consider an energetic landscape E_t at time t that is partitioned into regions—*metastable regions*—each separated by high energy barriers. For a system contained in any such region, the barriers severely limit the chance of escape over long timescales. Each metastable region therefore represents an information-storing mesostate, or *memory state*, that robustly constrains the system. Also, for any system state distribution that has support over exactly one metastable region m, the system will relax to a stable distribution l_t^m over the region much faster than the timescale of escape if the energetic landscape is left unperturbed. While not the true, global equilibrium distribution π_t, which generally has support over all metastable regions, we call l_t^m the local equilibrium distribution for m.

Now, prepare the system in arbitrary distributions over all of phase space and then allow the system to locally equilibrate in E_t. We call the system state distribution κ obtained after local equilibration a *metastable distribution*. Within each metastable region m, κ must match l_t^m up to normalization and, therefore, the equilibrium distribution in that region. Thus, we have:

$$\kappa(z) = \kappa(m(z)) \frac{\pi_t(z)}{\pi_t(m(z))} \ ,$$

where $m(z)$ is the metastable region for microstate z, the probability of the system being in the metastable region m is defined as $\kappa(m) \equiv \int_m dz\, \kappa(z)$, and $\pi_t(m) \equiv \int_m dz\, \pi_t(z)$ is the equilibrium probability of being in the region m.

A metastable process is then a forward process where (i) the initial and final energetic landscapes E_0 and E_τ can each be partitioned into metastable regions and (ii) the initial distribution ρ is metastable over E_0.

2. Metastable Free Energies

For a metastable distribution κ, all system states in a metastable region m have the same free energy. That is, for $z \in m$:

$$\begin{aligned} f(z; \kappa, E_t) &= E_t(z) + \beta^{-1} \ln \kappa(z) \\ &= E_t(z) + \beta^{-1} \ln \kappa(m(z)) \frac{\pi_t(z)}{\pi_t(m(z))} \\ &= E_t(z) + \beta^{-1} \ln \frac{\kappa(m(z))}{\pi_t(m(z))} e^{-\beta(E_t(z) - F_t^{\text{eq}})} \\ &= \beta^{-1} \ln \frac{\kappa(m(z))}{\pi_t(m(z))} + F_t^{\text{eq}} \ . \end{aligned}$$

Thus, the free energy for a locally equilibrated distribution over a metastable region is the free energy of any state in the region. We call such a free energy a *metastable free energy*.

This further simplifies if we identify the *memory-state free energy*:

$$F_t^{\text{mem}}(m) \equiv F_t^{\text{eq}} - \beta^{-1} \ln \pi_t(m) \ , \quad (27)$$

as the fixed contribution of a particular memory state to the free energy, regardless of the metastable distribution κ. The free energy is thus the free energy of the memory state plus the surprisal of that memory state:

$$f(z; \kappa, E_t) = F_t^{\text{mem}}(m(z)) + \beta^{-1} \ln \kappa(m(z)) \ .$$

When averaged over all metastable regions with distribution κ, this returns a familiar expression [15] for average nonequilibrium free energy:

$$\langle f \rangle_{\mathcal{Z}}(\kappa, E_t) = \sum_m \kappa(m) F_t^{\text{mem}}(m) - \beta^{-1} H_M(\kappa) \ ,$$

where $H_M(\kappa) \equiv -\sum_m \kappa(m) \ln \kappa(m)$ is the Shannon entropy of κ over memory states—the average amount of information they store.

This decomposition offers an entrée to the problem of evaluating work production of a process whose control protocol $\vec{\lambda}$ must start in a particular initial configuration $\lambda(0)$ and end in a particular final configuration $\lambda(\tau)$. If their respective energetic landscapes E_0 and E_τ are not well understood, we can still extract bounds on work production using the TCFT.

We wish to derive an ensemble average free-energy difference for such a metastable process so that we can bound the work invested in the forward experiment that exploits the simplicity of metastable free energies. Of course, different choices of the second privileged distributions σ result in different free-energy differences, but we will consider the metastable distribution that corresponds to the final-time distribution of the metastable process. That is, we choose σ to be the distribution of the system if, holding the energetic landscape fixed at E_τ at the end of the protocol, the system locally-equilibrated after the end of the forward process:

$$\sigma(z) = \rho_\tau(m(z)) \frac{\pi_\tau(z)}{\pi_\tau(m(z))} \ .$$

We say that σ is then the locally-equilibrated distribution of ρ_τ. The resulting free-energy difference Δf is then called the *metastable free energy change* for the metastable process:

$$\begin{aligned} \Delta f(\overrightarrow{z}) = F_\tau^{\text{mem}}(m'(\overrightarrow{z}(0))) - F_0^{\text{mem}}(m(\overrightarrow{z}(\tau))) \\ + \beta^{-1} \ln \frac{\sigma(m'(\overrightarrow{z}(\tau)))}{\rho(m(\overrightarrow{z}(0)))} \ , \end{aligned}$$

where $m(z)$ and $m'(z)$ are the memory states containing z in E_0 and E_τ, respectively. Using Eq. (12), we then have:

$$\begin{aligned} \langle W \rangle_{\overrightarrow{z}} &\geq \langle \Delta f \rangle_{\overrightarrow{z}} \\ &= \langle \Delta F^{\text{mem}} \rangle_{\overrightarrow{z}} - \beta^{-1} \Delta H_M \ . \end{aligned} \tag{28}$$

The first term captures the free energy contribution of each state and is specific to the particular physical instantiation of the memory. The second term is characteristic of how the system's distribution over metastable regions transformed. If the memory states all had the same free energy then the first term would be zero, giving Landauer's bound. Generally, Eq. (28) falls short of the free-energy change bound on the average work, Eq. (13), because the actual final-time free energy, that of ρ_τ, is generally higher than the free energy of the locally-equilibrated σ.

3. Obtaining Memory-State Free Energies

Consider a system and a pair of initial and final protocol configurations $\lambda(0)$ and $\lambda(\tau)$, respectively. If each of these configurations contains metastable regions, capable of storing useful information, it is worthwhile considering the family of thermodynamic experiments that execute computations, stochastically mapping between different metastable regions of these end-points. As we will show, any experiment that begins in a metastable distribution with the boundary conditions above obeys strong bounds on work production related to the metastable free energy.

This is useful since we can choose one or a small number of such processes to study in detail to obtain the memory state free energies and then, by Eq. (28), the average work for any computation implemented between these control points is simply determined by the initial and final memory state distributions.

Consider an initial metastable region m and final metastable region m' and the associated trajectory class that connects them:

$$C_{m,m'} = \{ \overrightarrow{z} \,|\, \overrightarrow{z}(0) \in m, \overrightarrow{z}(\tau) \in m' \} \ .$$

All trajectories within one class must all have the same free energy difference if we choose the privileged distributions ρ and σ of our forward and reverse experiment to be metastable. Practically, this can be experimentally implemented by allowing the system to relax to local metastable equilibria before executing the control protocol. If the metastable regions are appropriately chosen, this process can be much faster than relaxation to global equilibrium.

As a result, we can express the free-energy difference for this trajectory class in terms of the input and output memory states by considering $\overrightarrow{z} \in C_{m,m'}$:

$$\begin{aligned} \Delta f_{C_{m,m'}} &= f(\overrightarrow{z}(\tau); \sigma, E_\tau) - f(\overrightarrow{z}(0); \rho, E_0) \\ &= F_\tau^{\text{mem}}(m') - F_0^{\text{mem}}(m) + \beta^{-1} \ln \frac{\sigma(m')}{\rho(m)} \ . \end{aligned}$$

In the special case where $\rho(m) = \sigma(m')$, the change in memory state free energy is equal to the free-energy difference.

Regardless of the metastable privileged distributions used, we recover memory state free-energy differences via Eq. (26), using the work production probabilities and memory state probabilities:

$$\begin{aligned} &F_\tau^{\text{mem}}(m') - F_0^{\text{mem}}(m) \\ &\quad = -\beta^{-1} \ln \left\langle e^{-\beta W} \right\rangle_{C_{m,m'}} + \beta^{-1} \ln \frac{R(C_{m,m'}) \rho(m)}{P(C_{m,m'}) \sigma(m')} \ . \end{aligned}$$

If the protocol is cyclical, such that $E_\tau = E_0$, then we can fully obtain all memory state free-energy differences with $|\mathcal{M}| - 1$ trajectory classes, where $\mathcal{M}$ is the set of metastable regions of the initial-final energetic landscape. Otherwise, the memory state free-energies can be determined by considering $|\mathcal{M}_1| + |\mathcal{M}_2| - 1$ trajectory classes, where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the sets of metastable regions of the initial and final landscapes.

With the free energy landscapes determined, it becomes straightforward to set strong bounds not only on the process that resulted from the original experiment with $\vec{\lambda}$ and ρ, but on *any* experiment that has the same initial and final protocol configurations and begins in a metastable distribution. Suppose the initial memory state distribution of the process is q. Let the resultant final memory state distribution of the process be q':

$$q'(m') = \sum_m q(m) p_{m \to m'} \ ,$$

where $p_{m \to m'}$ is the probability of the system ending in m' given that it started in m. Then, by Eq. (28):

$$\begin{aligned} \langle W \rangle_{\vec{\mathcal{Z}}} \geq & \sum_{m'} q'(m') F_\tau^{\text{mem}}(m') - \sum_m q(m) F_0^{\text{mem}}(m) \\ & - \beta^{-1} (H_M(q') - H_M(q)) \ . \end{aligned}$$

With this section's results, it is possible to obtain strong work bounds on computations even when the memory-state free energies are unequal or unknown at the outset. This significantly strengthens the Second Law as applied to the thermodynamics of computation.

VI. STATISTICAL FREEDOM FROM THE TYRANNY OF THE RARE

When estimating statistical quantities from data, rare events can dominate sample averages [12, 13]. This can be particularly problematic when the events are associated with large resources. Consider the following case in point. By empirically estimating the exponential average work $\left\langle e^{-\beta W} \right\rangle_{\vec{\mathcal{Z}}}$ for a thermodynamic transformation, one can estimate the equilibrium free energy difference ΔF^{eq} via Eq. (15)'s Jarzynski's Equality. However, this can require thorough sampling of very rare events [12]. The ECFT of Eq. (20) can aid in solving this statistical challenge by removing consideration of these rare but work-dominant trajectories.

Specifically, if the privileged starting distributions of both the forward and reverse experiments are in equilibrium, then the free energy difference Δf_C of a class C (shown in Eq. (26)) is the change in free energy ΔF^{eq}, regardless of the chosen class. However, given a set of N forward and reverse experiments and associated work data for the forward experiment $\vec{W} = \{W_1, W_2, \cdots W_N\}$, statistical fluctuations in the data lead to fluctuations in trajectory class probability estimates $\tilde{P}(C)$ and $\tilde{R}(C)$, where the tilde indicates an estimate. This results in statistical fluctuations in free-energy difference estimates that depend on the trajectory class:

$$\Delta \tilde{f}_C \equiv -\beta^{-1} \ln \left(\frac{\sum_{i=1}^N \delta_{W_i \in C} e^{-\beta W_i}}{\sum_{i=1}^N \delta_{W_i \in C}} \right) + \beta^{-1} \ln \frac{\tilde{R}(C)}{\tilde{P}(C)}, \tag{29}$$

where $\delta_{W_i \in C}$ returns 1 if the ith work value is realized within the trajectory class C and 0 otherwise. If we had perfect statistics, these estimates would all be the actual change in free energy ΔF^{eq}. However, as the next section shows (and as App. C further explains), we can find better estimators of the free energy change by choosing more probable trajectory classes.

A. Tyranny of the Rare

Consider the following example 1D system. It is in contact with a thermal environment at inverse temperature β but otherwise obeys classical mechanics under a time-evolving potential energy landscape. There exist two regions in state space, A and B, each with potential energy that is constant over their regions. Arbitrarily high barriers separate and surround the regions so that all particles in a given region stay there. These two potential-energy wells start at energies:

$$E_0(A) = -\beta^{-1} \ln(1 - \epsilon)$$

and:

$$E_0(B) = -\beta^{-1} \ln(\epsilon) \ ,$$

respectively, where $0 < \epsilon \ll 1$ and the energy everywhere else is arbitrarily large.

Start the system in equilibrium over the two wells so that the probabilities of starting in the wells are:

$$P(A) = 1 - \epsilon$$

and:

$$P(B) = \epsilon \ .$$

Now, raise well A and lower well B to end at $E_\tau(A) = E_\tau(B) = \beta^{-1} \ln 2$ energy. Consider a trajectory class for trajectories that are wholly in the A well during the

process and similarly a class for the B well. We refer to the classes synonymously with their associated wells.

The work invested for either class is simply the change in energy of the corresponding well since the energy barrier is high enough that system states do not cross between wells during the control protocol:

$$\begin{aligned} W_A &= \beta^{-1} \ln(2(1-\epsilon)) \\ W_B &= \beta^{-1} \ln(2\epsilon) \ . \end{aligned}$$

The resulting integral fluctuation theorem yields:

$$\left\langle e^{-\beta W} \right\rangle_{\overrightarrow{\mathcal{Z}}} = P(A) \left\langle e^{-\beta W} \right\rangle_A + P(B) \left\langle e^{-\beta W} \right\rangle_B \ ,$$

where:

$$\begin{aligned} P(A) \left\langle e^{-\beta W} \right\rangle_A &= (1-\epsilon) \frac{1}{2(1-\epsilon)} \\ &= \frac{1}{2} \ , \end{aligned}$$

and:

$$\begin{aligned} P(B) \left\langle e^{-\beta W} \right\rangle_B &= \epsilon \frac{1}{2\epsilon} \\ &= \frac{1}{2} \ . \end{aligned}$$

So, the total exponential average work is:

$$\left\langle e^{-\beta W} \right\rangle_{\overrightarrow{\mathcal{Z}}} = 1 \ ,$$

and, thus, by Eq. (15) the equilibrium free energy change vanishes.

In this situation, the probabilities of the two classes are highly uneven with class B having only probability ϵ. However, B accounts for $1/2$ of the total exponential average work. This means that an accurate statistical estimate of the change in equilibrium free energy via sampling such a process and using Eq. (15) is highly dependent on rare events. Specifically, even though the true free energy change is 0, it would likely be estimated as $\Delta \tilde{F}^{\text{eq}} = \beta^{-1} \ln 2$ with small enough ϵ, as can be seen by the green histogram in Fig. 1. It converges to the true value only with very large samples of the rare class B. Thus, the variance of estimated values for the change in equilibrium free energy is large for finitely sampled experiments. Typically, estimates will be misleading.

B. Circumventing Tyranny

The TCFT solves this problem using appropriate trajectory classes. In principle, we may consider a process with arbitrary privileged distributions for both the forward and reverse processes. Thus, we can estimate arbitrary free-energy differences for a process, as long as we restrict consideration to a trajectory class C of constant free-energy difference, as described in Sec. V A. However, constant free-energy differences are automatically guaranteed for any class when we choose equilibrium privileged distributions for both the forward and reverse processes. The privileged distribution for the forward process described above was indeed equilibrium and we choose an equilibrium distribution for the corresponding reverse process as well.

Now, focus on Eq. (26), moving away from Jarzynski's Equality in Eq. (15). This expands the required estimators from simply the exponential average work to also include the forward and reverse process probabilities of class C. Moreover, the exponential average work is now conditioned on C. Thus, the estimator is a function of sampled data that comes from class C in both the forward and reverse experiments. However, we need C to be such that the class average of the exponential work, the forward probability of the class, and the reverse probability of the class are statistically easy to estimate.

This is the case when sampling trajectories from class A in the example above. First, note that A has very high probability $1-\epsilon$, so estimating its log-probability from data is statistically easy. Second, its class average exponential work is also easily estimated since the class itself is highly likely and its work distribution is narrow.

This leaves estimating the reverse class probability. In this, we choose our second privileged distribution to be the equilibrium distribution for the final-time energy landscape and so solve for the equilibrium free-energy change, as desired. This, then, fully specifies the reverse process. Since the true equilibrium free-energy change vanishes, Eq. (26) says that the reverse-process probability of class A is $R(A) = 1/2$. In this way, since A is likely in the reverse process, it too is easily estimated.

To verify that this reasoning is sound, consider estimates obtained from various numbers N of forward and reverse process trajectories for the three trajectory classes $\overrightarrow{\mathcal{Z}}$, A, and B. (See App. C for an explanation of the base calculations.) Figure 1 shows that the distribution of free energy estimates when using all trajectories $C = \overrightarrow{\mathcal{Z}}$ is heavily weighted towards $\Delta \tilde{f}_{\overrightarrow{\mathcal{Z}}} = \beta^{-1} \ln 2$ for $\epsilon = 0.01$ and a small data set of work values. However, if we restrict to the trajectory class that starts and ends in the most likely state A, then we see that free energy estimates are more closely centered around the correct value of $\Delta F^{\text{eq}} = 0$. This suggests that restricting to high probability regions of the work distribution improves free energy estimates.

Figure 2 further quantifies this advantage by plotting the average difference with the correct free energy $\langle \Delta \tilde{f}_C -$

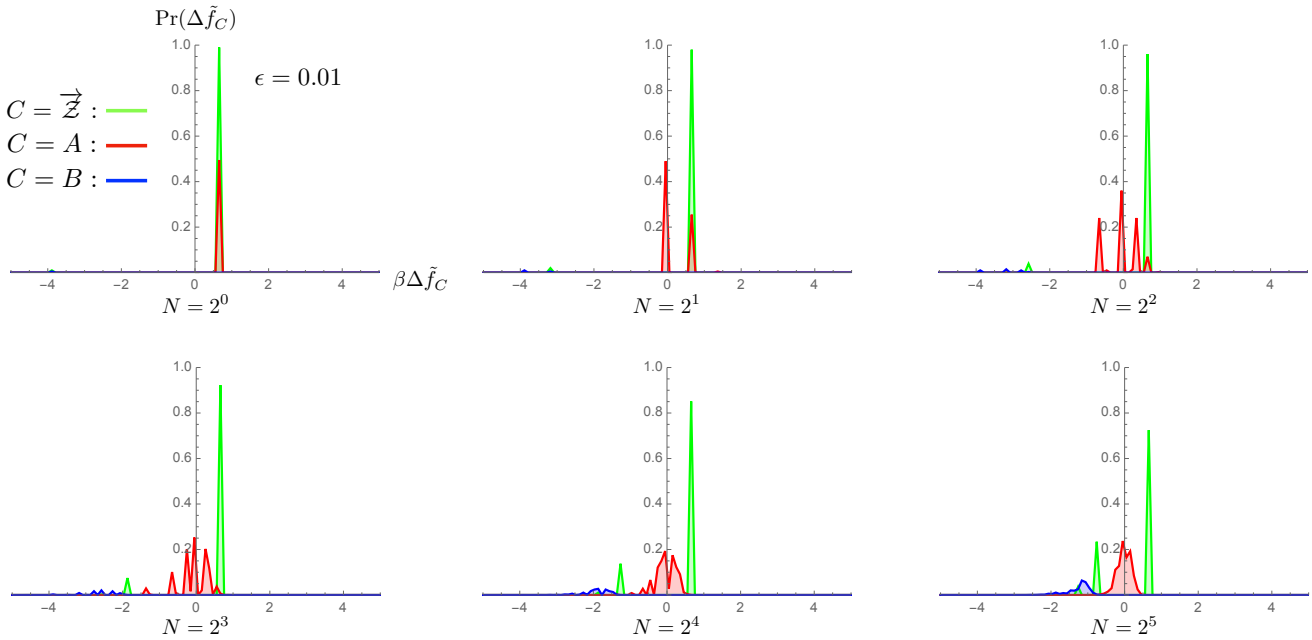

FIG. 1. Tyranny of the rare: Distribution of free energy estimates $\Delta\tilde{f}_C$ arising from N forward and N reverse experiments with $\epsilon = 0.01$ depends sensitively on the trajectory class C. Above, plots of free energy estimates for the trajectory class that includes all paths (green: $C = \text{all}$), the trajectory class that starts and ends in A (red: $C = A$), and the trajectory class that starts and ends in B (blue: $C = B$). Note that we do not plot divergent free energy estimates, for which we estimate $\tilde{P}(C) = 0$ or $\tilde{R}(C) = 0$. Both trajectory classes A and B can yield divergent estimates, but A often provides better estimates than $C = \text{All}$.

$\Delta F^{\text{eq}}\rangle$ and mean square deviation $\langle(\Delta\tilde{f}_C - \Delta F^{\text{eq}})^2\rangle$. We see a marked advantage to our restricted trajectory class $C = A$ over the full set of trajectories in both cases.

By contrast, restricting to the rare event $C = B$, Fig. 1 shows that the majority of the free energy estimates are divergent for a small data set of work values. However, even when considering only the nondivergent estimates, Fig. 2 shows that the rare trajectory class $C = B$ leads to the worst free energy estimates.

The resulting estimate of the free-energy change via the TCFT is indeed much more statistically robust and parsimonious. Reference [13] gives another analysis focusing on improvements in reducing the bias. The approach detailed above shows how we can reduce the bias as well reduce the variance.

VII. RELATED RESULTS

We now turn to the burgeoning collection of previously established fluctuation theorems. As noted, many existing fluctuation theorems are special cases of the TCFT.

For most of the results and setups considered in the following, the trajectory classes constructed can have zero probability. Of course, the TCFT is then not to be used directly. Instead, the methods of Sec. III E can be used to find either approximate trajectory classes or a limiting procedure that will then yield the exact result in question. To avoid redundancy, we simply assume a limiting procedure when the trajectory class has zero probability.

A. Measurement and Feedback for Free Energy Estimation

Reference [13] focuses on the problem of estimating free energy differences in the face of rare yet resource-dominant events. This was the subject of Sec. VI. By taking measurements of the system state at some number of times and rejecting or accepting data samples based on those measurement results, statistical estimation of free energies can be improved with the use of their Eq. (9).

Specifically, suppose that $A = \{A_1, \ldots, A_N\}$ is a set of regions of system state space. For times $T' = \{t_1, \ldots, t_N\}$ during the protocol, we measure whether the system occupies region A_i at time t_i for each i. Let the class C be trajectories that occupy A_i at time t_i for each i. Applying Eq. (20)'s ECFT to C, we derive their Eq. (9). They show that an estimator of the free-energy difference over a process can have a smaller bias when using their Eq. (9) and an appropriate set of measurement choices compared to using the Jarzynski Equality.

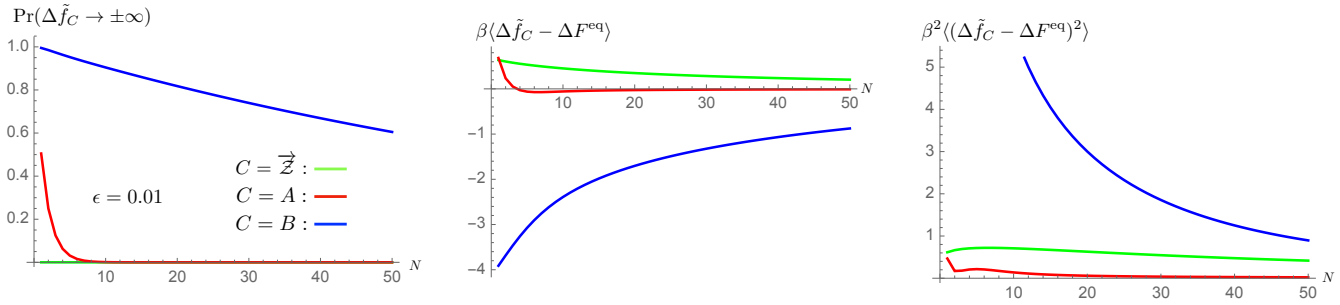

FIG. 2. Degree to which energy estimates $\Delta\tilde{f}_C$ diverge from the actual change in free energy ΔF^{eq} depends on the trajectory class: For $\epsilon = 0.01$, the probability of an infinitely divergent free energy estimate $\Pr(\Delta\tilde{f}_C \to \pm\infty)$ is significant for the trajectory class B, nonzero and swiftly decreasing for A with larger data, and zero for the set of all trajectories. Despite this advantage in using all trajectories to estimate free energy, both the average difference $\langle\Delta\tilde{f}_C - \Delta F^{\text{eq}}\rangle$ and the mean squared deviation $\langle(\Delta\tilde{f}_C - \Delta F^{\text{eq}})^2\rangle$ are minimized by the class $C = A$ by excluding the rare event B. These plots come from excluding infinitely divergent estimates, which represent an exponentially small likelihood for A as N increases. Looking at a reduced trajectory subspace A gives consistently improved estimates for small amounts of data.

Since C incorporates an arbitrary number of arbitrary system state measurements, their Eq. (9) is a rather broad fluctuation theorem. The key difference between Eq. (20) and their Eq. (9) is that Eq. (20) allow (i) an infinite number of specifications on the trajectories and (ii) more general types of trajectory specifications (e.g., work values) than instantaneous descriptions of the system state.

B. Phase Space Perspective on Dissipated Work

Reference [27] achieves several important results that can be understood as special cases of the TCFT. They consider a system that only interacts with a control device during the forward and reverse experiments. Note that this is a special case of our assumptions where the system has negligible or no interaction with the thermal environment. Also note that, as their examples illustrate, this still allows a system to be composed of two subsystems with one acting as a thermal environment for the other. In these cases, the TCFT can be applied to either the entire system or the latter subsystem. We apply it to the entire system to follow their core development.

Set the state space $\mathcal{Z}$ to be the microstates of the system and the modeled times T to be $[0, \tau]$. This ensures a deterministic, Hamiltonian evolution of the system, as they describe. Set ρ and σ to the initial and final Boltzmann distributions at the same inverse temperature β. So, the resulting free-energy differences of all trajectories are equal to the equilibrium free-energy difference ΔF^{eq}. The dissipated work $\langle W\rangle_{\overrightarrow{\mathcal{Z}}} - \Delta F^{\text{eq}}$ is then the average heat that would be dissipated from the system to a thermal environment at inverse temperature β if the system was equilibrated with the environment after the forward protocol while held at energetic landscape E_τ.

Then consider a partition $\{\chi_1, \ldots \chi_K\}$ of $\mathcal{Z}$ and a time t. For each j in $\{1, \ldots, K\}$, let C_j be the set of system state trajectories that occupy χ_j at time t. Then let ρ_j be the probability the system is in χ_j at time t in the forward experiment. Then:

$$\rho_j = P(C_j) \ .$$

Set $\widetilde{\chi}_j = \chi_j^\dagger = \{z^\dagger \mid z \in \chi_j\}$ and let $\widetilde{\rho}_j$ be the probability the system is in $\widetilde{\chi}_j$ at time $\tau - t$ in the reverse experiment. (They denote this time t, keeping all references of time to be relative to the forward experiment.) Then:

$$\begin{aligned}\widetilde{\rho}_j &= R'(C_j^\dagger) \\ &= R(C_j) \ .\end{aligned}$$

Applying Eq. (20)'s ECFT to C_j yields Ref. [27]'s Eq. (6). Eq. (22)'s TCSL applied to C_j yields their Eq. (7). And, Eq. (25)'s TPSL applied to the partition of trajectories $Q = \{C_1, \ldots, C_K\}$ yields their Eq. (8).

These results, especially their Eq. (8), were used to provide concise expressions involving work and free energies for simple but instructive processes, as well as to derive Landauer's bound.

C. Work Dissipation as the Distance from Equilibrium

Reference [28] obtains an inequality between the dissipated work up to any time t during a protocol and how far the system's state density at time t must be from equilibrium. They assume that the system dynamics are Markovian and that the system equilibrates, at any time

t, towards the Boltzmann distribution for E_t and β, if the protocol is suddenly interrupted at time t and the system is held under energetic landscape E_t. These assumptions are met in our model if the thermal environment has such a high relaxation rate that it is effectively memoryless as far as the influence on the system is concerned. They also assume that ρ and σ are the initial and final Boltzmann distributions. Let $\overrightarrow{\lambda}$ be the forward protocol, which runs from time 0 to τ. Let $\mathcal{Z}$ be the system microstates and $T = [0, \tau]$.

We derive their results from the TCFT by considering a separate protocol $\overrightarrow{\lambda}^t$ for any time t in $[0, \tau]$. $\overrightarrow{\lambda}^t$ runs from time 0 to time $2t$. For $t' \leq t$, $\overrightarrow{\lambda}^t(t') = \overrightarrow{\lambda}(t')$. For $t' > t$, $\overrightarrow{\lambda}^t(t') = \overrightarrow{\lambda}(t)$. Thus $\overrightarrow{\lambda}^t$ follows $\overrightarrow{\lambda}$ until time t, at which point $\overrightarrow{\lambda}^t$ holds fixed until it ends at $2t$. The forward process privileged distribution ρ^t for the protocol $\overrightarrow{\lambda}^t$ is simply set to the Boltzmann distribution ρ. At time t, denote the probability of being in state z as $\rho(z, t)$, which must be shared between both protocols $\overrightarrow{\lambda}$ and $\overrightarrow{\lambda}^t$ since the two protocols do not differ until time t.

The reverse process privileged distribution for $\overrightarrow{\lambda}^t$ is set to be Boltzmann with respect to E_{2t}. Since the protocol $\overrightarrow{\lambda}^t$ is fixed between times t and $2t$, this is also the physical-reverse state distribution at all times between 0 and t during the physical-reverse process. In particular, the state distribution for the physical-reverse process of protocol $\overrightarrow{\lambda}^t$ at time t is:

$$\rho^{\text{eq}}(z, \overrightarrow{\lambda}(t)) = e^{-\beta(E_t(z) - F_t^{\text{eq}})} \ .$$

Let C_z^t be the set of trajectories that occupy microstate z at time t. Then:

$$\rho(z, t) = P(C_z^t) \ ,$$

where P refers to the forward process of protocol $\overrightarrow{\lambda}^t$. And:

$$\begin{aligned} \rho^{\text{eq}}(z, \overrightarrow{\lambda}(t)) &= R'(C_z^t) \\ &= R(C_z^t) \ , \end{aligned}$$

where R' and R refer to the reverse process (physical and formal representations, respectively) of protocol $\overrightarrow{\lambda}^t$.

Let $W(t)$ denote the work conducted up to time t. The exponential average work up to time t conditioned on the system occupying state z at time t is $\left\langle e^{-\beta W(t)} \right\rangle_{z,t}$, during either protocol $\overrightarrow{\lambda}$ or $\overrightarrow{\lambda}^t$. Since no additional work is conducted under the protocol $\overrightarrow{\lambda}^t$ after time t, this quantity is then:

$$\left\langle e^{-\beta W(t)} \right\rangle_{z,t} = \left\langle e^{-\beta W} \right\rangle_{C_z^t} \ ,$$

where the second average is taken over the forward process for protocol $\overrightarrow{\lambda}^t$.

Then, applied to the protocol $\overrightarrow{\lambda}^t$ and using trajectory class C_z^t and the above equalities, Eq. (20)'s ECFT yields Ref. [28]'s Eq. (6). Equation (22)'s TCSL yields their Eq. (8) and Eq. (25)'s TPSL yields their Eqs. (2) and (9).

D. Work and State Fluctuation Theorem

Consider a process where ρ and σ are the initial and final Boltzmann distributions. Reference [9] establishes a fluctuation theorem that relates a work value, the equilibrium free-energy difference, and forward and reverse process probabilities for obtaining the work value and particular values for two functions of state. The two functions of state are evaluated at opposite ends of the trajectory. In their notation, this is written as:

$$P_{\text{F}}(\widetilde{W}, a \to b) e^{-\beta \widetilde{W}} = P_{\text{R}}(-\widetilde{W}, b^* \to a^*) e^{-\beta \Delta F^{\text{eq}}} \ ,$$

which is their Eq. (1). (Except that we use $\widetilde{W}$ for a specific work value.) Here, a and b are some output values of the two respective functions of state and a^* and b^* are the output values of the time reverse of system states that output a and b. P_{F} and P_{R} are the forward and reverse processes.

Let C be all trajectories (i) that obtain a work value W, (ii) whose state evaluates the first state function to a, and (iii) whose end state evaluates the second state function to b. Then applying the ECFT Eq. (20) to C gives their Eq. (1).

The equality was used to efficiently estimate the conformational free energy change of a simulated alanine dipeptide.

E. Landauer's Bound on Erasure Dissipation

Finally, Ref. [29] considers the process of erasing information implemented using a Brownian silica bead trapped in an optical tweezer. The laser initially induces an effective double-well potential that is symmetric across the center. They call the two wells 0 and 1. Manipulating the laser and moving the platform containing the bead, a protocol is realized that, while ending with the effective potential in the initial form, shifts the bead to the 0 well with high probability. This was obtained with a variety of specific initial conditions and protocol variations. Since the bead always starts in either well with 50% probability, this then demonstrates erasing one bit of information via a variety of protocols.

They also demonstrate that the average work required to conduct these protocols was always near $\beta^{-1} \ln 2$, verifying Landauer's Bound. To explain why Landauer's bound should hold, they utilize Ref. [28]'s Eq. (6) to produce the following:

$$\left\langle e^{-\beta W} \right\rangle_{\to 0} = \frac{1/2}{P_S} \ , \tag{30}$$

and:

$$\left\langle e^{-\beta W} \right\rangle_{\to 1} = \frac{1/2}{P_S} \ , \tag{31}$$

where P_S is the probability of a trajectory successfully ending in the 0 state, and $\to 0$ ($\to 1$) denotes an average conditioned on ending in the 0 (1) state. In particular, applying Jensen's inequality to Eq. (30) yields:

$$\langle W \rangle_{\to 0} \geq \beta^{-1} (\ln 2 + \ln P_S) \ ,$$

which is a generalization of Landauer's Bound for imperfect erasure.

We deduced Ref. [28]'s Eq. (6) from the TCFT already, but Eqs. (30) and (31) can be obtained from the TCFT directly and quickly. Since the bead starts off equilibrated over each well and has equal probability to start in either, the initial distribution ρ is equilibrium. Choose σ to also be equilibrium. Then the free energy difference is ΔF^{eq}, which must be zero since the effective potential ends as it begins. Then let C_0 be all trajectories that end in 0 and C_1 be all that end in 1. Applying the ECFT Eq. (20) first to C_0 and then to C_1 yields Eqs. (30) and (31), respectively.

VIII. DISCUSSION

The preceding provided guidance when selecting appropriate trajectory classes for a given process of interest. To achieve a much stronger bound on entropy difference than the traditional ensemble-average Second Law, Sec. IV C proposed choosing a partition of all trajectories into classes that resulted in narrow entropy difference distributions for each class. And, to avoid dominating rare events when calculating free-energy differences (Sec. VI), we recommended classes that are common in both the forward and reverse processes and that have narrow entropy-difference distributions. That said, the most effective classes for these tasks appear to be specific to the particular processes of interest. Developing procedures to identify these classes for arbitrary processes remain an open problem.

What does this look like in practice? Reference [14] experimentally investigated efficient bit erasure in a nanoscale flux qubit device. We found a natural partition of trajectories into classes that served well both to strengthen the Second Law and to dissect the entire process work distribution. The latter identified simple components characterizing the full work distribution's features—features functionally critical to efficient bit erasure.

Helpfully, the TCFT can be derived under less severe restrictions on the system than Sec. II A assumed. For example, the nature of the external influence that we called the control device can be allowed to instantiate nonconservative forces on the system. Moreover, the thermal environment does not need to be fixed at inverse temperature β for the duration of the protocol. That is, we need not require that the steady state of the system is in equilibrium nor that all forces acting on the system are conservative. Relaxing other assumptions is possible too. The essential equality needed for a version of the TCFT to also hold is one like the DFT:

$$\frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} = g(\overrightarrow{z}) \ ,$$

where g is some function of trajectory. Then a version of the TCFT holds where the exponential entropy difference $e^{-\Sigma}$ is replaced with the function g. This follows by the same logic as presented in Sec. III. However, how such a generalization of our presentation can be used remains to be explored.

IX. CONCLUSIONS

We presented the TCFT's core theory. With it, we detailed a path to solving for free-energy differences more efficiently than before. We also showed how to strengthen the Second Law in the presence of impoverished knowledge of any nonequilibrium and dissipative process. This led to a suite of new results that further advanced our understanding of how fluctuations underpin nonequilibrium thermodynamics.

We also showed how the TCFT fits more broadly within the ranks of fluctuation theorems. It unifies many previously known, but distinct results, spans the detailed and integral fluctuation theorems, and is rooted in the same conceptual foundation of time symmetries on small dynamical systems.

Follow-on efforts will expand on the way that the TCFT breaks free energy estimation from the tyranny of rare events. This will also clarify the role of metastable free energies in describing experimentally inaccessible free energies of interest and related thermodynamic costs. Via

explicit examples, this will also showcase how the TCFT solves for these metastable free energies.

X. ACKNOWLEDGMENTS

We thank Jacob Hastings, Kyle Ray, Paul Riechers, and Mikhael Semaan for helpful discussions. The authors thank the Telluride Science Research Center for hospitality during visits and the participants of the Information Engines Workshops there. JPC acknowledges the kind hospitality of the Santa Fe Institute, Institute for Advanced Study at the University of Amsterdam, and California Institute of Technology. This material is based upon work supported by, or in part by, FQXi Grant number FQXi-RFP-IPW-1902 and U.S. Army Research Laboratory and the U.S. Army Research Office under grants W911NF-21-1-0048 and W911NF-18-1-0028.

Data Availability Statement: Neither data nor programming code is necessary to support this work's results.

Appendix A: Alternative TCFT Derivations

Equation (20)'s Exponential Class Fluctuation Theorem (ECFT) can be derived from at least two prior results—from path ensemble averaging and from a master fluctuation theorem.

1. Path Ensemble Average

We first give a derivation from a generalization of Crooks' Path Ensemble Average to arbitrary privileged distributions. The Path Ensemble Average result is expressed in Eq. (15) of Ref. [30]:

$$\langle \mathcal{F} e^{-\Sigma(\overrightarrow{z})} \rangle_F = \langle \widehat{\mathcal{F}} \rangle_{R'} \ .$$

Here, $\mathcal{F}$ is an arbitrary trajectory functional, $\widehat{\mathcal{F}}$ is its time reverse, defined by $\widehat{\mathcal{F}}(\overrightarrow{z}) = \mathcal{F}(\overrightarrow{z}^\dagger)$, and $\langle \cdot \rangle_y$ denotes a trajectory ensemble average over the forward ($y = F$) or physical reverse representation of the reverse ($y = R'$) processes.

We first convert to the formal reverse representation for convenience:

$$\begin{aligned}
\left\langle \widehat{\mathcal{F}} \right\rangle_{R'} &= \int d\overrightarrow{z} R'(\overrightarrow{z}) \mathcal{F}(\overrightarrow{z}^\dagger) \\
&= \int d\overrightarrow{z} R(\overrightarrow{z}^\dagger) \mathcal{F}(\overrightarrow{z}^\dagger) \\
&= \int d\overrightarrow{z} R(\overrightarrow{z}) \mathcal{F}(\overrightarrow{z}) \\
&= \langle \mathcal{F} \rangle_R \ ,
\end{aligned}$$

where R denotes that the average is taken over the formal reverse representation of the reverse process. This gives:

$$\langle \mathcal{F} e^{-\Sigma(\overrightarrow{z})} \rangle_F = \langle \mathcal{F} \rangle_R \ . \tag{A1}$$

Consider an arbitrary trajectory class $C \in \mathcal{C}$. Then let $\mathcal{F}(\overrightarrow{z}) = [\overrightarrow{z} \in C]$—$C$'s characteristic function—for all $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$. Then the LHS of Eq. (A1) becomes:

$$\begin{aligned}
&\int d\overrightarrow{z} P(\overrightarrow{z}) [\overrightarrow{z} \in C] e^{-\Sigma(\overrightarrow{z})} \\
&\quad = \int_C d\overrightarrow{z} P(\overrightarrow{z}|C) P(C) e^{-\Sigma(\overrightarrow{z})} \\
&\quad = P(C) \int_C d\overrightarrow{z} P(\overrightarrow{z}|C) e^{-\Sigma(\overrightarrow{z})} \\
&\quad = P(C) \left\langle e^{-\Sigma} \right\rangle_C \ .
\end{aligned}$$

Equation (A1)'s RHS is simply $R(C)$. Combining yields Eq. (20), showing that the exponential class fluctuation theorem is the path or trajectory ensemble average of a characteristic function $[\overrightarrow{z} \in C]$.

2. Master Fluctuation Theorem

For Langevin dynamics, invoke Ref. [17]'s Master Fluctuation Theorem:

$$\left\langle g(\{S_\alpha\}) e^{-\Sigma} \right\rangle_F = \left\langle g(\{\epsilon_\alpha S_\alpha^\dagger\}) \right\rangle_{R'} \ .$$

This is Eq. (78) there. $\{S_\alpha\}$ is a set of functions of the system microstate trajectories for the forward process. $\{S_\alpha^\dagger\}$ is a corresponding set of functions of the trajectories for the reverse process with the following relationship: $S_\alpha^\dagger(\overrightarrow{z}^\dagger) = \epsilon_\alpha S_\alpha(\overrightarrow{z})$, where $\epsilon_\alpha = \pm 1$. And, g is any function of the set $\{S_\alpha\}$.

To derive the ECFT, we consider the singleton $\{S_\alpha(\overrightarrow{z})\} = \{[\overrightarrow{z} \in C]\}$. We then define $S_\alpha^\dagger(\overrightarrow{z}) = [\overrightarrow{z}^\dagger \in C]$, giving $S_\alpha^\dagger(\overrightarrow{z}^\dagger) = [\overrightarrow{z} \in C] = S_\alpha(\overrightarrow{z})$ and $\epsilon_\alpha = 1$. Then, set g to be the identity, obtaining:

$$\left\langle [\overrightarrow{z} \in C] e^{-\Sigma} \right\rangle_F = \left\langle [\overrightarrow{z}^\dagger \in C] \right\rangle_{R'} \ .$$

This is now in the form of Crooks' Path Ensemble Average for $\mathcal{F}(\overrightarrow{z}) = [\overrightarrow{z} \in C]$.

Appendix B: Irreversibility as Entropy-Difference Variability

The following shows that the average class irreversibility Ψ_C tracks the variability of entropy difference when small. Moreover, Ψ_C and variability both necessarily go to zero together. And so, the TCFT shows that finding a class with a narrow entropy-difference distribution is tantamount to minimizing the class irreversibility and class-average entropy difference.

First, translate Σ into $x = \Sigma - \langle\Sigma\rangle_C$, its difference from its average:

$$\langle e^{-\Sigma}\rangle_C = \langle e^{-x}\rangle_C \, e^{-\langle\Sigma\rangle_C} \ .$$

Then, Taylor expand:

$$\begin{aligned}\langle e^{-x}\rangle_C &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}\langle x^n\rangle_C \\ &= 1 + a \ ,\end{aligned}$$

with:

$$\begin{aligned}a &\equiv \sum_{n=2}^{\infty} \frac{(-1)^n}{n!}\langle x^n\rangle_C \\ &\geq 0 \ .\end{aligned}$$

When Σ's variability is small, the x are typically small and the second order term $\langle x^2\rangle_C$ dominates in a. a is, then, the variance of Σ over C.

Then, using Eqs. (20) and (16), we have:

$$e^{-\Theta_C} = (1 + a)e^{-\Theta_C - \Psi_C} \ .$$

This gives:

$$\Psi_C = \ln(1 + a) \ .$$

Since a goes as the variance, Ψ_C is also a measure of Σ's variability in C in the small variability limit.

Appendix C: Free Energy Estimate Distribution

If we wish to estimate the change in free energy from the work distributions of a collection of forward and reverse experiments that start in equilibrium, the TCFT provides a relation for the free-energy differences of each trajectory class C:

$$e^{-\Delta f_C} = \langle e^{-\beta^{-1}W}\rangle_C \frac{P(C)}{R(C)},$$

which each equal the change in free energy $\Delta F^{\text{eq}} = \Delta f_C$. Let us consider a particular experiment with two energy levels that start at:

$$\begin{aligned}E_0(A) &= -\beta^{-1}\ln(1-\epsilon) \\ E_0(B) &= -\beta^{-1}\ln\epsilon,\end{aligned}$$

with the corresponding equilibrium distribution:

$$\begin{aligned}\pi_0(A) &= 1 - \epsilon \\ \pi_0(B) &= \epsilon.\end{aligned}$$

Note that, because $E_t(s) \equiv F_t^{\text{eq}} - \beta^{-1}\ln\pi_t(s)$, the free energy in this case is zero initially: $F_0^{\text{eq}} = 0$. We then change the energy level instantaneously to the final energy landscape:

$$\begin{aligned}E_\tau(A) = E_\tau(B) &= -\beta^{-1}\ln\pi_\tau(A) \\ &= -\beta^{-1}\ln\pi_\tau(B) \\ &= \beta^{-1}\ln 2,\end{aligned}$$

which also has zero free energy $F_\tau^{\text{eq}} = 0$. If the initial state is s, then it remains s and the work investment is:

$$\begin{aligned}W(A) &= E_\tau(A) - E_0(A) \\ &= \beta^{-1}\ln(2(1-\epsilon)) \\ W(B) &= E_\tau(B) - E_0(B) \\ &= \beta^{-1}\ln(2\epsilon) \ .\end{aligned}$$

To evaluate the free-energy difference estimate, we note $\Delta\tilde{f}_C$ is itself a function of the estimated probability of realizations of the experiment that starts in A:

$$\begin{aligned}\Delta\tilde{f}_C(\tilde{P}(A), \tilde{R}(A)) &= -\beta^{-1}\ln\left(\frac{\delta_{A\in C}\tilde{P}(A)e^{-W(A)} + \delta_{B\in C}(1-\tilde{P}(A))e^{-W(B)}}{\delta_{A\in C}\tilde{P}(A) + \delta_{B\in C}(1-\tilde{P}(A))}\frac{\delta_{A\in C}\tilde{P}(A) + \delta_{B\in C}(1-\tilde{P}(A))}{\delta_{A\in C}\tilde{R}(A) + \delta_{B\in C}(1-\tilde{R}(A))}\right) \\ &= -\beta^{-1}\ln\left(\frac{\delta_{A\in C}\tilde{P}(A)e^{-W(A)} + \delta_{B\in C}(1-\tilde{P}(A))e^{-W(B)}}{\delta_{A\in C}\tilde{R}(A) + \delta_{B\in C}(1-\tilde{R}(A))}\right) \ .\end{aligned}$$

We use frequentist statistics to estimate the probabilities of our initial states and resultant works. Given N forward experiments and N reverse experiments, the probability of realizing free energy ΔF is determined by evaluating the number n_A of times the forward experiment starts in A and the number n_A^R of times the reverse experiment starts in A:

$$\begin{aligned}\Pr(\Delta\tilde{f}_C = \Delta F) &= \sum_{n_A, n_A^R} \Pr(\Delta\tilde{f}_C = \Delta F, \tilde{P}(A) = n_A/N, \tilde{R}(A) = n_A^R/N) \\ &= \sum_{n_A, n_A^R} \delta_{\Delta F, \Delta\tilde{f}_C(n_A/N, n_A^R/N)} \Pr(\tilde{P}(A) = n_A/N, \tilde{R}(A) = n_A^R/N) \ .\end{aligned}$$

For N experiments, we can combinatorially evaluate the probability of realizing n_A and n_A^R as a function of N:

$$\Pr(\tilde{P}(A) = n_A/N) = (1-\epsilon)^{n_A} \epsilon^{N-n_A} \binom{N}{n_A}$$

and:

$$\Pr(\tilde{R}(A) = n_A^R/N) = 2^{-N} \binom{N}{n_A^R} \ .$$

Assuming that the forward and reverse experiments are performed independently, the joint probability of realizing n_A and n_A^R is:

$$\begin{aligned}&\Pr(\tilde{P}(A) = n_A/N, \tilde{R}(A) = n_A^R/N) \\ &\quad = \Pr(\tilde{R}(A) = n_A^R/N) \Pr(\tilde{P}(A) = n_A/N) \\ &\quad = (1-\epsilon)^{n_A} \epsilon^{n_A - N} \binom{N}{n_A} 2^{-N} \binom{N}{n_A^R}.\end{aligned}$$

We then compute the probability of our free energy estimate $\Pr(\Delta\tilde{f}_C = \Delta F)$ for N experiments with $\pi_0(A) = \epsilon$:

$$\begin{aligned}&\Pr(\Delta\tilde{f}_C = \Delta F) \\ &\quad = \sum_{n_A, n_A^R} \delta_{\Delta F, \Delta\tilde{f}_C(n_A/N, n_A^R/N)} \\ &\qquad \times (1-\epsilon)^{n_A} \epsilon^{n_A - N} \binom{N}{n_A} 2^{-N} \binom{N}{n_A^R} \ .\end{aligned}$$

[1] D. J. Evans, E. G. D. Cohen, and G. P. Morriss. Probability of second law violations in shearing steady flows. *Phys. Rev. Lett.*, 71:2401–2404, 1993.

[2] G. Gallavotti and E. G. D. Cohen. Dynamical ensembles in nonequilibrium statistical mechanics. *Phys. Rev. Lett.*, 74:2694–2697, 1995.

[3] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.

[4] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721, 1999.

[5] C. Jarzynski. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Ann. Rev. Cond. Matter Physics*, 2(1):329–351, 2011.

[6] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, 90(5/6):1481–1487, 1998.

[7] C. Jarzynski. Nonequilibrium work theorem for a system strongly coupled to a thermal environment. *J. Stat. Mech.: Theor. Exp.*, 2004(09):P09005, 2004.

[8] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science*, 296:1832, 2002.

[9] P. Maragakis, M. Spichty, and M. Karplus. A differential fluctuation theorem. *J. Phys. Chem. B*, 112(19):6168–6174, 2008.

[10] I. Junier, A. Mossa, M. Manosas, and F. Ritort. Recovery of free energy branches in single molecule experiments. *Phys. Rev. Lett.*, 102(7):070602, 2009.

[11] S. Deffner and E. Lutz. Nonequilibrium work distribution of a quantum harmonic oscillator. *Phys. Rev. E*, 77(2):021128, 2008.

[12] C. Jarzynski. Rare events and the convergence of exponentially averaged work values. *Phys. Rev. E*, 73(4):046105, 2006.

[13] S. Asban and S. Rahav. Nonequilibrium free-energy estimation conditioned on measurement outcomes. *Phys. Rev. E*, 96(2):022155, 2017.

[14] G. Wimsatt, O.-P. Saira, A. B. Boyd, M. H. Matheny, S. Han, M. L. Roukes, and J. P. Crutchfield. Harnessing fluctuations in thermodynamic computing via time-reversal symmetries. *Phys. Rev. Res.*, 3(3):033115, 2021.

[15] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015.

[16] G. E. Crooks. *Excursions in statistical dynamics*. PhD thesis, University of California, Berkeley, 1999.

[17] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75:126001, 2012.

[18] G. Wimsatt, A. B. Boyd, and J. P. Crutchfield. Measure-theoretic fluctuation theorems. *in preparation*, 2024.

[19] C. Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. *J. Stat. Phys.*, 98(1-2):77–102, 2000.

[20] U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.*, 95(4):040602, 2005.

[21] A. Gomez-Marin, J. M. R. Parrondo, and C. Van den Broeck. Lower bounds on dissipation upon coarse graining. *Phys. Rev. E*, 78(1):011107, 2008.

[22] P. M. Riechers, A. B. Boyd, G. W. Wimsatt, and J. P. Crutchfield. Balancing error and dissipation in computing. *Physical Review Research*, 2(3):033524, 2020.

[23] G. W. Wimsatt, A. B. Boyd, P. M. Riechers, and J. P. Crutchfield. Refining Landauer's stack: Balancing error and dissipation when erasing information. *J. Stat. Physics*, 183(1):1–23, 2021.

[24] É. Roldán and J. M. R. Parrondo. Entropy production and Kullback-Leibler divergence between stationary trajectories of discrete systems. *Phys. Rev. E*, 85(3):031129, 2012.

[25] I. A. Martínez, G. Bisker, J. M. Horowitz, and J. M. R. Parrondo. Inferring broken detailed balance in the absence of observable currents. *Nature Comm.*, 10(1):1–10, 2019.

[26] D. J. Skinner and J. Dunkel. Estimating entropy production from waiting time distributions. *Phys. Rev. Let.*, 127(19):198101, 2021.

[27] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck. Dissipation: The phase-space perspective. *Phys. Rev. L*, 98(8):080602, 2007.

[28] S. Vaikuntanathan and C. Jarzynski. Dissipation and lag in irreversible processes. *Euro. Phys. Let.*, 87(6):60005, 2009.

[29] A. Berut, A. Petrosyan, and S. Ciliberto. Detailed Jarzynski equality applied to a logically irreversible procedure. *Euro. Phys. Let.*, 103:60002, 2013.

[30] G. E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61(3):2361–2366, 2000.