Fluctuations and Information Processing in Nonequilibrium Thermodynamics

Wimsatt, Gregory William

Fluctuations and Information Processing in Nonequilibrium Thermodynamics

By

GREGORY WILLIAM WIMSATT
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Physics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

James P. Crutchfield, Chair

David Doty

Michael L. Roukes

Committee in Charge

2024

To Jeffrey and Sheryl Wimsatt

Contents

Abstract

We present results connecting the fluctuations of small-scale thermodynamics with information processing and computation. To begin, we experimentally demonstrate that highly structured distributions of work emerge during even the simple task of erasing a single bit. These are signatures of a refined suite of time-reversal symmetries in distinct functional classes of microscopic trajectories. As a consequence, we introduce the Trajectory Class Fluctuation Theorem (TCFT), a deep fluctuation theorem that the component work distributions must satisfy. Since they identify entropy production, the component work distributions encode both the frequency of various mechanisms of success and failure during computing as well as giving improved estimates of the total irreversibly-dissipated heat. This new diagnostic tool provides strong evidence that thermodynamic computing at the nanoscale can be constructively harnessed. We experimentally verify this functional decomposition and the new class of fluctuation theorems by measuring transitions between flux states in a superconducting circuit.

The TCFT provides broader insights. It substantially strengthens the Second Law of Thermodynamics and its consequences. Practically, the TCFT improves empirical estimates of free energies, a task known to be statistically challenging. It 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).

We further utilize fluctuation theory to construct new thermodynamic bounds for systems controlled with a time-symmetric protocol, again studying bit erasure in detail. We demonstrate that the bounds are tight and show that the costs overwhelm those implied by Landauer's energy bound on information erasure. Moreover, in the limit of perfect computation, the costs diverge. A takeaway is that time-asymmetric protocols should be developed for efficient, accurate thermodynamic computing. And, that Landauer's Stack—the full suite of theoretically-predicted thermodynamic costs—is ready for experimental test and calibration.

Acknowledgments

It is staggering to consider all the people who helped me reach this dissertation. I unfortunately must leave many people unmentioned and as many barely so. But I trust they all know, confidently, of their lasting effects, however transient our time together may have been.

Professor Manuel Calderón de la Barca Sánchez welcomed me to his research group on Heavy Ion Collisions when I was just finishing my undergraduate degree. With him and his group I gained invaluable insight into the nature of academic research. I thank him and everyone there for showing me the broad path of scientific work.

I ultimately found my particular path with my graduate advisor, Professor James P. Crutchfield. I was frequently awed by his ability to observe large and important reprecussions in what I found to be small results or details. He taught me to always remember and conceive the big picture in whatever we do. He provided careful attention and guidance throughout my many-year trek. He has relentlessly supported my work since I first took wing and has always been a great champion of our ideas and results. For this profound mentorship, I am deeply indebted to him.

Professor Crutchfield also provided the environment and mentality to attract a wide variety of bright and fascinating minds to his group. Doctor Alexander B. Boyd requires special mention. When I began studying thermodynamics in the group, Doctor Boyd was still a graduate student and I like to think that we taught each other many of the fundamentals of the field. In fact, he served much more like a second advisor, providing generous patience, interest, and long hours. As we began writing papers together, a broader friendship emerged that holds strong today. His astoundingly creative mind has always inspired me, and I'm lucky that he has numerously offered me to join in working on his ideas. I owe great thanks to Doctor Boyd for partnering with me through one of the greatest journeys of my life.

Doctor Paul M. Riechers was also a crucial guide on my journey. I was blessed that he too became another major collaborator and great friend. Also a brilliantly creative researcher, he, Doctor Boyd, Professor Crutchfield and I spent many days and nights not just working on what we expected to achieve, but exploring the boundaries of what was barely conceivable. These explorations led, I believe, to some of our most important work, hopefully on display in this dissertation. Because of these friends, I learned how creative, diligent contemplation can materialize the nebulous.

Doctor Korana Burke was another member of Professor Crutchfield's group that I owe particular thanks to. I'm not sure I've met an individual so truly caring and giving as Doctor Burke. She frequently went above and beyond to help so many of us in the group, from academic advising to organzing extended research workshops. I believe that due to her, everyone's experience, capability, and successes were improved, mine certainly included.

Professor Michael L. Roukes and Doctor Olli Pentti-Saira were terrific collaborators. They grounded my work in the real world by providing first-class resources and expertise. My experience with them has been invaluable to my understanding of the nano-scale world and how we perceive it.

There are many others I have to thank for their many forms of help. In particular, I thank Doctor Christopher J. Ellison, Doctor Ryan G. James, Professor Sarah Marzen, Doctor Christopher C. Strelioff, Doctor Dowman Varn, and Doctor John Mahoney. These people were a part of or strongly associated with Professor Crutchfield's group when I arrived. The atmosphere of creativity that they presented, even in casual conversation, was astounding. Because of them, I knew that I was in the right place.

I also thank Doctor Cina Aghamohammadi, Doctor Dmitry Shemetov, Doctor Ariadna Venegas-Li, Doctor Alexandra M. Jurgens, Doctor Mikhael Semaan, Doctor Samuel Loomis, Doctor Adam Rupe, Doctor Jeffrey Emenheiser, Adam Kunesh, and Doctor Kyle Ray. I am grateful to have developed as a scientist alongside them.

Finally, I thank my family. I must specifically mention my parents, Jeffrey and Sheryl Wimsatt. They worked to provide me with all a child could hope for. They taught me a love of science and reason from a young age. Along the road to this dissertation, they provided constant and unconditional respites. Few people receive this kind of support. I thank them dearly.

Related Works

Chapters 2, 3, and 4 of this dissertation are based on the following three works, respectively:

G. W. 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. *Physical Review Research*, 3(3):033115, 2021.

G. W. Wimsatt, A. B. Boyd, and J. P. Crutchfield. Trajectory class fluctuation theorem. *arXiv preprint*, arXiv:2207.03612, 2022.

G. W. Wimsatt, A. B. Boyd, P. M. Riechers, and Crutchfield, J. P. Refining Landauer's stack: balancing error and dissipation when erasing information. *Journal of Statistical Physics*, 183, 2021.

I am also an author of the following related works.

O.-P. Saira, M. H. Matheny, R. Katti, W. Fon, G. W. Wimsatt, S. Han, J. P. Crutchfield, and M. L. Roukes. Nonequilibrium thermodynamics of erasure with superconducting flux logic. *Physical Reviews Research*, 2:013249, 2020

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.

K. J. Ray, A. B. Boyd, G. W. Wimsatt, and J. P. Crutchfield. Non-Markovian momentum computing: Thermodynamically efficient and computation universal. *Physical Review Research*, 3:023164, 2021.

A. B. Boyd, P. M. Riechers, G. W. Wimsatt, J. P. Crutchfield, and M. Gu. Time symmetries of memory determine thermodynamic efficiency. *arXiv preprint*, arXiv:2104.12072, 2021.

Of these four papers, the first provides a more complete analysis of the flux qubit device used in the material of Chapter 2. The second presents a generalized picture of the tradeoff between computational error and dissipated heat in time-symmetrically controlled systems that is explored in Chapter 4. The final two papers investigate phenomena of a more significant departure from the material in this dissertation, but are grounded in the core fluctuation and information processing theory that we take part in here.

CHAPTER 1

Introduction

In 1961, Landauer showed that conducting logically irreversible operations implies fundamental thermodynamic costs [1]. These costs are irrespective of the specific devices used or even of the speed at which at operations are performed.

Landauer focused on a simple yet fundamental irreversible operation. He conceived of a device which, depending on its specific configuration, encoded one of two logical states: a 0 or 1. In the operation, the device starts equally likely in the 0 or 1 states. The device is said to have a bit of *information*, since specifics of the device's past may be stored into the device as either a 0 or 1. The device is then is manipulated externally to end in the 0 state. Because any previous information that the device may have been storing in its informational bit is now irretrievable, we say that the operation is that of *bit erasure*. Using standard statistical mechanical assumptions, Landauer argued that the entropy of the device has to be reduced by $k_\mathrm{B} \ln 2$ to perform this operation, where $k_\mathrm{B} \approx 1.4 \cdot 10^{-23} J/K$ is Boltzmann's constant.

By the second law of thermodynamics, the entropy in some other object must increase to compensate. If no other devices take on this entropy, the thermal environment must take up energy in the form of heat. Let the temperature of the environment be T, the heat expelled into the environment during the process be Q, and the resultant change in environmental entropy be ΔS_env. If the environment is nearly an ideal heat bath, it obeys the Clausius equality, $Q = T\Delta S_\mathrm{env}$. The heat expelled from the device into the environment in order to conduct bit erasure must then be $k_\mathrm{B} T \ln 2$. This minimal heat cost is known as *Landauer's Bound*.

By the conservation of energy, another source of energy must balance the heat cost. The energy it provides to the device may be deemed work in contrast to the heat that is exchanged with the environment. Landauer's Bound is therefore also the minimum work required by the conductor of the operation. Heat and work costs for more complicated irreversible operations can be similarly derived.

Some three decades later, important progress was made in fluctuation theory. In 1997, Jarzynski presented an equality between the change in free energy and the exponential average work for a potentially highly nonequilibrium process on a system in contact with a thermal environment [2]. Crooks proved the equation under a Markovian framework before presenting further fundamental fluctuation theorems [3, 4]. Especially important is an equality due to Crooks which we refer to as the Detailed Fluctuation Theorem (DFT) [4]. In contrast to Jarzynski's Equality, which is centered around an ensemble average, the DFT is a very detailed equality which holds for each and every system state trajectory.

We use the DFT to derive our Trajectory Class Fluctuation Theorem (TCFT). The TCFT describes the thermodynamics of arbitrary subsets of system state trajectories that may evolve in a process—trajectory classes—in terms of their probabilities in the process and in a time reversed version of the process. Because these subsets are arbitrary, the TCFT spans a wide space between the trajectory-centric DFT and the ensemble-centric Jarzynski Equality.

We give the first exposition of the TCFT in Chapter 2 where we analyze the process of bit erasure in both simulation and experiment. The TCFT is used to set strong bounds on work requirements as well as help analyze and tailor a process to one's needs. Besides applying the TCFT, Chapter 2 provides a detailed look at the process of bit erasure. We develop a close relationship between trajectory classes of successful and unsuccessful erasure and the surprisingly structured work distribution uncovered. The chapter should serve to ground the remainder of the dissertation.

We then place the TCFT on center stage in Chapter 3 and give it a thorough theoretical treatment. We use the TCFT to find inequalities analagous to but stronger than the thermodynamic second law, a method of mapping out free energy changes in complex computational systems, and strong statistical estimators for free energy changes.

We conclude with Chapter 4, looking at systems that are driven with time-symmetric protocols. One motivation is that typical computing systems are powered with time-symmetric sources, say by alternating current received from the mains or the direct current provided by batteries. We use the DFT to find significant and unavoidable energetic costs on time-symmetrically conducted operations that vastly outgrow Landauer's bound with increasingly stringent demands on the computational error.

CHAPTER 2

Thermodynamic Structure in Bit Erasure

2.1. Introduction

Physics dictates that all computing is subject to spontaneous error. These days, this truism repeatedly reveals itself: despite the once-predictable miniaturization of nanoscale electronics, computing performance increases have dramatically slowed in the last decade or so. In large measure, this is due to the concomitant rapid decrease in the number of information-bearing physical degrees of freedom, rendering information storage and processing increasingly susceptible to corruption by thermal fluctuations. Said simply, all physical computing is thermodynamic.

Controlling the production of fluctuations and removing heat pose key technological challenges to further progress. Practically, the challenge remains of how to probe and diagnose information processing in overtly noisy systems. The following introduces trajectory class fluctuation theorems to do this by identifying the thermodynamic signature of successful and failed information processing. It then experimentally demonstrates how this is practically implemented in a new microscale platform for thermodynamic computing.

Only recently have tools appeared that precisely describe what trade-offs exist between thermodynamic resources and useful information processing—these are highly reminiscent of the centuries-old puzzle of how Maxwell's "very observant and neat-fingered" demon uses its "intelligence" to convert disorganized heat energy to useful work [5]. In our modern era, his demon has led to the realization that information itself is physical [6,7,8]—or, most constructively, that information is a thermodynamic resource [9, and references therein]. This opened up the new paradigm of *thermodynamic computing* [10] in which fluctuations play a positive role in efficient information processing on the nanoscale. We now conceptualize this via *information engines*: physical systems that are driven by, manipulate, store, and dissipate energy, but simultaneously generate, store, lose, communicate, and transform information. In short, information engines combine traditional engines comprised of heat, work, and other familiar reservoirs with, what we now call, *information reservoirs* [11,12].

Reliable thermodynamic computing requires detecting and controlling fluctuations in informational and energetic resources and in engine functioning. To do so requires a new generation of diagnostic tools. For these, we appeal to fluctuation theorems that capture exact time-reversal symmetries and predict entropy production leading to irreversible dissipation [**2**,**13**,**14**,**15**,**16**,**17**,**18**]. As the following demonstrates, these place us on the door-step of the very far-from-equilibrium thermodynamics needed to understand the physics of computing. And, in turn, the physical principles of how nature processes information in the service of biological functioning and survival have begun to emerge.

Proof-of-concept experimental tests have been carried out in several substrates: probing biomolecule free energies [**19**,**20**,**21**], work expended during elementary computing (bit erasure) [**22**,**23**,**24**,**25**,**26**,**27**], and Maxwellian demons [**28**]. That said, the suite of theoretical predictions and contemporary principles (App. 2.A) far outstrips experimental validation to date.

To close the gap, we show how to diagnose thermodynamic computing on the nanoscale by explaining the signature structures in work distributions generated during information processing. Previous efforts explored features in work and heat distributions that track the mesoscale evolution of a system's *informational states*; see Refs. [**29**, **30**] and App. 2.A. Here, we show that functional and nonfunctional informational-state evolutions can be identified by appropriate conditioning and that their thermodynamics obey a suite of trajectory-class fluctuation theorems. As such, the latter give accurate bounds on work, entropy production, and dissipation for computing subprocesses. The result is a practical tool that employs mesoscopic (work) measurements to diagnose microscopic thermodynamic computing. For simplicity and to make direct contact with previous efforts, we demonstrate the tools on Landauer erasure [**6**] of a bit of information in a superconducting flux qubit.

2.2. Model System

As a reference, we first explore the thermodynamics of bit erasure in a simple model: a particle with position and momentum in a double-well potential $V(x,t)$ and in contact with a heat reservoir at temperature T. (Refer to Fig. 2.1.) An external controller adds or removes energy from a work reservoir to change the form of the potential $V(\cdot,t)$ via a predetermined *erasure protocol* $\{(\beta(t),\delta(t)) : 0 \leq t \leq \tau\}$. $\beta(t)$ and $\delta(t)$ change one at a time piecewise-linearly through four protocol substages: (1) *drop barrier*, (2) *tilt*, (3) *raise barrier*, and (4) *untilt*. (See App. 2.B.) The system starts at time $t = 0$ in the equilibrium distribution for a double-well $V(x, 0)$ at temperature T.

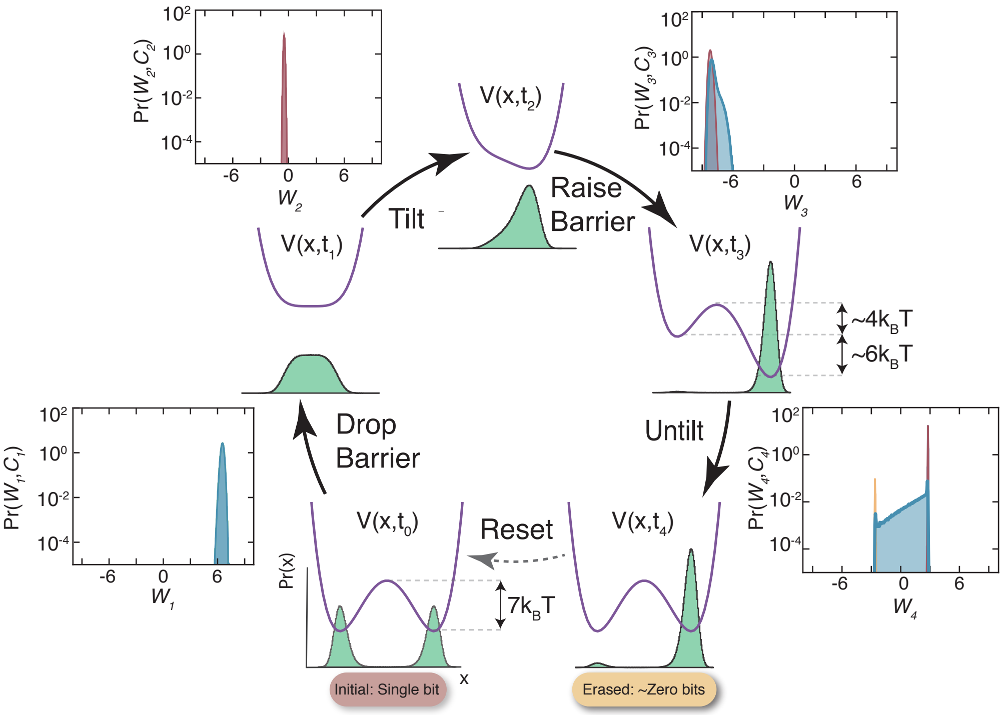

FIGURE 2.1. Inner plot sequence: Erasure protocol (Table 2.1) evolution of position distribution $\Pr(x)$ from simulation. Potential $V(x, t_s)$ at substage boundary times $t_s, s = 0, 1, 2, 3, 4$. Starting at $t = t_0$, the potential evolves clockwise, ending at $t = t_4$ in the same configuration as it starts: $V(x, t_0) = V(x, t_4)$. However, the final position distribution $\Pr(x)$ predominantly indicates the R state. The original one bit of information in the distribution at time $t = t_0$ has been erased. Outer plot sequence: Substage work distributions from simulation $\Pr(W_s, C_s)$ during substages s: (1) Barrier Drop, (2) Tilt, (3) Barrier Raise, (4) Untilt. During each substage s, distributions are given for up to three substage trajectory classes C_s: red consists of trajectories always in the R state, orange trajectories always in the L state, and blue the rest, spending some time in each state.

We use *underdamped* Langevin dynamics to simulate this model:

$$dx = v dt$$

$$m\, dv = \sqrt{2k_{\mathrm{B}}T\gamma}\, r(t)\sqrt{dt} - \left(\frac{\partial}{\partial x}V(x,t) + \gamma v\right) dt \ , \tag{2.1}$$

where k_{B} is Boltzmann's constant, γ is the coupling between the heat reservoir and system, m is the particle's mass, and $r(t)$ is a memoryless Gaussian random variable with $\langle r(t) \rangle = 0$ and $\langle r(t)r(t') \rangle = \delta(t - t')$.

The default potential, $V(\cdot, 0) = V(\cdot, \tau)$, has two symmetric wells separated by a barrier. Following common practice we call the two wells, from negative to positive position, the *Left* (L) and *Right* (R) *informational states*, respectively. Initially being equiprobable, the informational states associated with each of the two wells thus contain 1 bit of information [**31**].

The erasure protocol is designed so that the particle ends in the R state with high probability, regardless of its initial state. Simulating the protocol 3.5×10^6 times, 96.2% of the particles were successfully erased into the R state. Thus, as measured by the Shannon entropy, the initial 1 bit of information was reduced to 0.231 bits. We intentionally designed the protocol to fail frequently at erasure to better illustrate our main results on diagnosing success *and* failure. But, crucially, the results we present hold for arbitrarily-successful erasure protocols.

At all other times t, $V(\cdot, t)$ has either one or two local minima, naturally defining metastable regions for a particle to be constrained and gradually evolve towards local equilibrium. We therefore define the informational states at time $0 \leq t \leq \tau$ to be the metastable regions, labeling them R and, if two exist, L—from most positive to negative in position.

Since the protocol is composed of four simple substages, we coarse-grain the system's response by its activity during each substage at the level of its informational state. Specifically, for each substage, we assign one of three *substage trajectory classes*: the system (i) was always in the R state, (ii) was always in the L state, or (iii) spent time in each. Sometimes there is only one informational state and so the latter two classes are not achievable for all substages.

2.3. Work Characterization

We then focus on a single mesoscopic observable—the thermodynamic work expended during erasure. An individual realization generates a trajectory of system microstates, with $W(t, t')$ being the work done on the system between times $0 \leq t < t' \leq \tau$; see App. 2.H. Let $W_s = W(t_{s-1}, t_s)$ denote the work generated during substage s and C_s the substage trajectory class. Figure 2.1 (Outer plot sequence) shows the corresponding substage work distributions $\Pr(W_s, C_s)$ obtained from our simulations. (See App. 2.I.)

The drop-barrier and tilt substage work distributions are rather simple, being narrow and unimodal. The raise-barrier distributions have some asymmetry, but are also similarly simple. However, the untilt work distributions (farthest right in Fig. 2.1) exhibit unusual features that are significant for understanding the intricacies of erasure. Trajectories that spend all of the untilt substage in either the R state or L state form peaks at the most positive (red) and negative (orange) work values, respectively. This is because the R-state well is always increasing in potential energy while the L-state well is always decreasing during untilt. In contrast, the other trajectories contribute a log-linear ramp of work values (blue) dependent on the time spent in each. The ramp's positive slope signifies that more time is typically spent in the R state in this last set of trajectories.

Looking at the total work $W_{\text{total}} = W(0, \tau)$ generated for each trajectory over the course of the entire erasure protocol, we observe the strikingly complex and structured distribution $\Pr(W_{\text{total}})$ shown in Fig. 2.2(Rear). There are two clear peaks at the most positive and negative work values separated by a ramp. This highly structured work distribution, generated by bit erasure, contrasts sharply with the unimodal work distributions common in previous studies of fluctuation theorems; see, for example, Fig. 2.2(Inset) for the work distribution generated by a thermodynamically-driven simple harmonic oscillator translated in space or Fig. 2 in Ref. [**16**]. The total average work was $0.634\ k_{\text{B}}T$, satisfying Landauer's bound by being greater than the informational-state Shannon entropy decrease of $0.769 \times \ln 2\ k_{\text{B}}T = 0.533\ k_{\text{B}}T$.

We can understand the mechanisms behind this structure when decomposing Fig. 2.2(Rear)'s total work distribution under the untilt substage trajectory classes C_4. We label trajectories that spend all of the untilting substage in the R state *Success* since, via the previous substages, they reach the intended R state by the untilting substage and remain there until the protocol's end. Similarly, trajectories that spend all of the untilt substage in the L state are labeled *Fail.* The remaining trajectories are labeled *Transitional*, since they transition between the two informational states during untilt, potentially succeeding or failing to end in the R state.

Figure 2.2's three front plots show the work distribution for each of these three trajectory classes. Together they recover the total work distribution over all trajectories shown in Fig. 2.2(Rear). Though, now the thermodynamic contributions to the total from the functionally-distinct component trajectories are made apparent.

2.4. Trajectory-Class Fluctuation Theorem

Exploring the mesoscale dynamics of erasure revealed signatures of a "thermodynamics" for each trajectory that is uniquely associated with successful or failed information processing. We now

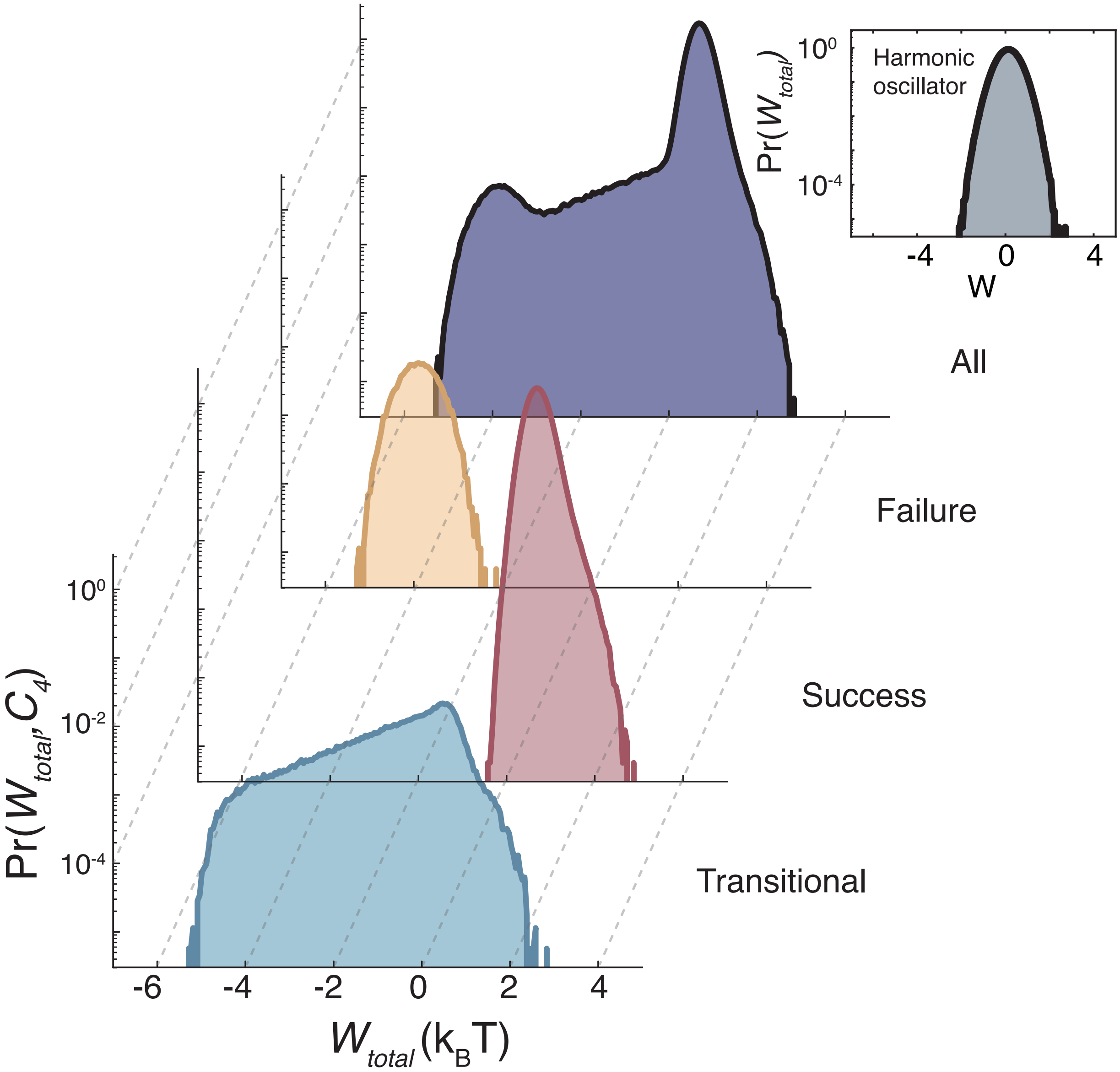

FIGURE 2.2. (Rear, purple) Total work distribution of all trajectories $\Pr(W_{\text{total}})$ during erasure simulation: A histogram generated from 3.5×10^6 trials for $W_{\text{total}} \in [-6, 4]$ over 201 bins. (Inset, gray) Typical unimodal work distribution illustrated for spatially-translated thermally-driven simple harmonic oscillator. (Three front plots) Work distributions $\Pr(W_{\text{total}}, C_4)$ for the trajectory classes C_4 determined by the untilt trajectory partition in simulation: The red work distribution (middle) is that of Success trajectories, the orange (rear) is that of Fail trajectories, and the blue (front) is that of the remaining, Transitional trajectories.

introduce the underlying fluctuation theory from which the trajectory thermodynamics follow. Key to this is comparing system behaviors in both forward and reverse time [**2**,**13**,**14**,**15**,**16**,**17**,**18**]. (See Apps. 2.C and 2.F.)

This suite of trajectory-class fluctuation theorems (TCFTs) applies to arbitrary classes of system microstate trajectories obtainable during a thermodynamic transformation. Importantly, they interpolate between Jarzynki's equality [**2**] and Crooks' detailed fluctuation theorem [**16**] as the trajectory class varies from the entire ensemble of trajectories to a single particular trajectory, respectively. Accordingly, they unify a wide range of other previously-established fluctuation theorems. (See App. 2.C.)

One TCFT presents a lower bound on the average work $\langle W \rangle_C$ over any measurable subset C of the ensemble of system microstate trajectories $\overrightarrow{\mathcal{Z}}$, where W is the total work for a trajectory:

$$\begin{aligned}\langle W \rangle_C &\geq \Delta F + k_{\mathrm{B}} T \ln \frac{P(C)}{R^{\mathrm{eq}}(C^{\dagger})} \\ &= \langle W \rangle_C^{\min} \ , \end{aligned} \tag{2.2}$$

with ΔF^{eq} the change in equilibrium free energy over the protocol, $P(C)$ the probability of realizing the class C during the protocol, and $R^{\mathrm{eq}}(C^{\dagger})$ the probability of obtaining the time reverse of class C under the time-reverse protocol. (k_{B} is Boltzmann's constant.) As detailed in App. 2.G, this allows accurate estimation of the work generated for trajectory classes with narrow work distributions, such as the Success and Fail classes of erasure, even with limited knowledge (low sampling) of system response under the protocol and its time reverse.

The TCFTs lead to additional consequences. First, they more strongly bound the average work over all trajectories compared to the equilibrium free energy change ΔF^{eq}. Second, they provide a new expression for obtaining equilibrium free-energy changes:

$$\Delta F^{\mathrm{eq}} = -k_{\mathrm{B}} T \ln \left(\frac{P(C)}{R^{\mathrm{eq}}(C^{\dagger})} \langle e^{-W/k_{\mathrm{B}}T} \rangle_C \right) \ . \tag{2.3}$$

Remarkably, this only requires statistics for any particular class C and its reverse $C^{\dagger}$ to produce the system's free energy change. Since rare microstate trajectories may generate sufficiently negative works that dominate the average exponential work, this leads to a substantial statistical advantage over direct use of Jaryznski's equality $\Delta F^{\mathrm{eq}} = -k_{\mathrm{B}} T \ln \langle e^{-W/k_{\mathrm{B}}T} \rangle_{\overrightarrow{\mathcal{Z}}}$ for estimating free energies [**32**]. (See App. 2.C.)

The erasure protocol's 96.2% success rate is reflected in the Success class's dominance in the work distributions of Fig. 2.2. We can adjust the protocol to exhibit arbitrarily-higher success rate while still maintaining high efficiency; i.e., keeping the total average work close to Landauer's bound. When done, we still observe the same qualitative features of the work distribution: two peaks separated by a log-linear ramp; each associated with the Success, Fail, and Transitional classes. Though, of course, the probabilities of the Fail and Transitional classes become arbitrarily small.

That said, the contributions of the Fail and Transitional classes to various fluctuation theorems—such as, Jarzynski's Equality to mention one—remain significant since the works generated by those classes compensate by becoming increasingly negative. In fact, the contribution to the exponential average work of the Success class only approaches the value 1/2 out of the required value of 1 when averaging over all trajectories! Thus, while the probabilities of the Fail and Transitional classes can become arbitrarily small by considering Erasure protocols with higher success rates, we cannot ignore the existence of the rare events due to Transitional and Fail trajectories unless we employ particular fluctuation theorems; in particular, a TCFT. (Again, see App. 2.C.) In this way, one sees that the TCFT provides a detailed diagnosis of successful and failed information processing and of the associated energetics.

2.5. Realizing Thermodynamic Computing

To explore these predictions, we selected a superconducting flux qubit composed of paired Josephson junctions (Fig. 2.3(A)), resulting in a double-well nonlinear potential that supports information storage and processing (Fig. 2.3(B)). Appendix 2.J.1 explains the physics underlying their nonlinear equations of motion, comparing the similarities and differences with our model's idealized Langevin dynamics.

Despite control protocols for double-well potentials that perform accurate and efficient bit erasure [**33**], we run the flux qubit in a mode that yields imperfect erasure (Fig. 2.3(C)). As with the simulations, our intention is to illustrate how trajectory classes and the TCFT can be used to diagnose and interpret success and failure in microscopic information processing using only mesoscopic measurements of work, which is done more clearly by increasing the probabilities of rare events.

Interplay between the geometric, linear magnetic, and the nonlinear Josephson inductances gives rise to a potential landscape that can be controlled with external bias fluxes. It is natural to call the ϕ_{x} and ϕ_{xdc} fluxes, threading the differential mode and the small SQUID loop, respectively, the *tilt* and *barrier controls*. (See (Fig. 2.3(A) caption.) Appendix 2.J presents a derivation of the flux

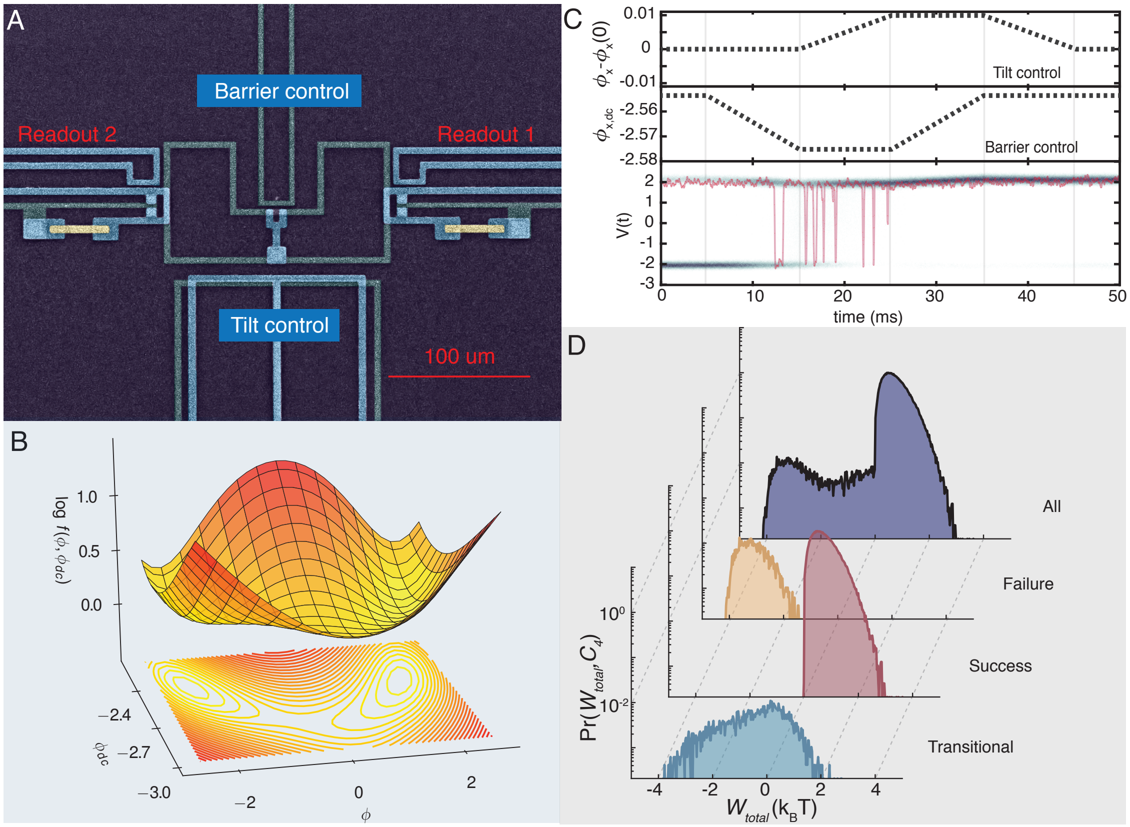

FIGURE 2.3. Superconducting implementation of metastable memory and bit erasure driven by thermal fluctuations: (A) Optical micrograph of a gradiometric flux qubit with control lines and local magnetometers for state readout. The flux ϕ_{x}, threading the large U-shaped differential-mode loop, controls the potential's tilt and flux ϕ_{xdc}, threading the small SQUID loop, controls the potential barrier height. Currents in the *barrier control* and *tilt control* lines modulate those fluxes. (B) Calculated potential energy landscape at the beginning of the erasure protocol; see Eqs. (2.24) and (2.25). (C top) Sequence of tilt and barrier control waveforms implementing bit erasure and (C bottom) sample of resulting magnetometer traces tracking the system's internal state. Note that for this experiment, in contrast to the simulation, two possible informational states were present at all times during the protocol. So, though the barrier was reduced sufficiently to allow transitions to the target R state, the trace is attracted to either a positive or negative value at all times. (D) Work distributions $\Pr(W_{\mathrm{total}}|C_4)$ over trajectories conditioning on the Success, Fail, and Transitional classes. Experimental distributions obtained from 10^5 protocol repetitions.

qubit potential and details its calibration. All experiments presented here were carried out at a temperature of 500 mK.

To execute an erasure protocol, we first choose an information-storage state with a tall barrier and two equal-depth wells. The two-dimensional potential for this at the calibrated device parameters is depicted in Fig. 2.3(B). We implement the bit erasure protocol as a time-domain deformation imposed by the two control fluxes that starts and ends at the storage configuration. In contrast to the simulation, the flux qubit maintains two metastable regions and, hence, two informational states L and R at all times, though they are shallow enough to allow transitions as the barrier drops. The amplitudes of the control waveforms in reduced units are small; see Fig. 2.3(C). Due to this, the microscopic energetics change linearly as a function of the control fluxes.

We use a local dc-SQUID magnetometer to continuously monitor the trapped flux state in the device—*Readout* 1 in Fig. 2.3(A). The digitized signal has a rise time of 100 μs, after which the two logical states are discriminated virtually without error. A typical magnetometer trace $V(t)$ acquired during the execution of the erasure protocol is shown in Fig. 2.3(C). We operate the magnetometer with a low-amplitude AC current bias at 10 MHz to avoid an increase in the effective temperature during continuous readout of the flux state due to wideband electromagnetic interference.

To collect work statistics, we repeat the erasure protocol 10^5 times. We identify the logical-state transitions from the magnetometer traces as zero-crossings, recording the direction δ_i—sign convention: $+1$ (-1) for a L-to-R (R-to-L) transition—and the time t_i relative to the start of the protocol. We evaluate a single-shot work estimate $W = \sum_i \delta_i U_{LR}(t_i)$, where $U_{LR}(t) = U_R(t) - U_L(t)$ is the biasing of the potential minima at time t_i. Making use of the linearity of the system energetics and the choice of offsets and compensation coefficients, we find $U_{LR}(t) = A\left(\phi_x(t) - \phi_x(0)\right)$, with the coefficient $A = 210\text{K} \times k_\text{B}$ evaluated from the calibrated potential. The above work estimate based on the logical-state transitions is an accurate estimate of the true microscopic work assuming that the timescales for the state transitions and for changes in the control parameters are much slower than the intra-well equilibration. (See App. 2.H.)

The total work distribution estimated from the flux qubit experiments is shown as the rear-most distribution in Fig. 2.3(D). Using the previous microstate trajectory partitioning into the *Success, Fail,* and *Transitional* trajectory classes reveals a decomposition of the total work distribution given by Fig. 2.3(D)(Three front panels). The close similarity with our simulations (Fig. 2.2) is notable. Especially so, given the rather substantial differences between the simulated system (idealized double-well potential, thermal noise, exactly one-dimensional system, ...) and the experimental system (complex potential in two dimensions, nonideal fluctuations, ...). *A priori* it is not clear that the TCFT predictions should apply so directly and immediately to the real-world qubit. That is, until one recalls that trajectory-class membership is a topological property and that trajectories carry their probabilities and so the thermodynamics.

In point of fact, these differences serve to emphasize the descriptive power and robustness of the mesoscopic-work TCFT: Despite substantial differences in system detail they successfully diagnose the information-processing classes of microscopic trajectories.

Indeed, looking to thermodynamic transformations beyond bit erasure, the essential requirement of our analysis is for the protocol to be slow enough compared to the time-scales of oscillations due to the potential and of the thermal fluctuations. With this, the system is always near metastable equilibrium—what we call *metastable quasistationarity*. This ensures that in an amount of time, small on the time-scale of the protocol, the system visits every point in the potential in proportion to the metastable distribution for the metastable region it occupies. Since the work rate for a particle is determined by the rate of change in potential at its location, the work rate must then be that of the metastable distribution's. We then need only describe which metastable region a particle is in as a function of time to characterize its total work. The specific shape and dimensionality of these metastable regions are then insignificant for determining the shape and qualitative features of the total work distribution. Under these conditions, there is substantial robustness. This will be especially helpful when using the TCFT to monitor thermodynamic computing in biological systems where, in many cases, information-bearing degrees-of-freedom cannot be precisely modeled.

2.6. Conclusion

We experimentally demonstrated that work fluctuations generated by information engines are highly structured. Nonetheless, they strictly obeyed a suite of time-reversal symmetries—the trajectory-class fluctuation theorems introduced here. The latter are direct signatures of how a system's informational states evolve and they identify functional and nonfunctional microscopic trajectory bundles. We showed that the trajectory-class fluctuation theorems naturally interpolate between Jarzynski's integral and Crooks' detailed fluctuation theorems, providing a unified diagnostic probe of nonequilibrium thermodynamic transformations that support information processing.

The trajectory-class fluctuation theorems gave a detailed thermodynamic analysis of the now-common example of erasing a bit of information as an external protocol manipulated a stochastic particle in a double-well potential (simulation) and the stochastic state of a flux qubit (experiment).

To give insight into the new level of mechanistic analysis possible, we briefly discussed the untilt trajectory-class partitioning. Though ignoring other protocol stages, this was sufficient to capture the basic trajectory classes that generate the overall work distribution's features. Partitioning on informational-state occupation times during barrier raising and untilting—an alternative used in follow-on studies—yields an even more incisive decomposition of the work distributions and diagnosis of informational functioning. Practically, the corresponding bounds on thermodynamic resources obtained via the TCFT also improve on current estimation methods. The net result is that trajectory-class fluctuation analysis can be readily applied to debug thermodynamic computing by engineered or biological systems.

2.A. Principles of Thermodynamic Computing: A recent synopsis

A number of closely-related thermodynamic costs of computing have been identified, above and beyond the *house-keeping heat* that maintains a system's overall nonequilibrium dynamical state. First, there is the *information-processing Second Law* [**34**] that extends Landauer's original bound on erasure [**6**] to dissipation in general computing and properly highlights the central role of information generation measured via the physical substrate's dynamical Kolmogorov-Sinai entropy. It specifies the minimum amount of energy that must be supplied to drive a given amount of computation forward. Second, when coupling thermodynamic systems together, even a single system and a complex environment, there are transient costs as the system synchronizes to, predicts, and then adapts to errors in its environment [**35**,**36**,**37**]. Third, the very modularity of a system's organization imposes thermodynamic costs [**38**]. Fourth, since computing is necessarily far out of equilibrium and nonsteady state, there are costs due to driving transitions between information-storage states [**39**]. Fifth, there are costs to generating randomness [**40**], which is itself a widely useful resource. Finally, by way of harnessing these principles, new strategies for optimally controlling nonequilibrium transformations have been introduced [**33**, **41**, **42**, **43**].

2.B. Microscopic Stochastic Thermodynamical System

For concreteness, we concentrate on a one-dimensional system: a particle with position and momentum in an external potential $V(x,t)$ and in contact with a heat reservoir at temperature T. An external controller adds or removes energy from a work reservoir to change the form of the potential $V(\cdot, t)$ via a predetermined *erasure protocol* $\{(\beta(t), \delta(t)) : 0 \leq t \leq \tau\}$. (See App. 2.H for details on the alternative definitions of work.) The potential takes the form:

$$V(x,t) = ax^4 - b_0\beta(t)x^2 - c_0\delta(t)x \ ,$$

with constants $a, b_0, c_0 > 0$. During the erasure protocol, $\beta(t)$ and $\delta(t)$ change one at a time piecewise-linearly through four protocol substages: (1) *drop barrier*, (2) *tilt*, (3) *raise barrier*, and (4) *untilt*, as shown in Table 2.1. The system starts at time $t = 0$ in the equilibrium distribution for a double-well $V(x, 0)$ at temperature T. Being equiprobable, the informational states associated with each of the two wells thus contain 1 bit of information [**31**]. The effect of the control protocol on the system potential and system response is graphically displayed in Fig. 2.1.

Stage	Drop Barrier	Tilt	Raise Barrier	Untilt	
t_s	t_0	t_1	t_2	t_3	t_4
$\beta(t)$	$\frac{t_1-t}{t_1-t_0}$	0	$\frac{t-t_2}{t_3-t_2}$	1	
$\delta(t)$	0	$\frac{t-t_1}{t_2-t_1}$	1	$\frac{t_4-t}{t_4-t_3}$	

Table 2.1. Erasure protocol.

We model the erasure physical information processing with *underdamped* Langevin dynamics:

$$\begin{aligned} dx &= v dt \\ m\, dv &= \sqrt{2k_\mathrm{B}T\gamma}\, r(t)\sqrt{dt} - \left(\frac{\partial}{\partial x}V(x,t) + \gamma v\right) dt \ , \end{aligned}$$

where k_B is Boltzmann's constant, γ is the coupling between the heat reservoir and system, m is the particle's mass, and $r(t)$ is a memoryless Gaussian random variable with $\langle r(t) \rangle = 0$ and $\langle r(t)r(t') \rangle = \delta(t - t')$.

For comparison to experiment, we simulated erasure with the following parameters, sufficient to fully specify the dynamics: $\gamma\tau/m = 500$, $2k_\mathrm{B}T\tau^2 a/(mb_0) = 2.5 \times 10^5$, $b_0^2/(4ak_\mathrm{B}T) = 7$, and $\sqrt{8a/b_0^3}c_0 = 0.4$. The resulting potential, snapshotted at times during the erasure substages, is shown in Fig. 2.1(Inner plot sequence).

Reliable information processing dictates that we set time scales so that the system temporarily, but stably, stores information. To support metastable-quasistatic behavior at all times the relaxation rates of the informational states are much faster than the rate of change of the potential, keeping the system near metastable equilibrium throughout. The entropy production for such protocols tends to be minimized.

2.C. Trajectory-Class Fluctuation Theorem and Interpretation

Here, we describe the trajectory-class fluctuation theorems, explaining several of their possible implications and exploring their application to both the simulations and flux qubit experiment. Their derivations are given in the section following.

First, we treat each system trajectory $\overrightarrow{z}$ as a function from time between 0 and τ (the time interval of a control protocol) to the set of possible system microstates. Then consider a forward process distribution P, defined by the probabilities of the system microstate trajectories $\overrightarrow{\mathcal{Z}}$ due to an initial equilibrium microstate distribution evolving forward in time under the control protocol. Then, the reverse process distribution R^{eq} is determined by preparing the system in equilibrium in the final protocol configuration and running the reverse protocol. The reverse protocol is the original protocol conducted in reverse order but also with objects that are odd under time reversal, like magnetic fields, negated. The time-reversal of a trajectory is $\overrightarrow{z}^{\dagger}(t) = (\overrightarrow{z}(\tau - t))^{\dagger}$, where for a microstate z the time reverse $z^{\dagger}$ is simply z but with time odd components (e.g., momentum or spin) negated. In other words, time reversing a trajectory runs the trajectory backwards while also negating all time-odd components of the microstates. For a measurable subset of trajectories $C \subset \overrightarrow{\mathcal{Z}}$, called a *trajectory class*, let the class average $\langle \cdot \rangle_C$ denote an average over the ensemble of forward process trajectories conditioned on the trajectories being in the class C. Let $P(C)$ and $R^{\mathrm{eq}}(C^{\dagger})$ denote the probabilities of observing the class C in the forward process and the reverse class $C^{\dagger} = \{\overrightarrow{z}^{\dagger} | \overrightarrow{z} \in C\}$ in the reverse process, respectively.

We first introduce a *trajectory-class fluctuation theorem* (TCFT) for the *class-averaged exponential work* $\langle e^{-W/k_{\mathrm{B}}T} \rangle_C$:

$$\langle e^{-W/k_{\mathrm{B}}T} \rangle_C = \frac{R^{\mathrm{eq}}(C^{\dagger})}{P(C)} e^{-\Delta F^{\mathrm{eq}}/k_{\mathrm{B}}T} \ , \tag{2.4}$$

with ΔF^{eq} the system equilibrium free energy change. We also introduce a *class-averaged work* TCFT:

$$
\begin{aligned}
\text{(2.5)} \qquad \langle W \rangle_C &= \Delta F^{\text{eq}} \\
&+ k_{\text{B}} T \Big(D_{\text{KL}} \Big[P(\overrightarrow{\mathcal{Z}}|C) || R^{\text{eq}}(\overrightarrow{\mathcal{Z}}^{\dagger}|C^{\dagger}) \Big] + \ln \frac{P(C)}{R^{\text{eq}}(C^{\dagger})} \Big) \ .
\end{aligned}
$$

This employs the Kullback-Liebler divergence $D_{\text{KL}}[\ \cdot\ ]$ taken between forward and reverse process distributions over all class trajectories $\overrightarrow{z} \in C$, conditioned on the forward class C and reverse class $C^{\dagger}$, respectively. If we disregard this divergence, which is nonnegative and can generally be difficult to obtain experimentally, we then find the lower bound $\langle W \rangle_C^{\text{min}}$ on the class-averaged work of Eq. (2.2).

If we choose the class C to consist of only a single trajectory, we recover detailed fluctuation theorems. For example, Eq. (2.4) then simplifies to Crooks' detailed fluctuation theorem [**16**]:

$$
\text{(2.6)} \qquad e^{-W(\overrightarrow{z})/k_{\text{B}}T} = \frac{R^{\text{eq}}(\overrightarrow{z}^{\dagger})}{P(\overrightarrow{z})} e^{-\Delta F^{\text{eq}}/k_{\text{B}}T} \ .
$$

If, however, we take C to be the entire set of trajectories $\overrightarrow{\mathcal{Z}}$, we recover integral fluctuation theorems. In this case, Eq. (2.4) simplifies to Jarzynski's Equality [**2**]:

$$
\text{(2.7)} \qquad \langle e^{-W/k_{\text{B}}T} \rangle_{\overrightarrow{\mathcal{Z}}} = e^{-\Delta F^{\text{eq}}/k_{\text{B}}T} \ ,
$$

exploiting the fact that $\overrightarrow{\mathcal{Z}}^{\dagger} = \overrightarrow{\mathcal{Z}}$ and $P(\overrightarrow{\mathcal{Z}}) = R^{\text{eq}}(\overrightarrow{\mathcal{Z}}) = 1$.

Furthermore, many other fluctuation theorems can be seen as special cases of the TCFT. In particular, Eq. (9) of Ref. [**44**] is closely related to Eq. (2.4). Having a nearly identical form, the former is a special case of the latter in that the corresponding classes consist of trajectories defined by restrictions on the visited microstates up to only a finite number of times. Similarly, Eqs. (6) and (7) of Ref. [**45**] derive from Eqs. (2.4) and (2.2); by considering a system with negligible contact with the thermal bath during the protocol and coarse-graining on features of the visited microstate at a single time. Along the same lines, Eqs. (7) and (8) of Ref. [**23**] are obtained from Eq. (2.4) by considering a bit erasure process and trajectory classes corresponding to ending in the target well and in the opposite well, respectively. And, letting the trajectory class be all trajectories that yield a particular value of obtained work during the forward process, Eq. (2.4) reduces to Crooks' Work

Fluctuation Theorem [16]:

$$\frac{P(W)}{R^{\mathrm{eq}}(-W)} = e^{(W-\Delta F^{\mathrm{eq}})/k_{\mathrm{B}}T} \ , \tag{2.8}$$

where $P(W)$ and $R^{\mathrm{eq}}(-W)$ are the probabilities of obtaining values W and $-W$ for the work when running the forward and reverse process, respectively. Finally, yet other fluctuation theorems arise directly from the TCFT by particular class choices [19, 30, 46, 47].

In this way, one sees the TCFT is a suite that spans the space of fluctuation theorems between the extreme of the detailed theorems, that require very precise information about an individual trajectory, and the integral theorems, that describe the system's entire trajectory ensemble. It thus unifies a wide range of existing (and future) fluctuation theorems. Appendix 2.F below provides proofs.

2.D. Empirical Use in Statistical Estimation

Beyond the synthesis of distinct fluctuation theorems, the TCFT is empirically useful in greatly improving sampling and errors in statistical estimation. And, this is its primary role here—a diagnostic tool for thermodynamic computing. We can rearrange Eq. (2.4) to obtain Eq. (2.3)—an expression for estimating equilibrium free energy changes:

$$\Delta F^{\mathrm{eq}} = -k_{\mathrm{B}}T \ln\left(\frac{P(C)}{R^{\mathrm{eq}}(C^{\dagger})}\langle e^{-W/k_{\mathrm{B}}T}\rangle_C\right) \ . \tag{2.9}$$

Thus, to estimate free energy one sees that statistics are needed for only one particular class and its reverse. Generally, this gives a substantial statistical advantage over direct use of Jaryznski's equality:

$$\Delta F^{\mathrm{eq}} = -k_{\mathrm{B}}T \ln\langle e^{-W/k_{\mathrm{B}}T}\rangle_{\overrightarrow{\mathcal{Z}}} \ ,$$

since rare microstate trajectories may generate negative work values that dominate the average exponential work [32]. The problem is clear in the case of erasure. Recall from Fig. 2.2(Three front panels) that Fail trajectories generate the most-negative work values. In the limit of higher success-rate protocols that maintain low entropy production, failures generate more and more negative works, leading them to dominate when estimating average exponential works despite becoming negligible in probability.

In contrast, to efficiently determine the change in equilibrium free energy from Eq. (2.3), its form indicates that one should choose a class that (i) is common in the forward process, (ii) has a reverse class that is common in the reverse process, and (iii) generates a narrow work distribution. This maximizes the accuracy of statistical estimates for the three factors on the RHS. For example, while the equilibrium free energy change in the case of our erasure protocol is theoretically simple (zero); the Success class fits the criteria.

We can then monitor the class-averaged work in excess of its bound:

$$\begin{aligned}E_C &= \langle W \rangle_C - \langle W \rangle_C^{\mathrm{min}} \\ &= k_B T D_{\mathrm{KL}} \left[P(\overrightarrow{\mathcal{Z}}|C) || R^{\mathrm{eq}}(\overrightarrow{\mathcal{Z}}^{\dagger}|C^{\dagger}) \right] \\ &\geq 0 \ .\end{aligned}$$

The inequality in Eq. (2.2) is a refinement of the equilibrium Second Law and therefore the bound $\langle W \rangle_C^{\mathrm{min}}$ generally provides a more accurate estimate of the average work of trajectories in a class compared to the equilibrium free energy change ΔF^{eq}. More precisely, as we will see below, an average of the excess E_C over all classes C in a partition of trajectories must be smaller than the dissipated work $\langle W \rangle - \Delta F^{\mathrm{eq}}$. For trajectory classes with narrow work distributions, this can be a significant improvement. We can see this by Taylor expanding the LHS of Eq. (2.4) about the mean dimensionless work $\langle W/k_{\mathrm{B}}T \rangle_C$. This shows that Eq. (2.2) becomes an equality when the variance and higher moments vanish. Appendix 2.G below delves more into moment approximations. In any case, trajectory classes with narrow work distributions have small excess works E_C.

2.E. Fluctuations in Thermodynamic Computing: The Case of Erasure

Before applying the TCFT to analyze thermodynamic fluctuations during erasure, we first explore both Jarzynski's Equality Eq. (2.7) and Crooks' Work Fluctuation Theorem Eq. (2.8).

Since the erasure protocol is cyclic, the change in equilibrium free energy ΔF^{eq} vanishes. Jarzynski's Equality then predicts that the average exponential work $\langle e^{-W/k_{\mathrm{B}}T} \rangle_{\overrightarrow{\mathcal{Z}}}$ must be 1. From simulation, we obtain a value of $1.0025 \pm 5 \times 10^{-5}$; which is very close to the predicted value. From experiment, we obtain a value of $0.89 \pm 5 \times 10^{-5}$—within 10% of the prediction, but falls somewhat outside the expected error.

Crooks' Work Fluctuation Theorem predicts that the quantity $\ln(P(W)/R^{\text{eq}}(-W))$ must equal βW at each W. We verify this experimentally by building probability histograms for $P(W)$ and $R^{\text{eq}}(-W)$, taking their log ratios, and plotting against their binned work values expressed in units of $k_{\text{B}}T$. Figure 2.4 shows that the experiment follows the theoretical prediction quite closely. Though, as for the case of Jarzynski's Equality, the experimental results are not all within expected errors. The discrepancies appear to arise in the statistical errors in estimating the obtained work values for each trajectory. Due to the complicated nature of the approximations (see App. 2.H.3), estimating the corresponding error is challenging and will be left for future detailed investigation.

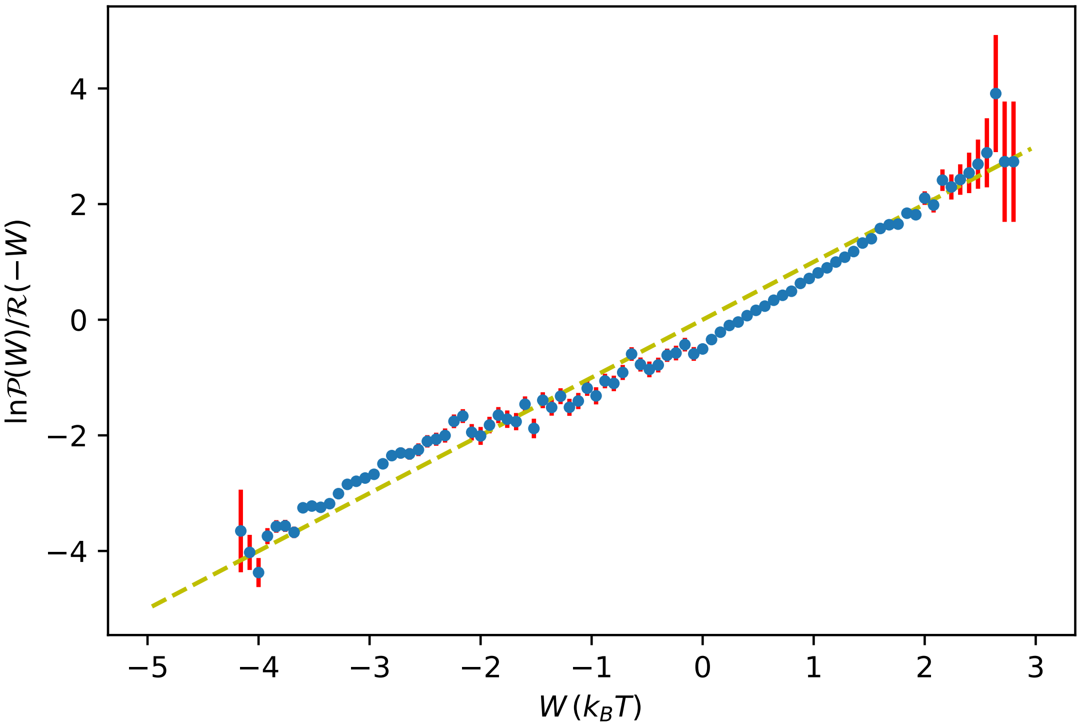

FIGURE 2.4. Flux qubit experiment work fluctuations: Crooks Work Fluctuation Theorem prediction (green dashed line), measured values (blue), and 1-σ statistical errors (red).

We now turn to analyze work fluctuations in various trajectory classes during the erasure operation, demonstrating that the TCFT allows analyzing the thermodynamics of trajectories falling between the Jarzynski and Crooks extreme classes. The main point here being that the classes between these extremes consist of "functionally" interpretable trajectories—e.g., successful and failed erasure.

In this way, one can diagnose the energetics and general thermodynamics of this functioning in the physical computing device.

To estimate $R^{\mathrm{eq}}(C^{\dagger})$ for the three chosen classes $C \in \{\text{Success}, \text{Failure}, \text{Transitional}\}$ of our simulated Erasure process, we ran 3.5×10^6 simulations of the reverse process. Table 2.2 shows that the Success and Fail classes have small excesses and, as seen in Fig. 2.2(Three front panels), these classes indeed have narrow work distributions. Elsewhere, we explore these and additional partition schemes, finding that the Transitional trajectories can be further partitioned to yield narrow work distributions so that all trajectory classes have small excesses E_C. In short, this demonstrates how well-formulated trajectory classes allow accurate estimates of the works for all trajectories.

Class C	$\langle W \rangle_C$	$\langle W \rangle_C^{\min}$	E_C
All $\overset{\rightrightarrows}{\mathcal{Z}}$	0.634	0.0	0.634
Success	0.713	0.683	0.030
Fail	-3.885	-3.951	0.066
Transitional	-0.546	-1.650	1.170

TABLE 2.2. Class-average works and bounds for different trajectory classes during erasure: *All* trajectories $\overset{\rightrightarrows}{\mathcal{Z}}$, *Success* trajectories, *Fail* trajectories, and *Transitional* trajectories. These are identified in Fig. 2.2 (Four front panels). From left to right, columns give the estimated class-average work $\langle W \rangle_C$, TCFT lower bound $\langle W \rangle_C^{\min}$, and their difference E_C. 3.5×10^6 simulations were run for each of the forward and reverse processes, with 96.2% trajectories successfully ending in the R informational state under the forward process.

Partition Q	$\langle W \rangle_{\overset{\rightrightarrows}{\mathcal{Z}}}$	$\langle W \rangle_Q^{\min}$	E_Q
Trivial $\left\{\overset{\rightrightarrows}{\mathcal{Z}}\right\}$	0.634	0.0	0.634
Untilt-Centric I	0.634	0.560	0.074
Untilt-Centric II	0.634	0.601	0.032

TABLE 2.3. Ensemble-average work and bounds due to different partitions: *Trivial* partition; *Untilt-Centric I* partition, composed of Success, Fail, and Transitional; and *Untilt-Centric II* partition, described in follow-on work. From left to right, columns give the estimated ensemble-average work, the partition bound $\langle W \rangle_Q^{\min}$, and their difference E_Q. All values in units of $k_{\mathrm{B}}T$.

To measure the efficacy of a given partition Q of trajectories into classes, we ask what the ensemble-average of class-average excess works is:

$$\begin{aligned} E_Q &= \sum_{C \in Q} P(C) E_C \\ &= \langle W \rangle_{\overrightarrow{\mathcal{Z}}} - \sum_{C \in Q} P(C) \langle W \rangle_C^{\min} \\ &= \langle W \rangle_{\overrightarrow{\mathcal{Z}}} - \langle W \rangle_Q^{\min} ~, \end{aligned}$$

with $\langle W \rangle_Q^{\min} = \sum_{C \in Q} P(C) \langle W \rangle_C^{\min}$.

From Eq. (2.2), we see that $\langle W \rangle_Q^{\min}$ is the coarse-grained lower bound on ensemble-average dissipation from Ref. [48]:

$$\langle W \rangle_Q^{\min} = \Delta F^{\mathrm{eq}} + k_{\mathrm{B}} T D_{\mathrm{KL}} \left[P(Q) || R^{\mathrm{eq}}(Q^\dagger) \right] ~,$$

where $D_{\mathrm{KL}} \left[\, \cdot \, \right]$ is the Kullback-Liebler divergence between forward and reverse process distributions over the trajectory classes $C \in Q$. Since Kullback-Liebler divergences are nonnegative, such a bound always provides an improvement over the equilibrium Second Law. Table 2.3 shows both $\langle W \rangle_Q^{\min}$ and E_Q for the trivial partition $\{\overrightarrow{\mathcal{Z}}\}$, where $\langle W \rangle_Q^{\min} = \Delta F^{\mathrm{eq}}$, our three-class partition, labeled *Untilt-Centric I*, and the improved partition described in follow-on work, labeled *Untilt-Centric II*. Application of the first partition simply implies the equilibrium Second Law. In this case, the latter two improve on the nonequilibrium Second Law that, calculated by assuming metastable starting and ending distributions, provides a lower bound on the average work equal to 0.533, the change in nonequilibrium free energy ΔF.

We can appeal to Landauer's erasure bound—$k_B T \ln 2 \approx 0.693 \, k_B T$—to calibrate the excesses E_C and E_Q. We see for the simulation data that our three-class partition Untilt-Centric I provides class-average work bounds that, on average, are only about 11% of $k_B T \ln 2$ from the actual class-average works. The more refined Untilt-Centric II partition reduces this excess to about 5% while the trivial partition fails by about 91% of $k_B T \ln 2$.

We also recover the equality of Ref. [48] for the ensemble-average work by averaging Eq. (2.5) over each class:

$$\begin{aligned}\langle W\rangle &= \sum_{C} P(C)\langle W\rangle_C \\ &= \Delta F^{\mathrm{eq}} + k_{\mathrm{B}}T\Big(\sum_{C\in Q} P(C) D_{\mathrm{KL}}\left[P(\overrightarrow{\mathcal{Z}}|C)||R^{\mathrm{eq}}(\overrightarrow{\mathcal{Z}}^{\dagger}|C^{\dagger})\right] + D_{\mathrm{KL}}\left[P(Q)||R^{\mathrm{eq}}(Q^{\dagger})\right]\Big) ~,\end{aligned}$$

which of course is lower bounded by $\langle W\rangle_Q^{\min}$.

These results suggest the criterion for optimal trajectory partitions: Select a partition sufficiently refined to yield tight bounds on class-average works, but no finer. Machine learning methods for model-order selection provide a basis for a natural classification scheme for trajectories that captures all relevant thermodynamics and information processing. Moreover, by changing the forward and reverse processes P and R^{eq} to begin in system microstate distributions other than equilibrium, a yet-broader class of TCFTs emerge. We can then find analogous results for heats and comparisons with works and nonequilibrium free-energy changes. We explore these extensions in depth elsewhere.

2.F. TCFT Derivations

We now present derivations for the two TCFTs introduced in Eqs. (2.4) and (2.5).

Assume that the system dynamics is described by a Hamiltonian specified in part by an external control protocol, as well as by a weak coupling to a thermal environment that induces steady relaxation to canonical equilibrium.

Start the system in equilibrium distribution π_0 for Hamiltonian $\mathcal{H}_0$ and run a protocol until time τ, causing the system Hamiltonian to evolve to $\mathcal{H}_\tau$. If we then hold the Hamiltonian at $\mathcal{H}_\tau$ for a long time, the system relaxes into the equilibrium distribution π_τ. The system's ensemble entropy change from $t = 0$ to $t = \infty$ is then:

$$\Delta S_{\mathrm{sys}} = \sum_{z} \left[-\pi_\tau(z)\ln \pi_\tau(z) + \pi_0(z)\ln \pi_0(z)\right] ~.$$

The trajectorywise system entropy difference is defined to be:

$$\Delta s_{\mathrm{sys}}(\overrightarrow{z}) = \ln \frac{\pi_0(z_0)}{\pi_\tau(z_\tau)} ~,$$

where z_0 and z_τ are the initial and final microstates of system microstate trajectory $\overrightarrow{z}$, respectively. Averaged over all trajectories $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$, this then becomes the ensemble entropy change.

Let $p(\overrightarrow{z}|z_0)$ denote the probability of obtaining system microstate trajectory $\overrightarrow{z}$ via the protocol conditioned on starting the system in state $z_0 = \overrightarrow{z}(0)$.

Now, start the system Hamiltonian at $\mathcal{H}_\tau$ and run the reverse protocol, ending the Hamiltonian at $\mathcal{H}_0$. We then obtain the trajectory $\overrightarrow{z}$ with a different conditional probability: $r(\overrightarrow{z}|z_0)$.

Assuming microscopic reversibility and given a system trajectory $\overrightarrow{z}$, the change in the heat bath's entropy is:

$$\Delta S_{\text{res}}(\overrightarrow{z}) = -\beta Q(\overrightarrow{z}) \tag{2.10}$$

$$= \ln \frac{p(\overrightarrow{z}|z_0)}{r(\overrightarrow{z}^\dagger|z^\dagger{}_\tau)} \ , \tag{2.11}$$

where $\beta = 1/k_{\text{B}}T$, $Q(\overrightarrow{z})$ is the net energy that flows out of the heat bath into the system given the trajectory $\overrightarrow{z}$, and $(\cdot)^\dagger$ denotes time-reversal. This holds for systems with strictly finite energies and Markov dynamics that induce the equilibrium distribution when control parameters are held fixed [49]. Both our simulated Duffing potential system and flux qubit obey these requirements at sufficiently short time scales. Then we can express the total trajectorywise change in entropy production due to a trajectory $\overrightarrow{z}$ as the sum of system and heat reservoir entropy changes:

$$\Delta S_{\text{tot}}(\overrightarrow{z}) = \Delta s_{\text{sys}}(\overrightarrow{z}) + \Delta S_{\text{res}}(\overrightarrow{z}) \ .$$

Since $\pi_t(z) = e^{-\beta(\mathcal{H}_t(z) - F^{\text{eq}}_t)}$, with F^{eq}_t the system's equilibrium free energy at time t, we can write:

$$\begin{aligned}
\Delta S_{\text{tot}}(\overrightarrow{z}) &= -\ln \pi_\tau(z_\tau) + \ln \pi_0(z_0) - \beta Q(\overrightarrow{z}) \\
&= \beta \left(\mathcal{H}_\tau(z_\tau) - F^{\text{eq}}_\tau\right) - \beta \left(\mathcal{H}_0(z_0) - F^{\text{eq}}_0\right) - \beta Q(\overrightarrow{z}) \\
&= \beta \left(\Delta\mathcal{H}(\overrightarrow{z}) - Q(\overrightarrow{z}) - \Delta F^{\text{eq}}\right) \\
&= \beta \left(W(\overrightarrow{z}) - \Delta F^{\text{eq}}\right) \ .
\end{aligned}$$

Using Eq. (2.10), we also have:

$$\begin{aligned}\Delta S_{\text{tot}}(\overrightarrow{z}) &= \Delta s_{\text{sys}}(\overrightarrow{z}) + \Delta S_{\text{res}}(\overrightarrow{z}) \\ &= \ln \frac{\pi_0(z_0)}{\pi_\tau(z_\tau)} \frac{p(\overrightarrow{z}|z_0)}{r(\overrightarrow{z}^\dagger | z^\dagger{}_\tau)} \\ &= \ln \frac{P(\overrightarrow{z})}{R^{\text{eq}}(\overrightarrow{z}^\dagger)}\end{aligned}$$

with:

$$\begin{aligned}P(\overrightarrow{z}) &= \pi_0(z_0) p(\overrightarrow{z}|z_0) \text{ and} \\ R^{\text{eq}}(\overrightarrow{z}^\dagger) &= \pi_\tau(z_\tau) r(\overrightarrow{z}^\dagger | z^\dagger_\tau) \ .\end{aligned}$$

Combining, we obtain Crooks' detailed fluctuation theorem [**16**]:

$$R^{\text{eq}}(\overrightarrow{z}^\dagger) = P(\overrightarrow{z}) e^{-\beta\left(W(\overrightarrow{z}) - \Delta F^{\text{eq}}\right)} \ . \tag{2.12}$$

From here, we derive our first TCFT by integrating each side of Eq. (2.12) over all trajectories $\overrightarrow{z}$ in a measurable set $C \subset \overrightarrow{\mathcal{Z}}$. Starting with the LHS and recalling the Iverson bracket $[\cdot]$, which is 1 when the interior expression is true and 0 when false, we have:

$$\begin{aligned}\int d\overrightarrow{z} [\overrightarrow{z} \in C] R^{\text{eq}}(\overrightarrow{z}^\dagger) &= \int d\overrightarrow{z}^\dagger [\overrightarrow{z} \in C] R^{\text{eq}}(\overrightarrow{z}^\dagger) \\ &= \int d\overrightarrow{z}^\dagger [\overrightarrow{z}^\dagger \in C^\dagger] R^{\text{eq}}(\overrightarrow{z}^\dagger) \\ &= \int d\overrightarrow{z} [\overrightarrow{z} \in C^\dagger] R^{\text{eq}}(\overrightarrow{z}) \\ &= R^{\text{eq}}(C^\dagger) \ .\end{aligned}$$

The first three steps used the unity of the Jacobian in reversing a microstate, the definition $C^\dagger = \{\overrightarrow{z}^\dagger | \overrightarrow{z} \in C\}$, and swapping all instances of $\overrightarrow{z}^\dagger$ with $\overrightarrow{z}$, respectively. Integrating the RHS of

Eq. (2.12) then gives:

$$\begin{aligned}
\int d\overrightarrow{z}[\overrightarrow{z} \in C]P(\overrightarrow{z})e^{-\beta\left(W(\overrightarrow{z})-\Delta F^{\mathrm{eq}}\right)} \\
&= e^{\beta\Delta F^{\mathrm{eq}}} \int d\overrightarrow{z} P(\overrightarrow{z}, C)e^{-\beta W(\overrightarrow{z})} \\
&= P(C)e^{\beta\Delta F^{\mathrm{eq}}} \int d\overrightarrow{z} P(\overrightarrow{z}|C)e^{-\beta W(\overrightarrow{z})} \\
&= P(C)e^{\beta\Delta F^{\mathrm{eq}}} \langle e^{-\beta W}\rangle_C \ .
\end{aligned}$$

Combining, we have our first TCFT, Eq. (2.4).

To obtain the second TCFT, we first change the form of Eq. (2.12):

$$W(\overrightarrow{z}) = \Delta F^{\mathrm{eq}} + \beta^{-1} \ln \frac{P(\overrightarrow{z})}{R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger})} \ .$$

Then we calculate the class-average. The equilibrium free energy change is unaffected while the rightmost term becomes:

$$\begin{aligned}
\beta^{-1} \left\langle \ln \frac{P(\overrightarrow{z})}{R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger})} \right\rangle_C &= \beta^{-1} \int_C d\overrightarrow{z} P(\overrightarrow{z}|C) \ln \frac{P(\overrightarrow{z})}{R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger})} \\
&= \beta^{-1} \int_C d\overrightarrow{z} P(\overrightarrow{z}|C) \ln \frac{P(\overrightarrow{z}|C)P(C)}{R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger}|C^{\dagger})R^{\mathrm{eq}}(C^{\dagger})} \\
&= \beta^{-1} \left(\int_C d\overrightarrow{z} P(\overrightarrow{z}|C) \ln \frac{P(\overrightarrow{z}|C)}{R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger}|C^{\dagger})} + \ln \frac{P(C)}{R^{\mathrm{eq}}(C^{\dagger})} \right) \\
&= \beta^{-1} \left(D_{\mathrm{KL}} \left[P(\overrightarrow{z}|C) || R^{\mathrm{eq}}(\overrightarrow{z}^{\dagger}|C^{\dagger}) \right] + \ln \frac{P(C)}{R^{\mathrm{eq}}(C^{\dagger})} \right) \ ,
\end{aligned}$$

which gives Eq. (2.5)'s TCFT.

2.G. Class-Averaged Work Approximation for Narrow Distributions

Here, we demonstrate that the class-averaged work $\langle W\rangle_C$ approaches its bound $\langle W\rangle_C^{\mathrm{min}}$ when the variance and higher moments of the class' distribution of works vanish. One concludes that $\langle W\rangle_C^{\mathrm{min}}$ is a good approximation for $\langle W\rangle_C$ when the class' work distribution is narrow.

We first express the LHS of Eq. (2.4) in terms of the unitless distance of work from its class-average:

$$\langle e^{-\beta W}\rangle_C = \langle e^{-x}\rangle_C \, e^{-\beta \langle W\rangle_C} \ ,$$

with $x = \beta(W - \langle W \rangle_C)$. Then, we Taylor expand the exponential inside the class-average:

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

with $a = \sum_{n=2}^{\infty} \frac{(-1)^n}{n!} \langle x^n \rangle_C$. Equation (2.4) then gives:

$$(1 + a) e^{-\beta \langle W \rangle_C} = \frac{R^{\mathrm{eq}}(C^\dagger)}{P(C)} e^{-\beta \Delta F^{\mathrm{eq}}} \ .$$

Since e^{-x} is convex,

$$(1 + a) = \langle e^{-x} \rangle_C \geq e^{-\langle x \rangle_C} = 1 \ ,$$

so $a \geq 0$. Then:

$$\begin{aligned}\langle W \rangle_C &= \Delta F^{\mathrm{eq}} + \beta^{-1} \ln \frac{P(C)}{R^{\mathrm{eq}}(C^\dagger)} + \beta^{-1} \ln(1 + a) \\ &\geq \Delta F^{\mathrm{eq}} + \beta^{-1} \ln \frac{P(C)}{R^{\mathrm{eq}}(C^\dagger)} \\ &= \langle W \rangle_C^{\mathrm{min}} \ .\end{aligned}$$

The second line becomes an equality when a goes to zero, which occurs as the variance and higher moments vanish.

2.H. Work Definitions and Experimental Estimation

Properly estimating the required works and devolved heats from experimental devices undergoing cyclic control protocols requires explicitly and consistently accounting for energy and information flows between the system, its environment, and the controlling laboratory apparatus. To this end, we construct a model Hamiltonian universe for common processes involving small systems interacting with laboratory apparatus and a thermal environment. After deriving key equalities for two definitions of work, the inclusive and exclusive works, we define a method of approximating them in appropriate cyclic protocols.

2.H.1. The Model Universe and Hamiltonian. To study a small system that exchanges energy with its environment in the forms of heat and work, we introduce a model universe: a *system* *of interest*, a *heat bath*, and a *lab* (laboratory apparatus) that controls the system and derives any needed energy from a *work reservoir*. The system directly interacts with both the heat bath and the lab, but the heat bath and lab are not directly coupled.

We assume that a Hamiltonian $\mathcal{H}$ describes the universe's evolution and that there is a set of generalized coordinates which can be sensibly partitioned into those for the system, heat bath, and lab. Then, we decompose the universe Hamiltonian into the following form:

$$\mathcal{H}(s,b,l) = H_{\mathrm{B}}(b) + h_{\mathrm{S,B}}(s,b) + H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s,l) + H_{\mathrm{L}}(l) ~,$$

where s, b, and l denote both the generalized coordinates and conjugate momenta for the system, bath, and lab, respectively. For any universe Hamiltonian $\mathcal{H}$, there can be many choices for this decomposition.

We also define the system Hamiltonian $\mathcal{H}'$ as the three components that depend on the system coordinates:

$$\mathcal{H}'(s;b,l) = h_{\mathrm{S,B}}(s,b) + H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s,l) ~.$$

First, consider the subset of lab coordinates l for which $h_{\mathrm{S,L}}$ has nontrivial dependence. These so-called *protocol parameters* λ are often simple and much fewer than the entire set of l. We often assume that we have total control of their evolution. More precisely, under an appropriate preparation for the lab at time $t = 0$, a specific trajectory for the protocol parameters $\{\lambda(t)\}_t$ for $0 \leq t \leq \tau$ is guaranteed for all preparations of the heat bath and system coordinates. We refer to the parameter trajectory as the *protocol*.

Suppose the heat-bath degrees of freedom that interact with the system change much faster than the system's. We can assume that the system response to the bath resembles Brownian motion. On the time scale of changes in the system coordinates, then, we ignore the system-bath interaction term $h_{\mathrm{S,B}}$ in writing the system Hamiltonian $\mathcal{H}'$:

$$\begin{aligned}\mathcal{H}'(s;\lambda) &= H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s,\lambda) \\ &= T(s) + V(s,\lambda) ~.\end{aligned}$$

The latter decomposition into kinetic energy T and potential energy V can be used to write Langevin equations of motion for the system. Furthermore, if the heat bath has a relaxation time sufficiently short that it is roughly in equilibrium at all times with fixed temperature, then its influence on the system will be memoryless.

2.H.2. Inclusive and Exclusive Works and Heats. The basic scenario for executing a protocol is as follows. The universe coordinates begin according to a given initial distribution $\Pr(s)$ at time $t = 0$ and they evolve in isolation until $t = \tau$. As above, we assume that a well-defined protocol $\{\lambda_t\}_t$ emerges due to our preparation of the lab coordinates.

We label all energy exchanged between the system and lab as *work* and all energy exchanged between the system and heat bath as *heat*. Since the lab is directly coupled only to the system, the work they exchange is given by the change in energy of the lab's work reservoir. Similarly, since the heat bath is directly coupled only to the system, the heat exchanged is given by the change in the heat bath's energy.

Note that this requires choices as to what constitute the energies of the three universe subsystems. While H_{B}, H_{S}, and H_{L} define energies for the heat bath, system, and work reservoir, respectively, what of $h_{\mathrm{S,L}}$ and $h_{\mathrm{S,B}}$? If all subsystems were macroscopic, these interaction terms would be negligible. While it may be desirable to assume that the system is only weakly coupled to the heat bath—so that $h_{\mathrm{S,B}}$ can be ignored—$h_{\mathrm{S,L}}$ can be significant in many important small systems.

And so, in general, we define the system energy to be H_{S} plus any portions of $h_{\mathrm{S,L}}$ and $h_{\mathrm{S,B}}$. Then the work reservoir energy is H_{L} plus the rest of $h_{\mathrm{S,L}}$, while the heat bath energy is H_{B} plus the rest of $h_{\mathrm{S,B}}$. To make these distinctions clear we label two types of works, each corresponding to the two extremes for allocation of $h_{\mathrm{S,L}}$ between the system and work reservoir: the *inclusive work* W and the *exclusive work* W_0 [50]. Specifically:

$$\begin{aligned}\frac{dW}{dt} &= -\frac{d}{dt}(H_{\mathrm{L}}) &&= \frac{d}{dt}(H_{\mathrm{B}} + h_{\mathrm{S,B}} + H_{\mathrm{S}} + h_{\mathrm{S,L}}) \\ \frac{dW_0}{dt} &= -\frac{d}{dt}(h_{\mathrm{S,L}} + H_{\mathrm{L}}) &&= \frac{d}{dt}(H_{\mathrm{B}} + h_{\mathrm{S,B}} + H_{\mathrm{S}}) \ .\end{aligned}$$

We can similarly define the *inclusive heat* Q and *exclusive heat* Q_0 depending on how we allocate $h_{\mathrm{S,B}}$ between the system and heat bath:

$$\begin{aligned}\frac{dQ}{dt} &= -\frac{d}{dt}(H_{\mathrm{B}}) &&= \frac{d}{dt}(h_{\mathrm{S,B}} + H_{\mathrm{S}} + h_{\mathrm{S,L}} + H_{\mathrm{L}}) \\ \frac{dQ_0}{dt} &= -\frac{d}{dt}(H_{\mathrm{B}} + h_{\mathrm{S,B}}) &&= \frac{d}{dt}(H_{\mathrm{S}} + h_{\mathrm{S,L}} + H_{\mathrm{L}}) ~.\end{aligned}$$

The inclusive work corresponds to fully including $h_{\mathrm{S,L}}$ in the system energy, while the exclusive work corresponds to excluding it. Inclusive and exclusive heat correspond similarly with respect to $h_{\mathrm{S,B}}$.

There is a key relation between the inclusive and exclusive works:

$$\frac{dW}{dt} = \frac{dW_0}{dt} + \frac{dh_{\mathrm{S,L}}}{dt} ~. \tag{2.13}$$

That is, the inclusive work for an interval of time equals the sum of the exclusive work and the change in the system-lab interaction term $h_{\mathrm{S,L}}$.

In the above expressions, calculating the rate of change of a work or heat requires the time derivative of one or more of H_{L} and H_{B}. This can be problematic. Fortunately, there are alternate forms that are amenable. One can show that the inclusive work rate is given by:

$$\begin{aligned}\frac{dW}{dt} &= -\frac{dH_{\mathrm{L}}}{dt} \\ &= \frac{\partial h_{\mathrm{S,L}}}{\partial \lambda}\frac{d\lambda}{dt} ~.\end{aligned} \tag{2.14}$$

This is a more common definition for the work rate in small-system nonequilibrium thermodynamics. And, it allows the work to be calculated as:

$$W(t,t') = \int_t^{t'} dt'' \frac{d\lambda}{dt''}\frac{\partial h_{\mathrm{S,L}}(s,\lambda)}{\partial \lambda}|_{s=s(t''),\lambda=\lambda(t'')} ~. \tag{2.15}$$

The exclusive work W_0 has a corresponding form:

$$\begin{aligned}\frac{dW_0}{dt} &= -\frac{d(h_{\mathrm{S,L}} + H_{\mathrm{L}})}{dt} \\ &= -\frac{\partial h_{\mathrm{S,L}}}{\partial s}\frac{ds}{dt} ~,\end{aligned} \tag{2.16}$$

For the case where $h_{\mathrm{S,L}}$ is a scalar potential for s, this is the product of the corresponding force with velocity. This makes the exclusive work equal to a familiar mechanics definition of work as the integral of the dot product of force and displacement:

$$W_0(t,t') = -\int_t^{t'} dt'' \frac{ds}{dt''} \frac{\partial h_{\mathrm{S,L}}(s,\lambda)}{\partial s}|_{s=s(t''),\lambda=\lambda(t'')} \ .$$

In this way, we write the inclusive and exclusive work rates in terms of the rates of change of the system and work-reservoir interaction term $h_{\mathrm{S,L}}$ with respect to either the system or work reservoir coordinates.

2.H.3. Approximating Inclusive Work Experimentally. For the flux qubit experimental system investigated here, we assume the following:

$$H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s,\lambda) = T(s) + V(s,\lambda) \ .$$

That is, as far as the flux qubit and work reservoir are concerned, the only relevant energies at least partially ascribable to the flux qubit are its kinetic energy and the potential energy with the work reservoir. $h_{\mathrm{S,L}}$ must then capture the change in the potential V due to changes in the protocol parameters. We could simply define $h_{\mathrm{S,L}}(s,\lambda) = V(s,\lambda)$ so that $H_{\mathrm{S}}(s) = T(s)$. However, it is more useful to allocate the initial potential energy to H_{S}. That is:

$$H_{\mathrm{S}}(s) = T(s) + V(s,\lambda_0) \text{ and}$$

$$h_{\mathrm{S,L}}(s,\lambda) = V(s,\lambda) - V(s,\lambda_0) \ .$$

For cyclic protocols where $V(\cdot,\lambda_0) = V(\cdot,\lambda_\tau)$, such as in our erasure operation, $h_{\mathrm{S,L}}(s(t),\lambda(t))$ vanishes for all trajectories at $t = 0, \tau$. By Eq. (2.13) we then have the useful equality $W = W_0$ between inclusive and exclusive works taken over the entire protocol.

Estimating W for a system trajectory is then equivalent to estimating W_0 for the cyclic protocols we consider. In the flux qubit, the form of $h_{\mathrm{S,L}}$ is known and the specific protocol $\{\lambda_t\}_{t\in[0,\tau]}$ is known. Unfortunately, we lack sufficient information about its instantaneous state s at all times to estimate the total inclusive work from Eq. (2.15), since the device's physics precludes precise measurements of system flux ϕ—the relevant part of s for determining the potential $h_{\mathrm{S,L}}$. Instead, we do have reliable measurement of large and stable changes in the flux ϕ. This specifically monitors when the system moves between wells in a double-well potential $V(\cdot,\lambda(t))$, if the rate of transition between wells is sufficiently slow.

And so, we can use information about the flux ϕ to approximate the exclusive work contribution at each moment in time. Then, adding up these contributions yields an approximation to the total exclusive work W_0 over the entire protocol and therefore of the inclusive work W over the entire protocol. Note that the protocols used here maintain two wells at all times for the system flux ϕ. We develop the approximation in two steps.

2.H.3.1. *First-Order Approximation.* We first partition the potential in flux space into three segments. Two segments constitute the wells for the flux in which that state spends all its time, except for very brief transitions between wells. Then, the third segment connects the two wells, capturing the dynamics arising from crossing the barrier that separates them.

We require that the partitioning allows the following two approximations. First, the particle spends negligible total duration in between the two wells. Second, the wells do not change shape over the protocol, but instead simply raise or lower in potential at different times, if they change at all. This means that the shape of the system-lab interaction term $h_{\mathrm{S,L}}(\cdot, \lambda(t))$ at any time t is very simple in the two wells—flat.

The result is that the exclusive work over any time duration is easily calculable from the experimental data. During times when the flux remains in a well, the exclusive work rate must be zero, since $h_{\mathrm{S,L}}$ does not change with s. During a transition, the shape of $h_{\mathrm{S,L}}$ does not change significantly due to the first approximation. Then, the exclusive work during a transition is the difference in heights of the two wells as measured by $h_{\mathrm{S,L}}$:

$$\begin{aligned} \Delta W_0^{\mathrm{trans}} &= \int d\phi \left(- \frac{\partial h_{\mathrm{S,L}}}{\partial \phi} \right) \\ &\approx h_{\mathrm{S,L}}(w_0, \lambda(t)) - h_{\mathrm{S,L}}(w_1, \lambda(t)) \ , \end{aligned} \tag{2.17}$$

where $\lambda(t)$ is the protocol parameter setting at any time during the transition and w_0 and w_1 are arbitrary flux values in the starting and ending wells, respectively.

Thus, the total inclusive work W over the protocol for a trajectory is simply the sum of the jump contributions above for each transition.

2.H.3.2. *Second-Order Approximation.* In point of fact, the potential wells do change shape. Fortunately, our method for calculating the inclusive work over the protocol remains valid under weaker constraints on the protocol.

We first require the protocol to maintain two metastable regions, the *informational states*, at all times; each possessing a unique local potential minimum continuously in time. We denote the flux value at the potential minima of informational state i at time t as ϕ_i^t. The protocol must also evolve slowly enough so that the potential landscape changes slowly compared to the system's relaxation rate in each metastable region. Both of these criteria are met by our erasure protocol.

Consider a short duration Δt during which the potential $V(\cdot, t)$ changes little but long compared to the relaxation rates of the informational states. Consider two cases: either the system crosses the barrier between the two informational states during this time or it remains in one informational state.

First, suppose that the system transitions from one informational state i to the other j. Denote the system flux at the beginning of the transition as ϕ^t and at the end as $\phi^{t+\Delta t}$. By Eq. (2.13), the exclusive work contribution $\Delta W_0^{\text{trans}}$ is the difference of the inclusive work contribution and the change in system-lab interaction term $\Delta h_{\text{S,L}}$. Our protocol ensures that the total number of transitions is so small and the time durations Δt so narrow that we can ignore the total contributions of inclusive works ΔW^{trans} due to these transition durations. The change $\Delta h_{\text{S,L}}$ can itself be broken down into two terms, one for the difference in $h_{\text{S,L}}$ between the informational-state minima and the other for the change in $h_{\text{S,L}}$ local to the respective minima. In other words:

$$\begin{aligned}
\Delta h_{\text{S,L}} &= h_{\text{S,L}}(\phi^{t+\Delta t}) - h_{\text{S,L}}(\phi^t) \\
&= \left[\left(h_{\text{S,L}}(\phi^{t+\Delta t}) - h_{\text{S,L}}(\phi_j^{t+\Delta t})\right) + h_{\text{S,L}}(\phi_j^{t+\Delta t})\right] - \left[\left(h_{\text{S,L}}(\phi^t) - h_{\text{S,L}}(\phi_i^t)\right) + h_{\text{S,L}}(\phi_i^t)\right] \\
&= \left[h_{\text{S,L}}(\phi_j^{t+\Delta t}) - h_{\text{S,L}}(\phi_i^t)\right] + \left[\left(h_{\text{S,L}}(\phi^{t+\Delta t}) - h_{\text{S,L}}(\phi_j^{t+\Delta t})\right) - \left(h_{\text{S,L}}(\phi^t) - h_{\text{S,L}}(\phi_i^t)\right)\right] \qquad (2.18) \\
&= \Delta m^t + \Delta l^t \ , \qquad (2.19)
\end{aligned}$$

where Δm^t, Eq. (2.18)'s first term, is the change in $h_{\text{S,L}}$ at the informational-state minima and Δl^t, Eq. (2.18)'s second term, is the change in $h_{\text{S,L}}$ of the system with respect to the informational-state minima. We can therefore approximate the exclusive work contribution during a transition as follows:

$$\begin{aligned}
\Delta W_0^{\text{trans}} &= -\Delta h_{\text{S,L}} \\
&= -\Delta m^t - \Delta l^t \ .
\end{aligned}$$

Suppose, now, that the system remains in one informational state i during a time interval Δt. Since the relaxation rate is fast compared to the duration Δt, we assume that the system visits all microstates in the informational state roughly in proportion to the local equilibrium distribution. Then, the inclusive work contribution ΔW^{stay} is approximately independent of the specific system trajectory during this time and, instead, is determined by the time duration and the informational state i. If during this time we had additionally simultaneously shifted the entire potential up or down by a given amount, we would have added an inclusive work contribution equal to the potential shift but the system trajectory would have remained unchanged since adding a term constant over position to the potential does not affect dynamics. Thus, the actual inclusive work contribution is equal to an amount due to the change in the system-lab interaction term at the informational-state minimum plus an amount due solely to the change in potential shape at the informational state with respect to its minimum. That is:

$$\Delta W^{\text{stay}} = \Delta W_s + \Delta m^t \ , \tag{2.20}$$

where ΔW_s is the inclusive work contribution due to the change in potential shape at the informational state and Δm^t is defined as above. Equation (2.19) applies equally well here in describing the change in system-lab interaction term. Thus, first using Eq. (2.13) once more, the exclusive work contribution ΔW_0^{stay} for a time interval where the particle remains in one informational state is as follows:

$$\Delta W_0^{\text{stay}} = \Delta W^{\text{stay}} - \Delta h_{\text{S,L}} \tag{2.21}$$

$$= \Delta W_s - \Delta l^t \ . \tag{2.22}$$

The result is that we have exclusive work contributions for both durations when the system transitions between informational states and when it remains in one.

To find the total exclusive work over the protocol for a given trajectory we add up these contributions. The sum of all local $h_{\text{S,L}}$ changes Δl^t over all durations is the net local change in $h_{\text{S,L}}$. Recall, though, that the minima of the informational states begin and end at the same values. And so, the total local change in $h_{\text{S,L}}$ reduces to the absolute change in $h_{\text{S,L}}$. However, since we chose

$h_{\mathrm{S,L}}(\cdot, 0) = h_{\mathrm{S,L}}(\cdot, \lambda(t)) = 0$, this must vanish:

$$\sum_t \Delta l^t = 0 \ .$$

We can now specify our final approximation: For any time interval Δt, the inclusive work contribution ΔW_s due to the change in potential shape is independent of the informational state. This is reasonable for our erasure protocol since the asymmetric contribution to the change in potential—the tilt—is slight. While it clearly breaks the symmetry of the double-well potential by changing the well heights, it has less effect on the well shapes and even less in making those shapes distinct.

Then, we can assume that the sum of ΔW_s for any trajectory is the same as that for a particle that stays in either informational state the entire time. But since the protocol is cyclic and very slow compared to the informational states' relaxation rates, a particle that stays in one informational state the entire time must receive approximately zero inclusive work W. Given that the sum over all ΔW_s must be equal to W for such a trajectory, the former must also be negligible.

Altogether, the total exclusive work is approximately given by the sum over all transitions between informational states of the difference in potential at the informational-state minima:

$$W_0(0, \tau) = - \sum_{\text{trans}} \Delta m^t \ . \tag{2.23}$$

To reiterate, since $\Delta h_{\mathrm{S,L}} = 0$, this is also the total inclusive work $W(0, \tau)$ for a trajectory over the entire protocol.

2.I. Substage Work Distributions Commentary

Here, we briefly interpret several features of the substage work distributions observed in Fig. 2.1(Outer left plots).

The distributions for barrier dropping and tilting are narrow, symmetric peaks; see Fig. 2.1 (Outer left plots). Barrier raising also has a rather narrow peak, composed primarily of trajectories always in the R state, but also exhibits a bulge toward positive work; see Fig. 2.1 (Top right). Note that the L state is created mid-way through barrier raising, allowing for trajectories that spend some time in either informational state, but disallowing trajectories that spend all time in the L state. The former induce the positive work bulge toward less negative works, which while notable will not be further explored here.

The substage work distributions for untilting presents the most striking picture; see Fig. 2.1 (Bottom right). Always-R trajectories induce a large positive work peak (red), always-L trajectories induce a large negative work peak (orange), and all other trajectories induce a ramp between them (blue).

These features can be directly interpreted by following the locations of the potential minima over time and noting how the shifting potential adds or removes energy from a particle. During barrier dropping, to take one example, the protocol raises both minima by over 7 $k_{\mathrm{B}}T$, resulting in a narrow, peaked work distribution with a mean near 7 $k_{\mathrm{B}}T$.

Most interesting is the untilt substage. Since most particles start and then stay in the R state for this substage, a large positive work is probable, due to the rising R-state well. However, it is also possible for the system to start in and then get stuck in the L-state well, resulting in a large negative work. The final possibility is transitioning between states during untilting, resulting in an intermediate range of less-likely work values. For trajectories that do transition between states during untilting, it is more likely to spend more time in the R state, since it is energetically favored, resulting in the rising probability with increasing work in their work distribution—giving rise to the log-linear ramp in the work distribution.

Note that there are small peaks on each end of this third class' distributions that require a more nuanced explanation. When a particle crosses a barrier—due to random thermal excitation—the surplus energy may quickly send the particle back to the previous well before it can be dissipated. Such particles then spend almost all of the substage in this first well, generating a work value accordingly. Statistics of the ramp proper are due to particles that have time to locally equilibrate before crossing any barriers.

Follow-on work develops the theory underlying this detailed mechanistic analysis and analyzes similar behavior in all metastable-quasistatic processes.

2.J. Flux Qubit Device, Calibration, and Measurement

The benefits of the flux qubit device are several-fold. First, their physics provide a genuine two-degree of freedom dynamics, while other comparable experiments on Maxwellian demons and bit erasure are very high dimensional, only indirectly providing an effectively few-degree of freedom dynamics [23, 25, 51]. Second, they operate at very high frequency and so one readily captures the substantial amounts of data required to accurately estimate rare-fluctuation statistics. Third, they leverage recent advances in superconducting-device manufacturing technology led by efforts in quantum computing. Fourth, being constructed via modern integrated circuit technology they form the basis of a technology that even-today is ready to scale to large, multicomponent circuit devices for more sophisticated thermodynamic computing. And, finally, in the near future flux qubits will facilitate experiments that probe the thermodynamics of the transition from classical to quantum information processing.

At the microscopic level, a fraction of the electrons in a superconducting metal form bosonic Cooper pairs—a quantum-coherent condensate. For designing superconducting electronic circuits, though, one can forgo the microscopic description and work with higher-level phenomena, such as flux quantization and the Josephson relations for weak links. Importantly, the circuit-level degrees of freedom are not coarse-grained quantities, but display a full range of quantum behavior, including quantized excitations, coherent superpositions, and entangled states in such circuits. For our purposes here, however, we run the device so that it exhibits only classical stochastic dynamics, reserving quantum information thermodynamic explorations for the future.

This section lays out the basic physics of the flux qubit device and details of the experimental implementation. A fuller discussion of the platform and its calibration is found in Ref. [52].

2.J.1. Flux qubit physics. Our experimental information processor is a special type of superconducting quantum interference device (SQUID) with two degrees of freedom—a gradiometric flux qubit or the variable-I_c rf SQUID introduced by Ref. [53]. Notably, the energies associated with the motion perpendicular to and along the escape direction differ substantially by about a factor of 12. Practically, this asymmetry reduces the two-dimensional potential to one dimension. The net result is a device with an effective double-well potential with barriers as low as $\Delta U \sim k_{\mathrm{B}}T$ that operates at frequencies in the GHz range. The potential shape is controlled by fluxes that are readily controlled by currents. SQUID device parameters, used to determine the potential shape and energy scales, were all independently determined.

The variable-I_c rf SQUID replaces the single Josephson junction in a standard rf SQUID with a symmetric dc SQUID with small inductance $\beta_{\mathrm{dc}} = 2\pi\ell I_{c0}/\Phi_0 \ll 1$, where 2ℓ is the loop inductance, $I_{c0} = i_{c1} + I_{c2}$ is the sum of critical currents of the two junctions, and Φ_0 is the flux quantum $h/2e$. This architecture gives a device whose parameters can be accurately measured and that can be selected to exhibit a range of phenomena including thermal activation, macroscopic quantum tunneling, incoherent relaxation, photon-induced transitions, and macroscopic quantum coherence. It also allows us to perform, as we demonstrate, nanoscale thermodynamic computing.

Its macroscopic dynamical variables are the magnetic flux Φ through the rf SQUID loop and Φ_{dc} through the dc SQUID loop. Based on the resistively-capacitively-shunted junction model of Josephson junctions, in the classical limit the variable-I_c rf SQUID's deterministic equations of motion are [53]:

$$\begin{aligned} 2C\ddot{\Phi} + \frac{\dot{\Phi}}{R/2} &= -\frac{\partial U(\Phi, \Phi_{\mathrm{dc}})}{\partial \Phi} \quad \text{and} \\ \frac{C}{2}\ddot{\Phi}_{\mathrm{dc}} + \frac{\dot{\Phi}}{2R} &= -\frac{\partial U(\Phi, \Phi_{\mathrm{dc}})}{\partial \Phi_{\mathrm{dc}}} ~. \end{aligned} \tag{2.24}$$

In units of $\Phi_0/2\pi$, the 2D potential for the variable-I_c rf SQUID is $U(\phi, \phi_{\mathrm{dc}}) = U_0 f(\phi, \phi_{\mathrm{dc}})$ with:

$$f(\phi, \phi_{\mathrm{dc}}) = \tfrac{1}{2}(\phi - \phi_x)^2 + \tfrac{\gamma}{2}(\phi_{\mathrm{dc}} - \phi_{\mathrm{xdc}})^2 - \beta_0 \cos \tfrac{\phi_{\mathrm{dc}}}{2} \cos \phi + \delta\beta \sin \tfrac{\phi_{\mathrm{dc}}}{2} \sin \phi ~, \tag{2.25}$$

where $U_0 = \Phi_0^2/(4\pi^2 L)$. Here, $\gamma = L/(2\ell)$ is the ratio of rf and dc SQUID inductances; ϕ_x (ϕ_{xdc}) is the external flux applied to the rf (dc) SQUID loop; ϕ (ϕ_{dc}) is the flux enclosed in the rf (dc) SQUID loop; $\beta_0 = 2\pi L I_{c0}/\Phi_0$; and $\delta\beta = 2\pi L(I_{c2} - I_{c1})/\Phi_0$.

For large-amplitude tuning of the external controls, the system response to ϕ_x (ϕ_{xdc}) is 2π (4π) periodic. We make use of the global features to accurately determine the coefficients of the potential.

In the experiment, cross-coupling between the barrier and tilt controls was canceled by an affine transformation $(\phi_{\mathrm{x}}, \phi_{\mathrm{xdc}}) \rightarrow (\phi_{\mathrm{x}} + \alpha\phi_{\mathrm{xdc}}, \phi_{\mathrm{xdc}})$, with the coefficient α chosen such that the equilibrium population of the left and right wells was unaffected to first order by the barrier control ϕ_{xdc}.

Operating the magnetometer generates wide-band local electromagnetic interference that can affect the dynamics of the flux qubit. A careful study of the back-action indicates that low-amplitude operation of the magnetometer can induce transitions in a manner that corresponds to a shift in the effective tilt and flux controls. Importantly, the effective temperature under magnetometer operations was not elevated from 500 mK.

The dynamical variable ϕ describes the in-phase motion of the two junctions that results in a current circulating in the rf SQUID loop. The dynamical variable ϕ_{dc} describes the out-of-phase motion, resulting in a current circulating in the dc SQUID loop. The shape of the effective potential is completely determined by the dimensionless function $f(\phi, \phi_{\mathrm{dc}})$ and the energy scale of the potential is determined by U_0. With suitable device parameters and applied fluxes (ϕ_{x} and ϕ_{xdc}) one obtains a smooth family of double-well potentials. The barrier height ΔU separating the two wells is readily adjusted by varying ϕ_{xdc}. The effective potential is plotted in Fig. 2.3(B) with parameters: $\beta_0 = 6.2$, $\gamma = 12$ and $\delta_b = 0.2$.

2.J.2. Experimental implementation. The junctions were $1 \times 1\mu\mathrm{m}^2 \mathrm{Nb}/\mathrm{Al}_2\mathrm{O}_3/\mathrm{Nb}$ tunnel junctions of very low subgap leakage, typically having a quality factor of $V_m \approx 70$ mV at 4.2 K.

We followed a standard procedure (see, e.g., Ref. [**53**]) for calibrating the flux qubit parameters. An outline of the steps is given below. A complete description of the measurements is presented in Ref. [**52**].

First, by executing wide-range sweeps of the coil currents I_{tilt} and I_{barrier}, parameter values corresponding to single-valued and bistable potential landscapes are recorded. A linear transformation from I_{tilt} and I_{barrier} to $(\phi_{\mathrm{x}}, \phi_{\mathrm{xdc}})$ is established by matching the experimental periodicity with the theoretical one $(2\pi, 4\pi)$. Linear cross-talk from I_{barrier} to I_{tilt} is calibrated by orthogonalizing the global response. Cross-talk from I_{tilt} to I_{barrier} can be assumed to be small due to the symmetry of the on-chip flux lines and is taken to be zero.

The parameter values $\beta_0 = 6.2$, $\gamma = 12$, and $\delta_b = 0.2$ are determined by equating the observed extent of hysteresis at $\phi_{\mathrm{xdc}} = 0$ and the differential flux response $d\langle\phi\rangle/d\phi_{\mathrm{x}}$ at $\phi_{\mathrm{xdc}} = 2\pi$ to theoretical predictions. The prefactor $U_0 = 56.3$ K is determined by equating the observed escape energy for inter-well transitions at high temperatures with $k_{\mathrm{B}}T$. The plasma frequency $\omega_p = 1/\sqrt{LC} = 2\pi \times 13.7$ GHz is determined from the observed low-temperature cross-over temperature $T_{cr} = 103$ mK to macroscopic quantum tunneling (MQT) dominated dynamics. We obtain an upper bound $Q = \omega_p RC < 130$ from the coupling to the passive shunt resistor of the magnetometer. Parameter calibration measurements are performed in such a way that the effect of magnetometer back-action is nulled through pulsing of the readout or otherwise minimized. The effective temperature under continuous magnetometer operation was determined by repeating the measurement for escape energy for interwell transitions and comparing the result to that obtained under pulsed magnetometer operation.

CHAPTER 3

The Trajectory Class Fluctuation Theorem

3.1. Introduction

The century-old study of thermodynamic fluctuations was rejuvenated with the discovery of fluctuation theorems in the very late twentieth century [2, 16, 54, 55, 56]. Jarzynski's Equality [2] and Crooks' Detailed Fluctuation Theorem [15], 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 [56]. In addition, equilibrium and nonequilibrium free energy changes are readily obtained from those same stronger equalities [2, 57]. 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 [19, 20, 58] and quantum harmonic oscillators [59]. However, obtaining free energies is generally quite challenging statistically due to the existence of rare events that dominate the exponential average work [32, 44].

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* [60]. While a restricted form of the TCFT had been noted previously [44], 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 [60]. 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) [15], 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. [**60**]'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 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 [**8**].

Theoretically, the TCFT unites a wide variety of prior fluctuation theorems. These include theorems that range systematically from Crooks' DFT [**16**] to Jarzynski's *integral fluctuation theorem* (IFT) [**2**]. That noted, the TCFT itself can be derived from broader theorems still [**17**,**49**]; see App. 3.A. The TCFT's strength then is in its balance of specificity and generality.

Developing the TCFT proceeds as follows. Section 3.2 presents fluctuation theorem building blocks, culminating in Crooks' (DFT) and the basic IFTs. Section 3.3 uses the DFT to introduce and prove the TCFT. Section 3.4 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. 3.5 shows how to use the TCFT to solve for free-energy differences beyond just the equilibrium free-energy difference. Section 3.6 demonstrates how the TCFT overcomes the tyranny of rare events when estimating free energies from data. Section 3.7 highlights related results and surveys how the TCFT encapsulates them. Finally, we briefly discuss several subtle aspects of applying the TCFT in Sec. 3.8, considering the application to a recent nanoscale flux qubit experiment [**60**]. Section 3.9 concludes.

3.2. Background

3.2.1. 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 $\overrightarrow{\mathcal{Z}}$ 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 [61].) 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.

3.2.2. 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 semiprocess* 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 semiprocess* 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:

$$\begin{aligned} \Delta S_{\mathrm{env}}(\overrightarrow{z}) &= -\beta Q(\overrightarrow{z}) \\ &= \ln \frac{p(\overrightarrow{z}|\overrightarrow{z}(0))}{r'(\overrightarrow{z}^{\dagger}|(\overrightarrow{z}(\tau))^{\dagger})} \ . \end{aligned} \tag{3.1}$$

Recall that microscopic reversibility can be derived, for example, under Markov [**15**], Hamiltonian [**32**, **62**], or Langevin [**63**] assumptions.

The *forward process* is a trajectory probability density P specified by the initial distribution ρ and the forward semiprocess. 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:

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

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

$$R'(\overrightarrow{z}) \equiv (\sigma^{\dagger})(\overrightarrow{z}(0))r'(\overrightarrow{z}|\overrightarrow{z}(0)) \ . \tag{3.2}$$

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:

$$R(\overrightarrow{z}) \equiv R'(\overrightarrow{z}^{\dagger}) \ . \tag{3.3}$$

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

3.2.3. 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 [63] 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:

$$\begin{aligned} \Delta s_{\text{sys}}(\overrightarrow{z}) &\equiv s_{\text{sys}}(\overrightarrow{z}(\tau);\sigma) - s_{\text{sys}}(\overrightarrow{z}(0);\rho) \\ &= \ln \frac{\rho(\overrightarrow{z}(0))}{\sigma(\overrightarrow{z}(\tau))} \ . \end{aligned} \tag{3.4}$$

This is similar, but more general than the *change in system entropy* [63] 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:

$$\Sigma(\overrightarrow{z}) = \Delta s_{\text{sys}}(\overrightarrow{z}) + \Delta S_{\text{env}}(\overrightarrow{z}) \ . \tag{3.5}$$

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. (3.5), Eqs. (3.1) and (3.4) yield an expression for the entropy difference:

$$\Sigma(\overrightarrow{z}) = \ln \frac{\rho(\overrightarrow{z}(0))p(\overrightarrow{z}|\overrightarrow{z}(0))}{\sigma(\overrightarrow{z}(\tau))r'(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger)} ~. \tag{3.6}$$

The numerator is $P(\overrightarrow{z})$. And, by Eqs. (3.2) and (3.3):

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

These two observations translate Eq. (3.6) into a DFT:

$$\Sigma(\overrightarrow{z}) = \ln \frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} ~. \tag{3.7}$$

This is a fluctuation theorem obtained previously in a Langevin setting [**63**] and a generalization of the Crooks fluctuation theorem [**15**] to the case of arbitrary privileged distributions.

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

$$R(\overrightarrow{z}) = 0 \text{ when } P(\overrightarrow{z}) = 0 ~. \tag{3.8}$$

The motivation is that the forward process needs to "cover" all the trajectories that are significant to the reverse process. The failure of Condition (3.8) 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. 3.3, such complications are avoided so long as Condition (3.8) 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. (3.1) guarantees $r(\overrightarrow{z}^\dagger|(\overrightarrow{z}(\tau))^\dagger) = 0$ wherever $p(\overrightarrow{z}|\overrightarrow{z}(0)) = 0$. Then Condition (3.8) 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.

3.2.4. 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 [8]. 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:

$$\begin{aligned}\Sigma(\overrightarrow{z}) &= -\ln \frac{\sigma(\overrightarrow{z}(\tau))}{\rho(\overrightarrow{z}(0))} - \beta Q(\overrightarrow{z}) \\ &= \beta(\Delta E(\overrightarrow{z}) - \Delta f(\overrightarrow{z}) - Q(\overrightarrow{z})) \\ &= \beta(W(\overrightarrow{z}) - \Delta f(\overrightarrow{z})) \ , \qquad (3.9)\end{aligned}$$

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

3.2.5. Ensemble Fluctuation Theorems and the Second Law. From Eq. (3.7)'s DFT, it is easy to derive two general fluctuation theorems. First, there is the nominal ensemble fluctuation theorem:

$$
\begin{aligned}
\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &= -\int d\overrightarrow{z}\, P(\overrightarrow{z}) \ln \frac{R(\overrightarrow{z})}{P(\overrightarrow{z})} \\
&= \mathrm{D}_{\mathrm{KL}} \left[P \;||\; R \right]_{\overrightarrow{\mathcal{Z}}} \;.
\end{aligned} \tag{3.10}
$$

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:

$$
\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} \geq 0 \;. \tag{3.11}
$$

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

Applying Eq. (3.9), the average free energy change bounds the work done in the forward experiment:

$$
\begin{aligned}
\langle W \rangle_{\overrightarrow{\mathcal{Z}}} &\geq \langle \Delta f \rangle_{\overrightarrow{\mathcal{Z}}} \\
&= \int d\overrightarrow{z}\, P(\overrightarrow{z}) \Delta f(\overrightarrow{z}) \\
&= \int d\overrightarrow{z}\, P(\overrightarrow{z}) f(\overrightarrow{z}(\tau); \sigma, E_\tau) \\
&\qquad - \int d\overrightarrow{z}\, P(\overrightarrow{z}) f(\overrightarrow{z}(0); \rho, E_0) \\
&= \int dz \rho_\tau(z) f(z; \sigma, E_\tau) - \int dz \rho(z) f(z; \rho, E_0).
\end{aligned} \tag{3.12}
$$

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. (3.12) tightest? The answer is the final-time distribution ρ_τ:

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

where:

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

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:

$$\langle W \rangle_{\overrightarrow{\mathcal{Z}}} \geq \langle \Delta F^{\text{neq}} \rangle_{\overrightarrow{\mathcal{Z}}} \geq \langle \Delta f \rangle_{\overrightarrow{\mathcal{Z}}} ~. \tag{3.13}$$

Even though the nonequilibrium free-energy change for the forward process provides the tightest bound on work when it is used in Eq. (3.12), 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:

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

If we start the system in equilibrium $\rho = \pi_0$, a similar calculation shows that, from Eq. (3.12):

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

with:

$$\Delta F^{\text{eq}} = F^{\text{eq}}_{\tau} - F^{\text{eq}}_{0} .$$

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 3.5.2 discusses this shortly.

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

$$\begin{aligned} \left\langle e^{-\Sigma} \right\rangle_{\overrightarrow{\mathcal{Z}}} &= \int d\overrightarrow{z} P(\overrightarrow{z}) \frac{R(\overrightarrow{z})}{P(\overrightarrow{z})} \\ &= \int d\overrightarrow{z} R(\overrightarrow{z}) \\ &= 1 . \end{aligned} \tag{3.14}$$

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

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

giving Jarzynski's Equality:

$$\left\langle e^{-\beta W} \right\rangle_{\overrightarrow{\mathcal{Z}}} = e^{-\beta \Delta F^{\text{eq}}} . \tag{3.15}$$

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. (3.15) can lead to sampling issues due to rare but resource-dominant events. We use the TCFT in Secs. 3.5 and 3.6 to confront these two problems.

3.3. 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.

3.3.1. 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. 3.7 for more examples as they apply to known results.

3.3.2. 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 3.3.5 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:

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

3.3.3. 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 $\overrightarrow{z} \in C$:

$$\begin{aligned} \Sigma(\overrightarrow{z}) &= \ln \frac{P(\overrightarrow{z})}{R(\overrightarrow{z})} \\ &= \ln \frac{P(C)P(\overrightarrow{z}|C)}{R(C)R(\overrightarrow{z}|C)} \\ &= \Theta_C + \psi_C(\overrightarrow{z}) \ . \end{aligned} \tag{3.16}$$

Equation (3.16) 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. (3.16) 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. (3.16) 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:

$$\begin{aligned}\Psi_C &\equiv \langle \psi_C \rangle_C \\ &= \mathrm{D}_{\mathrm{KL}}\left[P \mid\mid R \right]_C \ , \end{aligned} \tag{3.17}$$

where:

$$\mathrm{D}_{\mathrm{KL}}\left[P \mid\mid R \right]_C \equiv \int d\overrightarrow{z} P(\overrightarrow{z}|C) \ln \frac{P(\overrightarrow{z}|C)}{R(\overrightarrow{z}|C)} \ , \tag{3.18}$$

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. (3.16) gives the following.

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

$$\begin{aligned}\langle \Sigma \rangle_C &= \Theta_C + \Psi_C \\ &= \ln \frac{P(C)}{R(C)} + \mathrm{D}_{\mathrm{KL}}\left[P \mid\mid R \right]_C \ . \end{aligned} \tag{3.19}$$

When Sec. 3.4 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. (3.16) 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):*

$$\begin{aligned}\langle e^{-\Sigma}\rangle_C &= e^{-\Theta_C} \\ &= \frac{R(C)}{P(C)} .\end{aligned} \tag{3.20}$$

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.

3.3.4. 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:

$$\Psi_{\overrightarrow{\mathcal{Z}}} = \mathrm{D}_{\mathrm{KL}}\left[P \mid\mid R\right]_{\overrightarrow{\mathcal{Z}}} .$$

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. (3.10) and (3.14).

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. 3.5 and 3.6. Also see Sec. 3.7 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. (3.7).

3.3.5. 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. (3.19) and (3.20) 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:

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

As ϵ decreases, the work distribution for C' necessarily narrows, so that $e^{-\beta(W-\Delta F^{\mathrm{eq}})}$ approaches $e^{-\beta(\widetilde{W}-\Delta F^{\mathrm{eq}})}$. 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 [**16**]:

$$e^{-\beta(W - \Delta F^{\mathrm{eq}})} = \frac{R(W)}{P(W)} \ , \tag{3.21}$$

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 $\overrightarrow{\lambda}^{\dagger}$ 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.

3.4. 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.

3.4.1. Trajectory Class Second Law. Discarding the class-average class irreversibility Ψ_C in Eq. (3.19)—a nonnegative quantity—gives the *trajectory class second law* (TCSL):

$$\begin{aligned} \langle \Sigma \rangle_C &\geq \Theta_C \\ &= \ln \frac{P(C)}{R(C)} \ . \end{aligned} \tag{3.22}$$

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. (3.11)'s Second Law that bounds the ensemble-averaged entropy difference to be nonnegative. However, Eq. (3.22) is more precise since it uses more information—the class probabilities in the forward and reverse processes. Section 3.4.3 elaborates on this advantage over the ensemble Second Law.

Equation (3.19) 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 3.B shows that having such a small class irreversibility is equivalent to the class C having a narrow entropy-difference distribution.

3.4.2. Trajectory Partition Fluctuation Theorem. Now partition all trajectories $\overrightarrow{\mathcal{Z}}$ into trajectory classes forming a collection Q. Then averaging Eq. (3.19) over all classes in Q gives an equality obtained in Ref. [48] that is generalized to arbitrary privileged distributions:

$$\begin{aligned} \langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &= \sum_{C \in Q} P(C) \langle \Sigma \rangle_C \\ &= \sum_{C \in Q} P(C) (\Theta_C + \Psi_C) \\ &= \langle \Theta \rangle_Q + \boldsymbol{\Psi}_Q \ , \end{aligned} \tag{3.23}$$

where the *partition averaged class reverse surprisal* is:

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

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

$$\mathrm{D}_{\mathrm{KL}} \left[P \mid\mid R \right]_Q \equiv \sum_{C \in Q} P(C) \ln \frac{P(C)}{R(C)} ~. \tag{3.24}$$

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.

3.4.3. Trajectory Partition Second Law. Since divergences are nonnegative, Eq. (3.23) leads directly to the *trajectory partition second law* (TPSL) by discarding the partition-class averaged class irreversibility $\mathbf{\Psi}_Q$:

$$\begin{aligned}\langle \Sigma \rangle_{\overrightarrow{\mathcal{Z}}} &\geq \langle \Theta \rangle_Q \\ &= \mathrm{D}_{\mathrm{KL}} \left[P \mid\mid R \right]_Q ~,\end{aligned} \tag{3.25}$$

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 [**64**, **65**] and system state space [**8**, **64**, **65**], as well as many other possibilities [**66**, **67**, **68**].

The information that is left out in going from Eq. (3.23) to Eq. (3.25) 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. 3.4.1, 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. (3.25) is a tight-bound.

Contrast this with Eq. (3.11)'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. (3.25)'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. (3.11) 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 [**60**] 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.

3.5. Free Energies via Constant-Difference Classes

Paralleling the bounds on work production derived from the Second Law, Eq. (3.9) converts the bounds of Sec. 3.4 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.

3.5.1. 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. (3.15) 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. (3.20) gives:

$$\Delta f_C = -\beta^{-1} \ln \langle e^{-\beta W} \rangle_C + \beta^{-1} \ln \frac{R(C)}{P(C)} . \tag{3.26}$$

This equality relates free-energy differences to statistics on works and class probabilities. Jarzynski's Equality Eq. (3.15) is the special case where $C = \overrightarrow{\mathcal{Z}}$, $\rho = \pi_0$, and $\sigma = \pi_\tau$.

3.5.2. Metastable Process Work Bounds.

Having established Eq. (3.26), 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 [**20**, **58**]. However, our procedure comes with the added possibility of treating distributions over multiple possible coarsened states.

3.5.2.1. *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.

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

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

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*:

$$F_t^{\text{mem}}(m) \equiv F_t^{\text{eq}} - \beta^{-1} \ln \pi_t(m) \ , \tag{3.27}$$

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 [8] 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:

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

where $m(z)$ and $m'(z)$ are the memory states containing z in E_0 and E_τ, respectively. Using Eq. (3.12), we then have:

$$\begin{aligned}\langle W \rangle_{\overrightarrow{\mathcal{Z}}} & \geq \langle \Delta f \rangle_{\overrightarrow{\mathcal{Z}}} \\ & = \langle \Delta F^{\mathrm{mem}} \rangle_{\overrightarrow{\mathcal{Z}}} - \beta^{-1} \Delta H_M ~ .\end{aligned} \tag{3.28}$$

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. (3.28) falls short of the free-energy change bound on the average work, Eq. (3.13), because the actual final-time free energy, that of ρ_τ, is generally higher than the free energy of the locally-equilibrated σ.

3.5.2.3. *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. (3.28), 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 $\overrightarrow{z} \in C_{m,m'}$:

$$\begin{aligned} \Delta f_{C_{m,m'}} &= f(\overrightarrow{z}(\tau); \sigma, E_\tau) - f(\overrightarrow{z}(0); \rho, E_0) \\ &= F^{\mathrm{mem}}_\tau(m') - F^{\mathrm{mem}}_0(m) + \beta^{-1} \ln \frac{\sigma(m')}{\rho(m)} \ . \end{aligned}$$

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. (3.26), using the work production probabilities and memory state probabilities:

$$\begin{aligned} &F^{\mathrm{mem}}_\tau(m') - F^{\mathrm{mem}}_0(m) \\ &\quad = -\beta^{-1} \ln \left\langle e^{-\beta W} \right\rangle_{C_{m,m'}} + \beta^{-1} \ln \frac{R(C_{m,m'})\rho(m)}{P(C_{m,m'})\sigma(m')} \ . \end{aligned}$$

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. (3.28):

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

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.

3.6. Statistical Freedom from the Tyranny of the Rare

When estimating statistical quantities from data, rare events can dominate sample averages [**32**, **44**]. 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 $\left\langle e^{-\beta W} \right\rangle_{\overrightarrow{\mathcal{Z}}}$ for a thermodynamic transformation, one can estimate the equilibrium free energy difference ΔF^{eq} via Eq. (3.15)'s Jarzynski's Equality. However, this can require thorough sampling of very rare events [**32**]. The ECFT of Eq. (3.20) 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. (3.26)) 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:

$$\Delta \tilde{f}_C \equiv -\beta^{-1} \ln \left(\frac{\sum_{i=1}^N \delta_{W_i \in C} e^{-\beta W_i}}{\sum_{i=1}^N \delta_{W_i \in C}} \right) + \beta^{-1} \ln \frac{\tilde{R}(C)}{\tilde{P}(C)}, \tag{3.29}$$

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. 3.C further explains), we can find better estimators of the free energy change by choosing more probable trajectory classes.

3.6.1. 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. (3.15) the equilibrium free energy change vanishes.

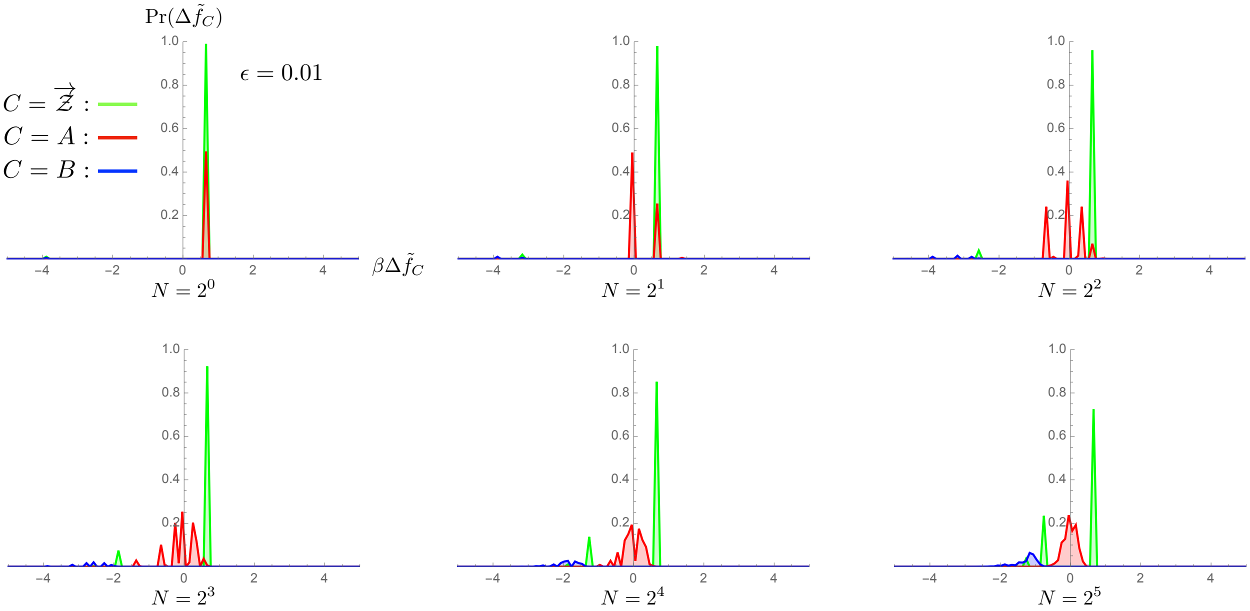

FIGURE 3.1. Tyranny of the rare: Distribution of free energy estimates $\Delta\tilde{f}_C$ arising from N forward and N reverse experiments with $\epsilon = 0.01$ depends sensitively on the trajectory class C. Above, plots of free energy estimates for the trajectory class that includes all paths (green: $C =$ 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 $C =$ All.

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. (3.15) 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. 3.1. 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.

3.6.2. 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. 3.5.1. 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. (3.26), moving away from Jarzynski's Equality in Eq. (3.15). 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. (3.26) 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. 3.C for an explanation of the base calculations.) Figure 3.1 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 3.2 further quantifies this advantage by plotting the average difference with the correct free energy $\langle \Delta \tilde{f}_C - \Delta F^{\mathrm{eq}} \rangle$ and mean square deviation $\langle (\Delta \tilde{f}_C - \Delta F^{\mathrm{eq}})^2 \rangle$. We see a marked advantage to our restricted trajectory class $C = A$ over the full set of trajectories in both cases.

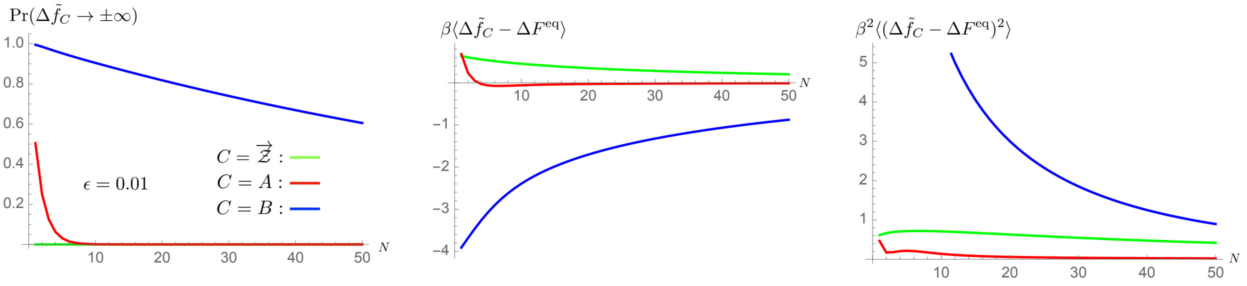

Figure 3.2. Degree to which energy estimates $\Delta \tilde{f}_C$ diverge from the actual change in free energy ΔF^{eq} depends on the trajectory class: For $\epsilon = 0.01$, the probability of an infinitely divergent free energy estimate $\Pr(\Delta \tilde{f}_C \to \pm\infty)$ is significant for the 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 $\langle \Delta \tilde{f}_C - \Delta F^{\mathrm{eq}} \rangle$ and the mean squared deviation $\langle (\Delta \tilde{f}_C - \Delta F^{\mathrm{eq}})^2 \rangle$ 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.

By contrast, restricting to the rare event $C = B$, Fig. 3.1 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. 3.2 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 [44] 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.

3.7. 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. 3.3.5 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.

3.7.1. Measurement and Feedback for Free Energy Estimation. Reference [44] focuses on the problem of estimating free energy differences in the face of rare yet resource-dominant events. This was the subject of Sec. 3.6. 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. (3.20)'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.

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. (3.20) and their Eq. (9) is that Eq. (3.20) 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.

3.7.2. Phase Space Perspective on Dissipated Work. Reference [45] 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 $\langle W \rangle_{\overrightarrow{\mathcal{Z}}} - \Delta F^{\mathrm{eq}}$ 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 $\{\chi_1, \dots \chi_K\}$ of $\mathcal{Z}$ and a time t. For each j in $\{1, \dots, K\}$, 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 $\widetilde{\chi}_j = \chi_j^\dagger = \{z^\dagger \,|\, z \in \chi_j\}$ 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. (3.20)'s ECFT to C_j yields Ref. [45]'s Eq. (6). Eq. (3.22)'s TCSL applied to C_j yields their Eq. (7). And, Eq. (3.25)'s TPSL applied to the partition of trajectories $Q = \{C_1, \dots, C_K\}$ 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.

3.7.3. Work Dissipation as the Distance from Equilibrium. Reference [46] 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:

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

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:

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

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^t_z and the above equalities, Eq. (3.20)'s ECFT yields Ref. [**46**]'s Eq. (6). Equation (3.22)'s TCSL yields their Eq. (8) and Eq. (3.25)'s TPSL yields their Eqs. (2) and (9).

3.7.4. Work and State Fluctuation Theorem. Consider a process where ρ and σ are the initial and final Boltzmann distributions. Reference [**19**] 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:

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

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. (3.20) to C gives their Eq. (1).

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

3.7.5. Landauer's Bound on Erasure Dissipation. Finally, Ref. [**23**] 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. [46]'s Eq. (6) to produce the following:

$$\left\langle e^{-\beta W} \right\rangle_{\to 0} = \frac{1/2}{P_S} \ , \tag{3.30}$$

and:

$$\left\langle e^{-\beta W} \right\rangle_{\to 1} = \frac{1/2}{P_S} \ , \tag{3.31}$$

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. (3.30) 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. [46]'s Eq. (6) from the TCFT already, but Eqs. (3.30) and (3.31) 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. (3.20) first to C_0 and then to C_1 yields Eqs. (3.30) and (3.31), respectively.

3.8. 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. 3.4.3 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. 3.6), 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 [60] 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. 3.2.1 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. 3.3. However, how such a generalization of our presentation can be used remains to be explored.

3.9. 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.

3.A. Alternative TCFT Derivations

Equation (3.20)'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.

3.A.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. [69]:

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

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:

$$\langle \mathcal{F} e^{-\Sigma(\overrightarrow{z})} \rangle_F = \langle \mathcal{F} \rangle_R \ . \tag{3.32}$$

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. (3.32) becomes:

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

Equation (3.32)'s RHS is simply $R(C)$. Combining yields Eq. (3.20), showing that the exponential class fluctuation theorem is the path or trajectory ensemble average of a characteristic function $[\overrightarrow{z} \in C]$.

3.A.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 $\mathcal{F}(\overrightarrow{z}) = [\overrightarrow{z} \in C]$.

3.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 $\left\langle x^2 \right\rangle_C$ dominates in a. a is, then, the variance of Σ over C.

Then, using Eqs. (3.20) and (3.16), 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.

3.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:

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

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

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

CHAPTER 4

Costs of Time Symmetric Control

4.1. Introduction

In 1961, Landauer identified a fundamental energetic requirement to perform logically-irreversible computations on nonvolatile memory [6]. Focusing on arguably the simplest case—erasing a bit of information—he found that one must supply at least $k_{\mathrm{B}} T \ln 2$ work energy ($\approx 10^{-21} J$ at room temperature), eventually expelling this as heat. (Here, k_{B} is Boltzmann's constant and T is the temperature of the computation's ambient environment.)

Notably, though still underappreciated, Landauer had identified a thermodynamically-reversible transformation. And so, no entropy actually need be produced—energy is not irrevocably dissipated—at least in the quasistatic, thermodynamically-reversible limit required to meet Landauer's bound.

Landauer's original argument appealed to equilibrium statistical mechanics. Since his time, advances in nonequilibrium thermodynamics, though, showed that his bound on the required work follows from a modern version of the Second Law of thermodynamics [8]. (And, when the physical substrate's dynamics are taken into account, this is the *information processing Second Law* (IPSL) [34].) These modern laws clarified many connections between information processing and thermodynamics, such as dissipation bounds due to system-state coarse-grainings [48], nanoscale information-heat engines [70], the relation of dissipation and fluctuating currents [71], and memory design [72].

Additional scalings recently emerged between computation time, space, reliability, thermodynamic efficiency, and robustness of information storage [33, 73, 74]. In contrast to Landauer's bound, these tradeoffs involve thermodynamically-irreversible processes, implying that entropy production and therefore true heat dissipation is generally required depending on either practicality or design goals.

In addition to these tradeoffs, it is now clear that substantial energetic costs are incurred when using logic gates and allied information-processing modules to construct a computer. Especially so, when compared to custom designing hardware to optimally implement a particular computation [38].

Taken altogether these costs constitute a veritable *Landauer's Stack* of the information-energy requirements for thermodynamic computing. Figure 4.1 illustrates Landauer's Stack in the light of historical trends in the thermodynamic costs of performing elementary logic operations in CMOS technology. The units there are joules dissipated per logic operation. We take Landauer's Stack to be the overhead including Landauer's bound ($k_\mathrm{B}T \ln 2$ joules) up to the current (year 2020) energy dissipations *due to information processing.* Thus, the Stack is a hierarchy of energy expenditures that underlie contemporary digital computing—an arena of theoretically-predicted and as-yet unknown thermodynamic phenomena waiting detailed experimental exploration.

To account for spontaneous deviations that arise in small-scale systems, the Second Laws are now most properly expressed by exact equalities on probability distributions of possible energy fluctuations. These are the *fluctuation theorems* [**18**], from which the original Laws (in fact, inequalities) can be readily recovered. Augmenting the Stack, fluctuation theorems apply directly to information processing, elucidating further thermodynamic restrictions and structure [**51**,**52**,**84**,**85**].

The result is a rather more complete accounting for the energetic costs of thermodynamic computation, captured in the refined Landauer's Stack of Fig. 4.1. In this spirit, here we report new bounds on the work required to compute in the very important case of computations driven externally by time-symmetric control protocols [**75**]. In surprising contrast to the fixed energy cost of erasure identified by Landauer, here we demonstrate that the scaling of the minimum required energy *diverges as a function of accuracy* and so can dominate Landauer's Stack. This serves the main goal in the following to validate and demonstrate the tightness of Ref. [**75**]'s thermodynamic bounds and do so in Landauer's original setting of information erasure.

In essence, our argument is as follows. Energy dissipation in thermodynamic transformations is strongly related to entropy production. The fluctuation theorems establish that entropy production depends on both forward and reverse dynamics. Thus, when determining bounds on dissipation in thermodynamic computing, one has to examine both when the control protocol is applied in forward and reverse. By considering time-symmetric protocols we substantially augment Landauer and Bennett's dissipation bound on logical irreversibility [**7**] with dissipation due to logical nonselfinvertibility (aka nonreciprocity). Our results therefore complement recent work on the consequences of logical and thermodynamic reversibility [**86**]. Parallel work on thermodynamic bounds for information processing in finite time, and bit-erasure in particular, include the use

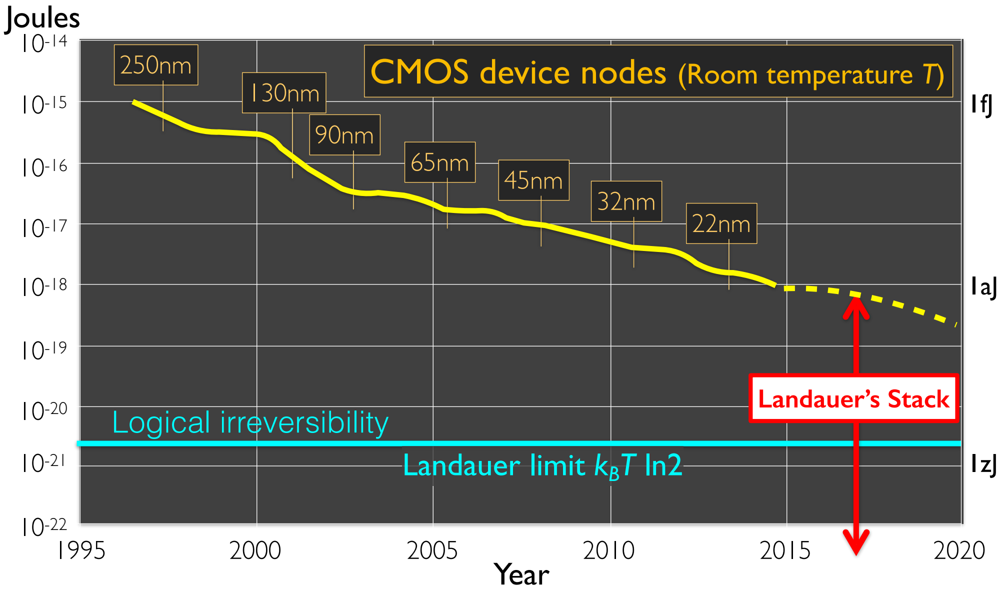

FIGURE 4.1. Historical trends in thermodynamic costs of performing elementary logic operations in CMOS technology quoted in energy dissipated (joules) per logic operation. Contemporary experimentally-accessible thermal resolution is approximately 10^{-24} joules. Landauer's Stack—Thermodynamic hierarchy of predicted "overhead" energy expenditures *due to information processing* that underlie contemporary digital computing, including Landauer's Principle of logical irreversibility [6] (which is now seen as a consequence of the broader information processing Second Law $\langle W \rangle \leq k_{\mathrm{B}} T \ln \Delta h_\mu$ [34]): (a) Nonreciprocity [75]; (b) Computation rate $1/\tau$ [33, 74]; (c) Accuracy: $-\ln \epsilon$ [75]; (d) Storage stability; (e) Circuit modularity [38]; (f) Mismatched expectations [76, 77]; (g) Transitions between nonequilibrium steady-state storage states [36, 39]; and (h) Quantum coherence [78]. (2015 and prior portion of figure courtesy M. L. Roukes, data compiled from [79, and citations therein]. Landauer's Stack cf. Table I in Ref. [10].) CMOS technology change to 3D device nodes around 2015 make linear feature size and its relation to energy costs largely incomparable afterwards [80, 81, 82, 83]. There are, of course, other sources of energy dissipation in CMOS such as leakage currents that arise when electrons tunnel from gate to drain through a thin gate-oxide dielectric. Thermodynamically, this source is a kind of "housekeeping heat", necessary to support the substrate's electronic properties but not directly due to information processing.

of optimized control in the linear response regime [87, 88, 89] and transport theory [90, 91, 92]. However, the cost of nonreciprocity necessarily goes beyond the cost of finite-time computing, because time-symmetrically driven computations incur this additional dissipation regardless of the rate at which they're executed.

Why time-symmetric protocols? Modern digital computers are driven by sinusoidal line voltages and square-wave clock pulses. These control signals function as control parameters, directly altering the energetics and therefore guiding dynamics of the computer components. Being time-symmetric control signals, modern digital computers must then obey Ref. [75]'s error-dissipation trade-off. Moreover, the costs apply to even the most basic of computational tasks—such as bit erasure. Here, we present protocols for time-symmetrically implementing erasure in two different frameworks and demonstrate that both satisfy the new bounds. Moreover, many protocols approach the bounds quite closely, indicating that they may in fact be broadly achievable.

After a brief review of the general theory, we begin with an analysis of erasure implemented with the simple framework of two-state rate equations, demonstrating the validity of the bound for different protocols of increasing reliability. We then expand our framework to fully simulated collections of particles erased in an underdamped Langevin double-well potential, seeing the same faithfulness to the bound for a wide variety of different erasure protocols. We conclude with a call for follow-on efforts to analyze even more efficient computing that can arise from *time-asymmetric* protocols.

4.2. Dissipation In Thermodynamic Computing

Consider a universe consisting of a computing device—the *system under study* (SUS), a *thermal environment* at fixed inverse temperature β, and a *laboratory device* (lab) that includes a *work reservoir*. The set of possible microstates for the SUS is denoted $\boldsymbol{\mathcal{S}}$, with s denoting an individual SUS microstate. The SUS is driven by a *control parameter* x generated by the lab. The SUS is also in contact with the thermal environment.

The overall evolution occurs from time $t = 0$ to $t = \tau$ and is determined by two components. The first is the SUS's Hamiltonian $\mathcal{H}_{SL}(s, x)$ that specifies its interaction with the lab device and determines (part of) the rates of change of the SUS coordinates consistent with Hamiltonian mechanics. We refer to the possible values of the Hamiltonian as the *SUS energies.* The second component is the thermal environment which exerts a stochastic influence on the system dynamics.

We prepare the lab to guarantee that a specific control parameter value $x(t)$ is applied to the SUS at every time t over the time interval $t \in (0, \tau)$. That is, the control parameter evolves deterministically as a function of time. The deterministic trajectory taken by the control parameter $x(t)$ over the computation interval is the *control protocol*, denoted by $\overrightarrow{x}$. The SUS microstate $s(t)$ exhibits a response to the control protocol, over the interval following a stochastic trajectory denoted $\overrightarrow{s}$.

For a given microstate trajectory $\overrightarrow{z}$, the net energy transferred from the lab to the SUS is defined as the *work*, which has the following form [**70**]:

$$W(\overrightarrow{z}) = \int_0^\tau dt\, \dot{x}(t) \frac{\partial \mathcal{H}_{SL}}{\partial x}\Big|_{(s(t),x(t))} ~.$$

This is the energy accumulated in the SUS directly caused by changes in the the control parameter.

Given an initial microstate s_0, the probability of a microstate trajectory $\overrightarrow{z}$ conditioned on starting in s_0 is denoted:

$$P(\overrightarrow{z}|s_0) = \underset{\overrightarrow{x}}{\Pr}(\overrightarrow{Z} = \overrightarrow{z}|\mathcal{S}_0 = s_0) ~.$$

With the SUS initialized in microstate distribution $\boldsymbol{\mu}_0$, the unconditioned *forward process* gives the probability of trajectory $\overrightarrow{z}$:

$$P(\overrightarrow{z}) = P(\overrightarrow{z}|s(0))\boldsymbol{\mu}_0(s(0)) ~.$$

Detailed fluctuation theorems (DFTs) [**16**, **62**] determine thermodynamic properties of the computation by comparing the forward process to the *reverse process*. This requires determining the conditional probability of trajectories under time-reversed control:

$$R(\overrightarrow{z}|s_0) = \underset{\boldsymbol{Я}\overrightarrow{x}}{\Pr}(\overrightarrow{Z} = \overrightarrow{z}|\mathcal{S}_0 = s_0) ~.$$

The reverse control protocol is $\boldsymbol{Я}x(t) = x(\tau - t)^\dagger$, where $x^\dagger$ is x, but with all time-odd components (e.g., magnetic field) flipped in sign. And, the *reverse process* results from the application of this dynamic to the final distribution $\boldsymbol{\mu}_\tau$ of the forward process with microstates conjugated:

$$R(\overrightarrow{z}) = R(\overrightarrow{z}|s(0))\boldsymbol{\mu}_\tau(s(0)^\dagger) ~.$$

The Crooks DFT [**16**] then gives an equality on both the dissipated work (or entropy production) that is produced as well as the required work for a given trajectory induced by the protocol:

$$\omega(\overrightarrow{z}) = \ln \frac{P(\overrightarrow{z})}{R(\boldsymbol{Я}\overrightarrow{z})}.$$

$\boldsymbol{Я}\overrightarrow{z}$ here is itself a SUS microstate trajectory with $\boldsymbol{Я}s(t) = s(\tau - t)^\dagger$.

Due to their practical relevance, we consider protocols that are symmetric under time reversal $\mathbf{Я}\overrightarrow{x} = \overrightarrow{x}$. That is, the reverse-process probability of trajectory $\overrightarrow{z}$ conditioned on starting in microstate s_0 is the same as that of the forward process:

$$R(\overrightarrow{z}|s_0) = P(\overrightarrow{z}|s_0) \ .$$

However, the unconditional reverse process probability of the trajectory $\overrightarrow{z}$ is then:

$$R(\overrightarrow{z}) = P(\overrightarrow{z}|s(0))\boldsymbol{\mu}_\tau(s(0)^\dagger) \ .$$

This leads to a version of Crook's DFT that can be used to set modified bounds on a computation's dissipation:

$$\omega(\overrightarrow{z}) = \ln \frac{P(\overrightarrow{z}|s(0))\boldsymbol{\mu}_0(s(0))}{P(\mathbf{Я}\overrightarrow{z}|s(\tau)^\dagger)\boldsymbol{\mu}_\tau(s(\tau))} \ . \tag{4.1}$$

Suppose, now, that the final and initial SUS Hamiltonian configurations $\mathcal{H}_{SL}(s, x(\tau))$ and $\mathcal{H}_{SL}(s, x(0))$ are both designed to store the same information about the SUS. The SUS microstates are partitioned into locally-stable regions that are separated by large energy barriers in these energy landscapes. On some time scale, a state initialized in one of these regions has a very low probability of escape and instead locally equilibrates to its locally-stable region. These regions can thus be used to store information for periods of time controlled by the energy barrier heights. Collectively, we refer to these regions as *memory states* $\boldsymbol{\mathcal{M}}$.

Then the probability of the system evolving to a memory state $y \in \boldsymbol{\mathcal{M}}$ given that it starts in a memory state $z \in \boldsymbol{\mathcal{M}}$ under either the forward or reverse process is:

$$P'(z \to y) = \frac{\int d\overrightarrow{z} \big[s(0) \in z \wedge s(\tau) \in y\big] P(\overrightarrow{z})}{\int d\overrightarrow{z} \big[s(0) \in z\big] P(\overrightarrow{z})} \ ,$$

where $\big[E\big]$ evaluates to one if expression E is true and zero otherwise.

To simplify the development, suppose that the energy landscape of each memory state looks the same locally. That is, up to translation and possibly reflection and rotation, each memory state spans the same volume in microstate space and has the same energies at each of those states. Further, suppose that the SUS starts and ends in a metastable distribution, differing from global equilibrium only in the weight that each memory state is given in the distribution. Otherwise the distribution looks identical to the global equilibrium at the local scale of any memory state. This ensures that the average change in SUS energy is zero, simplifying the change Δ in nonequilibrium free energy F [**75**]:

$$\Delta F = -\beta^{-1} \Delta H(\mathcal{M}_t) \ ,$$

where $H(\cdot)$ is the Shannon entropy (in nats), and $\mathcal{M}_t$ is the random variable for the memory state at time t. Finally, suppose that the time reversal of a microstate changes neither the memory state it exists in, nor its equilibrium probability, for any time during the protocol. This holds for memory states distinguished primarily by the positions of the system particles and system Hamiltonians that are unchanging under time reversal. See Ref. [**75**] for details behind these assumptions and generalized bounds without them.

Then we have the following inequality:

$$\beta \langle W_{\text{diss}} \rangle \geq \Delta H(\mathcal{M}_t) + \sum_{z \in \boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \sum_{y \in \boldsymbol{\mathcal{M}}} d(z,y), \tag{4.2}$$

where:

$$\begin{aligned} \boldsymbol{\mu}'_0(z) &= \int ds \big[s \in z \big] \boldsymbol{\mu}_0(s) && \text{and} \\ d(z,y) &= P'(z \to y) \ln \frac{P'(z \to y)}{P'(y \to z)} \ . \end{aligned}$$

See Appendix 4.A for a proof sketch.

Recalling that $\beta \left\langle W_{\text{diss}}(\overrightarrow{z}) \right\rangle = \beta(\langle W(\overrightarrow{z}) \rangle - \Delta \mathcal{F})$ and appealing to the inequality in Eq. (4.2), we find a simple bound on the average work over the protocol:

$$\begin{aligned} \beta \langle W \rangle &\geq \sum_{z \in \boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \sum_{y \in \boldsymbol{\mathcal{M}}} d(z,y) \\ &\equiv \beta \langle W \rangle^{\text{t-sym}}_{\text{min}} \ . \end{aligned} \tag{4.3}$$

This provides a bound on the work that depends solely on the logical operation of the computation, but goes beyond Landauer's bound.

Since we are addressing modern computing, we consider processes that approximate deterministic computations on the memory states. For such computations there exists a computation function

$C : \boldsymbol{\mathcal{M}} \to \boldsymbol{\mathcal{M}}$ such that the physically-implemented stochastic map approximates the desired function up to some small error. That is, $P'(z \to C(z)) = 1 - \epsilon_z$ where $0 < \epsilon_z \ll 1$. In fact, we require all relevant errors to be bound by a small error-threshold $\epsilon \ll 1$. That is, for all $C(z) \neq y$, let $P'(z, y) = \epsilon_{z \to y}$ so that $0 \leq \sum_{y \neq C(z)} \epsilon_{z \to y} = \epsilon_z \leq \epsilon \ll 1$.

We can then simplify Eq. (4.3)'s bound in the limit of small ϵ. First, we show that $d(z, y) \geq 0$ for any pair of z, y in the small ϵ limit, where we have:

$$\begin{aligned} d(z, y) &= P'(z \to y) \ln \frac{P'(z \to y)}{P'(y \to z)} \\ &\geq P'(z \to y) \ln P'(z \to y) \ . \end{aligned}$$

If $C(z) = y$, then $P'(z \to y) = 1 - \epsilon_z \geq 1 - \epsilon$, so that:

$$d(z, y) \geq (1 - \epsilon) \ln(1 - \epsilon) \ ,$$

which vanishes as $\epsilon \to 0$. And, if $C(z) \neq y$, then $P'(z \to y) = \epsilon_{z \to y}$, so that:

$$d(z, y) \geq \epsilon_{z \to y} \ln \epsilon_{z \to y}$$

which also vanishes as $\epsilon \to 0$. Setting this asymptotic lower bound on the dissipation of each transition facilitates isolating divergent contributions, such as those we now consider.

An *unreciprocated* memory transition $C(z) = y$ is one that does not map back to itself: $C(y) \neq z$. The contribution to the dissipation bound is:

$$\begin{aligned} d(z, y) &= (1 - \epsilon_z) \ln \frac{1 - \epsilon_z}{\epsilon_{y \to z}} \\ &\geq (1 - \epsilon) \ln \frac{1 - \epsilon}{\epsilon} \ . \end{aligned}$$

As $\epsilon \to 0$, this gives:

$$d(z, y) \geq \ln \epsilon^{-1} \ . \tag{4.4}$$

That is, as computational accuracy increases ($\epsilon \to 0$), $d(z, y)$ diverges. This means the minimum-required work (Eq. (4.3)) must then also diverge.

We then arrive at our simplified bound for the small-ϵ high-accuracy limit from Eq. (4.3)'s inequality on dissipation by only including the contribution from unreciprocated transitions $m' = C(m)$ for which $m \neq C(m')$:

$$\begin{aligned} \beta\langle W\rangle &\geq \ln(\epsilon^{-1}) \sum_{z \in \boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \big[C(C(z)) \neq z\big] \\ &\equiv \beta\langle W\rangle^{\text{approx}}_{\text{min}} \; . \end{aligned} \tag{4.5}$$

In this way, we see how computational accuracy drives a thermodynamic cost that diverges, overwhelming the Landauer-erasure cost. A similar logarithmic relationship between dissipated work and error was demonstrated in the context of the adaptation accuracy of Escherichia coli and other simple biological systems [**93**].

The bound in Eq. 4.5 also applies to digital computing such as that done with DRAM memory. We recognize that its operation places the device in a non-equilibrium steady state, appearing to negate the applicability of Crooks's fluctuation theorem in Eq. 4.1. However, the remedy for systems whose steady states are non-equilibrium is simply to replace the equality with an inequality, implying that more work must be dissipated than in the case of a local-equilibrium steady state [**94**]. Thus our derived bounds must still hold for these modern computing devices.

4.3. Erasure Thermodynamics

Inequalities Eqs. (4.3) and (4.5) place severe constraints on the work required to process information via time-symmetric control on memories. The question remains, though, whether or not these bounds can actually be met by specific protocols or if there might be still tighter bounds to be discovered.

To help answer this question, we turn to the case, originally highlighted by Landauer [**6**], of erasing a single bit of information. This remarkably simple case of computing has held disproportionate sway in the development of thermodynamic computing compared to other elementary operations. The following does not deviate from this habit, showing, in fact, that there remain fundamental issues. We explore this via two different implementations. The first, described via two-state rate equations, and the second with an underdamped double-well potential—Landauer's original, preferred setting.

Suppose the SUS supports two (mesoscopic) memory states, labeled L and R. The task of a time-symmetric protocol that implements erasure is to guide the SUS microscopic dynamics that starts with an initial $50-50$ distribution over the two memory states to a final distribution as biased as possible onto the L state. The logical function C of perfect bit erasure is attained when $C(\mathrm{L}) = C(\mathrm{R}) = \mathrm{L}$, setting either memory state to L. The probabilities of incorrectly sending an L state to R and an R state to R are denoted ϵ_{L} and ϵ_{R}, respectively.

Error generation is described by the binary asymmetric channel [**31**]—the *erasure channel* $\mathcal{E}$ with conditional probabilities:

$$\mathcal{E} = \begin{array}{c c} & \begin{array}{cc} \mathcal{M}_\tau{=}\mathrm{L} & \mathcal{M}_\tau{=}\mathrm{R} \end{array} \\ \begin{array}{c} \mathcal{M}_0{=}\mathrm{L} \\ \mathcal{M}_0{=}\mathrm{R} \end{array} & \left(\begin{array}{cc} 1-\epsilon_{\mathrm{L}} & \epsilon_{\mathrm{L}} \\ 1-\epsilon_{\mathrm{R}} & \epsilon_{\mathrm{R}} \end{array} \right) \end{array} .$$

For any erasure implementation, this Markov transition matrix gives the error rate $\epsilon_{\mathrm{L}} = \epsilon_{\mathrm{L}\to\mathrm{R}}$ from initial memory state $\mathcal{M}_0 = \mathrm{L}$ and the error rate $\epsilon_{\mathrm{R}} = \epsilon_{\mathrm{R}\to\mathrm{R}}$ from the initial memory state $\mathcal{M}_0 = \mathrm{R}$. Noting first that $d(z,z) = 0$ generically, we then have:

$$\begin{aligned} d(\mathrm{L},\mathrm{R}) &= \epsilon_{\mathrm{L}} \ln \frac{\epsilon_{\mathrm{L}}}{1-\epsilon_{\mathrm{R}}} \ , \\ d(\mathrm{R},\mathrm{L}) &= (1-\epsilon_{\mathrm{R}}) \ln \frac{1-\epsilon_{\mathrm{R}}}{\epsilon_{\mathrm{L}}} \ . \end{aligned}$$

So, the bound of Eq. (4.3) simplifies to:

$$\begin{aligned} \beta \langle W \rangle_{\min}^{\text{t-sym}} &= \frac{1}{2}\epsilon_{\mathrm{L}} \ln \frac{\epsilon_{\mathrm{L}}}{1-\epsilon_{\mathrm{R}}} + \frac{1}{2}(1-\epsilon_{\mathrm{R}}) \ln \frac{1-\epsilon_{\mathrm{R}}}{\epsilon_{\mathrm{L}}} \\ &= \Big(\frac{1}{2} - \langle \epsilon \rangle\Big) \ln \frac{1-\epsilon_{\mathrm{R}}}{\epsilon_{\mathrm{L}}} \ , \end{aligned} \tag{4.6}$$

where $\langle \epsilon \rangle = (\epsilon_{\mathrm{L}} + \epsilon_{\mathrm{R}})/2$ is the average error for the process.

Notice further that $C(C(\mathrm{L})) = \mathrm{L}$ but $C(C(\mathrm{R})) \neq \mathrm{R}$, indicating that only the computation on R is nonreciprocal. Therefore, the bound of Eq. (4.5) simplifies to

$$\beta \langle W \rangle_{\min}^{\text{approx}} = \frac{1}{2} \ln(\epsilon^{-1}) \ . \tag{4.7}$$

When we apply Eq. 4.7 to DRAM memory directly, it provides a quantitative comparison beyond a formal divergence of energy costs. Contemporary DRAM memory exhibits a range of "soft" error rates around 10^{-22} failures per write operation [95]. In fact, each write operation is effectively an erasure. (The quoted statistic is an average of $4,000$ correctable errors per 128 MB DIMM per year.) Using Eq. 4.7, this gives a thermodynamic cost of 25 $k_{\mathrm{B}}T$, which is markedly larger than Landauer's $k_{\mathrm{B}}T \ln 2$ factor. It is also, just as clearly, smaller by a factor of roughly 10 than the contemporary energy costs per logic operation displayed in Fig. 4.1. These numerical results on the ability to meet our bounds for the case of bit erasure support the idea that modern computers can be still improved in efficiency, despite that efficiency being limited by the bounds we have introduced. This is further reenforced by the numerical simulations in the following sections that nearly achieve our formal bounds.

4.3.1. Erasure with Two-state Rate Equations. A direct test of time-symmetric erasure requires only a simple two-state system that evolves under a rate equation:

$$
\begin{aligned}
&\frac{d\,\Pr(\mathcal{M}_t = m)}{dt} \\
&\quad = \sum_{m'} \left[r_{m' \to m}(t) \Pr(\mathcal{M}_t = m') - r_{m \to m'}(t) \Pr(\mathcal{M}_t = m) \right],
\end{aligned}
\tag{4.8}
$$

obeying the Arrhenius equations:

$$
\begin{aligned}
r_{\mathrm{R} \to \mathrm{L}}(t) &= A e^{-\Delta E_{\mathrm{R}}(t)/k_{\mathrm{B}}T} \text{ and} \\
r_{\mathrm{L} \to \mathrm{R}}(t) &= A e^{-\Delta E_{\mathrm{L}}(t)/k_{\mathrm{B}}T} \ ,
\end{aligned}
$$

where the states are labeled $\{\mathrm{L}, \mathrm{R}\}$ and the terms $\Delta E_{\mathrm{R}}(t)$ and $\Delta E_{\mathrm{L}}(t)$ in the exponentials are the activation energies to transit over the energy barrier at time t for the Right and Left wells, respectively.

These dynamics can be interpreted as a coarse-graining of thermal motion in a double-well potential energy landscape $V(q,t)$ over the positional variable q at time t. Above, A is an arbitrary constant, which is fixed for the dynamics. q^*_{R} and q^*_{L} are the locations of the Right and Left potential well minima, respectively. Thus, assuming that $q = 0$ is the location of the barrier's maximum between them, we see that the activation energies can be expressed as $\Delta E_{\mathrm{R}}(t) = V(0,t) - V(q^*_{\mathrm{R}}, t)$ and $\Delta E_{\mathrm{L}}(t) = V(0,t) - V(q^*_{\mathrm{L}}, t)$. By varying the potential energy extrema $V(q^*_{\mathrm{R}}, t)$, $V(q^*_{\mathrm{L}}, t)$, and $V(0,t)$ we control the dynamics of the observed variables $\{\mathrm{L}, \mathrm{R}\}$ in much the same way as is done with physical implementations of erasure where barrier height and tilt are controlled in a double-well [25].

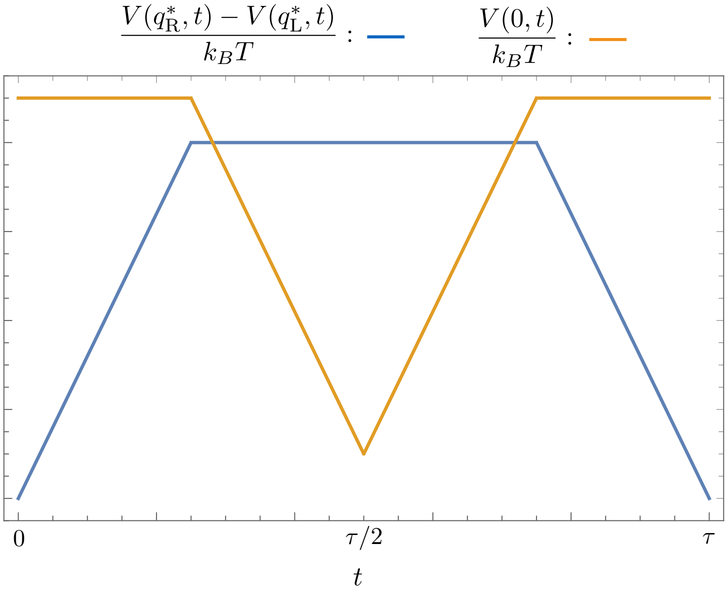

Figure 4.2. Time-symmetric control protocol for implementing moderately-efficient erasure. This should be compared to Landauer's original time-symmetric protocol [6]. Our protocol starts by tilting—increasing the difference in potential energy $(V(q^*_{\mathrm{R}},t)-V(q^*_{\mathrm{L}},t))/k_{\mathrm{B}}T$ between L and R. We increase this value such that transitions are more likely to go from R to L. Then we reduce the barrier height $V(0,t)$ to increase the total flow rate. Finally, we reverse the previous steps, cutting off the flow by raising the barrier, then untilting.

Deviating from previous investigations of efficient erasure, where Landauer's bound was nearly achieved over long times [25, 27], here the constraint to time-symmetric driving over the interval $t \in (0,\tau)$ results in additional dissipated work. As Landauer described [6], erasure can be implemented by turning on and off a tilt from R to L—a time-symmetric protocol. However, to achieve higher accuracy, we also lower the barrier while the system is tilted energetically towards the L well.

Consider a family of control protocols that fit the profile shown in Fig. 4.2. First, we increase the energy tilt from R to L via the energy difference $V(q^*_{\mathrm{R}},t)-V(q^*_{\mathrm{L}},t)$ measured in units of $k_{\mathrm{B}}T$. This increases the relative probability of transitioning R to L. However, with the energy barrier at it's maximum height, the transition takes quite some time. Thus, we reduce the energy barrier $V(0,t)$ to its minimum height halfway through the protocol $t=\tau/2$. Then, we reverse the protocol, raising the barrier back to its default height to hold the probability distribution fixed in the well and untilt so that the system resets to its default double-well potential.

Increasing the maximum tilt—given by $V(q^*_{\mathrm{R}},\tau/2)-V(q^*_{\mathrm{L}},\tau/2)$ at the halfway time—increases erasure accuracy. Figure 4.3 shows that the maximum error $\epsilon=\max\{\epsilon_{\mathrm{R}},\epsilon_{\mathrm{L}}\}$ decreases nearly exponentially with increased maximum energy difference between left and right, going below 1 error in every 1000 trials for our parameter range. Note that ϵ starts at a very high value (greater than 1/2) for zero tilt, since the probability $\epsilon_{\mathrm{R}} = \epsilon$ of ending in the R well starting in the R well is very high if there is no tilt to push the system out of the R well.

Figure 4.3 also shows the relationship between the work and the bounds described above. Given that our system consists of two states $\{\mathrm{L}, \mathrm{R}\}$ and that we choose a control protocol that keeps the energy on the left $V(q^*_{\mathrm{L}}, t)$ fixed, the work (marked by green +s in the figure) is [70]:

$$\begin{aligned}\langle W \rangle &= \int_0^\tau dt \sum_s \Pr(\mathcal{S}_t = s)\partial_t V(s,t) \\ &= \int_0^\tau dt \Pr(\mathcal{M}_t = \mathrm{R})\partial_t V(q^*_{\mathrm{R}}, t) \ .\end{aligned}$$

This work increases almost linearly as the error reduces exponentially.

As a first comparison, note that the Landauer bound $\langle W \rangle^{\mathrm{Landauer}}_{\mathrm{min}} = -k_{\mathrm{B}} T \Delta H(\mathcal{M}_t)$ (marked by orange ×s in the figure) is still valid. However, it is a very weak bound for this time-symmetric protocol. The Landauer bound saturates at $k_{\mathrm{B}} T \ln 2$. Thus, the dissipated work—the gap between orange ×s and green +s—grows approximately linearly with increasing tilt energy.

In contrast, Eq. (4.6)'s bound $\langle W \rangle^{t\text{-sym}}_{\mathrm{min}}$ for time-symmetric protocols is much tighter. The time-symmetric bound is valid: marked by blue circles that all fall below the calculated work (green +s). Not only is this bound much stricter, but it almost exactly matches the calculated work for a large range of parameters, with the work only diverging for higher tilts and lower error rates.

Finally, the approximate bound $\langle W \rangle^{\mathrm{approx}}_{\mathrm{min}} = \frac{k_{\mathrm{B}} T}{2} \ln \epsilon^{-1}$ (marked by red +s) of Eq. (4.7), which captures the error scaling, behaves as expected. The error-dependent work bound nearly exactly matches the exact bound for low error rates on the right side of the plot and effectively bounds the work. For lower tilts, this quantity does not bound the work and is not a good estimate of the true bound, but this is consistent with expectations for high error rates. This approximation should only be employed for very reliable computations, for which it appears to be an excellent estimate. Thus, the two-level model of erasure demonstrates that the time-symmetric control bounds on work and dissipation are reasonable in both their exact and approximate forms at low error rates.

4.3.2. Erasure with an Underdamped Double-well Potential.

The physics in the rate equations above represents a simple model of a bistable thermodynamic system, which can serve

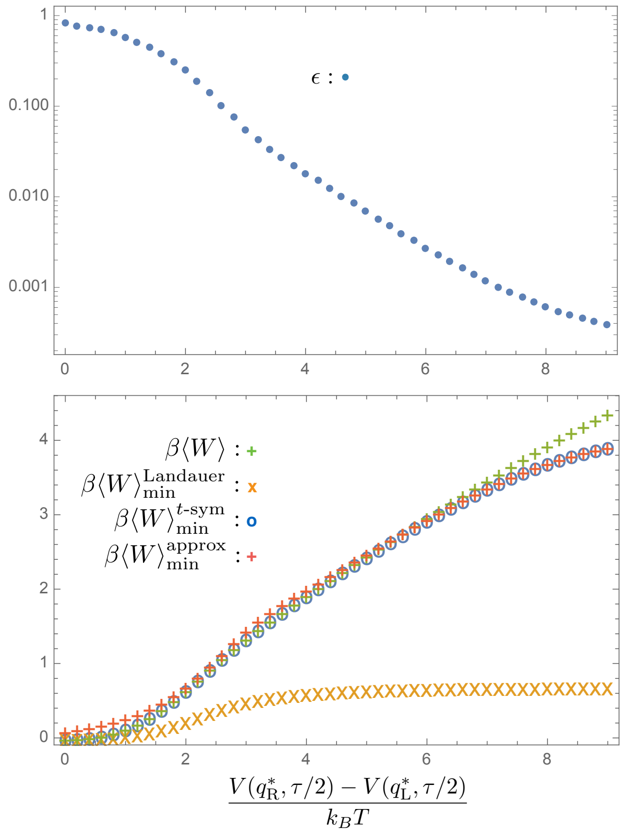

FIGURE 4.3. (Top) Maximum error ϵ (blue dots) decreases approximately exponentially with increasing maximum tilt. The latter is given by the maximum energy difference between the right and left energy well $V(q^*_{\text{R}}, \tau/2) - V(q^*_{\text{L}}, \tau/2)$. (Bottom) Work $\langle W \rangle$ (green +s), scaled by the inverse temperature $\beta = 1/k_{\text{B}}T$, increases with increasing maximum tilt and decreasing error. The Landauer work bound $\langle W \rangle^{\text{Landauer}}_{\text{min}}$ (orange ×s) is a very weak bound, asymptoting to a constant value rather than continuing to increase, as the work does. The bound $\langle W \rangle^{t\text{-sym}}_{\text{min}}$ (blue circles) on time-symmetrically driven protocols, on the other hand, is a very tight bound for lower values of maximum tilt. The work deviates from the time-symmetric bound for higher tilts. Finally, the approximate bound $\langle W \rangle^{\text{approx}}_{\text{min}}$ (red +s), which scales as $\ln \epsilon^{-1}$, is not an accurate bound over the entire range, but it very closely matches the exact time-symmetric bound $\langle W \rangle^{t\text{-sym}}_{\text{min}}$ for small ϵ, as expected.

as an approximation for many different bistable systems. One possible interpretation is a coarse-graining of the Langevin dynamics of a particle moving in a double-well potential. To explore the broader validity of the error–dissipation tradeoff, here we simulate the dynamics of a stochastic particle coupled to a thermal environment at constant temperature and a work reservoir via such a 1D potential. Again, we find that the time-symmetric bounds are much tighter than Landauer's, reflecting the error–dissipation tradeoff of this control protocol class.

Consider a one-dimensional particle with position and momentum in an external potential and in thermal contact with the environment at temperature T. We consider a protocol architecture similar to that of Sec. 4.3.1, but with additional passive substages at the beginning middle and end: (i) hold the potential in the symmetric double-well form, (ii) positively tilt the potential, (iii) completely drop the potential barrier between the two wells, (iv) hold the potential while it is tilted with no barrier, (v) restore the original barrier, (vi) remove the positive tilt, restoring the original symmetric double-well, and (vii) hold the potential in this original form.

As a function of position q and time t, the potential then takes the form:

$$V(q,t) = aq^4 - b_0 b_f(t) q^2 + c_0 c_f(t) q \ ,$$

with constants $a, b_0, c_0 > 0$. The protocol functions $b_f(t)$ and $c_f(t)$ evolve in a piecewise linear and time-symmetric manner according to Table 4.1, where

$$t_0 = 0 \ , \quad t_1 = \frac{1}{12}\tau \ , \quad t_2 = \frac{3}{12}\tau \ , \quad \ldots \ , \quad t_6 = \frac{11}{12}\tau \ , \quad t_7 = \tau \ .$$

The potential thus begins and ends in a symmetric double-well configuration with each well defining a memory state. During the protocol, though, the number of metastable regions is temporarily reduced to one. Figure 4.4 (top three panels) shows the protocol functions over time as well as the resultant potential function at key times for one such set of protocol parameters; see nondimensionalization in App. 4.B. At any time, we label the metastable regions from most negative position to most positive the L state and, if it exists, the R state.

t	t_0		t_1		t_2		t_3		t_4		t_5		t_6		t_7
$b_f(t)$		1		1		$\frac{t_3-t}{t_3-t_2}$		0		$\frac{t-t_4}{t_5-t_4}$		1		1	
$c_f(t)$		0		$\frac{t-t_1}{t_2-t_1}$		1		1		1		$\frac{t_6-t}{t_6-t_5}$		0	

TABLE 4.1. Erasure protocol.

We simulate the motion of the particle with underdamped Langevin dynamics:

$$\begin{aligned} dq &= v dt \\ m dv &= -\left(\frac{\partial}{\partial q} V(q,t) + \lambda v\right) dt + \sqrt{2 k_{\mathrm{B}} T \lambda}\, r(t) \sqrt{dt} \ , \end{aligned}$$

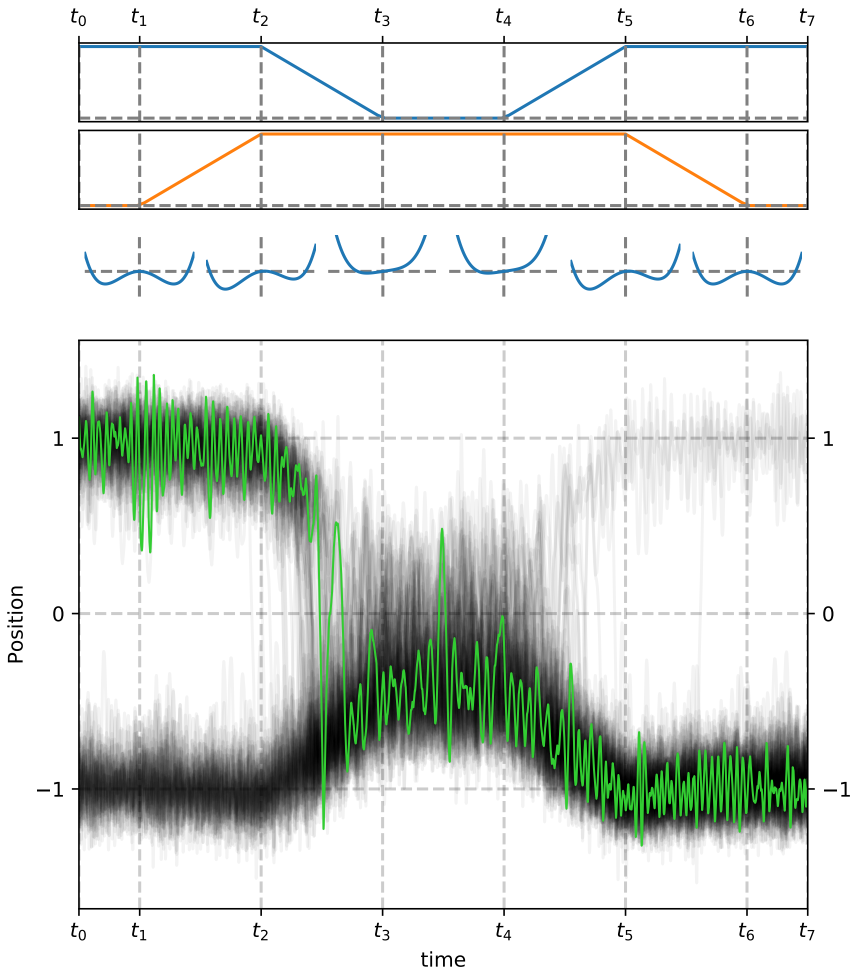

FIGURE 4.4. Erasure via an underdamped double-well potential: Protocol functions $b(t)$ (top panel, blue) and $c(t)$ (second panel, orange) are symmetric in time, guaranteeing the potential function (third panel) to evolve symmetrically in time. Due to the spatial asymmetry in the potential over the majority of the protocol, however, erasure to state L ($x < 0$) typically occurs, evidenced by the evolution of the system position for 100 randomly-chosen trajectories (bottom panel, black). The L and R states merge into one between times t_2 and t_3 and separate again between times t_4 and t_5. A single trajectory (bottom panel, green) shows the typical behavior of falling into the $x < 0$ region by time t_3 and remaining there when the L state is reintroduced for the rest of the protocol.

where λ is the coupling between the thermal environment and particle, m is the particle's mass, and $r(t)$ is a memoryless Gaussian random variable with $\langle r(t) \rangle = 0$ and $\langle r(t)r(t') \rangle = \delta(t - t')$. The particle is initialized to be in global equilibrium over the initial potential $V(\cdot, 0)$. Figure 4.4 (bottom panel) shows 100 randomly-chosen resultant trajectories for a choice of process parameters.

The work done on a single particle over the course of the protocol with trajectory $\left(q(t)\right)_t$ is [70]:

$$W = \int_0^\tau dt \frac{\partial V(q,t)}{\partial t}\Big|_{q=q(t)} .$$

Figure 4.5 shows the net average work over time for an erasure process, comparing it to (i) the Landauer bound, (ii) the exact bound of Eq. (4.6), and (iii) the approximate bound of Eq. (4.7). Notice that the final net average work lies above all three, as it should and that the time-symmetric bounds presented here are tighter than Landauer's.

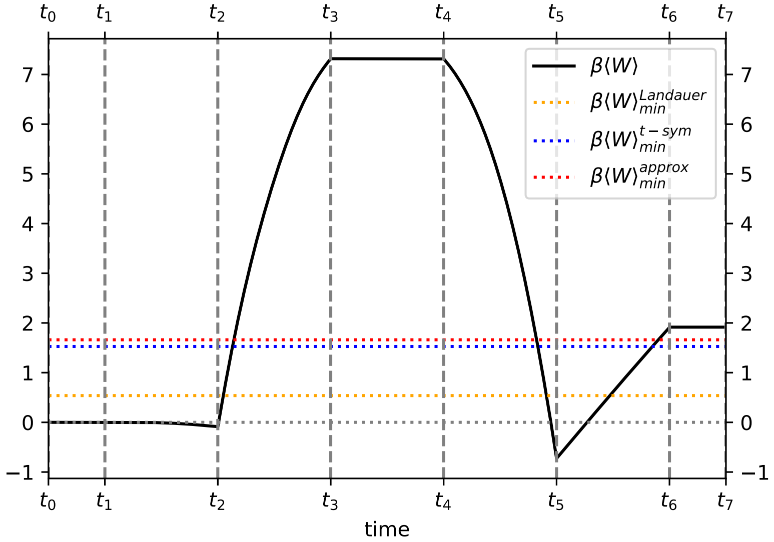

FIGURE 4.5. Average work in $k_{\mathrm{B}}T$ over time for an erasure (black). Calculated from the simulation-estimated values ϵ_L and ϵ_R, Landauer's bound is given by the dashed yellow line and our approximate and exact bounds (Eqs. (4.7) and (4.6)) are given in dashed red and blue lines, respectively.

We repeat this comparison for an array of different parameters for the erasure protocol. As described in App. 4.B, we vary features of the dynamics—including mass m, temperature T, coupling to the heat bath λ, duration of control τ, maximum depth of the potential energy wells, and maximum tilt between the wells. Nondimensionalization reduces the relevant parameters to just four, allowing us to explore a broad swathe of possible physical erasures with 735 different protocols. For each protocol, we simulate 100,000 trajectories to estimate the work cost and errors ϵ_{R} and ϵ_{L} of the operation.

Figure 4.6 compares the work spent for each of the 735 erasure protocols to the sampled maximum error $\epsilon = \max(\epsilon_{\mathrm{L}}, \epsilon_{\mathrm{R}})$. Each protocol corresponds to a green cross, whose vertical position corresponds

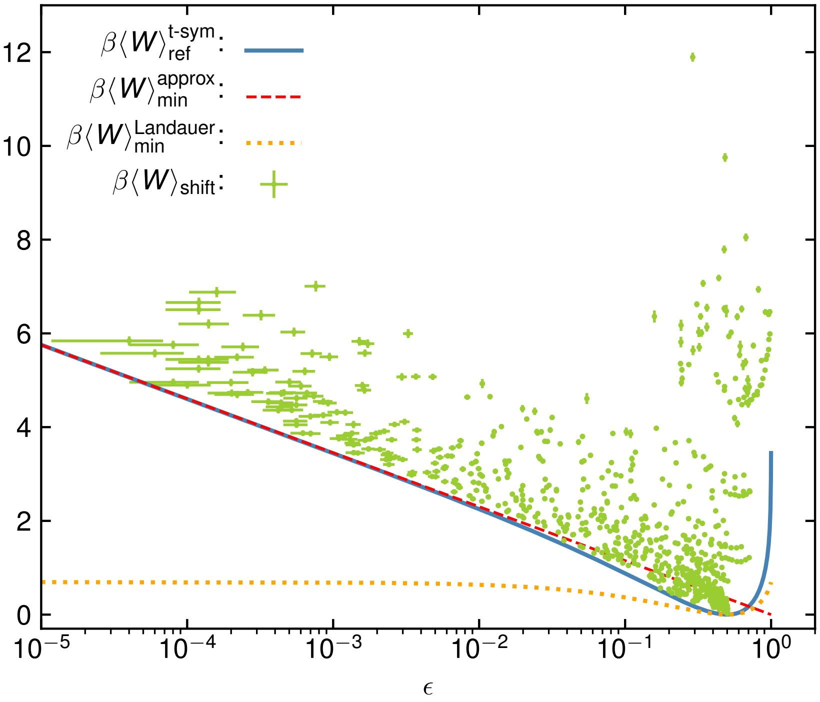

FIGURE 4.6. Reference bound $\langle W\rangle_{\text{ref}}^{t\text{-sym}}$ (blue line) lower bounds all of the shifted works $\langle W\rangle_{\text{shift}}$ (green markers), often quite tightly. The approximate bound $\langle W\rangle_{\text{min}}^{\text{approx}}$ (red dashed line) rapidly converges with decreasing error to $\langle W\rangle_{\text{ref}}^{t\text{-sym}}$. Time-asymmetric protocols can do better, needing only to satisfy Landauer's bound $\langle W\rangle_{\text{min}}^{\text{Landauer}}$ (orange dotted line).

to the shifted work $\langle W\rangle_{\text{shift}}$, which accounts for inhomogeneities in the error rate. Note that the exact bound $\langle W\rangle_{\text{min}}^{t\text{-sym}}$ from Eq. (4.6) reduces to a simple relationship between work and error tolerance ϵ when the errors are homogeneous $\epsilon_{\text{R}} = \epsilon_{\text{L}} = \epsilon$:

$$\langle W\rangle_{\text{ref}}^{t\text{-sym}} = \left(\frac{1}{2} - \epsilon\right) \ln \frac{1-\epsilon}{\epsilon} \ ,$$

which we plot with the blue curve in Fig. 4.6. The cost of inhomogeneities in the error is evaluated by the difference between this reference bound and the exact work bound. This cost is added to the calculated work for each protocol to determine the shifted work:

$$\langle W\rangle_{\text{shift}} = \langle W\rangle + \langle W\rangle_{\text{ref}}^{t\text{-sym}} - \langle W\rangle_{\text{min}}^{t\text{-sym}} \ ,$$

such that the vertical distance between $\langle W\rangle_{\text{shift}}$ and $\langle W\rangle_{\text{ref}}^{t\text{-sym}}$ in Fig. 4.6 gives the true difference $\langle W\rangle - \langle W\rangle_{\text{min}}^{t\text{-sym}}$ between the average sampled work and exact bound for the simulated protocol.

Figure 4.6 shows that the shifted average works for all of the simulated protocols in green, including error bars, all lay above the reference work bound in blue. Thus, we see that all simulated protocols satisfy the bound $\langle W \rangle \geq \langle W \rangle_{\text{min}}^{t\text{-sym}}$. Furthermore, many simulated protocols end up quite close to their exact bound. There are protocols with small errors, but they have larger average works. The error–dissipation tradeoff is clear.

The error–dissipation tradeoff is further illustrated in Fig. 4.6 by the red line, which describes the low-ϵ asymptotic bound $\langle W \rangle_{\text{min}}^{\text{approx}}$ given by Eq. (4.7). In this semi-log plot, it rather quickly becomes an accurate approximation for small error.

Finally, Fig. 4.6 plots the Landauer bound $\langle W \rangle_{\text{min}}^{\text{Landauer}}$ as a dotted orange line. It is calculated using the final probability of the R mesostate. The bound is weaker than that set by $\langle W \rangle_{\text{ref}}^{t\text{-sym}}$. As $\epsilon \to 0$, the gap between $\langle W \rangle_{\text{ref}}^{t\text{-sym}}$ and $\langle W \rangle_{\text{min}}^{\text{Landauer}}$ in Fig. 4.6 relentlessly increases. The stark difference in the energy scale of the time-symmetric bounds developed here and that of the looser Landauer bound shows a marked tightening of thermodynamic bounds on computation.

Notably, the protocol Landauer originally proposed to erase a bit requires *significantly more work* than his bound $k_{\text{B}} T \ln 2$ to reliably erase a bit. This extra cost is a direct consequence of his protocol's time symmetry. It turns out that time-*asymmetric* protocols for bit erasure have been used in experiments that more nearly approach Landauer's bound [**96**, **97**]. Although, it is not clear to what extent time asymmetry was an intentional design constraint in their construction, since there was no general theoretical guidance until now for why time-symmetry or asymmetry should matter. Figures 4.6 and 4.3 confirm that Ref. [**97**]'s time-asymmetric protocol for bit erasure—where the barrier is lowered before the tilt, but then raised before untilting—is capable of reliable erasure that is more thermodynamically efficient than any time-symmetric protocol could ever be.

These underdamped simulations drive home the point that our bounds are independent of the details of the dynamics used for computation. Our results are very general in that regard. As long as the system *starts* metastable and is then driven by a time-symmetric protocol, the error–dissipation tradeoff quantifies the minimal dissipation that will be incurred (for a desired level of computational accuracy) by the time the system relaxes again to metastability.

4.4. Conclusion

We adapted Ref. [75]'s thermodynamic analysis of time-symmetric protocols to give a detailed analysis of the trade-offs between accuracy and dissipation encountered in erasing information.

Reference [75] showed that time symmetry and metastability together imply a generic error–dissipation tradeoff. The minimal work expected for a computation $\mathcal{C}$ is the average nonreciprocity. In the low-error limit—where the probability of error must be much less than unity ($\epsilon \ll 1$)—the minimum work diverges according to:

$$\beta \langle W \rangle_{\min}^{\text{approx}} = \left\langle \left[\mathcal{C}(\mathcal{C}(\mathcal{M}_0)) \neq \mathcal{M}_0 \right] \right\rangle_{\mathcal{M}_0} \ln(\epsilon^{-1})$$

Of all of this work, only the meager Landauer cost $\Delta \mathrm{H}(\mathcal{M}_t)$, which saturates to some finite value as $\epsilon \to 0$, can be thermodynamically recovered in principle. Thus, irretrievable dissipation scales as $\ln(\epsilon^{-1})$. The reciprocity coefficient $\left\langle \left[\mathcal{C}(\mathcal{C}(\mathcal{M}_0)) \neq \mathcal{M}_0 \right] \right\rangle_{\mathcal{M}_0}$ depends only on the deterministic computation to be approximated. This points out likely energetic inefficiencies in current instantiations of reliable computation. It also suggests that time-asymmetric control may allow more efficient computation—but only when time-asymmetry is a free resource, in contrast to modern computer architecture.

The results here verified these general conclusions for erasure, showing in detail how tight the bounds can be and, for high-reliability thermodynamic computing, how they overwhelm Landauer's. It may be fruitful to explore the ideas behind our results in explicitly quantum, finite, and even zero-temperature systems. Refined versions of Landauer's bound and other thermodynamic results can be obtained for such models [98, 99]. Also, explicit consideration of finite-time protocols can reveal efficiency advantages when treating ensembles of systems under majority-logic decoding [100, 101, 102]. Perhaps analagous refinements of the results presented here can be found as well. Despite the almost universal focus on information erasure as a proxy for all of computing, we now see that there is a wide diversity of costs in thermodynamic computing. Looking to the future, these costs must be explored in detail if we are to design and build more capable and energy efficient computing devices. Beyond engineering and sustainability concerns, explicating Landauer's Stack will go a long way to understanding the fundamental physics of computation—one of Landauer's primary goals [103]. In this way, we now better appreciate the suite of thermodynamic costs—what we called Landauer's Stack—that underlies modern computing.

4.A. Proof of Exact Bound for Time-Symmetric Protocols

Here we prove Eq. 4.2, an exact bound under the following conditions: the protocol is time-symmetric; it operates on systems that start and end in metastable equilibrium; the corresponding metastable regions look the same in state space vs. energy up to translation, rotation, and reflection in state space; the memory states are symmetric under time reversal; and the equilibrium probability of a microstate at any time does not change under time reversal of the microstate. These conditions are all met in a wide variety of computational processes, including the bit erasure processes we study here. For a more general bound that applies under fewer restrictions, see Ref. [**75**].

We start with Eq. 4.1, Crooks' DFT applied to time-symmetric protocols:

$$\begin{aligned}\omega(\overrightarrow{z}) &= \ln \frac{P(\overrightarrow{z}|s(0))\boldsymbol{\mu}_0(s(0))}{P(\boldsymbol{Я}\overrightarrow{z}|s(\tau)^\dagger)\boldsymbol{\mu}_\tau(s(\tau))} \\ &= \ln \frac{P(\overrightarrow{z})}{R(\boldsymbol{Я}\overrightarrow{z})} \ .\end{aligned}$$

To help with the algebra, we introduce a second way to define reverse process probabilities:

$$Q(\overrightarrow{Z} = \overrightarrow{z}) = Q(\overrightarrow{z}) = R(\boldsymbol{Я}\overrightarrow{z}) = P(\boldsymbol{Я}\overrightarrow{z}|s(\tau)^\dagger)\boldsymbol{\mu}_\tau(s(\tau)) \ . \tag{4.9}$$

We also allow both P and Q to take as arguments system microstate trajectories $\overrightarrow{z}$, pairs of initial and final memory states $\overrightarrow{m} = (z(0), z(\tau))$, or a combination of the two, where $\overrightarrow{m}$ may even appear as a condition. We define the time reverse of the pair of initial and final memory states as $\boldsymbol{Я}\overrightarrow{m} = \boldsymbol{Я}(z(0), z(\tau)) = (z(\tau)^\dagger, z(0)^\dagger) = (z(\tau), z(0))$. (Note that $m^\dagger \equiv \{s^\dagger : s \in m\}$, and we explicitly assume in the following that the each memory state is time-reversal invariant; i.e., $m^\dagger = m$ for all $m \in \boldsymbol{\mathcal{M}}$.) For example, the probability of observing the reverse trajectory $\boldsymbol{Я}\overrightarrow{z}$ in the reverse process conditioned on observing the reverse of the pair of initial and final memory states $\boldsymbol{Я}\overrightarrow{m}$ is:

$$\begin{aligned}Q(\overrightarrow{z}|\overrightarrow{m}) = Q(\overrightarrow{Z} = \overrightarrow{z}|\overrightarrow{M} = \overrightarrow{m}) &= \frac{Q(\overrightarrow{Z} = \overrightarrow{z}, \overrightarrow{M} = \overrightarrow{m})}{Q(\overrightarrow{M} = \overrightarrow{m})} = \frac{Q(\overrightarrow{z}, \overrightarrow{m})}{Q(\overrightarrow{m})} \\ &= \frac{\Pr_{\boldsymbol{Я}\overrightarrow{x}}(\overrightarrow{Z} = \boldsymbol{Я}\overrightarrow{z}, \overrightarrow{M} = \boldsymbol{Я}\overrightarrow{m}|\mathcal{S}_0 = s(\tau)^\dagger)\boldsymbol{\mu}_\tau(s(\tau))}{\int d\overrightarrow{z}' \Pr_{\boldsymbol{Я}\overrightarrow{x}}(\overrightarrow{Z} = \boldsymbol{Я}\overrightarrow{z}', \overrightarrow{M} = \boldsymbol{Я}\overrightarrow{m}|\mathcal{S}_0 = s'(\tau)^\dagger)\boldsymbol{\mu}_\tau(s'(\tau))} \ ,\end{aligned}$$

where $\overrightarrow{M}$ is the random variable for the pair of initial and final memory states.

To derive Eq. 4.2, we first show that the average dissipated work is bounded by a Kullback–Leibler divergence on forward process memory state distributions compared with reverse process memory state distributions. Letting $\overrightarrow{\mathcal{M}}$ be the set of all possible pairs of initial and final memory states,

$$\begin{aligned}
\langle \omega \rangle &= \int d\overrightarrow{z}\, P(\overrightarrow{z}) \ln \frac{P(\overrightarrow{z})}{Q(\overrightarrow{z})} \\
&= \sum_{\overrightarrow{m} \in \overrightarrow{\mathcal{M}}} \int_{\overrightarrow{m}} d\overrightarrow{z}\, P(\overrightarrow{z}, \overrightarrow{m}) \ln \frac{P(\overrightarrow{z}, \overrightarrow{m})}{Q(\overrightarrow{z}, \overrightarrow{m})} \\
&= \sum_{\overrightarrow{m} \in \overrightarrow{\mathcal{M}}} \int_{\overrightarrow{m}} d\overrightarrow{z}\, P(\overrightarrow{z}|\overrightarrow{m}) P(\overrightarrow{m}) \ln \frac{P(\overrightarrow{z}|\overrightarrow{m}) P(\overrightarrow{m})}{Q(\overrightarrow{z}|\overrightarrow{m}) Q(\overrightarrow{m})} \\
&= \sum_{\overrightarrow{m} \in \overrightarrow{\mathcal{M}}} P(\overrightarrow{m}) \ln \frac{P(\overrightarrow{m})}{Q(\overrightarrow{m})} + \sum_{\overrightarrow{m} \in \overrightarrow{\mathcal{M}}} P(\overrightarrow{m}) \int_{\overrightarrow{m}} d\overrightarrow{z}\, P(\overrightarrow{z}|\overrightarrow{m}) \frac{P(\overrightarrow{z}|\overrightarrow{m})}{Q(\overrightarrow{z}|\overrightarrow{m})} \\
&= \mathrm{D}_{\mathrm{KL}}\left[P \,||\, Q\right]_{\overrightarrow{\mathcal{M}}} + \sum_{\overrightarrow{m} \in \overrightarrow{\mathcal{M}}} P(\overrightarrow{m})\, \mathrm{D}_{\mathrm{KL}}\left[P(\cdot|\overrightarrow{m}) \,||\, Q(\cdot|\overrightarrow{m})\right]_{\overrightarrow{m}} \ ,
\end{aligned}$$

where $\mathrm{D}_{\mathrm{KL}}\left[\cdot \,||\, \cdot\right]_{\overrightarrow{\mathcal{M}}}$ is the Kullback–Leibler divergence for distributions over all $\overrightarrow{m} \in \overrightarrow{\mathcal{M}}$, and $\mathrm{D}_{\mathrm{KL}}\left[\cdot \,||\, \cdot\right]_{\overrightarrow{m}}$ is that for all $\overrightarrow{z}$ consistent with $\overrightarrow{m}$. Because Kullback–Leibler divergences are always nonnegative, we arrive at the following inequality:

$$\langle \omega \rangle \geq \mathrm{D}_{\mathrm{KL}}\left[P \,||\, Q\right]_{\overrightarrow{\mathcal{M}}} \ .$$

Second, we show that the above Kullback–Leibler divergence equals the right hand side of Eq. 4.2.

To help us with this second task, we start by establishing that $Q(\mathcal{M}_0 = z | \mathcal{M}_\tau = y) = P(\mathcal{M}_\tau = z | \mathcal{M}_0 = y)$. We have:

$$Q(\mathcal{M}_0 = z | \mathcal{M}_\tau = y) = \frac{Q(\mathcal{M}_0 = z, \mathcal{M}_\tau = z')}{Q(\mathcal{M}_\tau = y)} \ . \tag{4.10}$$

The denominator is simply the probability of observing the memory state y at the end of the forward process:

$$\begin{aligned}
Q(\mathcal{M}_\tau = y) &= \int_{\overrightarrow{m}} d\overrightarrow{z}\, \Pr_{\overrightarrow{x}}(\overrightarrow{Z} = \boldsymbol{Я}\overrightarrow{z}, \mathcal{M}_0 = y^\dagger | \mathcal{S}_0 = s_\tau^\dagger) \boldsymbol{\mu}_\tau(s_\tau) && (4.11) \\
&= \int_{\overrightarrow{m}} d\overrightarrow{z}\, \Pr_{\overrightarrow{x}}(\overrightarrow{Z} = \boldsymbol{Я}\overrightarrow{z} | \mathcal{S}_0 = s_\tau^\dagger) \boldsymbol{\mu}_\tau(s_\tau) && (4.12) \\
&= \int_y ds_\tau \boldsymbol{\mu}_\tau(s_\tau) && (4.13) \\
&= P(\mathcal{M}_\tau = z') \ . && (4.14)
\end{aligned}$$

Because the process begins and ends in a metastable distribution, the probability of observing a particular microstate given some memory state is the same as the probability of observing that microstate given that we have locally equilibrated to that memory state. And because the process is time symmetric, the starting and ending local equilibrium distributions are the same. That is, letting $\pi_t^{(y)}(s)$ denote the local equilibrium probability of observing microstate s given memory state y at time t:

$$P(\mathcal{S}_\tau = s_\tau | \mathcal{M}_\tau = y) = \pi_\tau^{(y)}(s_\tau) = \pi_0^{(y)}(s_\tau) \ . \tag{4.15}$$

Additionally,

$$P(\mathcal{S}_\tau = s_\tau | \mathcal{M}_\tau = y) = \frac{P(\mathcal{S}_\tau = s_\tau)\big[s_\tau \in m'\big]}{P(\mathcal{M}_\tau = y)} = \frac{\boldsymbol{\mu}_\tau(s_\tau)\big[s_\tau \in m'\big]}{P(\mathcal{M}_\tau = y)} \ . \tag{4.16}$$

We then have

$$Q(\mathcal{M}_0 = z | \mathcal{M}_\tau = y) = \frac{\int_{s_0 \in m, s_\tau \in m'} d\overrightarrow{z} \, \mathrm{Pr}_{\overrightarrow{x}}(\overrightarrow{Z} = \boldsymbol{Я} \overrightarrow{z}, \mathcal{M}_0 = y^\dagger, \mathcal{M}_\tau = z^\dagger | \mathcal{S}_0 = s_\tau^\dagger) \boldsymbol{\mu}_\tau(s_\tau)}{P(\mathcal{M}_\tau = y)} \tag{4.17}$$

$$= \int_{s_0 \in m, s_\tau \in m'} ds_0 \, ds_\tau \, \Pr_{\overrightarrow{x}}(\mathcal{S}_\tau = s_0^\dagger | \mathcal{S}_0 = s_\tau^\dagger) \, \pi_0^{(y)}(s_\tau) \tag{4.18}$$

$$= \int_{s^\dagger \in m, s'^\dagger \in m'} ds \, ds' \, \Pr_{\overrightarrow{x}}(\mathcal{S}_\tau = s | \mathcal{S}_0 = s') \, \pi_0^{(y)}(s'^\dagger) \tag{4.19}$$

$$= \int_{s \in m^\dagger, s' \in m'^\dagger} ds \, ds' \, \Pr_{\overrightarrow{x}}(\mathcal{S}_\tau = s | \mathcal{S}_0 = s') \, \pi_0^{(y)}(s'^\dagger) \ . \tag{4.20}$$

Note that, in Eq. (4.19), we have simply changed the symbols denoting the variables of integration.

We now invoke our assumption that the memory partitions are chosen to be time-reversal invariant, such that $m^\dagger = m$ for all $m \in \boldsymbol{\mathcal{M}}$. If we furthermore assume that the equilibrium probability of a microstate does not change under time reversal, then time-reversal symmetry of the memory states implies that the local equilibrium probability of a microstate does not change either:

$$\pi_0^{(m)}(s) = \pi_0^{(m)}(s^\dagger) \ . \tag{4.21}$$

We therefore have

$$Q(\mathcal{M}_0 = z|\mathcal{M}_\tau = y) = \int_{s\in m, s'\in m'} ds\, ds' \Pr_{\overrightarrow{x}}(\mathcal{S}_\tau = s|\mathcal{S}_0 = s')\, \pi_0^{(y)}(s') \tag{4.22}$$

$$= \int_{\overrightarrow{m}} d\overrightarrow{z} \Pr_{\overrightarrow{x}}(\overrightarrow{Z} = \overrightarrow{z}, \mathcal{M}_\tau = z|\mathcal{S}_0 = s_0)\pi_0^{(y)}(s_0) \tag{4.23}$$

$$= P(\mathcal{M}_\tau = z|\mathcal{M}_0 = y)\ . \tag{4.24}$$

This allows us to rewrite the memory state trajectory probability under the reverse process in terms of forward process probabilities:

$$Q(\mathcal{M}_0 = z, \mathcal{M}_\tau = y) = Q(\mathcal{M}_\tau = y)Q(\mathcal{M}_0 = z|\mathcal{M}_\tau = y) \tag{4.25}$$

$$= P(\mathcal{M}_\tau = y)P(\mathcal{M}_\tau = z|\mathcal{M}_0 = y) \tag{4.26}$$

$$= \boldsymbol{\mu}'_\tau(y)P(\mathcal{M}_\tau = z|\mathcal{M}_0 = y)\ . \tag{4.27}$$

We can now show that $\mathrm{D_{KL}}\left[P \mid\mid Q\right]_{\overrightarrow{\mathcal{M}}}$ equals the right hand side of Eq. 4.2, finishing the proof.

$$\mathrm{D_{KL}}\left[P \mid\mid Q\right]_{\overrightarrow{\mathcal{M}}} = \sum_{\overrightarrow{m}\in\overrightarrow{\mathcal{M}}} P(\overrightarrow{m}) \ln \frac{P(\overrightarrow{m})}{Q(\overrightarrow{m})} \tag{4.28}$$

$$= \sum_{z,y\in\boldsymbol{\mathcal{M}}} P(z,y) \ln \frac{P(z,y)}{Q(z,y)} \tag{4.29}$$

$$= \sum_{z,y\in\boldsymbol{\mathcal{M}}} P(z,y) \ln \frac{\boldsymbol{\mu}'_0(z)P(y|z)}{\boldsymbol{\mu}'_\tau(y)P(z|y)} \tag{4.30}$$

$$= \sum_{z,y\in\boldsymbol{\mathcal{M}}} P(z,y) \ln \boldsymbol{\mu}'_0(z) - \sum_{z,y\in\boldsymbol{\mathcal{M}}} P(z,y) \ln \boldsymbol{\mu}'_\tau(y) + \sum_{z,y\in\boldsymbol{\mathcal{M}}} P(z,y) \ln \frac{P(y|z)}{P(z|y)} \tag{4.31}$$

$$= \sum_{z\in\boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \ln \boldsymbol{\mu}'_0(z) - \sum_{y\in\boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_\tau(y) \ln \boldsymbol{\mu}'_\tau(y) + \sum_{z\in\boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \sum_{y\in\boldsymbol{\mathcal{M}}} P(y|z) \ln \frac{P(y|z)}{P(z|y)} \tag{4.32}$$

$$= -H(\mathcal{M}_0) + H(\mathcal{M}_\tau) + \sum_{z\in\boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \sum_{y\in\boldsymbol{\mathcal{M}}} d(z,y) \tag{4.33}$$

$$= \Delta H(\mathcal{M}_\tau) + \sum_{z\in\boldsymbol{\mathcal{M}}} \boldsymbol{\mu}'_0(z) \sum_{y\in\boldsymbol{\mathcal{M}}} d(z,y)\ , \tag{4.34}$$

where $H(\mathcal{M}_t)$ is the entropy of the memory state at time t.

4.B. Langevin Simulations of Erasure

To help simulate a wide variety of protocols, we first nondimensionalize the equations of motion, using variables:

$$q' = \sqrt{\frac{2a}{b_0}}q \ , \ t' = \frac{2ak_\mathrm{B}T}{b_0\lambda}t \ , \ v' = \frac{\lambda}{k_\mathrm{B}T}\sqrt{\frac{b_0}{2a}}v \ , \ \text{and}$$
$$V' = \frac{1}{k_\mathrm{B}T}V \ .$$

Note that the position scale $\sqrt{b_0/2a}$ is the distance from zero to either well minima in the default potential $V(\cdot,0) = V(\cdot,\tau)$. Substitution then provides the following nondimensional equations:

$$dq' = v'dt'$$
$$m'dv' = -\left(\frac{\partial}{\partial q'}V'(q',t') + v'\right)dt' + \sqrt{2}r(t')\sqrt{dt'} \ ,$$

with:

$$m' = \frac{2amk_\mathrm{B}T}{b_0\lambda^2} \ ,$$

which is the first nondimensional parameter to specify an erasure process.

The nondimensional potential can be expressed as:

$$V'(q',t') = \alpha\Big(q'^4 - 2b'_f(t')q'^2 + \zeta c'_f(t')q'\Big) \ ,$$

where:

$$\alpha = \frac{b_0^2}{4ak_\mathrm{B}T} \text{ and } \zeta = c_0\sqrt{\frac{2a}{b_0^3}}$$

are two more nondimensional parameters to specify and:

$$b'_f(t') = b_f\left(\frac{2ak_\mathrm{B}T}{b_0\lambda}t\right) \text{ and } c'_f(t') = c_f\left(\frac{2ak_\mathrm{B}T}{b_0\lambda}t\right)$$

simply express b_f and c_f with the nondimensional time as input. The fourth and final nondimensional parameter is the nondimensional total time:

$$\tau' = \frac{2ak_\mathrm{B}T}{b_0\lambda}\tau \ .$$

To explore the space of possible underdamped erasure dynamics, we simulate 735 different protocols, determined by all combinations of the following values for the four nondimensional parameters: $m' \in \{0.25, 1.0, 4.0\}$, $\alpha \in \{2, 4, 7, 10, 12\}$, $\zeta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$, and $\tau' \in \{4, 8, 16, 32, 64, 128, 256\}$. $100,000$ trials of each parameter set were simulated. For the simulations of Figs. 4.4 and 4.5, we set $m' = 1$, $\alpha = 7$, $\zeta = 0.4$, and $\tau' = 100$. Figure 4.6 shows that the (error, work) pairs obtained for these various dynamics fill in the region allowed by our time-symmetric bounds. These bounds can indeed be tight, but it is quite possible to waste more energy if the computation is not tuned for energetic efficiency.

To update particle position and velocity each time step, we used the fourth-order Runge–Kutta integration for the deterministic portion of the equations of motion and a simple Euler method in combination with a Gaussian number generator for the stochastic portion. To determine the time step size, we considered a range of possible time steps for 81 of the possible 735 parameter sets and looked for convergence of the sampled average works and maximum errors ϵ, again using $100,000$ trials per parameter set.

The maximum errors were stable over the whole range of tested step sizes. Looking with decreasing step size, the final step size of 0.0025 was chosen when the average works stopped fluctuating within 5σ of their statistical errors for all 81 parameter sets. The error bars presented for the average works in Fig. 4.6 were then generously set to be 5 times the estimated statistical errors, which were each obtained by dividing the sampled standard deviation by the square root of the number of trials. Error bars for the maximum errors were set to be the statistical errors of ϵ_{L} or ϵ_{R}, depending on which was the maximum, whose statistical errors were obtained by assuming binomial statistics.

Bibliography

[1] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.*, 5(3):183–191, 1961.

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

[3] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. eistical hysics*, 90(5-6):1481–1487, 1998.

[4] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721–2726, 1999.

[5] C. G. Knott. *Life and Scientific Work of Peter Gurthrie Tait.* Cambridge University press, Cambridge, United Kingdom, 1911. Letter from Maxwell to Tait, 11 December 1867, quoted herein pp. 213-214.

[6] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5(3):183–191, 1961.

[7] C. H. Bennett. Thermodynamics of computation—A review. *Intl. J. Theo. Phys.*, 21:905, 1982.

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

[9] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *J. Stat. Phys.*, 167(6):1555–1585, 2016.

[10] T. Conte et al. Thermodynamic computing. *arxiv:1911.01968*.

[11] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of Maxwell's demon. *Proc. Natl. Acad. Sci. USA*, 109(29):11641–11645, 2012.

[12] A. B. Boyd and J. P. Crutchfield. Maxwell demon dynamics: Deterministic chaos, the Szilard map, and the intelligence of thermodynamic systems. *Phys. Rev. Lett.*, 116:190601, 2016.

[13] G. N. Bochkov and Y. E. Kuzovlev. Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics: I. generalized fluctuation-dissipation theorem. *Physica A: Stat. Mech. App.*, 106(3):443–479, 1981.

[14] D. J. Evans and D. J. Searles. Equilibrium microstates which generate second law violating steady states. *Phys. Rev. E*, 50(2):1645–1648, 1994.

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

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

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

[18] R. Klages, W. Just, and C. Jarzynski, editors. *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond*. Wiley, New York, 2013.

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

[20] 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.

[21] A. Alemany, A. Mossa, I. Junier, and F. Ritort. Experimental free-energy measurements of kinetic molecular states using fluctuation theorems. *Nature Physis*, 8:688–694, 2012.

[22] B. Lambson, D. Carlton, and J. Bokor. Exploring the thermodynamic limits of computation in integrated systems: Magnetic memory, nanomagnetic logic, and the Landauer limit. *Phys. Rev. Lett.*, 107:010604, 2011.

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

[24] M. Madami, M. d'YAquino, G. Gubbiotti, S. Tacchi, C. Serpico, and G. Carlotti. Micromagnetic study of minimum-energy dissipation during Landauer erasure of either isolated or coupled nanomagnetic switches. *Phys. Rev. B*, 90:104405, 2014.

[25] Y. Jun, M. Gavrilov, and J. Bechhoefer. High-precision test of Landauer's principle. *Phys. Rev. Lett.*, 113:190601, 2014.

[26] A. Berut, A. Petrosyan, and S. Ciliberto. Information and thermodynamics: Experimental verification of Landauer's erasure principle. *J. Stat Mech: Theory and Experiment*, 2015(6):P06015, 2015.

[27] J. Hong, B. Lambson, S. Dhuey, and J. Bokor. Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. *Sci. Adv.*, 2:e1501492, 2016.

[28] J. V. Koski, A. Kutvonen, I. M. Khaymovich, T. Ala-Nissila, and J. P. Pekola. On-chip Maxwell's demon as an information-powered refrigerator. *Phys. Rev. Lett.*, 115:260602, 2015.

[29] É. Roldán, I. A. Martinez, J. M. R. Parrondo, and D. Petrov. Universal features in the energetics of symmetry breaking. *Nature Physics*, 10(6):457–461, 2014.

[30] M. Gavrilov, R. Chétrite, and J. Bechhoefer. Direct measurement of weakly nonequilibrium system entropy is consistent with Gibbs-Shannon form. *Proc. Natl. Acad. Sci.*, 114(42):11097–11102, 2017.

[31] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, second edition, 2006.

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

[33] A. B. Boyd, A. Patra, C. Jarzynski, and J. P. Crutchfield. Shortcuts to thermodynamic computing: The cost of fast and faithful erasure. arXiv:1812.11241.

[34] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18:023049, 2016.

[35] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks. Thermodynamics of prediction. *Phys. Rev. Lett.*, 109:120604, 2012.

[36] A. B. Boyd, D. Mandal, P. M. Riechers, and J. P. Crutchfield. Transient dissipation and structural costs of physical information transduction. *Phys. Rev. Lett.*, 118:220602, 2017.

[37] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017.

[38] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Thermodynamics of modularity: Structural costs beyond the Landauer bound. *Phys. Rev. X*, 8(3):031036, 2018.

[39] P. M. Riechers and J. P. Crutchfield. Fluctuations when driving between nonequilibrium steady states. *J. Stat. Phys.*, 168(4):873–918, 2017.

[40] C. Aghamohammdi and J. P. Crutchfield. Thermodynamics of random number generation. *Phys. Rev. E*, 95(6):062139, 2017.

[41] P. R. Zulkowski and M. R. DeWeese. Optimal control of overdamped systems. *Phys. Rev. E*, 92(5):032117, 2015.

[42] T. R. Gingrich, G. M. Rotskoff, G. E. Crooks, and P. L. Geissler. Near optimal protocols in complex nonequilibrium transformations. *Proc. Natl. Acad. Sci. U.S.A.*, 113(37):10263–10268, 2016.

[43] A. Patra and C. Jarzynski. Classical and quantum shortcuts to adiabaticity in a tilted piston. *J. Phys. Chem. B*, 121:3403–3411, 2017.

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

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

[46] S. Vaikuntanathan and C. Jarzynski. Dissipation and lag in irreversible processes. *EPL*, 87(6):60005, 2009.

[47] T. M. Hoang, R. Pan, J. Ahn, J. Bang, H. T. Quan, and T. Li. Experimental test of the differential fluctuation theorem and a generalized jarzynski equality for arbitrary initial states. *Phys. Rev. L*, 120(8):080602, 2018.

[48] 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.

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

[50] C. Jarzynski. Comparison of far-from-equilibrium work relations. *Comptes Rendus Physique*, 8(5-6):495–495, 2007.

[51] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz. Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, 483:187–190, 2012.

[52] O.-P. Saira, M. H. Matheny, R. Katti, W. Fon, G. Wimsatt, S. Han, J. P. Crutchfield, and M. L. Roukes. Nonequilibrium thermodynamics of erasure with superconducting flux logic. *Phys. Rev. Res.*, 2:013249, 2020.

[53] S. Han, J. Lapointe, and J. E. Lukens. Effect of a two-dimensional potential on the rate of thermally induced escape over the potential barrier. *Phys. Rev. B*, 46:6338, 1992.

[54] 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.

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

[56] 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.

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

[58] 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.

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

[60] 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.

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

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

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.unications*, 10(1):1–10, 2019.

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

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

[70] S. Deffner and C. Jarzynski. Information processing and the second law of thermodynamics: An inclusive, Hamiltonian approach. *Phys. Rev. X*, 3:041003, 2013.

[71] J. Li, J. M. Horowitz, T. R. Gingrich, and N. Fakhri. Quantifying dissipation using fluctuating currents. *Nature Comm.*, 10(1):1–9, 2019.

[72] S. Still. Thermodynamic cost and benefit of memory. *Phys. Rev. Let.*, 124(5):050601, 2020.

[73] M. Gopalkrishnan. A cost/speed/reliability tradeoff to erasing. In C. S. Calude and M. J. Dinneen, editors, *Unconventional Computation and Natural Computation*, volume 9252 of *Lecture Notes in Computer Science*, pages 192–201, Berlin, 2015. Springer-Verlag.

[74] S. Lahiri, J. Sohl-Dickstein, and S. Ganguli. A universal tradeoff between power, precision and speed in physical communication. *arXiv:1603.07758*.

[75] 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.

[76] A. Kolchinsky and D. H. Wolpert. Dependence of dissipation on the initial distribution over states. *J. Stat. Mech. Th. Exp.*, 2017(8):083202, 2017.

[77] P. M. Riechers. Transforming metastable memories: The nonequilibrium thermodynamics of computation. In D. Wolpert, C. Kempes, P. Stadler, and J. Grochow, editors, *The Energetics of Computing in Life and Machines*. SFI Press, 2019.

[78] S. Loomis and J. P. Crutchfield. Thermodynamically-efficient local computation and the inefficiency of quantum memory compression. *Phys. Rev. Res.*, 2(2):023039, 2019.

[79] Technology Working Group. The International Technology Roadmap for Semiconductors 2.0: 2015, Executive Summary. Technical report, Semiconductor Industry Association, 2015.

[80] Technology Working Group. The International Roadmap for Devices and Systems: 2020, Executive Summary. Technical report, Institute of Electrical and Electronics Engineers, 2020.

[81] Technology Working Group. The International Roadmap for Devices and Systems: 2020, More Moore. Technical report, Institute of Electrical and Electronics Engineers, 2020.

[82] Technology Working Group. The International Roadmap for Devices and Systems: 2020, Beyond CMOS. Technical report, Institute of Electrical and Electronics Engineers, 2020.

[83] J. Shalf. The future of computing beyond Moore's law. *Phil. Trans. Roy. Soc.*, 378:20190061, 2020.

[84] T. Sagawa. Thermodynamics of information processing in small systems. *Prog. Theo. Phys.*, 127(1):1–56, 2012.

[85] J. L. England. Dissipative adaptation in driven self-assembly. *Nature Nanotech.*, 10(11):919, 2015.

[86] T. Sagawa. Thermodynamic and logical reversibilities revisited. *J. Stat. Mech. Th. Exp.*, 2014(3):P03025, 2014.

[87] P.R. Zulkowski and M.R. DeWeese. Optimal finite-time erasure of a classical bit. *Physical Review E*, 89(5):052140, 2014.

[88] P. R. Zulkowski and M. R. DeWeese. Optimal protocols for driven quantum systems. 2014. arXiv:1506.03864.

[89] P. R. Zulkowski and M. R. DeWeese. Optimal control of overdamped systems. *Phys. Rev. E*, 92(3):032117, 2015.

[90] E. Aurell, K. Gawędzki, C. Mejía-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi. Refined second law of thermodynamics for fast random processes. *Journal of statistical physics*, 147(3):487–505, 2012.

[91] K. Proesmans, J. Ehrich, and J. Bechhoefer. Finite-time landauer principle. *Physical Review Letters*, 125(10):100602, 2020.

[92] K. Proesmans, J. Ehrich, and J. Bechhoefer. Optimal finite-time bit erasure under full control. *Phys. Rev. E*, 102(3):032105, 2020.

[93] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu. The energy–speed–accuracy trade-off in sensory adaptation. *Nature Physics*, 8(5):422, 2012.

[94] P. M. Riechers and J. P. Crutchfield. Fluctuations when driving between nonequilibrium steady states. *J. Stat. Physics*, 168(4):873–918, 2017.

[95] B. Schroeder, E. Pinheiro, and W.-D. Weber. Dram errors in the wild: A large-scale field study. *SIGMETRICS/Performance'09, Seattle, WA*, June:1–12, 2009.

[96] R. Dillenschneider and E. Lutz. Memory erasure in small systems. *Phys. Rev. Lett.*, 102:210601, May 2009.

[97] Y. Jun, M. Gavrilov, and J. Bechhoefer. High-precision test of Landauer's principle in a feedback trap. *Phys. Rev. Lett.*, 113:190601, Nov 2014.

[98] D. Reeb and M. M. Wolf. An improved landauer principle with finite-size corrections. *New Journal of Physics*, 16(10):103011, 2014.

[99] A. M. Timpanaro, J. P. Santos, and G. T. Landi. Landauer's principle at zero temperature. *Physical Review Letters*, 124(24):240601, 2020.

[100] H. J. D. Miller, G. Guarnieri, M. T. Mitchison, and J. Goold. Quantum fluctuations hinder finite-time information erasure near the landauer limit. *Physical Review Letters*, 125(16):160602, 2020.

[101] S. Sheng, T. Herpich, G. Diana, and M. Esposito. Thermodynamics of majority-logic decoding in information erasure. *Entropy*, 21(3):284, 2019.

[102] K. Proesmans and J. Bechhoefer. Erasing a majority-logic bit. *arXiv preprint arXiv:2010.15885*, 2020.

[103] R. Landauer. Private communication with J. P. Crutchfield, 1981.