Computing with Physical Systems

By

KYLE J. RAY
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

Daniel L. Cox

Michael L. Roukes

Committee in Charge

2023

Contents

Computing with Physical Systems

Abstract

The investigation of microscopic nonequilibrium thermodynamic systems is a wide a varied field of study. While hard to pin down with one particular frame of reference, the technological importance of understanding what happens when small systems are driven out of equilibrium is undeniable. While it is theoretically possible to do most control processes without driving a system very far from an equilibrium state, it generally takes an inordinate amount of time to do so. Computing itself is the process of preserving, transforming and translating a physical system's states though various nonequlibrium procedures in finite time; thus, the mechanical computers that we all rely on are, at a fundamental level, nanoscale nonequilibrium thermal systems. In the following, various properties of nonequilibrium systems are discussed with an eye towards useful operations in computing. In chapter 1 the issue is tackled from a historical perspective; we see that controlling information can have a cost even using only equilibrium considerations. Next, chapter 2 moves away from purely equilibrium considerations by considering the costs that come from operating in finite time. Chapter 3 reviews a suite of relatively recent equalities that extend arbitrarily far outside the regime of equilibrium, as well as introducing novel equalities and applications for these new results. Chapter 4 investigates the applicability and scope of the new results, and in doing so, reveals the connection between a class of highly nonequlibrium processes and the precision of currents within them. Finally, in chapter 5, this class of protocol is leveraged to design highly efficient devices that are capable of universal computation that operate on similar timescales of todays state of the art machines, but hold the promise of being 4 or 5 orders of magnitude more efficient energetically.

Acknowledgments

This endeavor would not have been possible without the unflagging support and enthusiasm of Jim Crutchfield, and the group that he has fostered at the Complexity Sciences Center. While working in Jim's group, I've had the pleasure of working very closely with Greg Wimsatt and Alec Boyd from day 1; the value of their ideas, insights, and creativity to my graduate work has been incalculable. I'd also like to express my gratitude to my friends and family, for their support over the past several years.

Last, but in no way least, to Willow: thank you.

Introduction

In his 1871 book *Theory of Heat*, Maxwell first formally introduced a seeming paradox: a "finite being" that could, in essence, capture individual thermal fluctuations to extract macroscopic amounts of work from a heat bath in violation of the Second Law [**1**]: rendering disordered heat energy into useful, ordered work. Over the following decades, many attempted resolutions addressed purely mechanical limitations imposed by how a given *Maxwell demon* (MD) acted on its observations to sort molecules [**2**].

Thomson makes this point quite explicitly in a lecture given before the Royal Institution in 1879, where he closes his abstract [**3**]:

> The conception of the 'sorting demon' is merely mechanical, and is of great value in purely physical science. It was not invented to help us to deal with questions regarding the influence of life and of mind on the motions of matter, questions essentially beyond the range of mere dynamics.

Thomson highlights two key distinctions made in early conceptions of the demon. First, the demon's intelligence serves primarily as a means to physically sort microscopic particles by their individual characteristics. Second, MD cannot shed light on the influence of "mind" on the behavior of matter. (This presumably addressed Maxwell's and others' repeated appeals to undefined notions such as "intelligent beings".)

Not until 1929, when Leo Szilard published his seminal work "On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings" [**4**], was a direct connection established between a thermodynamic cost and what Maxwell called "intelligence"—and what we now call "information".[1] In this, Szilard showed that both of Thomson's assertions could be relaxed. Notably, Szilard's constructions do not involve the direct manipulation of individual molecules, but always involve their observation (measurement) and control by an "unintelligent" protocol. The genius in this was to introduce a new, operational, and minimal definition of "mind" as storing information in physical states; thus, inextricably linking a demon with its physical instantiation. While Szilard acknowledged that the biological phenomena governing the working of a "finite being" were beyond the scope of physics, he delineated the minimal capabilities a mind needed to exhibit

[1] Notably, Szilard discussed his manuscript's development with Albert Einstein [**5**].

MD-like behavior and then created idealized machines with these abilities. Szilard's conclusion: if the Second Law is to hold, a physical memory's interaction with a thermodynamic system must entail entropy production.

This exchange marks the beginning of one of the most significant revolutions in science in technology of all time: the information age. Communication technology advanced and mechanical computers began to replace human ones, but these advances only exacerbated the confounding questions that come with the union of control theory and thermodynamic theory. Information theory was formalized to deal with the questions, but with the definitions of the information theoretic Shannon and Von Neumann entropies – the very idea that entropy always increases over time was thrown into question, as both are conserved globally. Over time, the central conceit of thermodynamics (the 'thermodynamic limit') has eroded. The smaller the system of interest is, the less well-founded it becomes to assume, for example, that one piece of it can act like an ideal thermal bath. Nanotechnology and the miniaturization of machines, however, necessitates the investigation of such systems. Even neglecting non-ideal bath treatments, questions as simple as 'what is equilibrium?' become nontrivial when the system in contact with the bath is, for example, a single molecule. In what way can we say that a single classical degree of freedom is in a Boltzmann distribution? Regardless of these nagging questions, the central goal remains the same: what are the consequences of controlling systems with inherent and uncontrollable noise? The case of equilibrium thermodynamics has been worked out, but nonequilibrium theory is quite another matter. The past few decades have seen leaps in theoretical and technological advances in nonequilibrium studies through fluctuation theorems, uncertainty relations, quantum information, and biomolecular processes but there is as of yet no unified picture. And, it seems entirely possible that no such unified picture exists because the scope of the field is so large. A field of study defined by what it is not is necessarily difficult to circumscribe. It brings to mind Stanislaw Ulam's (apocryphal?) comparison of nonlinear science to the study of 'non-elephant' animals.

Given predicted explosive growth in societal demands for information processing and that digital microelectronics is now approaching the physical limits of available architectures [**6**], exploring alternative computing paradigms is not only prudent but necessary. One alluring vision for the future involves hybrid devices, composed of a suite of computing modules—classical/quantum, digital/analog, deterministic/thermal—each with its own architecture and function that operate in concert. A hybrid architecture allows dynamically harnessing the processing node best suited for the task at hand. The underlying insight is that a computing device's physical substrate should match its desired processing function [**7**].

It can be shown that systems driven slightly out of equilibrium are more wasteful than those driven to stay in equilibrium, yet a system driven far enough out of equilibrium can regain a significant portion of this efficiency. The difference between these two processes, as will be discussed in the following chapters, is regime. Macroscale, microscale; underdamped, overdamped; as physicists we want one theory to connect these qualifiers that lie on opposite ends of their spectrums. This goal is worthy, and the results (if found) are elegant– but when it comes to controlling a system for a particular purpose the differences between the two extremes are not a problem to be solved, but indicative of two distinct and useful tools. An underdamped system can tolerate a nonequilibrium state for far longer than an overdamped system, and so will be useful when the preservation of such a state is favorable and will be pernicious otherwise. My research program at UC Davis has evolved towards the following approach when exploring nonequilibrium thermodynamic systems: Instead of thinking about what is possible given a set of constraints, think about what the goal is– and what type of dynamics match the goal effectively; I did not, however, begin there– so neither will we. We start by looking back the the inception of the idea of intelligent control of thermal systems, and the first attempts to rectify this intervention with the second law of thermodynamics. So, let us return to Szilard.

CHAPTER 1

Maxwell's Demon and Szilard's Engines

That Szilard's prescient analysis of measurement anticipated by two decades Claude Shannon's information theory has been often mentioned, with varying levels of credulity [**2**, **8**, **9**, **10**]. It is generally overlooked, however, that Szilard's 1929 work laid out *three different constructions* of thermodynamic machines. Taken together, they were his attempt to account more generally for how the flow of heat, work, and information (our modern word, not his[1]) drive each step of a thermodynamic process. In today's parlance we refer to these devices as *information engines* [**11**]. Since then, as history would have it, the descriptor "Szilard Engine" came to refer only to his first construction—the single-molecule engine. In light of recent experimental and theoretical developments allowing new treatments of information engines, it is pertinent to revisit Szilard's foundational work *en toto*. What additional insights can be gleaned from the other Szilard devices, if any? How do they compare to his first, oft-cited single-molecule engine?

Below, we retrace Szilard's steps in constructing his second device and investigate his reasoning using more contemporary ideas and techniques for analyzing deterministic chaotic systems, information flow, and the energetics of thermodynamic transformations. Once completed, we turn to his third construction: a thermodynamic analysis of the process of measurement itself. We show that the second one, though a markedly different implementation employing a population of distinct molecular species and semipermeable membranes, is informationally and thermodynamically equivalent to an ideal gas of the single-molecule engines. One concludes that (i) it reduces to a chaotic dynamical system—called the Szilard Map, a composite of three piecewise linear maps and associated thermodynamic transformations that implement measurement, control, and erasure; (ii) its transitory functioning as an engine that converts disorganized heat energy to work is governed by the Kolmogorov-Sinai entropy rate; (iii) the demon's minimum necessary "intelligence" for optimal functioning is given by the engine's statistical complexity; and (iv) its functioning saturates

[1] "Information" appears only once and, then, in a narrative sense.

thermodynamic bounds and so it is a minimal, optimal implementation. We show that Szilard's third construction is rather different and addresses the fundamental issue raised by the first two: the link between entropy production and the measurement task required to implement either of his engines. The analysis gives insight into designing and implementing novel nanoscale information engines by investigating the relationships between the demon's memory, the nature of the "working fluid", and the thermodynamic costs of erasure and measurement.

1.1. Demon Gas: Szilard's Second Engine

Consider an ensemble of *demon-particle* molecules contained in a long cylindrical tube in contact with a thermal reservoir at temperature T. See Fig. 1.1 (Top). Each demon-particle $i = 1, \ldots, N$ is defined by two variables: a particle-type variable $s_i \in \{\square, \bigcirc\}$ and a variable that relates to the demon's knowledge $y_i \in \{0, 1\}$ about the particle type. Demon i "knows" its molecule's type when y_i's value exactly correlates that of s_i. We refer to y_i as demon i's *memory.*

Particles spontaneously convert "monomolecularly"—Szilard's phrasing—from one type to the other at a given rate. This rate is chosen to maintain a particular desired equilibrium distribution $\rho_0(s)$ in which the probability of being one type is given by $\Pr(s_i = \square) = \delta$ and the other by $\Pr(s_i = \bigcirc) = 1 - \delta$. Total particle number N is conserved. This equilibrium distribution of types can be enforced by there being an energy difference $\Delta\epsilon$ between the particle types or, perhaps, by spin statistics—as in the case of ortho- and para-hydrogen [**12**]. Thus, it is not necessary that the particle-type energies differ significantly. We assume that the energies do differ for the sake of generality, but the masses do not for the sake of clarity. As such, we define the N-particle Hamiltonian:

$$H_0 = \epsilon_{\bigcirc} N_{\bigcirc} + \epsilon_{\square} N_{\square} + \sum_{i=1}^{N} \frac{p_i^2}{2m} , \tag{1.1}$$

where $\epsilon_{\bigcirc}$ and $\epsilon_{\square}$ are the particle-type energies ($\Delta\epsilon = \epsilon_{\bigcirc} - \epsilon_{\square} > 0$), the particle numbers $N_{\bigcirc}$ and $N_{\square}$ sum to the total N, m is the particle mass, and p_i the i^{th} particle's momentum.

The cylinder walls are impermeable to either type of particle. Inside the cylinder, there are four thin membranes set perpendicularly to the cylinder axis. Two are also impermeable and all of the molecules lie between them. Paralleling Szilard, we denote the impermeable membranes by $\bigcirc$ and

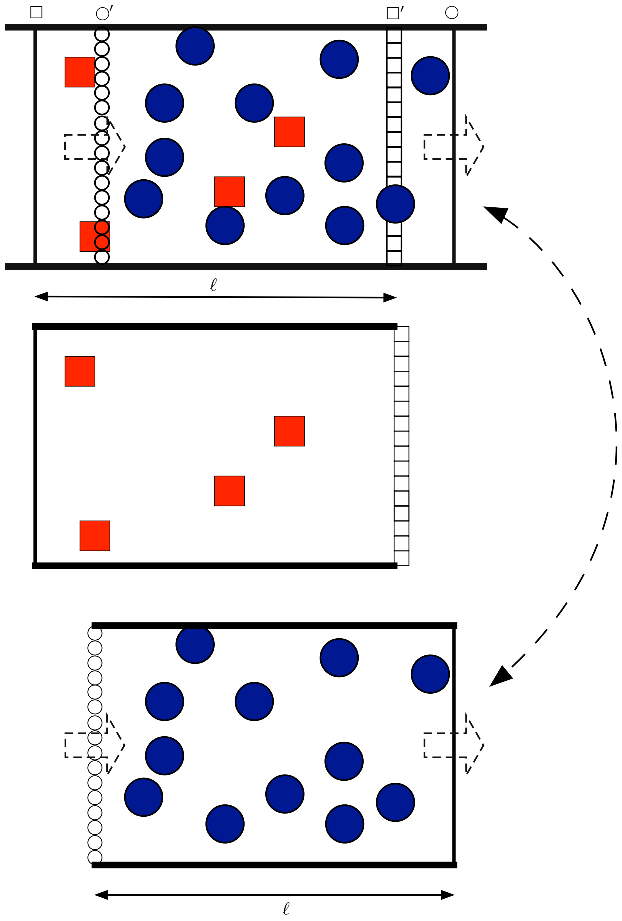


Figure 1.1. Second-engine components: (Top) Two movable overlapping compartments inside of the cylinder, each bounded by one of the two sets of membranes. The distance ℓ between ○ (□) and ○′ (□′) remains fixed as they move. (Bottom) Particle-type separation: Two volumes of constant length ℓ slide through each other, blue-circle particles ● are moved from the original left (L) volume to the right and red-square particles ■ are unaffected. A membrane permeable to ○s (□s) is depicted as a vertical line of squares (circles), as they are, in essence, walls for the □s (○s) only.

□. (Reusing type labels as membrane labels will become clear.) These membranes are initially set a distance ℓ apart. The other two membranes, denoted ○′ and □′, are permeable to only one of the two particle types, □ or ○ particle types, respectively. Each semipermeable membrane is initially set just inside of the impermeable membranes: ○′ being set next to and to the right of □ and □′ next to and to the left of ○.

The four membranes move along the cylinder axis, but are constrained to keep the distance between $\bigcirc$ and $\bigcirc'$ and between $\square$ and $\square'$ fixed at ℓ. In short, the system operates on two overlapping volumes of fixed length ℓ that slide relative to each other. Each volume has impermeable walls, an impermeable membrane ($\bigcirc$ or $\square$) at one end, and a semipermeable membrane ($\bigcirc'$ or $\square'$) at the other. Refer to Fig. 1.1 (Top).

Szilard's second construction is a protocol executed by translation and manipulation of these membranes. The protocol breaks down into three key transformations:

(1) *Measurement*: in which each particle's initial type s_i is stored in its memory y_i;

(2) *Control*: in which the system's thermodynamic resources are manipulated; and

(3) *Erasure*: in which the measurements are leveraged to return the overall system to its initial configuration.

These steps generally describe the behavior of information engines as they leverage information resources to gain thermodynamic advantage. Let's describe each of these in turn and in detail.

The first step of the protocol cycle is *measurement*. Initially, the ensemble's particle-type distribution is given by $\rho_0(s)$ and the distribution $f(y)$ of the memory variable y is uncorrelated to particle type: $\Pr(s, y) = \rho_0(s) f(y)$. We choose the parameter γ to represent the initial distribution over the memory state of the particles, so that $f(y)$ is initially distributed as $\Pr(y_i = 0) = \gamma$ and $\Pr(y_i = 1) = 1 - \gamma$. During measurement, the current type s_i of each particle is imparted to its memory y_i such that each type $\square$ ($\bigcirc$) particle has its y variable set to 0 (1). Here, the distribution $f(y)$ changes so that the conditional distribution $f(y_i|s_i)$ is deterministic or, equivalently, the joint distribution over s and y is given by nonzero elements $\Pr(y_i = 0, s_i = \square) = \delta$ and $\Pr(y_i = 1, s_i = \bigcirc) = 1 - \delta$. See Fig. 1.2, where particle type is depicted via shape and particle memory via color.[2] Szilard does not, at this point, give a physical mechanism that implements how each particle's memory y_i becomes correlated with its type s_i. However, this is addressed by his third engine—the subject of a later section below.

Next, the engine enters the *control* step. The volume bounded by $\bigcirc$ and $\bigcirc'$ slides to the right until the semipermeable membranes ($\bigcirc'$ and $\square'$) come into contact with each other. In doing so,

[2] Note that Szilard does not consider the energetic or entropic costs associated with manipulation of the memory state variable. He provides only a description of the correlations the memory must be able to create and sustain. Indeed, neglecting the cost of manipulating the memory is central to Szilard's point.

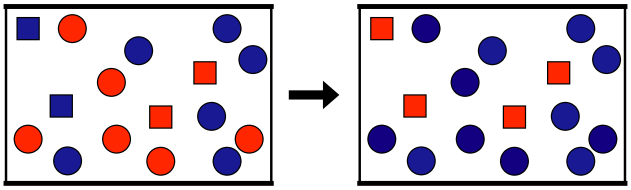


FIGURE 1.2. Measurement in Szilard's second engine: Particle type variable $s \in \{\square, \bigcirc\}$ is depicted by shape and memory variable $y \in \{0, 1\}$ by color. (Left) Initially-uncorrelated demon-particle states—particle type is not correlated with memory (shape is not correlated with color). (Right) Configuration of the gas after measurement. Tracking from the left diagram to the right, the measurement process establishes a correlation between color (y) and shape (s): $\square \to$ red and $\bigcirc \to$ blue. There are only ■s and ●s.

the semipermeable membranes separate the particles by type. This is done without any input of work or heat since, from the perspective of each particle, its container is merely being translated or held fixed; as demonstrated in Fig. 1.1 (Bottom). This transformation separates the particles into one of two compartments (L or R). See Fig. 1.3 (Top). Particles that are type $\square$ are all in the original volume (compartment L) bounded by the membranes $\square$ and $\square'$; those that are type $\bigcirc$ have been shifted to the right compartment (R) that is bounded by the membranes $\bigcirc$ and $\bigcirc'$.

Time scales are important here. Type separation must happen sufficiently slowly that the gas is always in equilibrium with respect to its compartment's spatial volume, but fast enough that no particle changes type during the process. This is not a generally prohibitive constraint, as we can assume the time-scale for a gas to fill its container uniformly is generally short. After the separation, each particle type exists independently in a container of the same size as the initial container.

Each compartment is no longer in equilibrium with respect to the type variable, though. Again, refer to Fig. 1.3 (Top). In principle, we can recover an equilibrium distribution with respect to H_0 within the individual containers either by waiting for the system to re-thermalize with the heat bath or by taking an active role by instituting a protocol involving the input or output of work. See Fig. 1.3 (Bottom). The latter is discussed in detail in Sec. 1.2 shortly.

At this point in the engine's operation, Szilard claims the "entropy has certainly increased". The entropy change ΔS from the initial macrostate to the macrostate in which the particles have re-achieved equilibrium can be found by the Sakur-Tetrode equation (detailed in App. 1.B), yielding:

$$\begin{aligned} \frac{\Delta S}{N} &= -k_{\mathrm{B}} \left(\delta \ln \delta + (1-\delta) \ln(1-\delta) \right) \\ &\equiv S(\delta) \ . \end{aligned} \tag{1.2}$$

The system's entropy has increased, as Szilard claimed. In the most efficient *control* scheme, we reach the equilibrium distribution reversibly and there must be a corresponding decrease $-S(\delta)$ in thermal reservoir entropy. Note that we cannot easily move the cylinders back into each other now, since there are particles of both types on each side of the semipermeable membranes.

Finally, the engine enters the *erasure* step of the protocol cycle. It erases by making clever use of its memory (y): the engine exchanges the type (shape, s) semipermeable membranes with memory (color, y) semipermeable membranes; see Fig. 1.4.

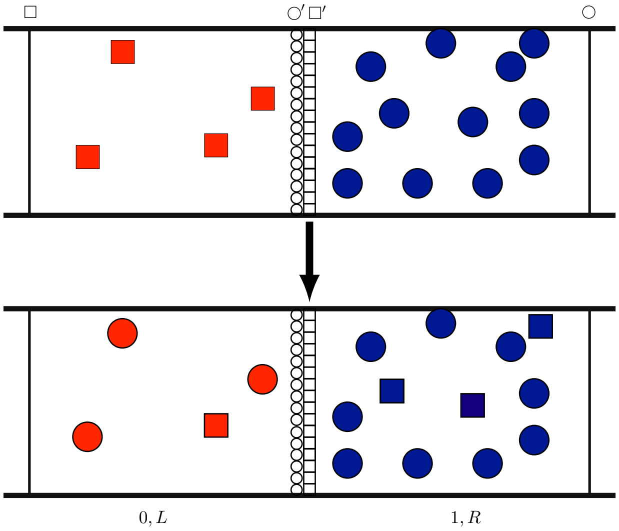


FIGURE 1.3. *Control* step particle-type equilibration: (Top) Deterministic distributions $\mathrm{Pr}_L(s = \square) = \mathrm{Pr}_L(y = 0) = 1$ and $\mathrm{Pr}_R(s = \bigcirc) = \mathrm{Pr}_L(y = 1) = 1$ at the end of sliding-separation of Fig. 1.1. (Bottom) Distributions after a period of particle-type conversion. Particles are no longer separated by shape type $s \in \{\square, \bigcirc\}$, but still are separated by memory state (color) $y \in \{0, 1\}$). That is, $\rho_L(s) = \rho_R(s) = \rho_0(s)$ but the memory state distribution in each compartment remains deterministic.

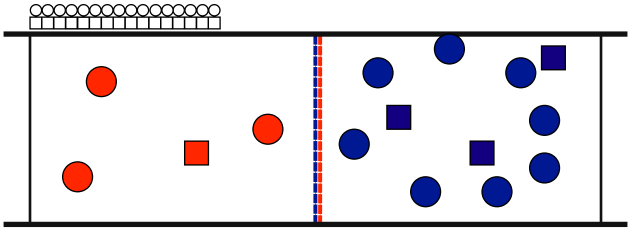


Figure 1.4. *Erasure*: First, replace the type (shape) semipermeable membranes with memory (color) semipermeable membranes.

The system is then ready to operate the reverse strategy of the particle-type separation of Fig. 1.1 to bring the particles back into the same volume (ℓ). Like the particle separation step, this action is effectively a translation, and can be accomplished work-free if the process unfolds on the proper time scale. See Fig. 1.5. Now that the particles are back within the original volume again, they are no longer separated by color or shape. Thus, the *erasure* step returns the system to its initial $\rho_0(s)$ macrostate,[3] without interacting with the heat bath. See Fig. 1.6.

The change in entropy for the system over the entire protocol cycle is, then, zero: the start macrostate is the final macrostate. The thermal reservoir, however, experienced a net decrease of entropy during the reversible control step. Here, Szilard appeals to the validity of the Second Law, stating that [**4**]:

> If we do not wish to admit that the Second Law has been violated, we must conclude that . . . the measurement of s by y, must be accompanied by a production of entropy.

That is, to resolve the apparent violation of the Second Law, Szilard associates measurement with a change in thermodynamic entropy and gives a functional form for this entropic cost in Eq. (1.2).[4]

The careful reader will notice several issues that require further investigation and refinement. First, Szilard does not specify an explicit mechanism for measurement, when the particle types s_i are stored in the memory variables y_i. Second, he does not determine the work required to drive the reversible *control* transformation he postulates. Third, one notes that the final distribution over the memory variable y, while not correlated with type variable s at the cycle's end, is necessarily

[3] This is not, strictly speaking, a full erasure without special tuning of δ and γ. However, it is an *apparent* macrostate reset as long as we ignore the energetics of the y dimension. Again, this "oversight" is crucial to Szilard's point.

[4] Here, Szilard anticipates Shannon's communication theory and its measure of information [**13**] by nearly two decades.

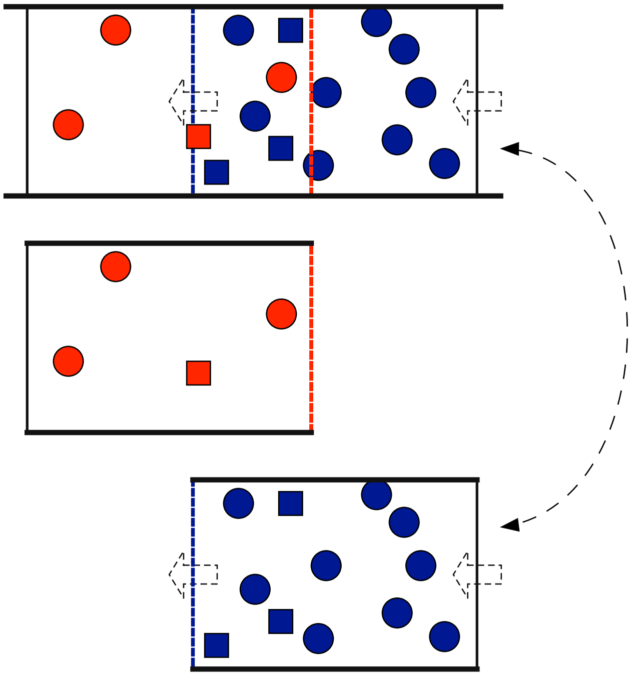


FIGURE 1.5. *Erasure*: Second, leveraging the memory variable y with the newly inserted memory (color) semipermeable membranes to reintegrate molecules, return to the initial macro-state.

distributed so that $N\delta$ particles are in the $y = 0$ memory state and $N(1-\delta)$ particles are in the $y = 1$ memory state; that is, unless we include an additional erasure step that resets y to some arbitrary initial distribution. In addressing these (and related) concerns we shall see that, while the selection of the initial distribution $f(y)$ over memory variables y_i is arbitrary, the choice impacts the thermodynamic costs of measurement and erasure. First, we investigate the bounds on the

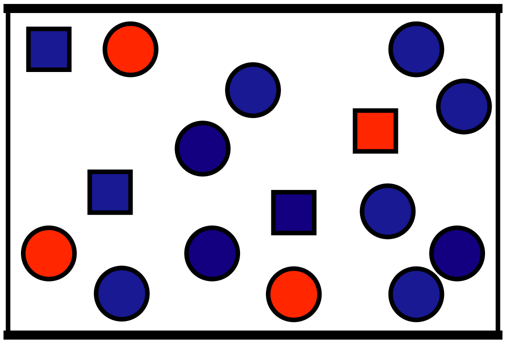


FIGURE 1.6. Reintegration with sliding the memory-state (color) semipermeable membranes recovers the original distribution over particle type in the initial container.

work required to perform Szilard's reversible control transformation. Then, we turn to analyze the information dynamics of the second engine as a thermodynamical system.

1.2. Engine Version 2.5

During the control step, each compartment begins in a nonequilibrium (completely deterministic) macrostate $\rho_L(s, y)$ (or $\rho_R(s, y)$) (Fig. 1.3 (Top)) and ends in the canonical equilibrium macrostate ρ_0 (Fig. 1.3 (Bottom)). To understand the effects of this transformation, we appeal to recent developments in information theory and stochastic thermodynamics [**14**, **15**, **16**] that allow us to connect the Gibbs *statistical entropy*:

$$\begin{aligned} S(\rho) &= -k_{\mathrm{B}} \sum_{s \in \{\bigcirc, \square\}} \rho(s) \ln \rho(s) \\ &= k_{\mathrm{B}} \langle - \ln \rho \rangle_\rho \end{aligned}$$

to the energetics of the isothermal equilibration process.

The two compartments (L and R) interact separately with the heat bath, so we take the following process to be executed independently within each compartment. As such, we drop the L and R subscripts for clarity and take the final extensive quantities to be of the form $S(\rho) \equiv S(\rho_L) + S(\rho_R)$. Moreover, since the memory state remains fixed for all particles within each compartment throughout the control step, the only relevant distribution is the marginal distribution—$\rho_L(s)$ or $\rho_R(s)$—over particle type.

Assuming perfect manipulation of the Hamiltonian at any point during the transformation allows us to design the most efficient protocol for the equilibration process. Consider the particle-cylinder system immediately after particle separation, in contact with a thermal bath at temperature T. Initially, the Hamiltonian is that given in Eq. (1.1). We break the process into two distinct steps. Both steps are executed in each compartment as follows.

First, we instantaneously shift the Hamiltonian from H_0 to $H_\rho = -k_{\mathrm{B}} T \ln \rho$. Tautologically, ρ is now the equilibrium distribution since $e^{-\beta H_\rho} = \rho$. Shifting the Hamiltonian requires a minimum amount $W_{\Delta H}$ of work given by the difference ΔH in the system's total energy under the two

Hamiltonians:

$$W_{\Delta H} = \langle H_\rho \rangle_\rho - \langle H_0 \rangle_\rho \ .$$

Next, we quasistatically shift the Hamiltonian back to H_0, which keeps the system in equilibrium by definition.

The transformation is now complete—the Hamiltonian returned to H_0 and the system's macrostate is given by ρ_0. Energy conservation in the second step implies that thermal reservoir and system energies change according to the work W_{qs} invested in the transformation:

$$W_{\mathrm{qs}} = \Delta U_{\mathrm{res}} + \Delta U_{\mathrm{sys}} \ .$$

Assuming the reservoir maintains constant volume, we write the W_{qs} in terms of initial and final free energies:

$$W_{\mathrm{qs}} = F(\rho_0) - F(\rho) \ .$$

(Appendix 1.C gives the details.) Then, the total work $W_{\mathrm{drive}} \equiv W_{\mathrm{qs}} + W_{\Delta E}$ to drive the two-step transformation is:

$$W_{\mathrm{drive}} = \langle H_0 \rangle_{\rho_0} - \langle H_0 \rangle_\rho + TS(\rho) - TS(\rho_0) \ .$$

Each term is readily interpreted in the present setting.

When considering the sum of both compartments—recall $\langle H_0 \rangle_\rho = \langle H_0 \rangle_{\rho_L} + \langle H_0 \rangle_{\rho_R}$—the energy expectation values for ρ and ρ_0 are the same. The average kinetic energy KE_{avg} will be the same since the whole system is thermalized to the same temperature, so we can neglect its contribution. For the initial nonequilibrium distribution ρ we have:

$$\langle H_0 \rangle_\rho = N \epsilon_{\bigcirc} \delta + N \epsilon_{\Box} (1 - \delta) \ .$$

And, under the ρ_0 distribution:

$$\begin{aligned}\langle H_0 \rangle_{\rho_0} =& N\delta(\epsilon_{\bigcirc}\delta + \epsilon_{\square}(1-\delta)) \\ &+ N(1-\delta)(\epsilon_{\bigcirc}\delta + \epsilon_{\square}(1-\delta)) ~.\end{aligned}$$

$\langle H_0 \rangle_{\rho_0}$ simplifies trivially to $\langle H_0 \rangle_{\rho}$. Together they make no contribution to W_{drive}. The $TS(\rho)$ term vanishes since the initial distribution of particle types within each compartment (L and R) is deterministic. The final term, the equilibrium state entropy, is $S(\rho_0) = NS(\delta)$. And so:

$$\begin{aligned}W_{drive} &= -TS(\rho_0) \\ &= -NTS(\delta) ~.\end{aligned}$$

It is now clear that the thermodynamic cycle is an engine. The work to drive the most efficient transformation that takes the engine from Fig. 1.3 (Top) to Fig. 1.3 (Bottom) is negative, signifying that there is an opportunity to extract work from the heat bath. Once again, we are faced with the reality that either measurement must involve compensating thermodynamic costs or admit that Szilard's second engine is a type of perpetual motion machine.

1.3. Demon Gas as a Thermodynamical System

To investigate the cost of measurement thermodynamically, we must choose a specific implementation of the device. We start with a 3-dimensional unit cube containing N particles and in contact with a heat bath at temperature T. The previous section established that the work extracted by Szilard's engine is independent of the energy difference $\Delta\epsilon$. We are, then, free to set this difference to zero—yielding a box of particles that are all identical according to H_0. The particles need not interact with each other to perform any of the necessary operations, so we take them to be noninteracting. Thus, our system is an ideal gas of N identical particles.

The membranes separating the particles into the L and R compartments slide along the box's x axis. We take all particles to start in the region $x < \ell$, with an ideal barrier inserted along $x = \ell$ to keep them from moving thermally into the region $x > \ell$. This defines two compartments L and R corresponding to $x_i < \ell$ and $x_i > \ell$, respectively.

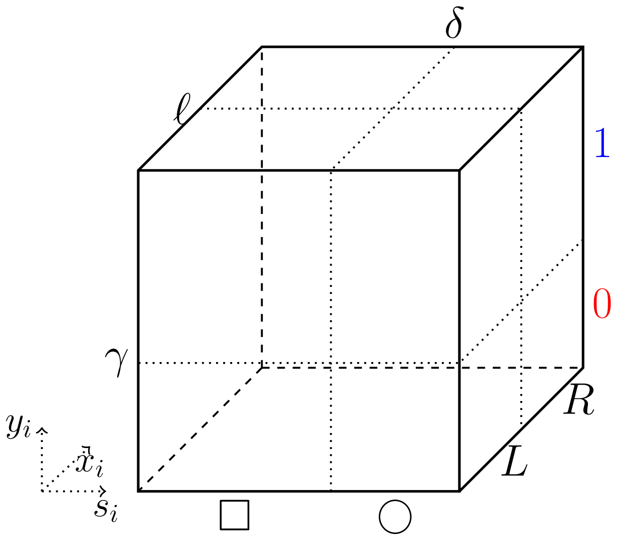


FIGURE 1.7. Markov partition of a demon-particle's state space—3D unit box. The i^{th} particle's position on the s axis corresponds to its particle type as $S_i = \square$ when $s_i < \delta$ (and $S_i = \bigcirc$ when $s_i > \delta$). The y-axis partition corresponds to the memory state as $Y_i = 0$ when $y_i < \gamma$ (and $Y_i = 1$ when $y_i > \gamma$). The depth dimension, parameterized by ℓ, corresponds similarly to particle position being in the left or right compartment. Note that ℓ must be equal to 1/2 for the $L \leftrightarrow R$ transition to always be work free, as Szilard noted.

We still need an operational definition of particle type which satisfies Szilard's requirements that there is a fixed particle-type equilibrium and that particles convert monomolecularly (N is constant) from one type to another. For our particle type, we use the position of a particle along the s dimension. If the coordinate of a particle is $s_i < \delta$ or $s_i > \delta$ we consider it to be particle type □ or ○, respectively. As the particles move about thermally, they cross back and forth across the line $s = \delta$, exactly modeling Szilard's monomolecular type conversion. Additionally, varying parameter δ sets the equilibrium distribution over particle type, relying on the gas' tendency to quickly fill its container uniformly.

This choice for particle type also allows us to define semipermeable particle-type membranes as ideally impermeable membranes that cover only the region associated with the relevant particle type. It is well known that the physical position of particles stores information [**17**]. And so, we choose a particle's memory states to be stored in its y coordinate with 0 (1) corresponding to $y_i < \gamma$ ($y_i > \gamma$). See Fig. 1.7 for the full symbolic partitioning of the demon-particle system.

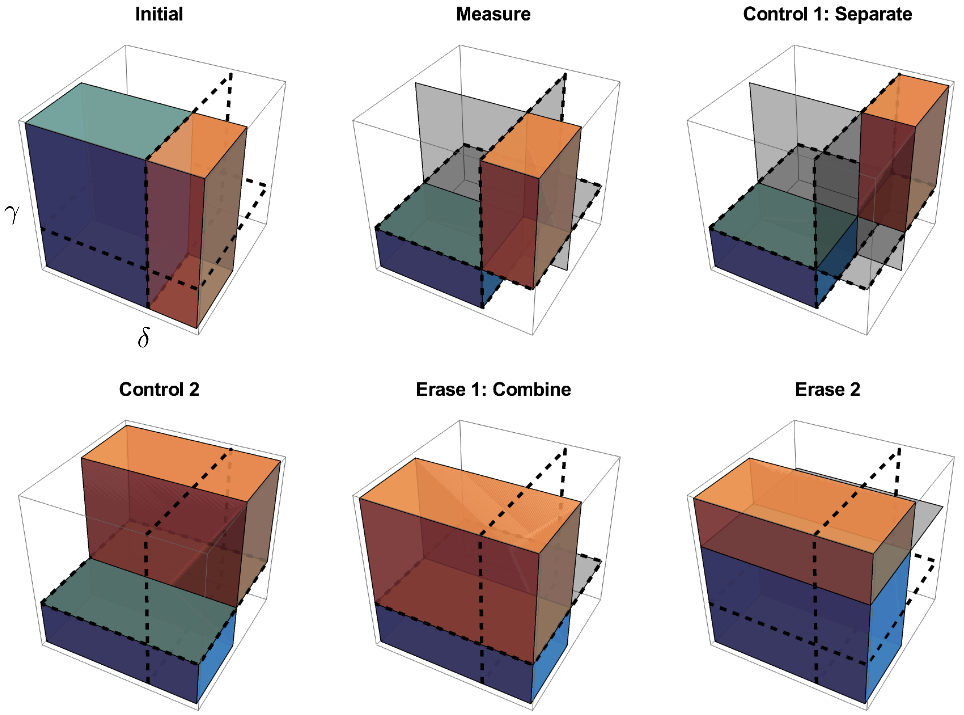


FIGURE 1.8. Second Szilard engine as a chaotic dynamical system: Its action on the unit-cube demon-particle-compartment state space decomposed into individual steps on an initially uniform distribution. Dashed outlines show the planes where $s = \delta$ and $y = \gamma$. The ideal barriers used to execute the protocol are depicted as low opacity gray planar partitions. Color is illustrative, to help track the separate manipulations of the particles that start in the regions $s < \delta$ and $s > \delta$.

This choice of memory state follows Ref. [**11**], and much of Sec. 1.1's analysis can be enhanced by comparing. As such, App. 1.A briefly summarizes the essential arguments from it and compares them to Sec. 1.1.

We are now ready to (i) analyze this engine's thermodynamics, (ii) set up the symbolic dynamics for the demon-particle gas, and (iii) analyze the engine's intrinsic computation.

1.3.1. Thermodynamics. This model allows us to easily probe the thermodynamics of each step in the Szilard Engine V. 2.5 operation, as just described in Sec. 1.1. Given this representation of Szilard's second engine, the overall thermodynamic cycle is the series of transformations shown in Fig. 1.8: *measure*, *control*, and *erase*. These operations are executed by inserting, sliding, and removing barriers.

The *measure* step, for example, involves three barriers. First, we insert a barrier along $s = \delta$. This is thermodynamically free, since the gas is identical on either side of the barrier. Next, we use a barrier perpendicular to the y axis that extends until $s < \delta$ to compress the particles that are in the □ partition to fit entirely within the 0 partition. Similarly, we use a barrier perpendicular to the y axis that covers $s > \delta$ to compress the particles that are in the ○ partition to fit entirely within the 1 partition. This establishes the necessary correlation between type and memory state: all particles are either ■ or ●.

For the *control* step, the first operation separates particles by type into either the L or R partition. This involves translating the ● particles to the R partition by inserting a barrier perpendicular to the x axis at $x = 0$ that covers from $s = \delta$ to $s = 1$. Then, along with the $s > \delta$ section of the initial barrier, this barrier translates the gas to the rear partition. This requires no interaction with the heat bath, since the volume of the ● gas remains constant. (See Control 1: Separate in Fig. 1.8.)

The second part of the control step expands along the particle-type dimension by allowing the two sections of the particle-type partition corresponding to $x < \ell$ and $x > \ell$ to slide independently of one another. The work W_{drive} the gas exerts on the barrier for an isothermal operation is calculated easily as $-\int P \mathrm{d}V$, with $P = N k_{\text{B}} T / V$:

$$
\begin{aligned}
W_{\text{drive}} &= -\int_{\ell\delta\gamma}^{\ell\gamma} \frac{N k_{\text{B}} T}{V} \mathrm{d}V - \int_{\ell(1-\delta)(1-\gamma)}^{\ell(1-\gamma)} \frac{N k_{\text{B}} T}{V} \mathrm{d}V \\
&= N k_{\text{B}} T \left(\ln \delta + \ln(1-\delta) \right) \\
&= -NTS(\delta) \ .
\end{aligned}
$$

This accords with the value calculated previously above. Thus, the model achieves the ideal efficiency bound. (See Control 2 in Fig. 1.8.)

We can also calculate the thermodynamic costs of the measurement and erasure transformations. In these, the gas' internal energy remains fixed and so $Q_{sys} = -W_{sys}$. To investigate the energy that is dissipated in the heat bath, we draw a relation between Q_{sys}, which is positive when heat flows into the system from the bath, and $Q_{diss} = -Q_{sys}$, which is positive when heat is being dissipated into the heat bath. For measurement we have:

$$\begin{aligned}Q_M = &-\int_{\ell\delta}^{\ell\delta\gamma}\frac{N\delta k_{\mathrm{B}}T}{V}\mathrm{d}V - \int_{\ell(1-\delta)}^{\ell(1-\delta)(1-\gamma)}\frac{N(1-\delta)k_{\mathrm{B}}T}{V}\mathrm{d}V\\ &= Nk_{\mathrm{B}}T\left(-\delta\ln\gamma-(1-\delta)\ln(1-\gamma)\right)\\ &= Nk_{\mathrm{B}}T\left(\delta\ln\frac{1-\gamma}{\gamma}-\ln(1-\gamma)\right)\ .\end{aligned}$$

Figure 1.9 (Top) displays a contour plot of the measurement heat Q_M as a function of the partition parameters γ and δ. We see that the measurement thermodynamics strongly depends on these parameters and that heat will always be dissipated during measurement. To implement an efficient engine, then, we would select a set of parameters that minimizes the heat dissipated in measurement.

Measurement is only part of the overall engine cycle, though. There is also the erasure transformation. The first erasure step in Fig. 1.8, which translates the particles back into the same $\{L, R\}$ partition, is similar to the first *control* operation. The current model makes it abundantly clear that this is not sufficient to return the gas to its initial state, though. The gas above and below the memory-state partition (inserted at the beginning of the *measurement* step) will not generally have the same pressure.

We require an additional step to return the gas to it's initial maximum-entropy state. Translating the boxes back to the L compartment does not require any thermodynamic input or output, so this final step is the source of the thermodynamic costs of erasure. The final step allows the gas to slide the partition that separates our memory states until the pressure on each side equalizes—until it rests at $y = \delta$. The barrier may then be removed at no cost or it may be left in the box and allowed to move freely along with the next cycles without affecting the thermodynamics. Calculating the energetic cost of this transformation is as simple as that preceding, yielding:

$$Q_E = Nk_{\mathrm{B}}T\left((1-\delta)\ln\frac{1-\gamma}{1-\delta}+\delta\ln\frac{\gamma}{\delta}\right)\ .$$

It is not surprising that the entropy cost of erasure vanishes when $\delta = \gamma$, since then the barrier at γ is already in the equal-pressure position before the final step. Figure 1.9 (Bottom) shows that erasure does not incur a cost: instead, the erasure provides yet another opportunity to extract energy from the heat bath. This is as expected, as the erasure process always increases the entropy of the system. However, examining $Q_M + Q_E$ we see that choosing the parameters to maximize the energy extraction in erasure increases the cost of measurement commensurately. Suggestively, the total thermodynamic cost of measurement and erasure is algebraically independent of the memory parameter γ:

$$\begin{aligned}\frac{Q_M + Q_E}{N k_{\mathrm{B}} T} &= \left((1-\delta) \ln \frac{1-\gamma}{1-\delta} + \delta \ln \frac{\gamma}{\delta} \right) \\ &\quad + \left(\delta \ln \frac{1-\gamma}{\gamma} - \ln(1-\gamma) \right) \\ &= -(1-\delta)\ln(1-\delta) - \delta \ln \delta\end{aligned}$$

Or:

$$\frac{Q_M + Q_E}{NT} = S(\delta) \ .$$

That is, the total combined cost of measurement and erasure depends only on δ, as in $NTS(\delta)$.

This is exactly the energy necessary to compensate for the work extracted from the heat bath during control. Since the choice of γ affects neither the total work extracted from the heat bath nor the total cost of the *measurement* and *erasure* processes together, one can set $\gamma = \delta$ so that erasure is cost neutral and all of the extracted work comes from the *control* process.

In this way, we need only consider the “cost” of measurement and the “revenue” from control. Of course, there is no net profit. Even in the most efficient system, the Second Law holds. This was one of Szilard’s main points—the point that resolved Maxwell’s paradox. By giving the demon (or control subsystem) a physical embodiment and properly accounting for its thermodynamics, there is no Maxwell demon paradox.

Interestingly, the distinction between measurement and erasure turns out to be, in a sense, arbitrary. We may increase or decrease the cost of one, but we do so at the expense of the other. This harkens back to Szilard’s original set-up, where he assigned entropy production to “the measurement” and then went on to demonstrate with a specific measurement apparatus that the erasure step increases the entropy. (Section 1.4 below discusses this apparatus in detail.) Szilard

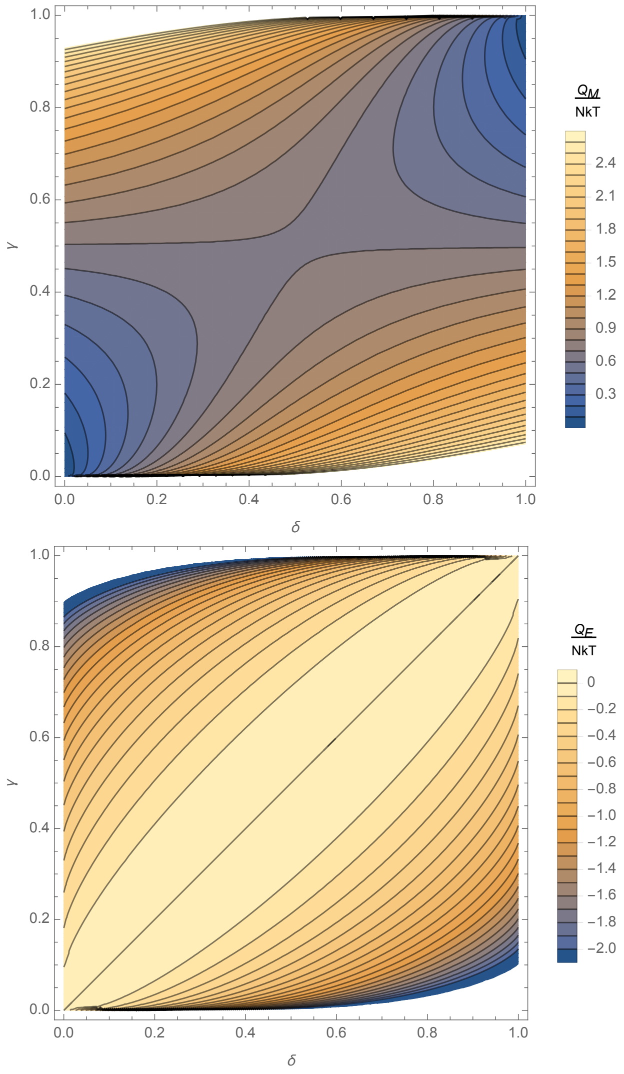


FIGURE 1.9. Measurement thermodynamic cost (heat dissipation) $Q_M/Nk_\text{B}T$ (Top) and erasure cost $Q_E/Nk_\text{B}T$ (Bottom) as a function of partition location parameters δ—particle-type—and γ—memory.

was not so much concerned about the details of whether erasure or measurement was the costly step. Instead, and presciently again, he points out that entropy production is associated with the entire process of establishing and destroying correlations between particle type and memory state as whole.

This contrasts with the view advocated by Landauer and Bennett half a century after—the logical irreversibility of erasure solely determines thermodynamic costs [**18**, **19**]. We now see, as others have recently emphasized [**11**,**20**,**21**,**22**,**23**], a more balanced view that there is a generalized principle bounding the total costs of measurement and erasure.

1.3.2. Computational Mechanics of Demon-Particle Gas Symbolic Dynamics. The coarse-graining of the microstate-space's unit box, depicted in Fig. 1.7, is a Markov partition [**24**] of the microstate dynamics under the macroscopic thermodynamic transformations that make up the Szilard engine. This immediately suggests defining a vector of binary-valued random variables $(S_i \in \{\square, \bigcirc\}, Y_i \in \{0, 1\}, X_i \in \{L, R\})$, sufficiently-long sequences of which accurately track of the engine's microscopic dynamical behavior.

At each protocol step a compound symbol SYX is generated according to the particle's location in the state-space box. For example, a particle that ends a protocol step and generates the compound symbol $\square 0L$ corresponds to a particle that is currently type $\square$, was particle type $\bigcirc$ when the most recent measurement was performed, and is in the Left compartment.

We must remind ourselves that the state space of this gas is large. There are N demon-particles each with three dimensions, so the full state space describes a $3N$-dimensional dynamical system. However, since the particles are noninteracting, Szilard's second engine is actually a collection (direct product) of N 3D particle state spaces. Applying computational mechanics' *predictive equivalence relation* collapses the $3N$-dimensional state-space to 3 dimensions of equivalent causal states [**25**]. Thus, we can use the symbolic dynamics of a single particle to find the engine's effective information processing behavior—and scale it to the demon-particle gas with a prefactor of N. (Appendix 1.E reviews the dimension reduction for an ideal gas.)

The problem simplifies even further since, having faithfully considered Szilard's initial problem statement, it is clear that the LR dimension of the state-space box is redundant in the current model. In Szilard's original construction, the LR dimension stops particles corresponding to the different memory states from intermixing during the control stage. However, the current engine already stores the memory and type information in positional coordinates, so the barrier used to compress the gas in the cycle's *measure* step already serves this purpose. Thus, we do not even need the full 3-dimensional state-space box to model the system's information and thermodynamic action. (Also see App. 1.E.)

Instead, we examine the action of Szilard's second engine on a 2-dimensional projection onto the sy plane of the box in Fig. 1.8. The resulting 2-dimensional map is nearly identical to the Szilard Map introduced in Ref. [**11**], constructed by considering Szilard's first (single-molecule) engine. The primary difference between the two being a difference in what is considered the "initial state". For more details, see App. 1.A.3.

We now track the probability density of a particle within the gas. Having abandoned tracking each particle's exact position within the box by using the coarse-graining into discrete symbols, we now consider the actions of a deterministic map on the probability density as a whole. Each step in the process depicted in Fig. 1.8 compresses or expands the probability density along a particular dimension. The composite map τ_{Szilard} that includes each step when $\delta = \gamma$ is given by:

$$\tau_{\text{Szilard}}(s, y) = \begin{cases} \left(\frac{s}{\delta}, y\delta\right) & s < \delta \\ \left(\frac{s-\delta}{1-\delta}, \delta + y(1-\delta)\right) & s > \delta \end{cases} .$$

Appendix 1.F gives the maps for each individual measure, control, and erase step.

Leaving in the memory-state partitions that are added each cycle, allows the map's action to build up the same self-similar interleaving within the particle's state-space probability distribution as seen in the Baker's Map [**26**]. While the probability density is not uniform throughout each component map step, we find that the distribution over the state space is uniform and constant for the composite map τ_{Szilard} that includes each step in the protocol.

We can again apply computational mechanics' predictive equivalence relation—now not to the gas' microscopic state space but to the symbolic dynamics induced by Markov partition of Fig. 1.7. This leads directly to an ϵ-transducer [**27**] that captures the information processing embedded in the engine's operation via causal states and their transitions. Appendix 1.A and 1.A.2 discusses these ϵ-transducers and compares them to Ref. [**11**]'s results. The agreement of the two analyses on various information-theoretic quantities establishes that Szilard's first and second engines are informationally and thermodynamically equivalent, though they arise from rather different implementations.

While the detailed analysis of the ϵ-transducers can be found in the Appendix, we close by simply noting that the computational-mechanics analysis gives physically- and informationally-interpretable results and identified functionally-relevant structure. This achieved the principle aims of: (i) determining the correct information generation (entropy) rate, (ii) demonstrating that this and the joint-machine information metrics align with those in Ref. [**11**], and (iii) identifying a signature of the choices made in constructing the map.

1.4. Szilard's Third Engine: Measurement as a Thermodynamic Cycle

The Szilard Map stores its memory state in an additional state-space dimension. Section 1.3 and Ref. [**11**] teased out the thermodynamic and information-processing consequences of this choice. The following introduces an alternative implementation of information storage—one introduced by Szilard himself, acknowledging that his first two engines did not completely specify physical measurement.

After concluding that the measurement process in his engines must generate entropy, Szilard introduces a bound on the entropy production from a binary measurement:

$$e^{-S_\square / k_\mathrm{B} T} + e^{-S_\bigcirc / k_\mathrm{B} T} \leq 1 ~,$$

where $S_\square$ and $S_\bigcirc$ are the entropies that a protocol generates when taking the measurement value $\square$ or $\bigcirc$, respectively. Investigating this limit further, he adopts a specific mechanical system that performs the minimal measurement tasks that his engines require. We now examine this implementation in detail, before returning to Szilard's thermodynamic bound on measurement.

The essential measurement tasks needed to implement either of Szilard's engines are as follows. First, establish a correlation between the instantaneous value of a fluctuating variable x and another variable y. Second, store that value in the "memory" of the second variable so that if x later changes, y remains fixed. Finally, return to a default state so that the system is ready to perform another measurement.

In this third construction of Szilard's, the variable to be measured is the position x of a pointer that moves back and forth according to a completely general protocol, either stochastic or deterministic. The variable y that stores the position is a function of the temperature of a body K that

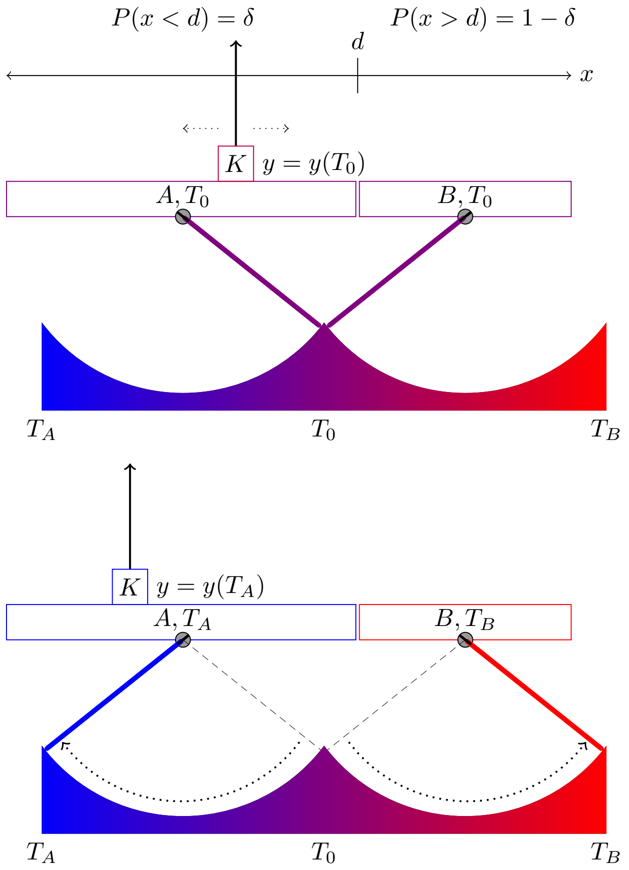


FIGURE 1.10. (Top) Default state before measurement: Variable x tracks the position of the pointer and y is a function $y(T)$ of the temperature T of body K. (Bottom) Measuring position of the pointer by temperature of K. The pointer location at the time of measurement determines if K is cooled to T_A or heated to T_B. Consequently, y is set to either $y(T_A)$ or $y(T_B)$.

is mechanically connected to the end of the pointer. As this pointer moves back and forth, it brings K in contact with one of two intermediate temperature reservoirs, A or B. These reservoirs are connected by movable heat-conducting rods to a continuum of temperature reservoirs that span from a cold temperature T_A to a hot temperature $T_B > T_A$. Initially, both rods are connected to an intermediate temperature T_0; see Fig. 1.10.

Coarse-graining the pointer position into two regions, A and B, the device can make a binary measurement. The measurement, which must happen over a timescale during which the pointer is stationary, involves moving the connecting rods through the continuum of heat reservoirs so that the intermediate reservoir A (B) is cooled (heated) to T_A (T_B). In this K either is heated or cooled depending on where the pointer was; see Fig. 1.10. This process can be done with arbitrarily small dissipation, if it is performed slowly enough that the rods, intermediate reservoirs, and K remain in thermal equilibrium at all times. This accomplishes a binary measurement: K is either at T_A or T_B, depending on the position of the pointer at the moment of measurement.

Next, the entire assembly of reservoirs is thermally isolated from the pointer and K so that, as the pointer continues to move, K maintains its temperature either at T_A or at T_B, even as the pointer leaves the interval it was in at the time of measurement. In this condition, the measurement value is stored in K's energy content.

Now, to be ready to make another measurement, the system must return to its initial state. If one knew with certainty K's temperature, the system could be returned to the default state without entropy cost: Simply wait until the pointer is in the region that corresponds to K's temperature, bring the system back into contact with the reservoirs, and institute the measurement protocol in reverse. This is, of course, actually two different protocols—and requires knowledge (measurement) of the result of each measurement to decide which to implement on each cycle.

There is no single protocol that can blindly return the system to its original state without producing entropy. Anticipating, by more than three decades, Landauer's well-known argument for resetting a particle in a bistable well [**19**], Szilard notes that an increase in entropy "cannot possibly be avoided" because [**4**]:

> After the measurement we do not know ... whether [K] had been in connection with T_A or T_B in the end. Therefore neither do we know whether we should use intermediate temperatures between T_A and T_0 or T_0 and T_B.

We create, then, a single protocol that returns the system to its original state—the "erasure" process—and measure its total entropy production. While the pointer is still uncoupled to the system, we return A and B to the equilibrium temperature T_0. Once again, this can be done reversibly on an appropriate timescale; see Fig. 1.11. Then, we bring K back into thermal equilibrium. This step cannot be done reversibly. This gives merit to the idea that erasure is the source of the entropic cost. Quantitative accounting for the entropy generation reveals additional insight.

All said, the body K undergoes a cyclic process—measurement itself is an engine.

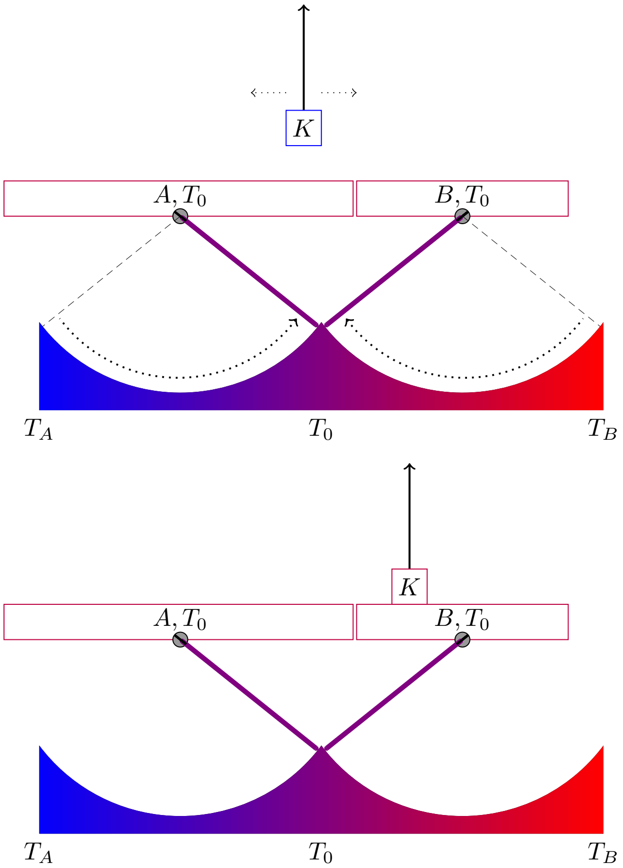


FIGURE 1.11. (Top) While K stores the location of the pointer at the time of measurement, A and B are returned to T_0. (Bottom) K is returned to T_0 by thermal contact with A or B, incurring an unavoidable entropic cost.

Quantitatively, over the measurement cycle the net change in system entropy is zero. Thus, we consider the entropy change only in the reservoirs. If the pointer was at a location that caused K to cool (heat) to T_A (T_B), then the reservoir's entropy increases (decreases) during the measurement period by $\int dQ/T$. Similarly, when K is returned to T_0 the reservoir's entropy decreases (increases) by $\Delta E/T_0$. We see that, while only the erasure process causes the entropy of the universe to increase, both the measurement and erasure processes play a role in increasing and decreasing the reservoir entropy. Szilard was unconcerned with keeping measurement and erasure as two different actions since he already concluded that it was possible for either to produce or consume the reservoirs' entropic resources. This is an insight that only recently received renewed attention

[**11**, **20**, **21**, **22**, **23**]. Furthermore, Szilard had also already concluded that the need to erase a binary random variable to a default state had unavoidable entropic costs.

To determine quantitatively the entropy gain from each measurement, Szilard adopts a 2-level system. The body K can be on one of two energy levels: a low energy state and a high energy state. Using standard canonical ensemble calculations, he shows that in the limiting process where the probability of the low (high) energy state at T_A (T_B) approaches unity the entropy generated by each process is:

$$S_A = -k_{\mathrm{B}} \ln p$$

$$S_B = -k_{\mathrm{B}} \ln q \ ,$$

where $p = p(T_0)$ and $q = q(T_0)$ are the probabilities that K is in the lower and upper energy state at temperature T_0, respectively. Szilard ended his analysis here. He does note that for this model:

$$e^{-S_A/k} + e^{-S_B/k} = 1$$

and that this represents the minimum amount of entropy generation necessary according to his bound:

$$e^{-S_A/k_{\mathrm{B}}T} + e^{-S_B/k_{\mathrm{B}}T} \leq 1 \ .$$

1.5. Szilard Measurement in Szilard Engines

To complete our analysis of Szilard's constructions, we couple the Szilard measurement device (SMD) of Sec. 1.4 to one of his engines. For a simple physical picture, we specialize to the more familiar single-particle Szilard engine [**4**, **11**, **18**], where a classical particle in a box is used to extract work from a temperature bath by inserting a partition and allowing the particle to move the partition. In essence, this engine leverages the measurement of a thermal fluctuation to do work. The considerations above show that the multi-particle second engine has the same thermodynamic and information processing behavior as the more-oft-quoted single-particle engine. And so, we lose nothing by specializing to his first, simpler model.

Now, take the SMD's pointer to be mechanically connected to the particle inside the Szilard engine, so that the position of the pointer tracks the particle's thermal motion. The SMD is calibrated so that the particle on the left- (right-) hand side of the partition corresponds to the pointer being at $x < \delta$ ($x > \delta$). The SMD is thermally isolated from the rest of the engine, as having its own set of reservoirs is crucial to its operation. In this example, the body K plays the demon's role. K changes length depending on its energy state, allowing K's state to select the engine's protocol; for example, by the position of a switch connected to K.

In this way, the entropy generated by one engine cycle is the sum of the entropy generated in the SMD's reservoirs and the particle-box system's reservoirs. The particle moves thermally through the entire box, so the probability that it falls in one or the other section of the box depends on the relative volume on either side of the inserted partition. During a cycle, the entropy generation in the system's reservoirs is proportional to either $\ln \delta$ or to $\ln(1-\delta)$ depending on if the particle starts on one or the other side of the partition, respectively. The mean entropy generated in the system reservoir over many cycles is then:

$$\begin{aligned} \langle \Delta S_{res_s} \rangle &\propto \delta \ln \delta + (1-\delta)\ln(1-\delta) \\ &= -H(\delta) \ . \end{aligned}$$

The mean entropy generation in the SMD's reservoirs over many cycles of the measurement process is:

$$\begin{aligned} \langle \Delta S_{res_m} \rangle &\propto \delta S_A + (1-\delta) S_B \\ &= -\delta \ln p - (1-\delta)\ln q \ . \end{aligned}$$

Adding the two contributions yields the average entropy generated in the universe per cycle:

$$\begin{aligned} \Delta S &\propto (1-\delta)\ln\frac{1-\delta}{1-p} + \delta \ln \frac{\delta}{p} \\ &= D_{KL}(\delta || \gamma) \ , \end{aligned}$$

where the relative information $D_{KL}(\cdot||\cdot)$ is positive for all values of $\delta \neq p$ and vanishes when $\delta = p$.

We see that, once again, there is no Maxwell demon paradox, since the total entropy generation is positive. In the case that $p = \delta$, the mean entropy produced in the SMD reservoirs during the measurement cycle is exactly enough to compensate the decrease of entropy in the system's reservoirs during work extraction. This encoding of the information yields the most efficient cycle. However, there is no physical requirement that $p = \delta$. The inner workings of body K need not match where the partition between the physical regions A and B lies.

It is tempting to directly compare p and the memory-state parameter γ discussed above, but there is a distinct difference between the two. Under the action of the transformations described in Sec. 1.3, there is inherent interaction between the parameters δ and γ that manifests itself in the density of the ideal gas that serves as the engine's "working fluid". When one part of the gas is compressed into a particular memory partition, the size of that partition determines the cost of the next step. If the partition is small, it is "more difficult" to squeeze in the same number of particles. Consequently, both δ and γ appear in the cost of both measurement and erasure. This coupling between the two dimensions becomes relevant when we take into account the total cost of measurement and erasure, finding that γ drops out of consideration. The result is an engine that is ideally efficient for every parameter setting.

When coupling the SMD to the first engine, the importance in the inherent interaction of δ and γ in the Szilard Map becomes even more apparent. Unlike γ, the engine is no longer ideally efficient for any choice of parameter p. Instead, we must choose the distribution of K's energy states at the equilibrium temperature to have the same distribution as the particle's position states. If not, the engine suffers additional dissipation from a mismatch of the system and the measurement device. (Reference [**28**] recently considered the energy costs of such mismatches.) Thus, the relationship between p and δ is qualitatively different than that between γ and δ.

Looking across the sweep of progress since his original results, we now see that Szilard's construction is a concrete example of Ashby's *Principle of Requisite Variety* [**29**]: the variety of actions available to a system controller must match the variety of perturbations it is able to compensate. Specifically, Szilard recognized that a minimal system controller for a binary measurement must have two states. (Yet again, Szilard predated the cybernetics era by several decades—a prescience also noted by Ref. [**2**].)

However, we also see a stricter requirement used to avoid unnecessary dissipation. The actual distribution over the controller's internal states must be the same as the system's. This also touches on the general arguments put forth in Ref. [**30**] that consider the efficiency of a thermodynamically-embedded *information ratchet* which interacts with an information reservoir to extract work. Finally, one sees a clear parallel between the information-theoretic concept of optimal encoding [**14**] in which minimizing the memory needed to store a particular message, the highest-probability events are given the shortest code-words. And so, it should not be surprising that it is optimal to match the controller to the system. However, it is gratifying to see such a clear and straightforward example—an example unfortunately ignored by Szilard's future colleagues.

The SMD model of measurement also provides a clear physical picture of adding memory to a Maxwell Demon engine. If we imagine that the SMD has two bodies K_1 and K_2 that store information, the single-particle Szilard engine can operate for two cycles without having to go through an erasure process. Instead of erasing the first body at the end of the first cycle, the SMD moves on to operate on K_2—leaving K_1 in whatever final state the first cycle determined. This avoids increasing the universe entropy while extracting work from the Szilard engine heat bath. This violation of the Second Law is only transient, though. To perform a third cycle, the SMD must erase K_1 or K_2 to store the next measurement. At the point immediately before erasure in each cycle, the joint system must pay the entropic cost for two fewer measurements than it has made.

It is easy to see how this construction generalizes to larger physical memories consisting of N memory "bits" $K_1, K_2, \ldots, K_N$. With more memory, the joint system of the Szilard engine and the SMD can continue to extract work from the engine's reservoirs for additional cycles before having to finally pay the cost of its first measurement. From that point forward, though, every new measurement must be associated with an erasure of a previous one. This restores the Second Law with respect to the erased measurement. Each measurement is eventually paid for and, as the number of cycles grows large, the transient leverage from having a large memory becomes less and less noticeable.

1.6. Conclusion

Since Szilard's day in 1929, the once-abstract conception of a molecular-scale "neat fingered and very observant" being [**1**] that interacts with heat and information reservoirs has only become more tenable, as modern computing emerged and micromanipulators were invented and then miniaturized through nanofabrication techniques. Thus, understanding the workings of information engines—microscopic machines interacting with such reservoirs—is now highly relevant, especially compared to the days when Maxwell first offered up the idea as a pedagogical absurdity. This is evinced by, if nothing else, a constant and increasing stream of recent efforts that take Szilard's original single-molecule engine as a jumping off point to investigate how measurement, information, thermodynamics, and energy interact with one another in support functional behaviors [**31**, **32**, **33**, **34**, **35**, **36**].

Szilard's early models grounded Maxwell's demon in physical embeddings. Since their introduction they provided the bedrock for much debate and occasional insights over the past century, largely through his first, brilliantly-simple single-particle engine. Here, we found that his second (multi-particle) engine, though more obtuse in construction, captures all of the same compelling consequences suggested by the first. Additionally, by setting the demon memory state to another positional coordinate, it maps exactly on to the first engine's operation as analyzed in Ref. [**11**]. In several important ways, though, his second engine is more physical and plausible. And so, the multiparticle-membrane engine is more robust to criticism arising from concerns about applying classical statistics to the behavior of the first engine's single particle. Thus, the second engine's relationship to its single-particle sibling supports the physicality of the limits on information costs as developed in Refs. [**16**, **18**, **19**, **30**, **37**].

Beyond the conceptual insights that arise from Szilard's various engines, we can even be somewhat literal-minded. Szilard's first engine has been the inspiration for a diverse set of models and experimental realizations [**38**, **39**, **40**, **41**, **42**]. Recent developments in nanofabrication suggest attempting to realize Szilard's multi-particle engine, as well. For example, the graphene membrane fabrication techniques discussed in Ref. [**43**] can provide mesoscale membranes with tunable pore size, pore density, and mechanical strength that are well suited to molecular gas separation. Recall, too, that only-moderately complex biomolecules can store information in their conformational states. Thus, with the right gas ensemble, it is possible these designable membranes are good candidates for the semipermeable membranes required by Szilard's second engine. When coupled with modern biomolecule synthesis and nanomechanical device design, a tantalizing engineering challenge to implement Szilard's second engine presents itself.

Szilard's engines are simple enough to be readily analyzed, as we showed, with all hitherto relevant thermodynamic and information calculations analytically solvable. We even have a measure of the required demon "intelligence" in the controller's memory—the statistical complexity. This thorough-going look at Szilard's original constructions gave reassuring results—results consistent with the fundamentals of both information theory and thermodynamics.

Appendix

1.A. Boyd and Crutchfield (2016)

The following summarizes components of Ref. [**11**] relevant to the main text's arguments and points out key comparisons between the two.

1.A.1. Thermodynamics. Reference [**11**] analyzes Szilard's famous single-molecule engine by choosing the demon's memory to be stored in an auxiliary positional dimension. From this perspective, the engine's operation is captured by a three-stage piecewise chaotic linear map of the unit square. Additionally, the thermodynamic cost of each stage is calculated by assuming the working fluid behaves as an ideal gas under isothermal expansion and compression.

Reference [**11**] calculated the costs of measurement and erasure, finding:

$$\begin{aligned} Q_M &= -k_\mathrm{B} T(1-\delta) \ln \frac{1-\gamma}{\gamma} \text{ and} \\ Q_E &= -k_\mathrm{B} T(1-\delta) \ln \frac{1-\gamma}{\gamma} + \frac{T}{k_\mathrm{B}} S(\delta) ~. \end{aligned}$$

As in Sec. 1.3.1's analysis, erasure and measurement appear as a conjugate pair and the sum total of the two is independent of γ and proportional to $S(\delta)$. Thus, the net thermodynamic cost of the second engine is nothing more than N particles worth of the single-molecule engine cost.

1.A.2. Transducers ϵ-transducers. We again apply computational mechanics' predictive equivalence relation—now not to the gas' microscopic state space but to the symbolic dynamics induced by Markov partition of Fig. 1.7. This leads directly to an ϵ-transducer [**27**] that captures the information processing embedded in the engine's operation via causal states and their transitions. Figures 1.A.1 and 1.A.2 show the transducer for each dimension—type and memory—separately and Fig. 1.A.3 shows the ϵ-machine for the joint process.

Composing Figs. 1.A.1 and 1.A.2 transducers with the period-3 input process—that specifies the measure-control-erase protocol—gives an ϵ-machine that generates the output process for particle type or for memory state. (In this case, this is trivially implemented by dropping the input symbols $\{M, C, E\}$ from the ϵ-transducer transitions.)

The two resulting ϵ-machines and that in Fig. 1.A.3 are *counifilar* [**44**]. Moreover, the processes are not *cryptic* and this greatly simplifies calculating various informational properties. For example, the entropy rate of the joint system's machine (Fig. 1.A.3) is $\frac{1}{3}\,\mathrm{H}(\delta)$ per step, consistent with the analytical result for the Baker's Map from Pesin's theorem. It is also immediately clear from Figs. 1.A.1 and 1.A.2 that the statistical complexities C_μ^x and C_μ^y are equal though, the calculations for C_μ are straightforward, nonetheless:

$$\begin{aligned} C_\mu^x &= C_\mu^y \\ &= -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\left(\delta\log_2\frac{\delta}{3} + (1-\delta)\log_2\frac{1-\delta}{3}\right)\right] \\ &= \log_2 3 + \frac{2}{3}\,\mathrm{H}(\delta)\ . \end{aligned}$$

This reflects two choices in the construction of the map: first, setting the memory and type parameters equal ($\gamma = \delta$); second, choosing a uniform distribution over the full state space as the initial and final state of the system. These choices symmetrize the stored information with respect to particle type and memory state.

Appendix 1.A discusses these ϵ-machines and compares them to Ref. [**11**]'s results. The agreement of the two analyses on various information-theoretic quantities establishes that Szilard's first and second engines are informationally and thermodynamically equivalent, though they arise from rather different implementations.

While more insights on the engine physics can be extracted in terms of the ϵ-machines and ϵ-transducers, we close by simply noting that the computational-mechanics analysis gave physically- and informationally-interpretable results and identified functionally-relevant structure. This achieved the principle aims of: (i) determining the correct information generation (entropy) rate, (ii) demonstrating that this and the joint-machine information metrics align with those in Ref. [**11**], and (iii)

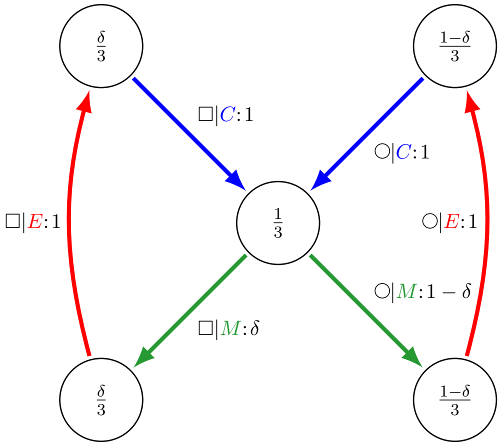


FIGURE 1.A.1. ϵ-Transducer for the particle type (x) subsystem. Protocol steps are designated by color: (control, measure, erase) $\Leftrightarrow$ (blue, green, red). Numbers inside states correspond to the asymptotic state probability. The transition notation $s|d : p$ corresponds to emitting the symbol s with probability p given the driving symbol d.

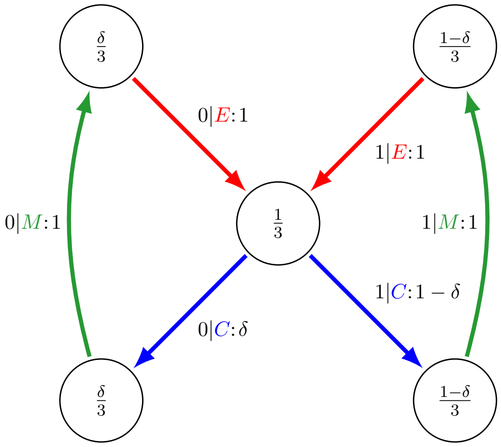


FIGURE 1.A.2. ϵ-Transducer for the memory state (y) subsystem. Notation as in previous figure.

identifying a signature of the choices made in constructing the map, which was reflected in the marginal machines' symmetrized C_μs.

1.A.3. Information and Intelligence. Reference [**11**] concludes (i) the single-molecule engine is represented by a chaotic map that converts disorganized heat energy to work at a rate governed by the Kolmogorov-Sinai entropy rate and (ii) the engine's minimum necessary "intelligence" for optimal functioning is given by the engine's statistical complexity. To investigate how

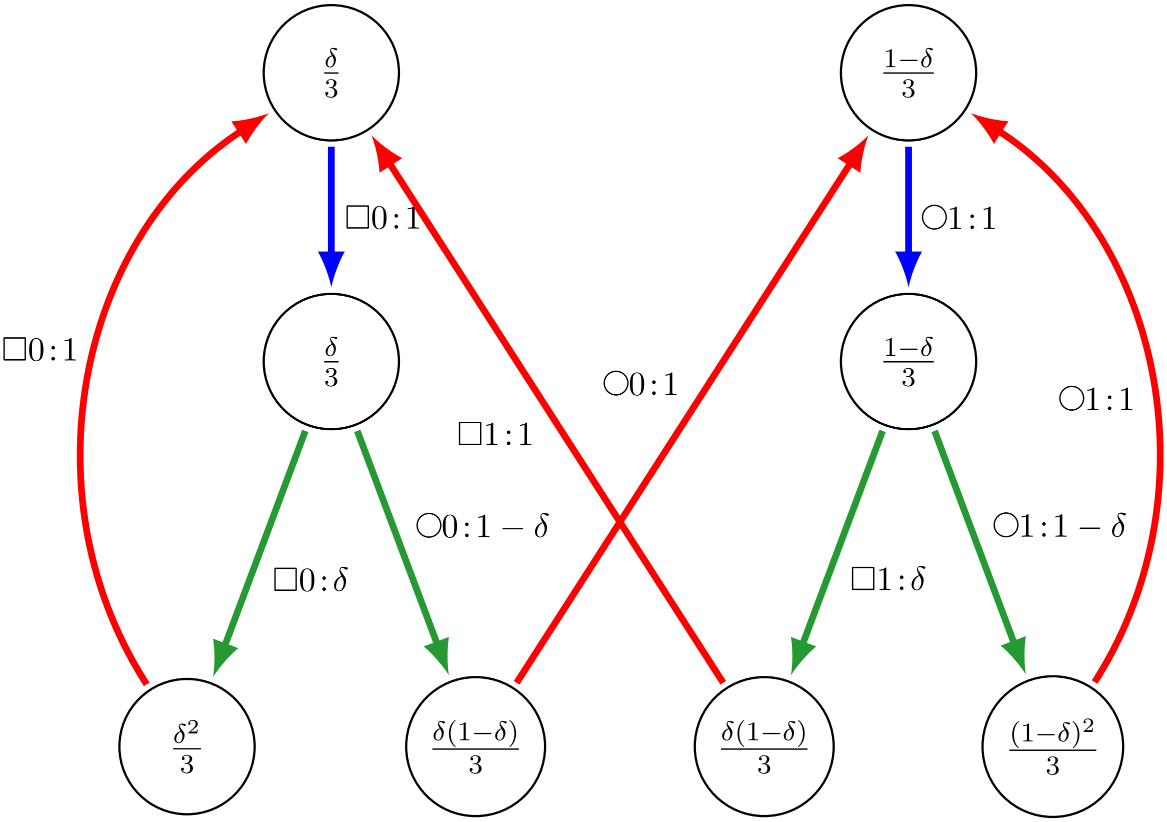


FIGURE 1.A.3. ϵ-Machine for the joint demon-particle system: Protocol steps designated by color as in Figs. 1.A.1 and 1.A.2. Driving symbols are suppressed in the transition notation for clarity—$s : p$ corresponds to emitting the symbol s with probability p.

these ideas play out in Szilard's second engine, we apply the predictive equivalence relation to find a minimal representation of the thermodynamic system introduced in Sec. 1.3.1, as well as to find the associated ϵ-machines that embody its information-processing; see Sec. 1.3.2).

The 2D Szilard Map introduced in Ref. [**11**] and that analyzed in Sec. 1.A.2 are both versions of the well-known Baker's Map. The primary difference between these maps arises from selecting different initial-state distributions. In Ref. [**11**], the system's initial state is chosen such that the working fluid is compressed into the region $y < \gamma$—rather than occupying the full range $y \in [0, 1]$ with uniform density. We could easily reconstruct the second Szilard engine to have the same initial state as the single-particle engine. For the purpose of illustration, though, we investigate the map under the new default memory state. At this point, though, one fully expects the results to agree with Ref. [**11**] on all quantities deriving from the thermodynamic cycle as a whole. Certainly, substage or marginal quantities may be shuffled around, but when examining the protocol cycle as a whole one anticipates agreement with Ref. [**11**].

Once having constructed the ϵ-machines (Figs. 1.A.2, 1.A.1, and 1.A.3), we retrace the steps in Ref. [**11**] to establish that Szilard's first and second engines are informationally and thermodynamically equivalent, though they arise from rather different implementations. First, we tackle the one place in which the machines in Sec. 1.A.2 and in Ref. [**11**] differ.

There is a slight variation from Ref. [**11**]'s analysis of the statistical complexity C_μ—the information in an ϵ-machine's causal-state distribution $\{\mathbf{S}\}$. Recall that:

$$C_\mu^x = C_\mu^y = \log_2 3 + \frac{2}{3}\,\mathrm{H}(\delta) \ .$$

In Ref. [**11**], it was not the case that the two marginal ϵ-machines had the same C_μ. Here, we see the different choice of initial state playing out: the choice symmetrizes the stored information with respect to particle type and memory state.

These results differ quantitatively from those in Ref. [**11**], but only due to an intentional change to what we consider the default distribution. This, however, is the extent of the differences. We can also calculate the join machine's statistical complexity, finding: $C_\mu^{\mathrm{joint}} = \frac{4}{3}\,\mathrm{H}(\delta) + \log_2 3$. If we consider the relationship between the second engine's three machines, we obtain $C_\mu^{\mathrm{joint}} = C_\mu^x + C_\mu^y - \log_2 3$. When analyzing the same relationship for the ϵ-machines in Ref. [**11**], we obtain the exact same expression.

So, we see that the first and second engines have the same information related to synchronization of their two respective subsystems—demon controller and particle "thermodynamic" subsystem. For the original single-molecule engine these subsystems were demon memory and molecule position; for Engine Version 2.5, they are shape and color subsystems.

We can explicitly check that all other informational measures for the ϵ-machines in Sec. 1.A.2 agree with those in Ref. [**11**]. In doing so, we recover the same values for the asymptotic communication rate (mutual information between type and memory processes):

$$\lim_{L \to 0} \frac{\mathrm{I}[X_{0:L}; Y_{0:L}]}{L} = \frac{1}{3}\,\mathrm{H}(\delta) \ ,$$

the correlation rate (also a mutual information, but between particle-type causal states and particle-memory causal states):

$$\lim_{L\to 0} \frac{\mathrm{I}[S^x_{0:L}; S^y_{0:L}]}{L} = \frac{1}{3}\,\mathrm{H}(\delta) ~,$$

and the interdependence of the correlation during the protocol steps (a conditional mutual information):

$$\mathrm{I}[X_0 : Y_0 | M] = \mathrm{H}(\delta) ~.$$

Note that the measurement step is where the single-symbol correlation is established. So, it stands to reason that the correlation dependence is localized there.

1.B. Entropy Change

Our goal is to determine ΔS in Szilard's second engine. The Sakur-Tetrode equation, the starting point, is:

$$S = N k_{\mathrm{B}} \ln \frac{V}{N} + \frac{3}{2} N k_{\mathrm{B}} \ln \frac{4\pi m U}{3h^2 N} + \frac{5}{2} N k_{\mathrm{B}} ~.$$

Terms that remain constant throughout an engine cycle can be neglected for our purposes. Cursory inspection reveals that ΔS will be determined by, at most:

$$N k_{\mathrm{B}} \ln \frac{V}{N} + \frac{3}{2} N k_{\mathrm{B}} \ln \frac{U}{N} ~.$$

In our case, the energy density term also drops out, since both the initial and final macrostates reach the equilibrium distribution $N_{\square} = \delta N$ and $N_{\bigcirc} = (1-\delta)N$. And so:

$$\frac{U}{N} = \delta\epsilon_A + (1-\delta)\epsilon_B + \mathrm{KE}_{\mathrm{avg}} ~,$$

for any number of particles, where $\mathrm{KE}_{\mathrm{avg}}$ is the average kinetic energy. Finally, since each container has the same volume, the volume term does not contribute. Calculating the resulting entropy change is straightforward, but requires attention. After some algebra, we have:

$$
\begin{aligned}
\frac{\Delta S}{k_\mathrm{B}} &= -\delta \ln N_{\square} - (1-\delta) \ln N_{\bigcirc} + \ln N \\
&= -\left(\delta \ln \delta + 1 - \delta \ln(1-\delta)\right) \\
&= S(\delta) \ .
\end{aligned}
$$

1.C. Free Energies

A similar need arises for obtaining the free energy in Szilard's second engine. We start observing that:

$$W_\mathrm{qs} = \Delta U_\mathrm{res} + \Delta U_\mathrm{sys} \ .$$

If reservoir volume remains constant, we note that $Q_\mathrm{res} = \Delta U_\mathrm{res}$. Then, using the expression above, we find $\Delta S_\mathrm{res} = Q_\mathrm{res}/T$ is given by:

$$T \Delta S_\mathrm{res} = W_\mathrm{qs} - \Delta U_\mathrm{sys} \ .$$

Thus, the total entropy change, including both the system and the reservoir must then be:

$$T \Delta S_\mathrm{res} + T \Delta S_\mathrm{sys} = W_\mathrm{qs} + T \Delta S_\mathrm{sys} - \Delta U_\mathrm{sys} \ .$$

The difference of $T \Delta S_\mathrm{sys} - \Delta U_\mathrm{sys}$ is nothing more than the change in the system's free energy $-\Delta F_\mathrm{sys}$. Following Szilard's statement, we choose a reversible process ($\Delta S_\mathrm{res} + \Delta S_\mathrm{sys} = 0$). This allows us to find the work to drive the quasistatic step:

$$W_\mathrm{qs} = F(\rho_0) - F(\rho) \ .$$

Adding the energy change $W_{\Delta H} = \langle H_\rho \rangle_\rho - \langle H_0 \rangle_\rho$ and the quasistatic work W_{qs} yields the the total driving work for both steps of the equilibration process:

$$\begin{aligned} W_{drive} &= \langle H_\rho \rangle_\rho - \langle H_0 \rangle_\rho + (\langle H_0 \rangle_{\rho_0} - TS(\rho_0)) \\ &\quad - (\langle H_\rho \rangle_\rho - TS(\rho)) \\ &= \langle H_0 \rangle_{\rho_0} - \langle H_0 \rangle_\rho + T\left(S(\rho) - S(\rho_0)\right) ~. \end{aligned}$$

1.D. An Erasure Alternative

Consider a different choice for the final erasure step in the Szilard Map.

We could simply remove the partition and allow the gas to spontaneously re-equilibrate. This gives the same relationship in terms of entropy, but we forego the advantage of a clear way to extract work that can be harnessed by the entropy increase.

Specifically, the change in entropy when mixing two identical gases at different densities depends only on that part of the entropy given by $\ln(V/N)$. Initially, there are two separate gases with the relevant entropy components:

$$S_A + S_B = Nk_{\mathrm{B}}\delta \ln \frac{\ell\gamma}{\delta N} + Nk_{\mathrm{B}}(1-\delta) \ln \frac{\ell(1-\gamma)}{(1-\delta)N} ~.$$

In the final state, we have a single gas:

$$S_F = Nk_{\mathrm{B}} \ln \frac{\ell}{N} ~.$$

The difference gives the entropy change:

$$\begin{aligned} \frac{\Delta S}{Nk_{\mathrm{B}}} &= \ln \frac{1}{N} - \delta \ln \frac{\gamma}{\delta N} - (1-\delta) \ln \frac{1-\gamma}{(1-\delta)N} \\ &= -\left((1-\delta) \ln \frac{1-\gamma}{1-\delta} + \delta \ln \frac{\gamma}{\delta} \right) ~. \end{aligned}$$

1.E. Ideal Gas Causal States and Szilard Engine Demon-Particle Gas

This section details the key steps showing the dynamics in the high-dimensional microstate space of the Szilard engine demon-particle gas reduces to the evolution of a distribution of demon-particles under a two-dimensional map of the unit square. One step argues that applying computational mechanics' predictive equivalence relation to a 1D ideal gas reduces its $2N$-dimensional microstate space to a space (of causal states) that describe a single-particle dynamics. That is, we need only track the evolution of a *distribution* of demon-particles under the single-particle dynamic. This happens since an ideal gas of N particles reduces to having microstates that evolve with no history; that is, the gas-particle trajectories individually evolve as independent, identically distributed random variables. Then we show how the predictive equivalence relation reduces the demon-particle 3D state space to a 2D state space.

Generally, the predictive equivalence $\sim_\epsilon$ relation reduces the high-dimensional microstate history $\chi_{-\infty:0}$ to only that information from the past needed to predict the microstate's future evolution $\chi_{0:\infty}$ [**25**]:

$$\chi_{-\infty:t} \sim_\epsilon \chi_{\infty:t'} \iff$$
$$\Pr(X_{0:}|X_{:0} = \chi_{-\infty:t}) = \Pr(X_{0:}|X_{:0} = \chi_{-\infty:t'}) \ ,$$

where $t \neq t'$. In brief, it groups histories that are equally-predictive of the future. Those groups are a process' *effective states*. Here, the semi-infinite past $\chi_{-\infty:t} = (\chi_{-\infty}, \chi_{-\infty+\tau}, \ldots, \chi_{t-\tau}, \chi_t)$ and, similarly, $\chi_{t:\infty}$ denotes the semi-infinite future. Note that time is discretized on a timescale τ.

Consider an ideal gas composed of N particles in $1+1$ dimensions, with time discretized on the scale of our Szilard map substages. The microstate trajectory $\chi_{-\infty:t}$ at time t holds *all* possible information about the system: it stores the position and velocity of each particle at each time up to time t: $\chi_{-\infty:t} = (x^1_{-\infty:t}, v^1_{-\infty:t}), (x^2_{-\infty:t}, v^2_{-\infty:t}), \ldots, (x^N_{-\infty:t}, v^N_{-\infty:t})$. The microstate χ is over burdened: It is not necessary to store the semi-infinite past if the goal is to predict future behavior. Instead, since the ideal-gas dynamics are Markov, we need only store the current state variables. Thus, the semi-infinite past is equivalent to the present under the predictive equivalence relation: $\chi_{-\infty:t} \sim_\epsilon \chi_t = \big((x^1, v^1), (x^2, v^2), \ldots, (x^N, v^N)\big)_t$. This is a familiar theoretical shorthand: identify states that have the same present with each other, regardless of their pasts.

Moreover, the gas demon-particles are also noninteracting, so the causal state can be expressed as a direct product: $\chi_t = (x^1, v^1)_t \times (x^2, v^2)_t \times \ldots \times (x^N, v^N)_t$.

This reduction is quite general. We can further reduce the dimensionality of our specific system, however, by applying the causal equivalence relation again. Since time is discretized on a scale τ that is much longer than the time scale on which the demon-particle equilibrates, we consider $(x^i, v^i)_{t+\tau}$ to be uncorrelated with $(x^i, v^i)_t$. This second application of the causal equivalence relation reduces our system to each particle having the same state: for any particle, $x_{t+\tau}$ is chosen by a uniform distribution over any position in the gas' container—regardless of its coordinates at time t. Thus, the causal state for each particle is only what container it is in at time t.

In the case of a single container, there is only one causal state σ—in which the future is decided by N flips of a "fair coin" over the gas' container; $\chi_t = \sigma_1 \times \sigma_2 \times \sigma_3 \times \ldots \sigma_N$.

Furthermore, the particles are indistinguishable, so there is also an equivalence between any states that are related by shuffling particle indices. There is no meaning to the labeling of the different particles, and they all are contained in the same causal state. Thus, the previous causal state reduces simply to N copies of the single particle dynamic: $\chi_t = N \cdot \sigma$.

The predictive equivalence relation also helps explain the dimensional reduction of our engine's positional coordinates under the Markov partitioning of Fig. 1.7's symbolic dynamics. Inspecting the transformations that make up the Szilard engine cycle (Fig. 1.8), we find two types of transformation. The first (measurement, for example) is of the form:

$$\begin{array}{c} (\square, 0, L) \\ (\square, 1, L) \end{array} \rightarrow (\square, 0, L)$$

and

$$\begin{array}{c} (\bigcirc, 0, L) \\ (\bigcirc, 1, L) \end{array} \rightarrow (\bigcirc, 1, L) \ .$$

This transformation includes state collapse, and so might prove to be thermodynamically relevant. However, the (L, R) dimension does not play a role here. The state $(\square, 0)$ is just as useful predictively as $(\square, 0, L)$. Truncating the third symbol, thus, gives a smaller (minimal) representation that is just as informative.

We also need to consider the other type of transformation, which is the first part of the control step:

$$(\square, 0, L) \to (\square, 0, L)$$

and

$$(\bigcirc, 1, L) \to (\bigcirc, 1, R) \ .$$

The (L, R) dimension does come into play here, but this step has no merging or expanding in state space—thus, it is thermodynamically mute. The (L, R) dimension is a bystander in the first example: present but not participating. In the second example, the (L, R) dimension does play a role—but it is the only thing changing and the transformation is a deterministic state translation. Each of the transformations in Fig. 1.8 falls into the two categories above, so the (L, R) dimension is not part of a minimal representation. It can be ignored and so the effective dimension of the symbolic dynamics is reduced further.

1.F. Szilard Engine Maps

There are three discrete-time maps of the unit-square state-space that correspond to *measurement*:

$$\tau_{\mathrm{M}}(s, y) = \begin{cases} (s, y\gamma) & s < \delta \\ (s, y(1-\gamma) + \gamma) & s > \delta \end{cases} ,$$

control:

$$\tau_{\mathrm{C}}(s, y) = \begin{cases} \left(\frac{s}{\delta}, y\right) & y < \gamma \\ \left(\frac{s-\delta}{1-\delta}, y\right) & y > \gamma \end{cases} ,$$

and *erasure*:

$$\tau_{\mathrm{E}}(s, y) = \begin{cases} \left(s, \frac{y}{\gamma}\delta\right) & y < \gamma \\ \left(s, \frac{(y-\gamma)(1-\delta)}{1-\gamma} + \delta\right) & y > \gamma \end{cases} ,$$

respectively.

Taken together, and specializing to the case where $\delta = \gamma$, we have the composite Szilard Map:

$$\begin{aligned}\tau_{\text{Szilard}}(s, y) &= \tau_{\text{E}} \circ \tau_{\text{C}} \circ \tau_{\text{M}} \\ &= \begin{cases} \left(\frac{s}{\delta}, y\delta\right) & s < \delta \\ \left(\frac{s-\delta}{1-\delta}, \delta + y(1-\delta)\right) & s > \delta \end{cases} .\end{aligned}$$

CHAPTER 2

Simulating Computers and Engines

While there is great didactic power in using an ideal gas undergoing quasi-static isothermal processes to rectify the second law in a Szilard engine, it is also profoundly unsatisfying. It leaves many questions (both practical and detailed) unanswered. First and foremost among these is what happens when we leave the idealized behavior behind: when our system has a finite size (less 'thermodynamic') and operates on a finite time scale (less quasi-static)? To answer this question, we turn to the perspective of thermodynamic computing [**45**], a framework in which degrees of freedom are treated as thermal particles evolving in a time dependent potential energy surface. In doing so, we also make an important conceptual shift from 'engines' to 'computations'. Instead of thinking about small physical systems as substrates for energy extraction, we can consider these systems as computational devices. The task becomes enacting a particular operation on the distribution over the system's states, and the energetic costs (heat, work) are byproducts of this goal; the work cost may be negative, and in this case we can consider the machine to be an engine. In this way, a larger range of devices are considered taking into account the full interplay of information, energy, and thermodynamic entropy.

2.1. What is a Computation?

In the most general sense, a computation can be thought of as any information processing operation: an operation that interfaces with Claude Shannon's notion of information. Taking Shannon's definition of information as the average value of the log probability over some distribution $H(\rho) = \langle -\ln \rho \rangle$. Here, information is a functional of a distribution, so a computation is a map between two distributions. We take a system from one state of knowledge to another through some dynamics. This is appealing due to its obvious breadth, but it is preferable to have a more specific and limited perspective, so we turn to modern ideas of digital computation. Consider, for example, a two state system $m \in \{0, 1\}$ defined by a distribution at time t' given by $\rho_{t'} =$

$(Pr(x=0|t=t'), Pr(x=1|t=t'))$. If we simply want to take this system from a distribution $\rho_0 = (.4, .6)$ to $\rho_\tau = (.2, .8)$ at time τ there are infinite ways to do so. One way is to take state 1 to itself, but takes state 0 to either 0 or 1 with even probability. Another method would be to map 0 to 1 with certainty and map 1 to 0 with probability 1/3. In a very real sense, we can consider these as different computations despite affecting the same change in Shannon entropy. In digital computing, the object of interest is not the overall distribution over inputs and output states but the conditional maps between input and output states. Thus, we consider a computation to be a conditional map $\mathbf{C} = Pr[m(\tau)|m(0)]$ over some set of states $\mathcal{M}$ from an initial memory state $m(0) \in \mathcal{M}$ to a final one $m(\tau) \in \mathcal{M}$. This map evolves the distribution of states according to the equation $\rho_\tau = \mathbf{C}\rho_0$.

On top of the mathematical definition above, we also consider that all computations must be physically embedded. Thus, when considering how to implement the map $\mathbf{C}$, limitations from the physical substrate are of incredible importance. In order to get a more focused picture, we again limit the scope of our interests. First, we will deal systems whose degrees of freedom are operating only in the classical regime. The restriction to the classical regime means that the two level system described above cannot be thought of as the microscopic degrees of freedom of some system. In the classical regime: the degrees of freedom are continuous– a microscopically two-level system is not possible. Instead, our memory states $\mathcal{M}$ represent a set of mesoscale states that are some coarse graining over the microscopic states $\mathcal{S}$. Second, we will look at systems that operate on energy scales that are comparable to the energy scales of the thermal environment. This choice is to allow us to consider the most efficient protocols possible, since information theoretic bounds on information processing are typically at or near the $k_\mathrm{B}T$ scale.

2.2. Memory and Metastability

In order for $\mathcal{M}$ to represent operationally useful states, the coarse graining needs to satisfy certain criteria. We first explain through an example by using the coarse graining used to define the informational states in the previous chapter (see section 1.3.2.) The 3D unit box contained impermeable membranes that could either allow or disallow transitions between the coarse grained states, so it was possible to define a set of discrete mesostates that faithfully captured the information processing in the protocol. In this case, the walls were ideal in that they were able to completely invalidate transitions between certain memory states. This means our coarse graining into informational states is useful on any timescale we choose. The downside is that such forbidden transitions require the energetic barrier between states to be infinitely large. Because we are looking to operate on $k_{\mathrm{B}}T$ scales, this needs to be relaxed. To do so, we need to maintain a concept of 'information lifetime'. This can be described approximately with a heuristic: the transitions between microstates that make up a particular $m \in \mathcal{M}$ must happen on a shorter time scale than the transitions that happen between microstates that belong to different m's. This transition rate between disparate m's is what defines the 'information lifetime', it tells us how long a microstate that starts in one memory state is likely to stay there. A coarse graining $\mathcal{M}$ is useful at timescales that lie between the two scales: long enough that there is good mixing of the microstates that lie with a memory state but short enough that there is negligible mixing between them.

Going back to at least Landauer's seminal paper on erasure [**19**], it was recognized that degrees of freedom that have multi-well potential energy surfaces were great candidates for information storing systems. Here, the energetic barriers that separate the wells needn't be infinite. Instead we can think of a barrier energy as setting the temperature and timescale at which the relevant wells serve as good memory mesostates. Assuming no stochastic thermal agitation, if a particle has energy less than the barrier height then it will stay within one well, and the well serves as a good memory state. For particles in contact with a thermal bath the energy of a particular particle changes over time, and is potentially subject to large fluctuations. Nevertheless, energy fluctuations from a thermal bath defined by inverse temperature β are exponentially damped in βE, so these energy fluctuations are rare provided that the energy barrier necessary to leave a well is larger than $k_{\mathrm{B}}T$.

Notationally, the N memory states along a particular dimension q will be written as $m_q \in \mathbb{N}_{0:N-1}$ with a joint memory state over both q_1 and q_2 given by $m_{q_1 q_2} \in \mathbb{N}_{0:N_1-1} \otimes \mathbb{N}_{0:N_2-1}$, different states along a particular dimension will simply be numbered so that a dimension broken into a binary(trinary) state space will have its states labeled $0, 1(0, 1, 2)$. This is a standard choice, and lines up well with contemporary digital computing where a continuous set of possible electrical currents are coarse grained into a low current 0 state and an high current 1 state. For a multi-well potential energy surface, it is the continuous set of positional degree of freedom that are coarse grained into the memory states. Provided that each positional dimension only stores a binary degree of freedom, it is simplest to coarse grain so that the information is stored in the sign of the position. This is Landauer's picture of the symmetric bi-stable well storing a 'bit' of information; a particle with a negative(positive) x value is likely to be near the bottom of the well located in the negative(positive) semi-plane. We will follow this sign coarse graining unless explicitly noted otherwise.

2.3. Langevin Dynamics

With a basic understanding of the neccesities, we begin by building a compelling and appropriate toy model. Assume a 'universe' governed by a hamiltonian $\mathcal{H}$ that obeys the deterministic equations of motion of classical physics. Our interest, however, lies in only a small subset of the entire universe– the computational system $H(x) = \frac{1}{2}m\dot{x}^2 + U(x,t)$– which is composed of some relatively small number of classical degrees of freedom that we are able to control by changing the potential energy over time. The dynamics of this partially observed system are no longer deterministic, but stochastic, due to unobserved differences in the initial state of the rest of the universe– which we call the environment. If the environment is a large weakly-coupled heat bath, with degrees of freedom that relax sufficiently quickly, the effective dynamics for our system of interest $\mathcal{S}$ are well modeled by Langevin dynamics:

$$dx = vdt \tag{2.1}$$

$$Mdv = -\gamma vdt - \partial_x U(x,t)dt + \sqrt{2\gamma k_{\mathrm{B}}T}\, r(t)\sqrt{dt}\ . \tag{2.2}$$

In these dynamics we can see the interplay between stochastic impulses from the bath in the third term, viscous damping in the first, and the deterministic control that we apply in the second. Simulating these equations of motion for different systems and protocols will serve as our primary way to investigate finite time nonequilibrium effects. Before going on to these simulations, we first cast the equation in terms of dimensionless equations of motion in order that they be suited for easy simulation. Our procedure for non-dimensionalization will include a redefinition of all dimensional variables (x, t, M, U, T) according to the prescription $x \to xx_c$ where the dimensional variable is now expressed as a demensional scaling x_c and a dimensionless number x. Applying this to the equations of motion above yields

$$dxx_c = vv_c dtt_c \tag{2.3}$$

$$dvv_c = -\frac{\gamma}{M_c} vv_c dtt_c - \partial_x \frac{1}{MM_c x_c} U_c U(x,t) dttc + \sqrt{\frac{2\gamma k_{\mathrm{B}} T T_c t_c}{M^2 M_c^2}}\, r(t)\sqrt{dt} \tag{2.4}$$

With a few 'common sense' definitions we can recover a Langevin dynamic for the dimensionless variables. First, we define our scales so that $v_c = \frac{x_c}{t_c}$. This recovers the relation $dx = vdt$ for the dimensionless variables. This leaves the second equation as

$$dv = -\left(\frac{\gamma t_c}{M_c}\right) vdt - \left(\frac{t_c^2 U_c}{MM_c x_c^2}\right) \partial_x U(x,t) dt + \left(\sqrt{\frac{\gamma k_{\mathrm{B}} T T_c t_c^3}{x_c^2 M^2 M_c^2}}\right) \sqrt{2} r(t) \sqrt{dt} \tag{2.5}$$

Direct inspection can show that each term in parenthesis is a dimensionless quantity in its own right. First we recognize that γ must have units of $\frac{\text{mass}}{\text{time}}$ from the original equations of motion. This makes the first term trivially dimensionless. The second term is clearly dimensionless because energies have the same units as $\text{mass} \cdot \text{velocity}^2$. The third term can be recognized as dimensionless by writing it as a product of two dimensionless terms:

$$\left(\sqrt{\frac{\gamma t_c}{MM_c}}\right) \left(\sqrt{\frac{k_{\mathrm{B}} T T_c t_c^2}{x_c^2 M M_c}}\right)$$

We can define a simulation parameter for each dimensionless prefactor in equation 2.5, yielding a simplified Langevin equation in terms of purely dimensionless variables and these parameters.

$$dx = vdt \tag{2.6}$$

$$dv = -\lambda v dt - \theta \partial_x U(x,t) dt + \eta \sqrt{2} r(t) \sqrt{dt}\ . \tag{2.7}$$

The definitions of the simulation parameters (λ, θ, η) can be inferred by comparison with equation 2.5. Of course, setting these parameters to particular values will represent certain relationships between the dimensional parameters– but interpreting these relationships will depend heavily on the system of interest because the problem is largely underdetermined at this level of abstraction. There are many different relationships between the parameters that will give you $\eta = 1$, for example.

2.4. The Szilard Engine as a time-dependent potential

Now that we are equipped with an appropriate framework and toolset, we can investigate the question at hand: what are the energetic consequences of implementing a Szilard engine in finite time, rather than the infinite time case considered in chapter 1. The process is straightforward. First, create a time dependent potential that will implement the conditional maps we want. Second, using the Langevin equation, simulate an ensemble of trajectories with the initial states sampled from the equilibrium distribution according to the potential energy landscape at $t = 0$. Lastly, analyze the results by looking at the ensemble of simulated trajectories.

The first order of business is to construct a time dependent potential that will implement the set of conditional maps that define the steps of the Szilard engine: measurement, control, and erasure. Recall that the engine is implemented in a 2D state space, with each dimension separated into two different memory states. Our Szilard implementing system must have 4 distinct states (potential energy minima) $m_{xy} \in (00, 01, 10, 11)$. This suggests a polynomial potential with at least 4th order terms in each dimension to allow for a bistable potential in each direction. Quadratic terms are also necessary, to provide an energy barrier between the wells. Recall that the Szilard engine begins with a uniform distribution over the state space. The analog here is a uniform distribution over the memory states $\rho_0 = (.25, .25, .25, .25)$ which can be achieved with an initial potential energy surface that is symmetric in x and y exchange (see the left panel of figure 2.4.1). To implement measurement, we want to implement the following map on the memory state space.

$$\mathbf{C}_{\text{measure}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{2.8}$$

This represents establishing a coupling between the x and y memory states. If the initial marginal x memory state was $m_x = 0(1)$, the full state will now be $m_{xy} = 00(11)$ with certainty. However, recognizing the stochasticity from the thermal environment, a perfect map is not assumed to be possible. Rather, we can consider that each zero in the map above is some small nonzero number quantifying an error rate. In a physically instantiated protocol, the first column of the matrix above will actually be $(1 - \epsilon_{00}, \epsilon_{00,01}, \epsilon_{00,10}, \epsilon_{00,11})$ where $\epsilon_{00} = \sum_{ij \neq 00} \epsilon_{00,ij}$ is defined such that the matrix is column stochastic and $\epsilon_{00,01}$ denotes the error rate of trials that started in state 00 and ended in state 01 instead of 00. The total error rate of the computation is given by $\epsilon = \rho_0 \cdot (\epsilon_{00}, \epsilon_{01}, \epsilon_{10}, \epsilon_{11})$. To avoid verbosity in the expression, these ϵ's will be suppressed in the transformation matrices.

While the operation is defined in terms of the coarse grained informational state space $\mathcal{M}$, the potential energy landscape and the Langevin dynamics operate on the underlying phase space variables of our system. In terms of these microstates, the operation can be implemented by removing the barrier between the wells that correspond to the same sign of x and then lowering the energy of the desired end state, depending on wether x is positive or negative (for a qualitative depiction, see the center panel of figure 2.4.1.) On some time-scale the particles will locally equilibrate to this new potential energy surface within their m_x memory state; hopping over the barrier that separates $m_x = 0$ and $m_x = 1$ will be surpressed by the energetic barrier provided that it does not take too long for the aforementioned local equilibration to occur. This can be thought of as two 'erasure' or 'reset' processes, where the reset state is different in the left and right semi-planes. The necessity of such a potential energy profile reveals the need for terms that are odd in both x and y, because the potential is not symmetric across the axes, and terms that couple the two variables together, because the need for the final m_y state to be conditioned on the m_x state.

The control stage of the engine, in which work is extracted from the system, is a re-equilibration along the x dimension. Raising the barriers between $m_y = 0$ and $m_y = 1$ again, we then lower the barriers that separate the wells that correspond to memory states with the same value for m_y, as depicted in the right panel of figure 2.4.1. The idealized map is

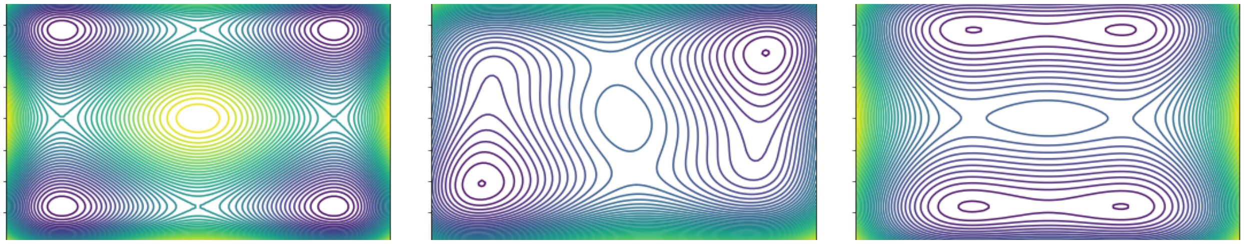


FIGURE 2.4.1. Qualitative potential energy landscapes that will implement the Szilard engine. Left is the initial potential energy profile, center is the potential energy profile during the middle of the measurement step, and right during the control step.

$$\mathbf{C}_{\text{control}} = \frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} . \tag{2.9}$$

Again, errors have been suppressed. A similar argument as above holds: because the barrier that separates m_y into 0 and 1 is still raised, the equilibration between 00 and 10 will happen on a much shorter timescale that equilibration that involves hopping over the energetic barrier. The final step is erasure, where we close the loop of the cycle. The potential is returned to its initial profile by once again raising a barrier between $m_x = 0$ and $m_x = 1$. Because the distribution over the four memory states is already approximately uniform at the end of the control step, the erasure step does not change the distribution and should be costless in the quasi-static limit. This is not surprising since the considerations in chapter 1.3.1 revealed that the cost of erasure vanished in the case that $\gamma = \delta$.

2.4.1. The explicit potential. Having laid out the overall qualitative shapes that the potential energy must take, an explicit form must be chosen. The simplest potential that can display the behavior would be of the form:

$$U(x, y) = Ax^4 + By^4 + Ex^2 + Fy^2 + Gxy + K \ .$$

Here, the capital letters represent time dependent parameters so the potential is implicitly time dependent. It is not evident merely from inspection how to change the parameters to yield the desired potential energy landscapes, so we turn to feature focused way of understanding the landscape.

Rather than the most general form, we look at the potential as a sum of pieces $U(x,y) = U^{\text{well}} + U^{\text{offset}} + U^{\text{well}}$. Each piece provides topological features in the potential energy surface that can be adjusted individually. The simplest is U^{well}, which represents four symmetric wells at $(x,y) = (\pm\ell, \pm\ell)$ of depth $s_w W$:

$$U^{\text{well}} = s_w W \left((x-\ell)^2(x+\ell)^2 + (y-\ell)^2(y+\ell)^2\right) \ . \tag{2.10}$$

Writing the well depth a dimensional scale s_w and non-dimensional W allows for the separation of the control protocol and the scale of the energetics involved, which is a useful abstraction because it allows the same time dependent signal ($W(t)$) to be used for multiple energetic scales. Importantly, the terms in U^{well} are the only 4^{th} order terms in $U(x,y)$; for coordinates that lie far from the four well region of state space it will always dominate.

This allows us to define the other two pieces of the potential energy to have a local impact without interrupting the global stability of the potential. The next term U^{offset} allows for localized offsets for each well. It is composed of functions of the form

$$g(L,x,y) = (L+x)(L+y) \ . \tag{2.11}$$

This function energetically biases the quadrant that contains the point $x = L$ and $y = L$ since the product will be maximized there while being minimized near any point for which $x = -L$ or $y = -L$ (which will happen near the minima of the wells in every other quadrant.) In full, the offset term is given by the following expression:

$$U^{\text{offset}} = s_L \left(D_{00} g(L,-x,-y) + D_{01} g(L,x,-y) + D_{10} g(L,-x,y) + D_{11} g(L,x,y)\right) \ . \tag{2.12}$$

The prefactors $D_{m_x m_y}$ determine the local offset of each well and the L parameter sets the intensity of how quickly the offset drops off over space as well as the location where the boost is maximized. The last component of the potential, and the most complicated are local barriers that can be used to control the probability of a particle jumping from one well to the other. The base unit for the barriers are functions of the form

$$h(L,x,y) = (L+x)(L+y)(L-y) \ . \tag{2.13}$$

Intuitively, this potential is a barrier between the wells located on either side of $y=0$ that both have a positive x coordinate. Again, L gives a length scale for the intensity of how quickly the barrier drops off over space. The full term for all four barriers is given by:

$$U^{\text{offset}} = s_B \left(D_{+,\pm} h(B,x,y) + D_{-,\pm} h(B,-x,y) + D_{\pm,+} h(B,y,x) + D_{\pm,-} h(B,-y,x) \right) \ . \tag{2.14}$$

Where the prefactors are named suggestively so that $D_{\pm,-}$, for example, parameterizes the height of the barrier between positive and negative x values for the wells that are both located at negative y values.

All said and done, the explicit form of this potential is very verbose– but having broken down the potential into operationally useful features, we are able to have an incredible amount of flexibility in control. In a general sense, this potential takes the form of

$$U(x,y) = Ax^4 + By^4 + Cx^2y + Dxy^2 + Ex^2 + Fy^2 + Gxy + Hx + Jy + K \ .$$

So the tradeoff of this readily programable potential is additional terms beyond the minimally required ones, specifically the terms that are linear in x and y and also the third order terms xy^2 and x^2y.

2.4.2. The Szilard Protocol. To implement the Szilard protocol outlined above, we will need to control 'knobs' that allow us to lower the energy level of the 00 and 11 wells (represented by the parameters D_{00} and D_{11}), and all four barriers (represented by the various $D_{\pm,\pm}$ terms). All other parameters are held fixed over time. The left panel of figure 2.4.1 shows the initial potential energy profile, which corresponds to setting $W = 0$, $D_{m_x m_y} = 0$, and $D_{\pm,\pm} = 1$. The protocol is implemented in six distinct steps:

(1) The barriers that separate wells that share an x coordinate are lowered by decreasing $D_{-,\pm}$ and $D_{+,\pm}$ from 1 to 0

(2) The 00 and 11 wells are made energetically favorable by lowering their energy values. This is accomplished by decreasing D_{00} and D_{11} from 0 to -1. The potential energy profile now resembles the middle panel of figure 2.4.1

(3) The barriers that separate wells that share an x coordinate are raised again by increasing $D_{-,\pm}$ and $D_{+,\pm}$ from 0 to 1

(4) The barriers that separate wells that share an y coordinate are lowered by decreasing $D_{\pm,-}$ and $D_{\pm,+}$ from 1 to 0

(5) The 00 and 11 wells are returned to their original energy values. This is accomplished by increasing D_{00} and D_{11} from -1 to 0. The potential energy profile now resembles the right panel of figure 2.4.1

(6) The barriers that separate wells that share an y coordinate are raised again by increasing $D_{\pm,-}$ and $D_{\pm,+}$ from 0 to 1. This returns the potential to its initial configuration, completing the cycle.

A graphical representation of the time dependent signal for each nontrivial parameter can be found in figure 2.4.2. It is worth noting that this implementation is, in no way, the unique solution. There are many different ways to implement the desired maps even assuming that we restrict our control knobs to the ones already described. For example, the order of steps 4 and 5 above could be swapped without ruining the qualitative integrity of the computation.

2.5. Finite time Cycle

Despite concerns of a lack of unique implementation, the behavior that we wish to showcase, the thermodynamic costs' behavior as we leave the quasistatic regime, is largely implementation agnostic. In infinite time, the process will take arbitrarily little work to accomplish (because the process is cyclic, and thus the free energy of the initial and final distribution is the same) and generate an arbitrarily small amount of entropy in the reservoir. As we diverge from this

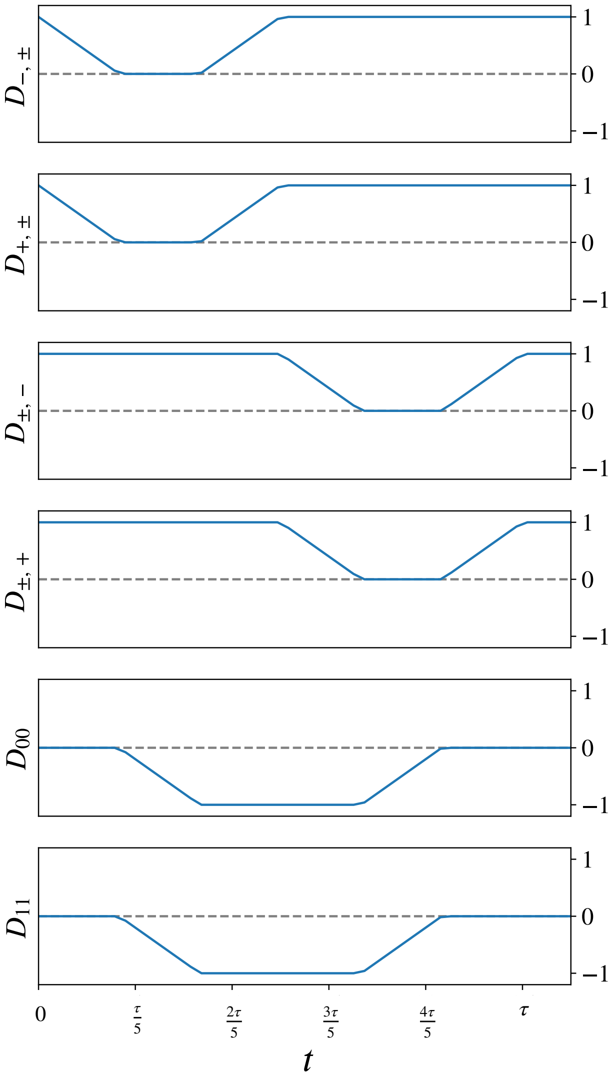


FIGURE 2.4.2. A graphical representation of the six steps of the Szilard protocol. This can be interpreted as time dependent control signals that are sent to the six tuning knobs we need to implement the protocol. Even from such a simple construction, it is easy to see that the protocol is not unique. The linear ramps could be sigmoids, or step functions, or even overshoot the target values. Additionally, the order of some of the steps– such as the first and second steps– can be changed or combined into a single step where two signals change at the same time. This is worth noting, because it hints at the high dimensional space that needs to be swept to find optimal implementations, but is not a pressing concern for the questions at hand. Animations of the protocol and a sample of simulated trajectories can be found at `https://kylejray.github.io/szilard/`

case and compute more quickly, the distribution is no longer always the equilibrium distribution associated with the instantaneous potential energy profile and thus heat begins to dissipate into the bath as the particles re-thermalize. General treatment of this nonequilibrium behavior is not trivial, though there are many results that give qualitative patterns within specific regimes ([**46**,**47**,**48**,**49**,**50**,**51**,**52**,**53**,**54**] to name only a tiny fraction). Armed with many such theoretical treatments of control protocols, simulation can serve as a testbed against which we judge the validity of these various results and see how their consequences play out in explicitly formed systems.

With the explicit dynamics, potential energy surface, and control signal given in the preceding sections, a suite of simulations can be carried out that show us what happens when we implement a cyclic protocol in finite time. Of primary importance is the net work cost of the protocol on average. The work cost represents the amount of energy exchanged with the work reservoir during a protocol when changing the energy level of the system's states, and can be calculated as a function of a particular system trajectory $\mathbf{x}(t) = (x(t), y(t))$ by

$$W(\mathbf{x}) = \int_0^\tau \partial_t U(\mathbf{x}(t), t) dt \ . \tag{2.15}$$

In the case of a discretized simulation the trajectories are defined over some finite set of times t_i, and the integral is well approximated by the following sum $W(\mathbf{x}) \approx \sum_i U(\mathbf{x}(t_i), t_{i+1}) - U(\mathbf{x}(t_i), t_i)$ provided a sufficiently fine time resolution. By sampling a large ensemble of N trajectories from the equilibrium distribution associated with $U(\mathbf{x}, 0)$ we can then estimate the average work necessary to implement the operation with a straight average$\langle W \rangle = \frac{1}{N} \sum_{j=1}^{N} W_j$.

Figure 2.5.1 shows the results of averaging over ensembles of 10,000 simulated trajectories. In each ensemble, the particles were exposed to the potential and time dependent control signal described above– twelve different values for the dimensionless duration τ were chosen from $\tau = 1$ to $\tau = 200$. A common result for the work cost of thermodynamic control protocols implemented in a finite time τ is what the work cost is expected to grow as $\langle W \rangle \sim \frac{1}{\tau}$ [**46**,**47**,**48**,**49**,**50**,**51**]. This result is borne out in the simulations, but there are also hints that the $\frac{1}{\tau}$ is only valid in certain regimes.

Looking from the longest to the shortest values of τ, three regimes can be identified. The first regime, indicated by a blue highlight, is what we might think of as the quasi-static regime. Here, the instantaneous distribution over the system states closely follows the instantaneous equilibrium distribution. In this regime, the $\frac{1}{\tau}$ scaling of the work is very reliable and τ can be considered large– which means that the work cost is not very sensitive to changes in the computation timescale and is very close to the infinite-time equilibrium result of $\langle W \rangle = 0$.

The next regime, which we will refer to as the 'barely-static' regime, is indicated by an orange highlight. Here, the $\frac{1}{\tau}$ is still valid– but τ is no longer large. The work cost becomes sensitive to changes in timescale; while the fidelity of the computation remains high, it costs more and more work to maintain. In this regime, the instantaneous distribution of the system states begin to depart markedly from the equilibrium distribution. It is somewhat remarkable that even as the distribution departs from the intuitive distribution used to create the protocol the $\frac{1}{\tau}$ scaling is carried over from the linear-reponse regime.

This cannot continue forever though, and a third regime emerges for short timescale protocols. In this case, the distribution becomes very different from the equilibrium distribution and a $\frac{1}{\tau}$ scaling of the work cost ceases to accurately describe the behavior. This regime is indicated by a red highlight, and is characterized by a significant decrease in the fidelity of the computation. The protocol is attempting to operate faster than the system can adequately respond to it, and so the protocol begins to have an effect that departs from the desired one. As the timescale becomes very short, the work cost drops– but so does the fidelity. Thus, while it is costing very little to do the protocol– this drop in cost merely means that the computation is failing. It costs very little, but also accomplished very little.

This glimpse into the world of nonequilibrium protocols only brings up further questions. For example: what mechanism explains the increased work values? what quantities determine the timescale over which the linear response approximations break down? To answer these more detailed questions, we need to appeal to more detailed simulations. Figure 2.5.2 shows the results of more detailed simulations in which 100,000 trajectories were simulated at four different values of τ. More trajectories allows us to look at the distribution of work values rather than rely on the average alone.

In the quasistatic case, where $\tau = 400$, we gain very little from looking at the distribution rather than the average. The average value is displaced somewhat from 0, the distribution over different work values is narrow, unimodal, and symmetric. As we proceed from the top to the

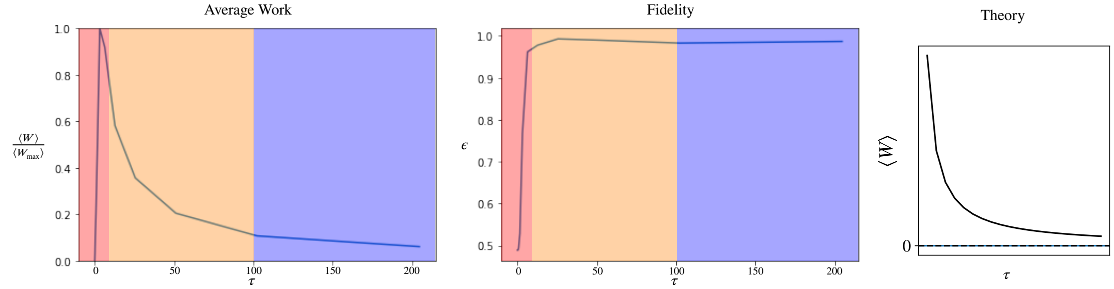


FIGURE 2.5.1. ((left, center) The results of implementing the Szilard protocol at twelve different timescales. Three regimes can be identified: the quasistatic regime (blue), the barely-static regime (orange), and the failure regime (red). (right) A qualitative curve showing the idealized $\frac{1}{\tau}$ behavior of the work cost's dependence on the timescale of the computation according to a wealth of different analyses (see references in text).

bottom of figure 2.5.2, the distributions are no longer well approximated by their average values alone. While the average work costs will display $\frac{1}{\tau}$ scaling regardless of the protocol implementation, the distributions over work outcomes and the timescale over which these distributions morph will depend on the specifics of the system. This is one reason that simulations of ensemble trajectories can yield powerful and unique insights about the nonequilibrium properties of thermodynamic control protocols. Armed with data for all the trajectories, we can even preform an even more detailed analysis that looks at only pieces of the full trajectory ensemble. This can be used to gain some insight about the mechanism by which the work distributions are deformed by nonequilibrium effects.

Consider that the difference between the first half of the protocol and the second half is that the first half mostly lowers the energy levels of the system states and the second half re-raises them to their initial values. Thus, if we separate each trajectory into its first and second half, we expect to see the first half to have a negative work value and the second to have a positive one. In the quasi-static case, because the net distribution is nearly symmetric, we would expect these positive and negative contributions will be nearly symmetric as well. What will happen, though, as we depart from the known case? Is the drift towards positive average work caused primarily by changes in the negative work parts of the trajectory, or the positive work parts? Figure 2.5.3 shows the same data as before, but with additional density plots for the work during the first and second half of he protocol.

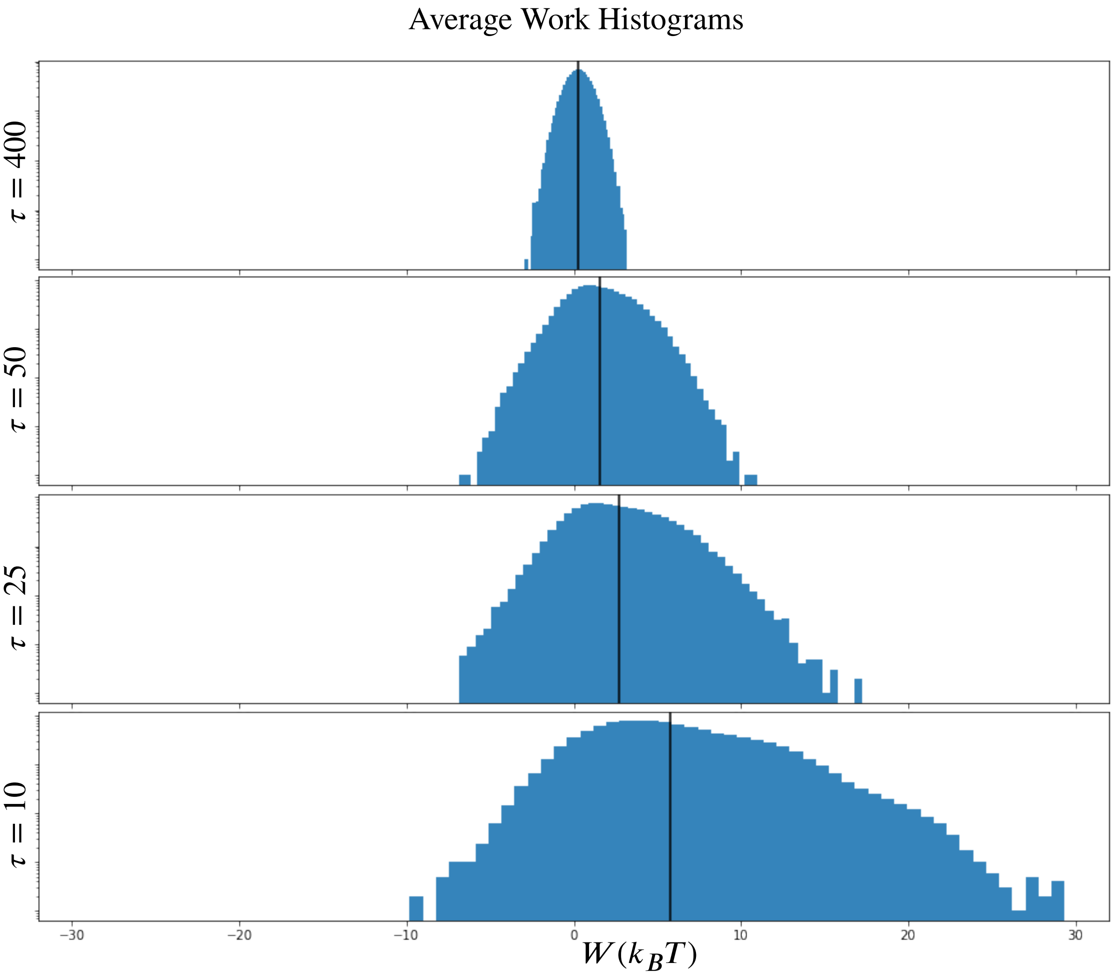


Figure 2.5.2. Work distributions for 100,000 at four different computation timescales τ. The top case is essentially a quasistatic protocol. While there are signatures of finite time in that the average value is displaced from 0, the distribution is simple: narrow, unimodal, and symmetric. As we proceed from the top to the bottom, the distributions are no longer well approximated by their average values alone. Each plot is a density histogram with a log scale in the y axis.

The plot shows that, while both distributions evolve interesting features and skew towards positive work, the part of the protocol that has negative work on average is more heavily affected by the finite time effects of quick computation.

2.6. Fluctuation Theorems

In order to understand this, and other nonequilibrium effects, a more fundamental theory than purely equilibrium or linear-response thermodynamics is required. This theory falls under the now

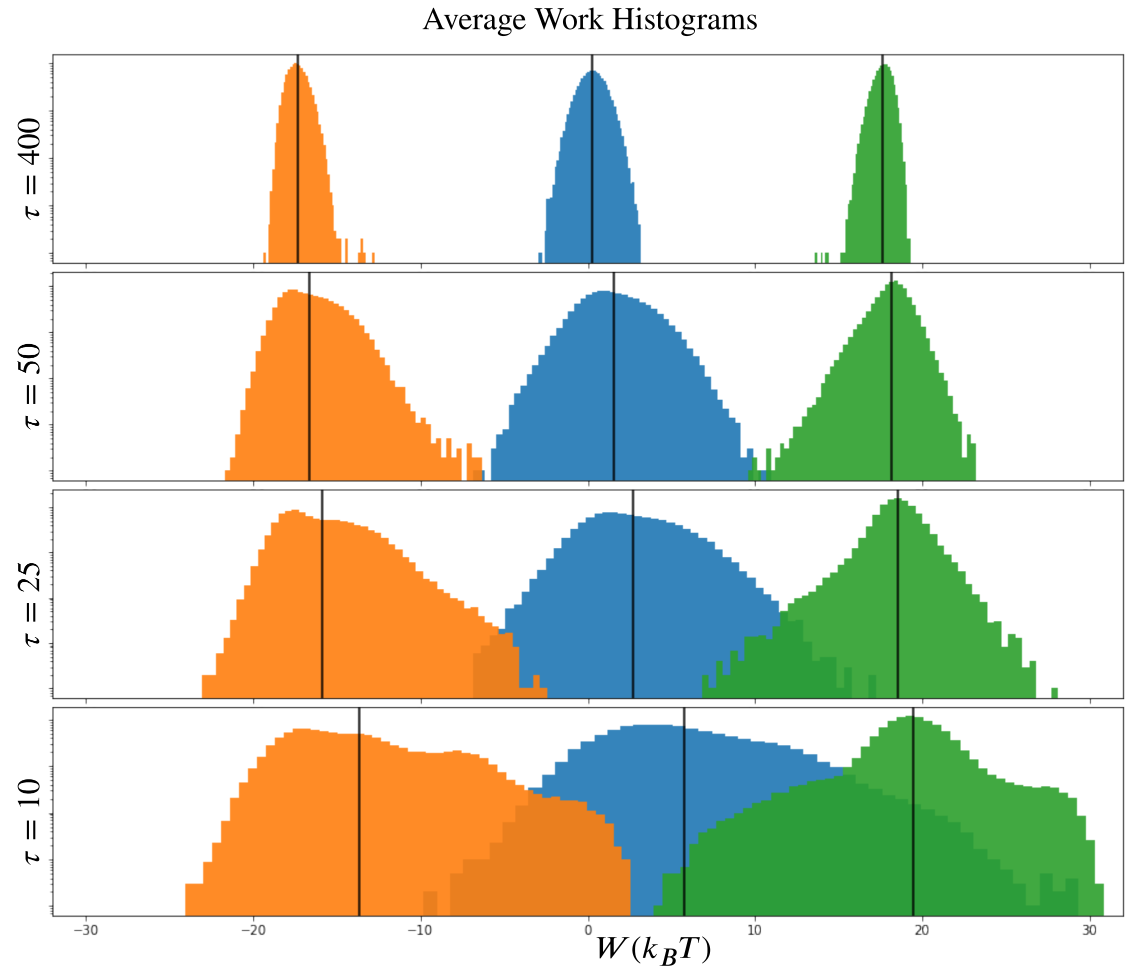


FIGURE 2.5.3. The same work distributions showed in figure 2.5.2 with sub-stage distributions overlayed on top. The orange(green) distributions include the first(second) half of the protocol only.

large umbrella of 'fluctuation theorems', which we will deal with in the next chapter. However, as a small preview, let's look at a well-founded thermodynamic equality that will hold true for all four timescales: the Jazynski equality. The equality, which applies in the case of cyclic protocols like the Szilard cycle, is given by the relation:

$$\langle e^{-\beta W} \rangle = 1 \tag{2.16}$$

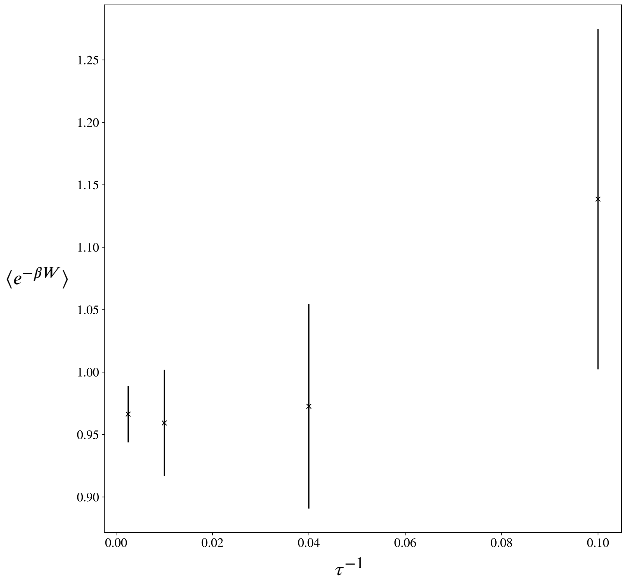


FIGURE 2.6.1. The Jarzynski equality applied to the full work distributions. The x axis is scaled as $\frac{1}{\tau}$, so that the point on the far right with the most variance is the fastest protocol– where equilibrium considerations matter the least. The error bars do not quite hit 1 for all cases, this is due to a known issue with sampling, which can cause large biases from a small subset of individual trajectories. This issue is addressed in section 3.2 in detail. The simulation was originally run to analyze work distributions, rather than the Jarzynski equality, and so did not involve enough samples or a small enough dt to overcome the issue. The fact that the exponential average is close regardless is actually quite remarkable.

Looking at figure 2.6.1 reveals that each of the four simulations obey this law. And, by breaking the law down, some insight is gained about how the lawfulness of work fluctuations leads to the increase in average work.

As a simple example, consider a bimodal distribution that can have only two potential work outcomes: $\pm\beta^{-1}W_0$ with probabilities $p(W_0)$ and $p(-W_0) = 1 - p(W_0)$. The Jazrkynski equality tells us that $p(W_0)e^{-W_0} + p(-W_0)e^{W_0} = 1$. They key point here is that the law that governs the probability of positive and negative work exponentially amplifies the term proportional to $p(-W_0)$. Another perspective says that for large values of W_0, even a very small probability negative event will have a significant impact on the distribution. The result of this, is that thermodynamics abhors a high probability, negative work event. Meanwhile, the average work value is given by $\langle W \rangle = \beta W_0(p(W) - p(-W))$. In this average, whatever term has the higher probability will dominate, which means that the the average skews towards the positive outcome as W_0 gets larger.

With this in mind, we return to the behavior of the plots in figure 2.5.3. While the exact profiles of the marginal distributions are beyond such a simple analysis, we can start to understand the behavior in a very general way. As the average work costs increase, it is impossible for the full distribution and its two marginal components to, in concert, move in the same direction and maintain their shapes because this would cause violations of the Jarzynski equality– in which negative work events must have their probabilities shifted differently then their corresponding positive counterparts. Negative work is treated differently than positive work, and the more negative it is– the more differently it is treated. However, we have only scratched the surface of fluctuation theorems at this point. The next chapter will serve as a brief review of modern fluctuation theorems, with a heavy focus on interesting applications rather than the mathematical underpinnings. We will also investigate some of the most recent advances in the field: trajectory class fluctuation theorems [**55**], and thermodynamic uncertainty relations [**56**].

CHAPTER 3

In the preceding chapter, we have implicitly departed from the historical equilibrium thermodynamic mode of thinking where the fundamental quantities of interest are macroscopic variables. The 'thermodynamic limit' assures that these variables, though really averages over a microscopically noisy and stochastic system of many individual pieces, are found with near certainty to be at their theoretical values. The sheer scale of the problems considered, and the restriction to equilibrium distributions suppresses fluctuations perfectly. The fundamental conceit underlying the 'thermodynamic limit' is that we do not need to consider the energetic or entropic consequence of a particular particle in the ensemble taking a particular trajectory; we care strictly about well defined averages over ensembles.

Stochastic thermodynamics [**15**] is a nonequilibrium theory that works in the other direction: the fundamental quantities of interest are trajectory-wise heats, works, or entropies and the historical macroscopic quantities are the mean values of a well defined distribution over these quantities. The transition into stochastic thermodynamics was made possible by both numerical and theoretical advancements. Numerically, simulations of trajectory ensembles that number in the hundreds of thousands of trajectories are accessible even without access to high performance computing hardware, allowing for easy numerical probing of system scales that were previously inaccessible. On the theory side, the distributions are bound by a new set of theoretical results known as 'fluctuation theorems'.

The study of stochastic thermodynamics goes all the way back to Einstein's treatment of Brownian motion [**57**], but the first result that has a direct line to the modern treatment was by Bochov and Kuzovlev in 1977 [**58**]. Despite the prescience of the result in hindsight, the work failed to elicit many immediate follow on results. Then, starting in the mid 1990s, rapid progress was made as results begin to pile up and connect with one another. To name a tiny sample of the full set of important works in this time period, we have: Evans' and Searle's 1994 paper that discusses the exponential suppression of negative entropy producing trajectories [**59**], Jarzynski's 1997 paper that introduces the Jarzynski equality [**60**], and Crooks' 1999 paper that expanded the realm applicability beyond reversible dynamics [**61**]. By this time, a consensus had been reached detailing two overarching types of fluctuation theorems: 'integral fluctuation theorems' that yield results about the moments of the distribution of a stochastic thermodynamic variable, and 'detailed fluctuation theorems' that yield results about the trajectory-wise values such a variable can take.

A recent development [**55**] introduces 'trajectory class fluctuation theorems', which defines a new suite of fluctuation theorems that allow us to define equalities for any arbitrary subset of trajectories thus interpolating between the 'detailed' and the 'integral' theorems. In the following, we will discuss new and historical uses of these theoretical results as well as provide both novel and familiar proofs of key results. We will conclude by using a fluctuation theorem to derive a new thermodynamic uncertainty relation [**56**], which is currently one of the most active subjects of study in stochastic thermodynamics.

3.1. The Jarzynski equality

One of the most prominent fluctuation theorems is the aforementioned Jarzynski equality (JE), which gives a simple relationship between the amount of work it takes to implement a control protocol for a thermostatted system,

$$\langle -e^{\beta W} \rangle = e^{-\beta \Delta F} \ . \tag{3.1}$$

The average on the left is an average over many repeated iterations of the same procedure, and the free energy change is the difference in free energy between the equilibrium states for the initial and final values of whatever control system is implementing the protocol. Below we outline a straightforward proof of the relation, provided by Jarzynski [**62**].

Assume a Hamiltonian of the form $H(\Gamma, \lambda) = H_E(y) + H_S(x, \lambda) + H_I(x, y)$. Here we have made a distinction between two subspaces of the Hamiltonian, breaking the full phase space Γ into the 'system' degrees of freedom x and the environment degrees of freedom y. Note also the parameter

λ which is the control parameter. It represents an external agent that manipulates the Hamiltonian of the system in a deterministic way. All the energy that this agent inputs or absorbs is useful work, generating no entropy. Once thermodynamically consistent way of thinking about this is that this 'work reservoir' is a heat bath with infinite temperature. The interaction $H_I(x, y)$ between the system and environment need not be small for the proof to work, but for simplicity of exposition we assume that it is.

Now, assume that we begin this composite system in an equilibrium state associated with the inverse temperature of the thermal environment β, $\rho(\Gamma_0) = Z_{\lambda_0}^{-1} e^{-\beta H(\Gamma_0, \lambda_0)}$. With the normalization given by $Z(\lambda_t) = \int d\Gamma e^{-\beta H(\Gamma, \lambda_t)}$. The system is then exposed to a protocol by changing lambda over time. Examining the LHS of equation 3.1, we have:

$$\langle e^{-\beta W} \rangle = \int d\Gamma_0 p(\Gamma_0) e^{-\beta W(\Gamma_0|\lambda)} \tag{3.2}$$

Note that the amount of work it will take to implement the protocol $\lambda = (\lambda_0, ..., \lambda_\tau)$ is a deterministic function of the initial point in the join phase space Γ_0 because the entire composite system evolves deterministically, so a function has been defined $W(\Gamma_0|\lambda)$. The work required to change the Hamiltonian of the system though a control parameter a small bit dW is given by $\frac{\partial H_S(x,\lambda)}{\partial \lambda} \mathrm{d}\lambda$. Importantly, because of the derivative with respect to λ, the only term in the Hamiltonian that matters for the work done is H_S, so $\partial_\lambda H_S(x, \lambda) = \partial_\lambda H(\Gamma, \lambda)$. This gives us an expression for the work:

$$W = \int dW = \int_0^\tau \frac{\partial H_S(x_t, \lambda_t)}{\partial \lambda} \dot{\lambda} dt \tag{3.3}$$

$$= \int_0^\tau \frac{\partial H(\Gamma_t, \lambda_t)}{\partial \lambda} \dot{\lambda} dt \tag{3.4}$$

$$= \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t} H(\Gamma_t, \lambda_t) dt \tag{3.5}$$

$$= H(\Gamma_\tau, \lambda_\tau) - H(\Gamma_0, \lambda_0) \tag{3.6}$$

This, in turn can be plugged into equation 3.2, which yields:

$$\langle e^{-\beta W} \rangle = \int d\Gamma_0 p(\Gamma_0) e^{-\beta(H(\Gamma_\tau,\lambda_\tau)-H(\Gamma_0,\lambda_0))} \tag{3.7}$$

$$= Z^{-1}(\lambda_0) \int d\Gamma_0 e^{-\beta H(\Gamma_\tau,\lambda_\tau)} \tag{3.8}$$

$$= \frac{Z(\lambda_\tau)}{Z(\lambda_0)} \tag{3.9}$$

The last equality is justified by a change of variables. The change of integration variable from $d\Gamma_0$ to $d\Gamma_\tau$ is trivial because Liouville's theorem ensures no state space contraction or expansion and so the phase space volume element corresponding to $d\Gamma_0$ and $d\Gamma_\tau$ have the same volume. With this in mind, the remaining integral is simply the normalization factor $Z(\lambda_\tau)$. In the case of weak coupling, $H_I << \min(H_E, H_S)$ and is neglected, yielding $Z(\lambda) = Z_x(\lambda)Z_y$. Because H_E is constant , the Z_y in both the top and bottom of the fraction cancels. The familiar result is what remains:

$$\langle e^{-\beta W} \rangle = \frac{Z_x(\lambda_\tau)}{Z_x(\lambda_0)} = e^{-\beta \Delta F} \ . \tag{3.10}$$

3.1.1. Jarzynski Equality in Practice. This integral fluctuation theorem can be put to immediate use. By invoking Jensen's inequality for convex functions ($\langle f(x) \rangle > f(\langle x \rangle)$), the Jarzynski equality recovers the second law of thermodynamics: $\langle W \rangle > \Delta F$. We also gain some insight as to why is it that trajectories that produce work ($W < 0$) cause such a problem, as briefly discussed in the previous section. In short: the work value associated with each realizations is exponentiated in our average, so even a small amounts of negative work can come to dominate the expression.

But, because the equation involves the entire distribution rather than just the average, it contains more information than just the second law. We can illustrate this with the case of a cyclic process, for which $\Delta F = 0$. In this case, $\langle e^{-\beta W} \rangle = 1$. Let's call this type of system and protocol an engine, for an engine must always return to its initial state in order to be reused. This theorem places limits on how an engine can be designed. Can we, for example, create an engine that most of the time produced a small chunk of work and occasionally would have a compensating work cost that would offset the transient gain? Could we create an engine that produces an arbitrarily large amount of work with an arbitrarily large probability, at the cost of an arbitrarily rare and arbitrarily catastrophic 'black swan' driving event that levies an enormous cost? Using the fact that the driving work must be no less than zero work on average this seems like a plausible scheme, but let's investigate further using the fluctuation theorem. Assume an engine with N possible work outcomes of value W_i and probability p_i. The yields the following equation:

$$\sum_{i=1}^{N} p_i e^{-\beta W_i} = 1 \ . \tag{3.11}$$

Singling out a special value of the work, W_j, gives:

$$\sum_{i \neq j}^{N} p_i e^{-\beta W_i} = 1 - p_j e^{-\beta W_j} \tag{3.12}$$

The LHS of the equality above is a sum of manifestly positive terms, so the RHS must itself be positive, this gives us

$$p_j e^{-\beta W_j} < 1 \tag{3.13}$$

$$-\beta W_j < -\ln p_j \tag{3.14}$$

$$\beta W_j > \ln p_j \tag{3.15}$$

There is, then, a limit to how negative we can make the work of any given W_j. It is bounded by the likelihood of our engine realizing a trajectory that produces that much work. This yields an emphatic "no" to the original query. It is impossible make an arbitrarily beneficial (work wise) trajectory happen arbitrarily often– no matter what the rest of the work distribution looks like. Each possible work value is bound by its own probability to happen, and only an incredibly rare event ($p_i \to 0$) can output a significant amount of work by having a large negative driving work value.

Additional insight can be gleaned from the Jarzynski equality as well. Consider, again, an arbitrary work distribution $\{W_i\}$ with probabilities $\{p_i\}$. Inspired by the connection above between the probability of an even and its work production, a set of numbers q_i can be defined as $\beta W_i = \ln p_i - \ln q_i$. This is, admittedly, an odd thing to do– but certainly can be accomplished for any set of $\{W_i\}$ and $\{p_i\}$. The result above that $\beta W_i > \ln p_i$ implies that $0 < q_i < 1$. If we plug this expression for work into the fluctuation theorem, the following is obtained:

$$\sum p_i e^{-\ln p_i + \ln q_i} = 1 \tag{3.16}$$

$$\sum q_i = 1 \ . \tag{3.17}$$

The $\{q_i\}$ satisfy the conditions of a distribution ($q_i > 0$ and $\sum q_i = 1$). Now, consider the average work $\langle \beta W \rangle = \sum_i p_i \ln \frac{p_i}{q_i}$. Because the q_i form a distribution, the average is the Kullback-Leibeler divergence between p and q: $\langle \beta W \rangle = D_{KL}(p||q) \geq 0$. The inequality is saturated only when the distributions $\{p_i\}$ and $\{q_i\}$ are identical, and so all $W_i = 0$. Values of W_i shift away from zero only create engines that are less efficient on the average. Trying to get more work out of one kind of cycle at the expense of the others being more costly will always be more inefficient.

An interesting example is of the single particle Szilard engine: ignoring the demon memory, it has two outcomes $\beta W_1 \sim \ln \delta$ and $\beta W_2 \sim \ln 1 - \delta$ with probabilities $p_1 = \delta$ and $p_2 = 1 - \delta$. Now what the result above tells us is that the demon memory operation is at its most efficient when it takes exactly $-kT \ln \delta$ work for the demon to synchronize and control the system during the first outcome and exactly $-kT \ln 1 - \delta$ work for the demon to synchronize and control the system during the second outcome. This result, is exactly that predicted by Szilard himself! Now, let us consider a more general case.

The considerations above ought not to be surprising. The engine was not provided any ‘fuel’ from which to extract energy. Relaxing the cyclic condition of our engine, and allowing the engine to consume a free energy resource, the fluctuation theorem becomes $\langle e^{-\beta W} \rangle = e^{-\beta \Delta F}$ where ΔF is the change in free energy associated with the thermal equilibrium states of the control Hamiltonian at the beginning and end of our process. In this case, we have

$$\sum p_i e^{-\ln p_i + \ln q_i} = e^{-\beta \Delta F} \tag{3.18}$$

$$\sum q_i = e^{-\beta \Delta F} \ . \tag{3.19}$$

The q_i no longer constitute a distribution. The q_i still encode how much a given work distribution changes the state of the machine used to enact it– but the encoding is not as straightforward. Using the same tactic as before, separating out a particular W_j, gives:

$$\sum_{i \neq j}^{N} q_i = e^{-\beta \Delta F} - p_j e^{-\beta W_j} \ . \tag{3.20}$$

Again, note that the LHS is strictly positive. This means that the RHS can be written as an inequality

$$p_j e^{-\beta W_j} < e^{-\beta \Delta F} \ , \tag{3.21}$$

which simplifies to

$$\beta W_j > \ln p_j + \beta \Delta F \ . \tag{3.22}$$

In the presence of ΔF, there is no bound that requires common trajectories have (at best) very small benefits. The inequality suggest a redefinition of our general work value expression from $\beta W_i = \ln p_i - \ln q_i$ to $\beta W_i = \ln p_i - \ln q_i + \beta \Delta F$ to include the dependence the work has on the change in equilibrium free energy by a redefinition of q_i. This subsumes the previous case, because it reduces to the previous expression when $\Delta F = 0$. Revisiting the fluctuation theorem with the re-scaled q_i gives us:

$$\sum p_i e^{-\ln p_i + \ln q_i - \beta \Delta F} = e^{-\beta \Delta F} \tag{3.23}$$

$$e^{-\beta \Delta F} \sum p_i e^{-\ln p_i + \ln q_i} = e^{-\beta \Delta F} \tag{3.24}$$

$$\sum q_i = 1 \tag{3.25}$$

Thus, we recover the property that q_i is a distribution. For illustrative purposes, consider a simple binary machine with outcomes W_1, W_2.

To build the desired 'black swan' machine, W_1 should be as negative as possible (recall that negative work in this context means work done on the environment), thus, ΔF must be negative as well. To make the event arbitrarily likely, p_1 can be set equal to $1-\gamma$, and to make the work as favorable as possible $\beta W_1 = \ln(p_1)+\beta\Delta F+\epsilon$ (where γ, ϵ are small parameters). Using the Jarzynski equality, the other work outcome is now completely constrained. First, note that $-\ln q_1 = \epsilon$ which yields $q_1 = e^{-\epsilon}$

$$\beta W_2 = \ln\gamma - \ln q_2 + \beta\Delta F \tag{3.26}$$

$$= \ln\gamma + \beta\Delta F - \ln\big(1 - e^{-\epsilon}\big) \tag{3.27}$$

Where the second equality used the fact that $q_2 = 1 - q_1$. Expanding the exponential in the log to order ϵ yields:

$$\beta W_2 = \ln\gamma + \beta\Delta F - \ln\epsilon \tag{3.28}$$

So we see that by being ϵ close to the best case for W_1 we pay for it in the W_2 outcome with a factor of $-\ln\epsilon$. Thus, the fluctuation theorem does not outright restrict the creation of a common useful cycle at the cost of a catastrophic result. In fact, the cost may not even be so catastrophic– if γ can appropriately compensate. The two parameters are actually working at odds with one another, and if ϵ scales the same as γ it appears we will have a cancellation. However, we must first make sure that these scalings are permitted by the second law of thermodynamics. The average work ought to satisfy $\langle\beta W\rangle > \beta\Delta F$. Our definition of work in therms of $\{q\}$ and $\{p\}$ allows us to write the equality:

$$\langle \beta W \rangle = (1-\gamma)\left(\ln(1-\gamma) - \beta|\Delta F| + \epsilon\right) + \gamma\left(\ln\gamma - \beta|\Delta|F - \ln\epsilon\right) \tag{3.29}$$

$$= (1-\gamma)\left(\ln(1-\gamma) + \epsilon\right) + \gamma\left(\ln\gamma - \ln\epsilon\right) + \beta\Delta F \tag{3.30}$$

$$= [(1-\gamma)\epsilon - \gamma\ln\epsilon - H(\gamma)] + \beta\Delta F \tag{3.31}$$

$$. \tag{3.32}$$

Where $H(\gamma)$ is the binary Shannon entropy associated with the value of γ. In order for the term in square brackets to be strictly positive, the following must be true:

$$\epsilon > \gamma\ln\epsilon + H(\gamma) + \gamma\epsilon \tag{3.33}$$

The inequality above puts bounds on how close to the lowest work cost possible can be to the ideal case of $\epsilon = 0$. Thus, there exists a nontrivial relationship here between the maximum fidelity we can permit ourselves (γ) and the 'idealness' of the common cycle (ϵ).

As an illustrative example, it is simple to consider the case where $\gamma \approx \epsilon$. This yields

$$\epsilon > \epsilon^2 - (1-\epsilon)\ln(1-\epsilon) \tag{3.34}$$

$$\epsilon > \epsilon^2 - (-\epsilon - \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}) \tag{3.35}$$

$$0 > \frac{3}{2}\epsilon^2 + \mathcal{O}(\epsilon^3) \tag{3.36}$$

Because this cannot be satisfied by any value of ϵ, the Jarzynski equality forbids the case where ϵ and γ take the same value. While these considerations show how the JE can provide analytic tools beyond those given by the second law of thermodynamics alone, they are of primarily theoretical interest. Fluctuation theorems also provide incredibly powerful practical tools useful in experimental physics.

3.2. Free Energy Estimation and Rare Events

One of the most immediately useful applications of the Jarzynski equality, lies in recognizing it provides an equality that relates a nonequilibrium quantity (the work done) to an equilibrium quantity (the free energy change). One of the first experimental verifications of the equality [**63**] put this to good use by repeatedly stretching a single RNA molecule and using the work done to do so to calculate the change in free energy between the stretched and unstretched state of the molecule. To this end the JE can be rearranged :

$$\Delta F = -\beta^{-1} \ln \langle e^{-\beta W} \rangle \ . \tag{3.37}$$

From which, it is clear that the important quantity to estimate is the exponential average of the work cost: $\langle e^{-\beta W} \rangle \approx \frac{1}{N} \sum_{i=0}^{N} e^{-\beta W_i}$. Crucially, the exponential means that some terms in the average will dominate. While previous discussion above danced around the topic, we now look at it head on– with an analysis that roughly follows the approach in reference [**64**] to explain convergence issues that pop up in experiment because of the dominance of certain terms.

First, we point out the problem by example. Returning to the simple cyclic engine ($\Delta F = 0$) of the previous section, imagine a work distribution that contains two work outcomes $\pm\beta^{-1}A$ with probabilities $p_{\pm}$. Expanding out the Jarzynski Equality gives

$$(1 - p_-)e^{-A} + p_- e^{A} = 1 \tag{3.38}$$

$$p_-(e^{A} - e^{-A}) = 1 - e^{-A} \tag{3.39}$$

$$p_- = \frac{1 - e^{-A}}{e^{A} - e^{-A}} \ , \tag{3.40}$$

And here we have a mathematical expression for the damping of negative work outcomes; for example, the probability to see negative work outcomes with a magnitude equal to even 10 $k_{\mathrm{B}}T$ are already approximately 1 in 100,000. Nevertheless, these outcomes are of critical importance to estimating free energy. To be specific, consider the case of $A = 2$ for which $p_- \approx .12$. The straight work average is

$$\langle W \rangle = p_+ A - p_- A \tag{3.41}$$

$$2 = .88 * 2 - .12 * 2 \ , \tag{3.42}$$

here, the common outcome accounts for the lions share of the average. For the average of the exponential work we have

$$\langle e^{-W} \rangle = p_+ e^{-A} + p_- e^{A} \tag{3.43}$$

$$1 = .12 - .88 \ , \tag{3.44}$$

which displays opposite behavior. The outcomes that are the least common are responsible for the majority of the average. In point of fact, the roles of the two work outcomes have swapped.

With this demonstration of how uncommon events can come to dominate averages in mind, we define two separate ideas of importance in a distribution. One is the usual sense of importance, called the 'typicality' of a realization, and is simply the probability of the event. The other sense, called the 'dominance', is how important the event is in a given average. This 'dominance' of a particular outcome in an average over $f(x)$ can be associated, in the discrete case, with a number between 0 and 1 that indicates how much of the average comes from that realization with the following equation:

$$D(X = x) = \frac{Pr(X = x) f(x)}{\langle f(x) \rangle} \ . \tag{3.45}$$

With this definition, the dominance of the event $W = -\beta^{-1} A$ in the straight average is .12 but its dominance in the exponential average is .88. This is a general effect, algebraically:

$$D(W = -\beta^{-1} A) = p_- e^{A} \tag{3.46}$$

$$= \frac{1 - e^{-A}}{1 - e^{-2A}} \tag{3.47}$$

$$= 1 - p_- \ . \tag{3.48}$$

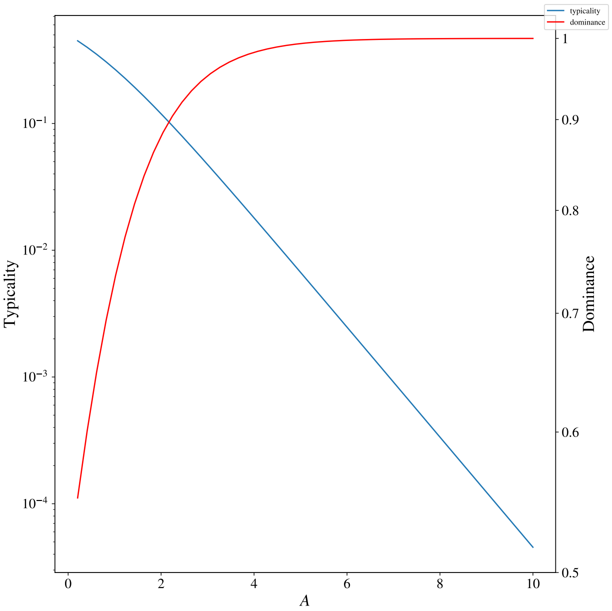


FIGURE 3.2.1. The typicality versus the dominance of the negative work outcome in the average of the exponential work. Frustratingly, the less likely you are to see the event, the more important it becomes in the average

See figure 3.2.1 for a graphical representation. The conclusion here is somewhat sobering: as the probability of an event goes to zero– it also becomes exclusively responsible for the average that allows an estimate of the free energy. While the math is a bit more tedious, the same issue holds for other distributions that are less trivial wether they are discrete or continuous.

In order to understand this, we shall turn to another analysis of the JE by means of a 'detailed fluctuation theorem'. In order to proceed, it will be prudent to discuss and define a set of 'time reversed' quantities and associated notation. Again, take a Hamiltonian system containing a system specific piece that depends only on a control parameter λ and the system's degrees of freedom. A 'process' $P(x) \equiv Pr(X = x|\Lambda = \lambda)$ will be defined as a probability distribution over the set of system trajectories $x \in X$, conditioned on the system being exposed to the control protocol $\lambda = \lambda(t)$. Similarly, the conditional process $P(x|a) \equiv Pr(X = x|X_0 = a, \Lambda = \lambda)$ is the probability that the system undergoes trajectory x conditioned on both starting in state a at time $t = 0$ and being exposed to the control protocol λ. A final useful piece of notation is $p_t(a) \equiv Pr(X_t = a|\Lambda = \lambda)$; combining these yields the following equality $P(x|a)p_0(a) = P(x)$ (which would be a bit ponderous to write without the shorthand).

The reverse protocol associated with P will be denoted as $R(x) \equiv Pr(X = x^\dagger | \Lambda = \lambda^\dagger)$. $R(x)$ gives the probability that the same system will undergo the trajectory $x^\dagger$ provided that it is exposed to the reverse control protocol $\lambda^\dagger = \lambda(\tau - t)$; similarly $R(x|a) \equiv Pr(X = x^\dagger | X_0 = a^*, \Lambda = \lambda^\dagger)$ can be defined as well as $r_t(a) \equiv Pr(X_t = a^* | \Lambda = \lambda^\dagger)$. For every trajectory x of the system of interest, which is defined by a set of values in the phase space of the system ($x = (x_0 ... x_\tau)$ for discrete time and $x = x(t)$ for continuous time), we can define also a reversed trajectory $x^\dagger = (x^*_\tau, ... x^*_0)$. The $*$ indicates sending any time odd quantity to its negative value, spin for example. This becomes relevant for the the phase space coordinates of the system, because they contain canonical momentum variables which do change sings under the time reversal conjugation.

Assume the same Hamiltonian as before: $H(\Gamma, \lambda) = H_E(y) + H_S(x, \lambda) + H_I(x, y)$, and that the system begins in thermal equilibrium with the environment. At $t = 0$, the system and environment are isolated from each-other by turning the interaction to 0 and the system is allowed to evolve under its own dynamics (driven by the time dependent λ) until $t = \tau$. Similarly, the reversed process begins in equilibrium with the thermal environment– but since $\lambda^\dagger_0 = \lambda_\tau$ the starting distribution is different.

A detailed fluctuation theorem deals with relationships between how a trajectory behaves in the forward process to how its reversed trajectory behaves in the reverse process, so we need expressions for both. Notably, because the system is allowed to evolve under its own Hamiltonian (deterministic) dynamics we can say that, for any valid trajectory that the system can undergo, $P(x|a) = \delta_{x_0, a}$ – meaning there is one and only one possible trajectory for each starting point $x_0 = a$, or $P(x) = p_0(x_0)$. Since both processes begin in equilibrium with their environments at some inverse temperature β, this gives:

$$P(x) = Z_x^{-1}(\lambda_0) e^{-\beta H_S(x_0, \lambda_0)} = e^{-\beta H_S(x_0, \lambda_0) + \beta F_{\lambda_0}} \tag{3.49}$$

$$R(x) = Z_x^{-1}(\lambda_\tau) e^{-\beta H_S(x^*_\tau, \lambda_\tau} = e^{-\beta H_S(x^*_\tau, \lambda_\tau) + \beta F_{\lambda_\tau}} \tag{3.50}$$

For the rest of this section, β is set equal to 1 to help with clarity. Assuming that the Hamiltonian has only time symmetric terms $H_S(x^*, \lambda_t) = H_S(x, \lambda_t)$ will allow a direct comparison of the two probabilities:

$$\frac{P(x)}{R(x)} = e^{\Delta H_S(x) - \Delta F(\lambda)} \tag{3.51}$$

$$\frac{R(x)}{P(x)} = e^{-\Delta H_S(x) + \Delta F(\lambda)} \ . \tag{3.52}$$

Above ΔH and ΔF represent quantities that are explicitly about the process and not the reverse process: the change in energy of the system when it undergoes trajectory x and the change in the free energy of the system as it undergoes the control protocol λ.

Following the same logic as in equation 3.7, the Hamiltonian evolution of the system implies that the change in the energy of the system is equal to the work done in the process during the trajectory x, $\Delta H_S(x) = W(x)$. The same can be said for the work done in the reverse process, which yields

$$W^R(x) = H(x_0, \lambda_0) - H(x_\tau, \lambda_\tau) = -W(x) \ . \tag{3.53}$$

Using the definition of work to rewrite the ration gives us the relation

$$\frac{P(x)}{R(x)} = e^{(W(x) - \Delta F)} = e^{-(W^R(x) + \Delta F)} \ . \tag{3.54}$$

This 'detailed fluctuation theorem' must hold true in order for the remaining conclusions to hold. It has been shown that the above holds for several classes of dynamics, notably including stochastic Markovian dynamics [**65**]; Therefore, the proof above should be considered not as restrictive but as illustrative of a particular regime in which the equality does hold true.

Introducing the 'dissipated work' in the process$W_d = W - \Delta F$, the JE tells us

$$\langle e^{-W_d} \rangle = 1 \tag{3.55}$$

$$\int dx P(x) e^{-W_d(x)} = 1 \tag{3.56}$$

$$\int dx Q(x) = 1 \ . \tag{3.57}$$

Where Q is defined implicitly from the preceding line. The function $Q(x)$ is a distribution with the same measure and support as the probability distribution $P(x)$– but is peaked where there are large contributions to the average of the exponential work. Returning to the concepts of 'typicality' and 'dominance', $Q(x)$ is a way to quantify the dominance the trajectory x in the continuous case in the same way $P(x)$ quantifies its typicality.

A particularly interesting result comes form considering how $Q(x)$ relates to the detailed fluctuation theorem in equation 3.54:

$$Q(x) = P(x)e^{-W_d(x)} = R(x)\ . \tag{3.58}$$

High dominance in the process is associated with high typicality in the reserve process. And, using the JE on the reverse process to obtain an analogous $Q^R(x)$ with yield an analogous expression $Q^R(x) = P(x)$. The reverse of typical trajectories in the forward process become dominant ones in the reverse process. The natural extension of this concept is to consider not just a single trajectory but sets, or 'classes', of them. Instead of doing an integral over the entire set of trajectories we can define an integral $\int_C dxf(x) = \int [x \in C]\, dxf(x)$ where $[expression]$ is equal to 0 unless the expression inside is true. More will come on trajectory classes, but this is sufficient for the question at hand. Given a set of trajectories C, the typicality and dominance of C are given by :

$$P(C) = \int_C dxP(x) \tag{3.59}$$

$$D(C) = \int_C dxQ(x)\ . \tag{3.60}$$

Defining the 'reverse set' of trajectories $C^\dagger$ composed of a $x^\dagger$ for every x in C, we get

$$R(C) = \int_C^\dagger dx^\dagger R(x^\dagger) \tag{3.61}$$

$$D^R(C) = \int_C^\dagger dx^\dagger Q^R(x^\dagger)\ . \tag{3.62}$$

Now, because the systems evolve with Hamiltonian dynamics, the Jacobian to transforms between the forward dx and reverse $dx^\dagger$ is 1. Through equation 3.58, the first equation above can be written as:

$$R(C) = \int \left[x^\dagger \in C^\dagger\right] dxQ(x) \tag{3.63}$$

$$= \int [x \in C]\, dxQ(x) \tag{3.64}$$

$$= \int_C dxQ(x) = D(C) \tag{3.65}$$

Thus, provided that the dynamics obey equation 3.54, the dominance of a trajectory class in a process is equal to the typicality of the class of reversed trajectories in the reversed process.

The exact same argument leads to the conclusion that $P(C) = D^R(C)$. The intuition here is actually quite informative. Trajectories that are dominant in the forward process are ones that will appear to move backwards in time: while the ensemble as a whole starts in an equilibrium distribution and then is driven out of it, these rare trajectories will starts as large fluctuations away from equilibrium and then appear to 'thermalize' by ending close to the equilibrium distribution associated with the final value of the driving parameter. The more out of equilibrium that the system is driven in the process, the less likely these kind of trajectories are. While it is possible to obtain equilibrium quantities from processes driven arbitrarily far from equilibrium through the JE, there exists a tradeoff: the probability of observing the outcomes that make the calculation possible becomes smaller as the system is driven further away from equilibrium.

3.3. Trajectory Class Fluctuation Theorems

The success of the treatment of arbitrarily subsets of trajectories at the end of section 3.2 is a hint that more can be done in this direction. Indeed, there exist detailed and integral fluctuation theorems for any class of trajectories. A rigorous proof will not be provided here, but one can be found in the work that introduced the trajectory class fluctuation theorem(TCFT) [**55**]. Instead, this section will provide a brief proof sketch and then focus on applications of the TCFT.

3.3.1. Trajectory Classes. While alluded to in the previous section, it is prudent at this point to be explicit about what a trajectory class C is. C is defined by a function $c(\mathbf{x})$ that takes a trajectory to a boolean: $c : \mathbf{X} \to \mathbb{B}$. Of all possible trajectories $\mathbf{x} \in \mathbf{X}$, C is the set of trajectories for which $c(\mathbf{x})$ is True. In standard mathematical notation, $C : \{\mathbf{x} \in \mathbf{X} | c(\mathbf{x})\}$. This notation allows us to define the reverse class $C^\dagger$ clearly, $C^\dagger : \{\mathbf{x} \in \mathbf{X} | c(\mathbf{x}^\dagger)\}$. Note that the $\mathbf{x}^\dagger$ 'reverse trajectories' come from the same space as the 'forward trajectories'. The dagger operator acting on a trajectory means conjugate time reversal, if we think of trajectories over a time τ as being composed of a bunch of instantaneous position and momentum coordinates $\mathbf{x}(t) = (x_t, v_t) \forall t \in [0, \tau]$ then $\mathbf{x}^\dagger(t) = (x_{\tau-t}, -v_{\tau-t})$.

As an example, consider a class of all trajectories with positive initial velocity, and so defined by the function $c(\mathbf{x}) = [v_0 > 0]$. Suggestively, the class can be called v_{0+}. The reverse class $v_{0+}^\dagger$ is

$$v_{0+}^\dagger : \{\mathbf{x} \in \mathbf{X} | c(\mathbf{x}^\dagger)\} \tag{3.66}$$

$$v_{0+}^\dagger : \{\mathbf{x} \in \mathbf{X} | -v_\tau > 0\} \tag{3.67}$$

$$v_{0+}^\dagger : \{\mathbf{x} \in \mathbf{X} | v_\tau < 0\} \tag{3.68}$$

We see that reverse class is the set of trajectories that end with a negative velocity; we might, for example, call it $v_{\tau-}$.

3.3.2. Proof Sketch. The integrated TCFT can be derived from the existence of a generalized detailed fluctuation theorem between a trajectory wise quantity $\Omega(\mathbf{x})$, a process $P(\mathbf{x})$ and a process conjugate to Ω called $P^\Omega(\mathbf{x})$. $P^\Omega(\mathbf{x})$ is the probability of the same trajectory in the conjugate process, though at this point the form of the conjugacy is left completely general. It is whatever it needs to be to satisfy the equality

$$\frac{P(\mathbf{x})}{P^\Omega(\mathbf{x})} = e^{\Omega(\mathbf{x})} \tag{3.69}$$

Assuming a conjugacy of the form above, deriving a TCFT is very straightforward: simply integrate over some measurable subset of trajectories (There are some details about integration that are important mathematically, but not necessary to understand the shape of the proof. For a rigorous proof, refer to [**55**].)

$$\int_C P(\mathbf{x}) e^{-\Omega(\mathbf{x})} = \int_C P^\Omega(\mathbf{x}) \tag{3.70}$$

$$\int_C P(C) P(\mathbf{x}|C) e^{-\Omega(\mathbf{x})} = \int_C P^\Omega(\mathbf{x}) \tag{3.71}$$

The second equality comes from the fact that $P(\mathbf{x}) = P(\mathbf{x}, C)$ because the integration is over trajectories in C only. Recognizing that $P(C)$ is independent of the integrand, the final line can be rearranged to read:

$$\langle e^{-\Omega} \rangle_C = \frac{P^{\Omega}(C)}{P(C)}$$

This is a very general statement of the TCFT, but under relatively common assumptions it can be easily interpreted alongside other common fluctuation theorems. If the system is coupled to a sufficiently ideal heat bath at inverse temperature β and the reverse process has the property that $r_0(x_\tau) = p_\tau(x_\tau)$– the Crooks detailed fluctuation theorem [**61**] takes the form of equation 3.69 in which $P^{\Omega}(\mathbf{x}) = R(\mathbf{x})$, and $\Omega = \Sigma$. Here, $R(\mathbf{x})$ is as described in the previous section, and Σ is the entropy production $\Sigma(\mathbf{x}) \equiv -\beta Q + \ln \frac{p_0(x_0)}{p_\tau(x_\tau)}$. Q is the heat absorbed by the system from the bath, so $-\beta Q$ is the entropy generated in the thermal bath. In this setting, the TCFT becomes:

$$\langle e^{-\Sigma} \rangle_C = \frac{R(C)}{P(C)}$$

3.3.3. Trajectory Class Jarzynski Equality. In the spirit of section 3.2, we set the initial distribution for P is the equilibrium distribution associated with λ_0, π_0 and the initial distribution for R is the equilibrium distribution for associated with λ_τ, π_τ. Using a first law $\Delta E = Q + W$ for the system, the entropy production breaks into

$$\Sigma(\mathbf{x}) = \beta(W - \Delta E) + \ln \frac{\pi_0(x_0)}{\pi_\tau(x_\tau)} \tag{3.72}$$

$$= -\beta(H_S(x_\tau, \lambda_\tau) - H_S(x_0, \lambda_0) - W) + \ln \frac{Z_\tau e^{-\beta H_S(x_0, \lambda_0)}}{Z_0 e^{\beta H_S(x_\tau, \lambda_\tau)}} \tag{3.73}$$

$$= \beta W + \ln \frac{Z_\tau}{Z_0} \tag{3.74}$$

$$= \beta(W(\mathbf{x}) - \Delta F) = W_d \tag{3.75}$$

The result is a class conditioned version of the Jarzynski equality estimation of the free energy, containing a correction term compensating for the incomplete integration:

$$\langle e^{-W_d} \rangle_C = \frac{R(C)}{P(C)} \tag{3.76}$$

$$e^{\beta\Delta F} \langle e^{-\beta W} \rangle_C = \frac{R(C)}{P(C)} \tag{3.77}$$

$$\beta\Delta F = -\ln\langle e^{-\beta W} \rangle_C - \ln\frac{P(C)}{R(C)}\ . \tag{3.78}$$

The takeaway is that we a fluctuation theorem can estimate the equilibrium free energy in a process where it is not possible to sample all trajectories. The tradeoff when compared to the JE is a need for additional information: the probabilities of the trajectories in both the forward and reverse process. This opens up the ability to avoid regions that might suffer from the existence of very rare but very dominant events, which will be discussed more in section 3.4

3.3.4. Nonequilibrium Free Energy.

The result in section 3.3.3 is actually a special case of a more general results. For a larger class of processes, the trajectory wise entropy generation can be written as $\Sigma(\mathbf{x}) = \beta(W(\mathbf{x}) - \Delta f(\mathbf{x}, \rho_\tau, \rho_0))$ where we have defined a pointwise free energy $f(x, \rho) = E(x) + \beta^{-1}\ln\rho(x)$ and a trajectory wise free energy difference $\Delta f(\mathbf{x}, \rho_\tau, \rho_0) = f(x_\tau, \rho_\tau) - f(x_0, \rho_0) = \Delta E(\mathbf{x}) + \beta^{-1}\ln\frac{\rho_\tau(x_\tau)}{\rho_0(x_0)}$. These free energies are often referred to as 'nonequilibrium free energy' because they define a free energy over arbitrary distributions. In the case above, the normal equilibrium free energy falls out because $F_\lambda = -\beta^{-1}\ln Z_\lambda = f(x, \pi_\lambda)$. However, we can also ask questions about the nonequilibrium free energy difference proper. From this perspective, as long as we choose a class for which all trajectories have the same nonequilibrium free energy difference $\Delta f(\mathbf{x}, \rho_f, \rho_0) = \Delta f_c \ \forall \ \mathbf{x} \in C$ the TCFT result from equation 3.76 will apply:

$$\beta\Delta f_c = -\ln\langle e^{-\beta W} \rangle_C - \ln\frac{P(C)}{R(C)}$$

An interesting extension assumes that we partition all trajectories in classes defined by their Δf, and then average over all classes:

$$\beta\langle\Delta f\rangle = \beta\sum_{c} P(C)\Delta f_C \tag{3.79}$$

$$= -\sum P(C)\ln\langle e^{-\beta W}\rangle_C + \sum P(C)\ln\frac{P(C)}{R(C)} \tag{3.80}$$

$$= -\ln\langle e^{-\beta W}\rangle + \sum P(C)\ln\frac{P(C)}{R(C)} \tag{3.81}$$

$$= \beta\Delta F + \sum P(C)\ln\frac{P(C)}{R(C)} \tag{3.82}$$

$P(C)$ and $R(C)$ are two distributions over the set of trajectory classes and so the last term can be written as $D_{KL}(P||R)$. Thus the equilibrium free energies can be written in terms of a weighted average over all possible non-equilibrum free energies and coarse-grained statistics on the partitioning. While it isn't, at this point, clear how to access f experimentally in a general way, it can at least be argued that this tells us the average $\langle\Delta f\rangle$ is bounded from below by the change in equilibrium free energy through the positivity of the D_{KL}.

3.3.5. Metastable Free Energy. A notable set of processes/classes that allow for easy partitioning into classes of constant and physically interpretable Δf's are processes that begin and end in metastable distributions. A metastable distribution means that the state space can be separated into metastable regions inside of which the distribution appears equilibrium– though it need not be globally in equilibrium. This is typical in a 'computational' physical setting– which requires the coexistence of multiple distinguishable, long lasting, informational states m. Recall that the memory states m are defined by a coarse graining on the space of X by a partitioning the full state space into subsets $\{X_m\}$. A metastable distribution can be described by a set of $\rho^m(x) = w^m\pi^m(x)$ where $\sum_m w^m = 1$. The $\pi^m(x) = \frac{1}{Z^m}e^{-\beta E(x)}$ are the local equilibrium distributions and are unique for a given process. w_i, the relative weights of the regions, are not unique; they are only subject to the normalization constraint. Note that these Z^m define 'local free energies' $F^m = -\beta^{-1}\ln Z^m$. The probability of a point in phase space, x, in a metastable distribution is $\rho(x) = \delta_{xm}\rho^m(x)$ (here δ_{xm} is the Iverson bracket $\delta_{xm} = [x \in \{X_m\}]$).

If a process begins in metastable distribution $\rho_{\lambda_0}(x,\{w\})$ and ends in a metastable distribution $\rho_{\lambda_\tau}(x,\{u\})$, the nonequilibrium free energy difference for a trajectory that starts in metastable region i and ends in metastable region j is:

$$\Delta f_{ij} = E(x_\tau) + \beta^{-1}\ln u^j \pi^j(x_\tau) - E(x_0) - \beta^{-1}\ln w^i \pi^i(x_0) \tag{3.83}$$

$$= \beta^{-1}\ln\frac{u^j}{Z^j} - \beta^{-1}\ln\frac{w^i}{Z^i} \tag{3.84}$$

Importantly, the trajectory dependence is gone so this is a case of a class of constant free energy difference.

$$\beta\Delta f_{ij} = \ln\frac{u^j}{w^i} + \beta\ln\frac{Z^i}{Z^j} \tag{3.85}$$

$$= \beta\Delta F^{ij} + \ln\frac{u^j}{w^i} \tag{3.86}$$

Here, the explicit definition of ΔF is $\Delta F_{ij} = F^j_{\lambda_\tau} - F^i_{\lambda_0}$, a difference of 'local free energies' for the memory states i and j for any process with the given control protocol. The quantity Δf_{ij} we call the difference in metastble free energy. Metastable free energy being defined as $f^m \equiv F^m + \beta^{-1}\ln w^m$. These metastable free energies are a special case of the nonequilibrium free energy– which means it tells you the minimum average work it takes to drive an ensemble of particles in the initial well with the given initial weight to the second well with the second weight. Plugging this case into equation 3.76 yields:

$$\beta\Delta F_{ij} = -\ln\langle e^{-\beta W}\rangle_{ij} - \ln\frac{P(C_{ij})}{R(C_{ij})} - \ln\frac{u^j}{w^i}\ . \tag{3.87}$$

An important thing to note here is that while $R(C_{ij})$ means the probability of $C^\dagger_{ij}$ in the reverse process,t the initial distribution of the reverse process is undetermined. Determining this is absolutely crucial and different choices will give different results. One such example is setting the initial weights of the forward and reverse process such that $\ln\frac{P(C_{ij})}{R(C_{ij})} + \ln\frac{u^j}{w^i} = 0$. In this case, averaging over the local free energies for all possible combinations of memory states will the equilibrium free energy change of the computation. This approach might be be useful to leverage the accuracy of several careful experiments that start in particular memory states rather than attempting to find the entire work distribution all at once.

3.4. Correcting Experimental Errors with the TCFT

In a typical experiment to estimate free energy using the nonequilibrium work distribution, one might perform the following:

(1) initialize a system in equilibrium
(2) drive the system out of equilibrium, measuring the work W needed to do so
(3) repeat the above using the same driving protocol N times
(4) use statistics to find $\bar{JE} = \frac{1}{N}\sum_{i=0}^{N} e^{-\beta W_i} \approx \langle e^{-\beta W}\rangle$ and its variance s_{JE}^2
(5) use $\bar{JE}$ and s_{JE}^2 as estimators for the free energy and the variance of the estimate using the JE .

In section 3.2, some of the pitfalls of this method were discussed. In this section, we compare the baseline approach of free energy estimation outlined above to an augmented method that requires also an experiment of the reverse process and uses the TCFT (equation 3.76):

$$\beta\Delta F = -\ln\langle e^{-\beta W}\rangle - \ln\frac{P(C)}{R(C)} \ , \tag{3.88}$$

to estimate the free energy. Rather than analyze the cases analytically, we will turn to numerics here. A simple protocol will be outlined, and then a large ensemble of realizations will be simulated.

Since the exact free energy is known, it will be a simple matter to estimate it through different methods and compare the methods. Before talking specifically about an example system, we briefly touch on error estimation in these systems. In a general sense, when using the TCFT, there are three sources of error instead of just one because $P(C)$ and $R(C)$ need to be estimated as well. However, since these are sample proportions rather than sample means– their error scaling is straightforward. Using the formula $\sigma^2_{\ln x+\ln y} = \left(\frac{\sigma_x}{\langle x\rangle}\right)^2 + \left(\frac{\sigma_y}{\langle y\rangle}\right)^2$ for two uncorrelated variables, we assume that

$$\sigma^2_{\Delta F} = \left(\frac{\sigma^2_{JE}}{\bar{JE}}\right)^2 + \left(\frac{\sigma^2_{P}}{\hat{P}(C)}\right)^2 + \left(\frac{\sigma^2_{R}}{\hat{R}(C)}\right)^2 \tag{3.89}$$

where the variances for each variable are calculated using the standard formulas for sample means and standard deviations for variables that satisfy the central limit theorem. Using the full JE is akin to choosing the class of all trajectories. The benefit is that there is no uncertainty in $P(C)$ or $R(C)$ as they both equal 1 with certainty; the downside is that you must take an average over all possible trajectories– which includes the exponentially rare, but dominant events.

3.4.1. The Untilt Process. In order to provide a simple example of a metastable process, we investigate an 'untilt' control protocol. Using 1D Langevin dynamics, as described in section 2.3, we set the potential energy to have only two potential energy minima located at $x = \pm x_0$. At $t = 0$, the well corresponding to the memory state $m_x = 1 (m_x = 0)$ (recall that our convention sets this memory state to be defined by the region $x > 0 (x < 0)$) will have a depth of $U_1 (U_0)$. Initially, $U_1 > U_0$ which means that the equilibrium distribution at this time favors being in the right well. As time progresses form $t = 0$ to $t = \tau$, the right well is raised linearly with time until $U_1 = U_0$ (see figure for the energy landscapes and equilibrium distributions 3.4.1). This is done on a timescale that is faster than the time it would take for the wells to equilibrate between each other, but slow enough that the local distributions conditioned on being in each well are not perturbed very far away from equilibrium. The reverse simulation, following section 3.3.3 will begin in the equilibrium distribution associated with the $U_1 = U_0$ energy landscape– and then follow the reverse protocol of lowering the energy of the right well.

Our first numerical experiment will go as follows:

(1) Sample a large ensemble of initial conditions from the equilibrium distribution for both the forward and reverse processes.
(2) Simulate the the system being driven out of equilibrium in both processes, and measure the work W needed to do so for each trajectory.
(3) Take a sample out of of a specific size n from the forward process trajectories
(4) Use the samples statistics to estimate the free energy change during the protocol using the JE and the forward process only, as described above
(5) Take a sample out of of a specific size n from the reverse process trajectories
(6) Use the samples' statistics to estimate the free energy change during the protocol using the TCFT (equation 3.76) under a suite of different trajectory classes

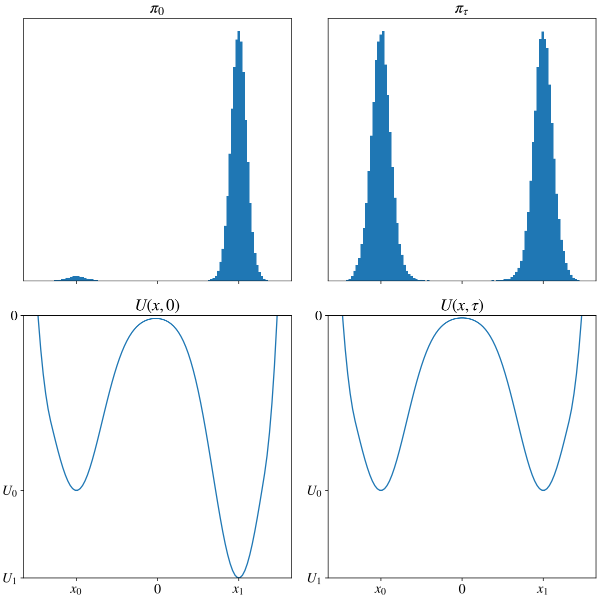


FIGURE 3.4.1. The initial and final equilibrium distributions and potential energy profiles for the untilt process; under the exchange of $\tau \leftrightarrow 0$, the same but for the reverse untilt process.

(7) Repeat steps 3-5 with another value of n to observe how the different estimates converge

The process above was carried out with an ensemble of 500,000 trajectories and n =1,000, 4,000, 20,000, and 100,000. Figure 3.4.2 displays the results. With the intent of showing how the various possible trajectory classes behave in a stripped down example, a suite of different trajectory classes are investigated:

(1) The most basic trajectory class 'jarz' is the set of all trajectories. This trajectory class is simply using the traditional JE, and gains nothing from reverse process.
(2) v_{i+} trajectories for which the initial velocity is positive, and its compliment v_{i-} for which initial velocities are negative. The conjugate classes for these two are $v_{i+}^\dagger = v_{f-}$ and $v_{i-}^\dagger = v_{f+}$. Because both processes begin and end in equilibrium, these classes are not expected to be advantageous– there is no coupling between position and velocity in the equilibrium distribution so we would expect these two classes to have the same attributes as the entire trajectory set– but with approximately half the sample size. These classes are expected to be outperformed by the traditional JE in every way. They should be similarly

unbiased as estimators, but there is nothing to be gained from using this class instead of the full set of outcomes

(3) $x_{i<}$ and its compliment $x_{i>}$ (and their conjugate classes $x_{f<}$ and $x_{f>}$) represent trajectories that start close(far) from the minimal of their local potential well. The idea of these classes is to separate initial conditions that are typical of the equilibrium distribution and those that represent large fluctuations in energy. In the figures that follow, the threshold was set so that the typical class contained 68% of initial conditions. It is not clear which of these should provide better estimates. On the one hand, one might think that the larger energy fluctuations explore more of the state space and so can provide more information about the energy landscape. On the other hand, the sample size is smaller and the proportions will be far from the case of $\hat{p} = .5$, for which the variance is the smallest.

(4) $x_{max<}$ and its compliment $x_{max>}$ have the same conditions as the previous two classes, but applied over the entire trajectory rather than just the initial point. Rather than separating into trajectories that start out as large fluctuations or not, the separation is into trajectories that don't experience any unusual fluctuation and those that do (note that the conjugate classes for these two are the same as the forward classes.) Keeping the same threshold as above yielded approximately 20% of trajectories in the in first and 80% in the second. Thus, the issue of the larger fluctuations having a small probability is ameliorated. Therefore, relative to the analogous classes above we can expect each to perform worse and better respectively.

(5) $m_{i \neq f}$: this trajectory class is the set of 'transitional' trajectories that end in a different memory state than they start in. The conjugate class is the same as the original class since the class contains two event $m_i = 0, m_f = 1$ and $m_i = 1, m_f = 0$ that map to each other under $C^{\dagger}$ which swaps the i and f but not the memory state (which is a coarse graining of a time symmetric variable.) Because of the protocol design, the probability of such a trajectory will be very low in the forward process. This should make it nearly impossible to get good statistics. This class implies two complementary classes 0 and 1, which are the two classes that stay in the well they begin in for the entire protocol.

(6) The final trajectory class, $W_>$ contains trajectories that have a work cost higher than a particular threshold, in this case $W > .8k_{\mathrm{B}}T$ which makes up about 96% of simulated trials. Its conjugate class is the set of trajectories that cost work less than the negative of the threshold $W < -.8k_{\mathrm{B}}T$ and makes up about 50% of the trajectories in the reverse process. This class intentionally neglects the exponentially damped work outcomes that have very negative values, in an attempt to avoid the problems of rare event dominance. Crucially, even though these events dominate the average– the TCFT should allow an 'un-biasing' of the estimate given proper statistics on the reverse protocol.

To really appreciate the TCFT, it is informative to focus in on a class like the 0 class. A remarkable feature of the 0 class, is that it is an incredibly unlikely set. In the forward process, the probability that a trajectory occupies only the 0 state for the entire protocol is only about 2%. However, provided that a good estimate can be made of the class probability in the reverse process (it is known to be $\approx .5$ from the setup of the untilt process, but it is estimated by the reverse process statistics in plots anyway), a mere 2% of trajectories can form an unbiased estimator of the free energy. This can be seen in figure 3.4.3 which shows a histogram of 100 different attempts at estimating the free energy using only 1000 trajectories each time. While using all trajectories will provide a lower variance between different attempts, the fact that such a highly biased and unlikely trajectory class can provide an unbiased estimate is staggering. While one might expect the TCFT to do little for the 1 class, which already makes up close to 98% of all trajectories, figure 3.4.3 shows the variance when using the TCFT is significantly lower, even though both estimates appear to be unbiased. A counter argument is that the TCFT uses twice as many trajectories to estimate (since information is needed about two different processes), but the reduction in variance is more than what would be expected by using twice as many trajectories in the forward process and ignoring the TCFT. In that case, the error would be expected to shrink by a factor of $\frac{1}{\sqrt{2}} \approx .707$; however, the use of the TCFT shrinks the variance by $\frac{.0907}{.0295} \approx .325$.

The analysis above is quite informative, but knowing which trajectory classes need the least statistics to be accurate does not on its own translate to a better experiment. For example, it might not be possible to select whichever class of trajectories is best. Instead, a more typical setup might be that a measurement device itself measures only a particular trajectory class which is not

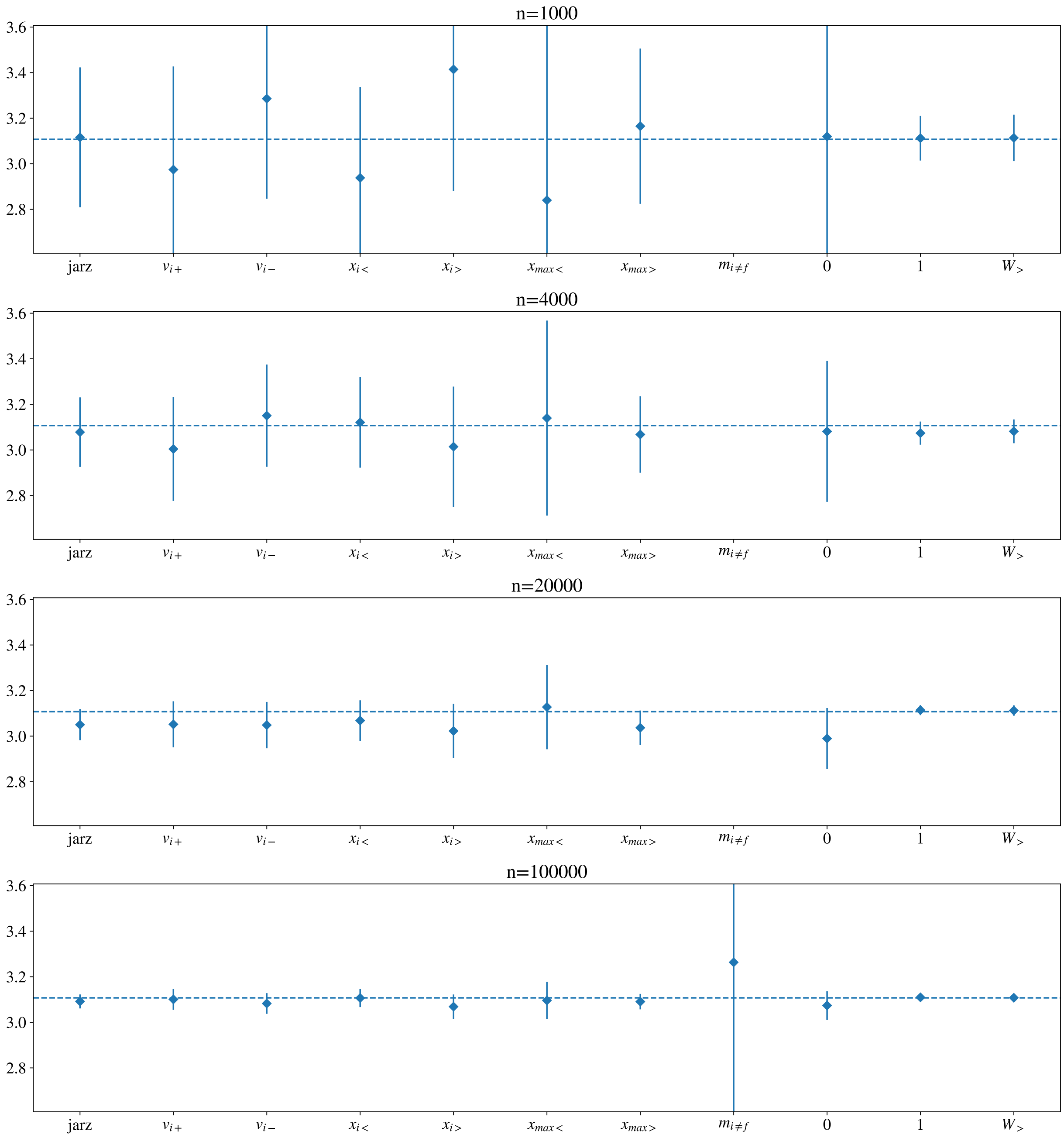
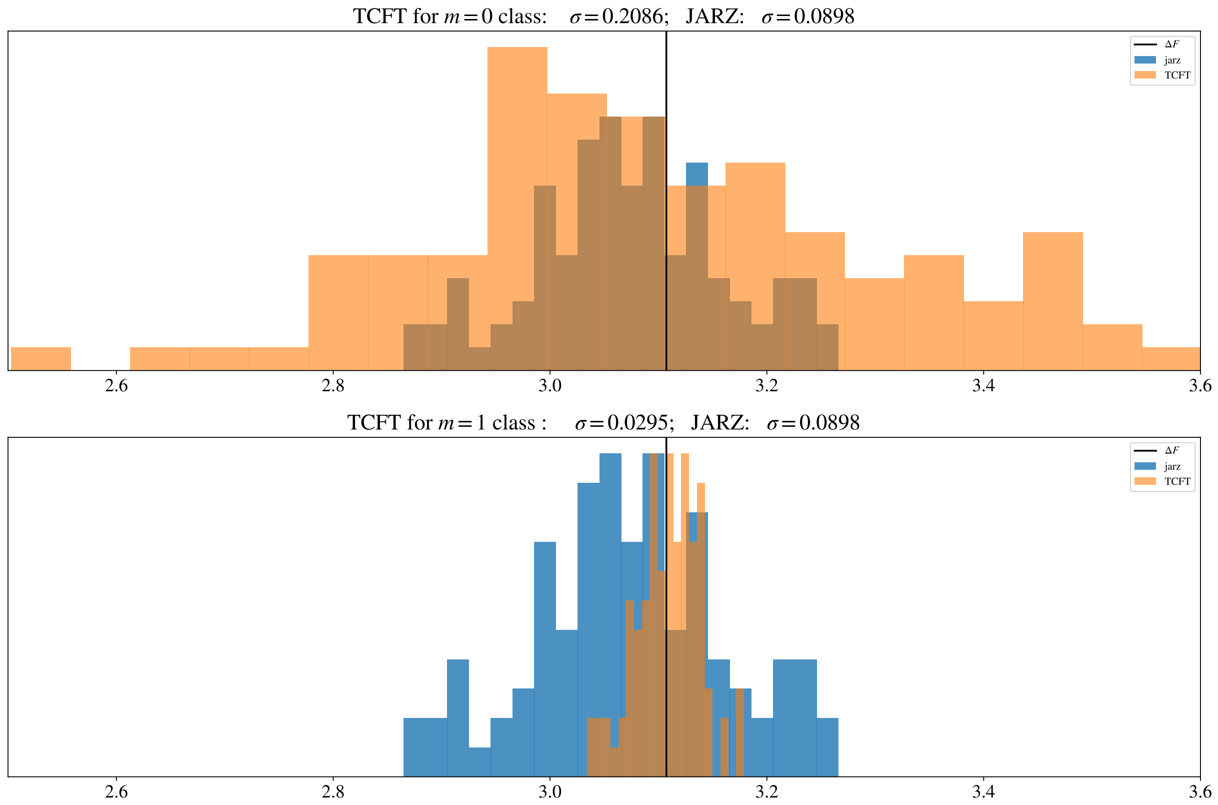


FIGURE 3.4.2. The results of applying the TCFT to the suite of trajectory classes described in the text. We can see that each class behaves in accordance with the expected behavior. Not also that in every single case, an unbiased estimator is formed, even if the trajectory class in question is very biased. Remarkably, there exist trajectory classes that work better to estimate the free energy than just using all trajectories. The two best trajectories are 1 and $W_>$, which have high probabilities in the forward process without being especially rare in the reverse process. This allows for the benefit of being able to ignore rare forward process events without offsetting the gain in uncertainty with a high uncertainty in $R(C)$. The transitional class $m_{i\neq f}$ is so rare that the a sample is not likely to even contain a single trajectory in both the forward and reverse processes until $n \approx 100,000$.



TCFT for $m = 0$ class: $\sigma = 0.2086$; JARZ: $\sigma = 0.0898$

TCFT for $m = 1$ class : $\sigma = 0.0295$; JARZ: $\sigma = 0.0898$

FIGURE 3.4.3. Detailed study of two classes from figure 3.4.2, (top) the class that always stays in the left well 0 and (bottom) the class that always stays in the right well 1. For each plot, 1000 trajectories were taking 100 different times from a pool of 500,000 trajectories and the resulting estimate of the free energy was plotted using both the JE and all trajectories as well as the TCFT and just the trajectories that belonged to the class in question.

controlled by an experimenter so much as an inherent bias of the machine. For example, perhaps work values that fall under a certain threshold cannot be measured well because the measurement instrument is not sensitive to a work value that s much less than $k_{\mathrm{B}}T$. Or, large energy fluctuations escape the detector and thus are ignored. These types of measurement shortcomings can cause additional bias uncertainty in the free energy estimate. From this perspective, the TCFT provides an invaluable tool take charge of these measurement errors– reversing the effect of missing some trajectories. The question is slightly different here, so the experiment will also differ slightly. Instead of focusing on the scaling behavior of TCFT estimates across different classes and sample sizes,

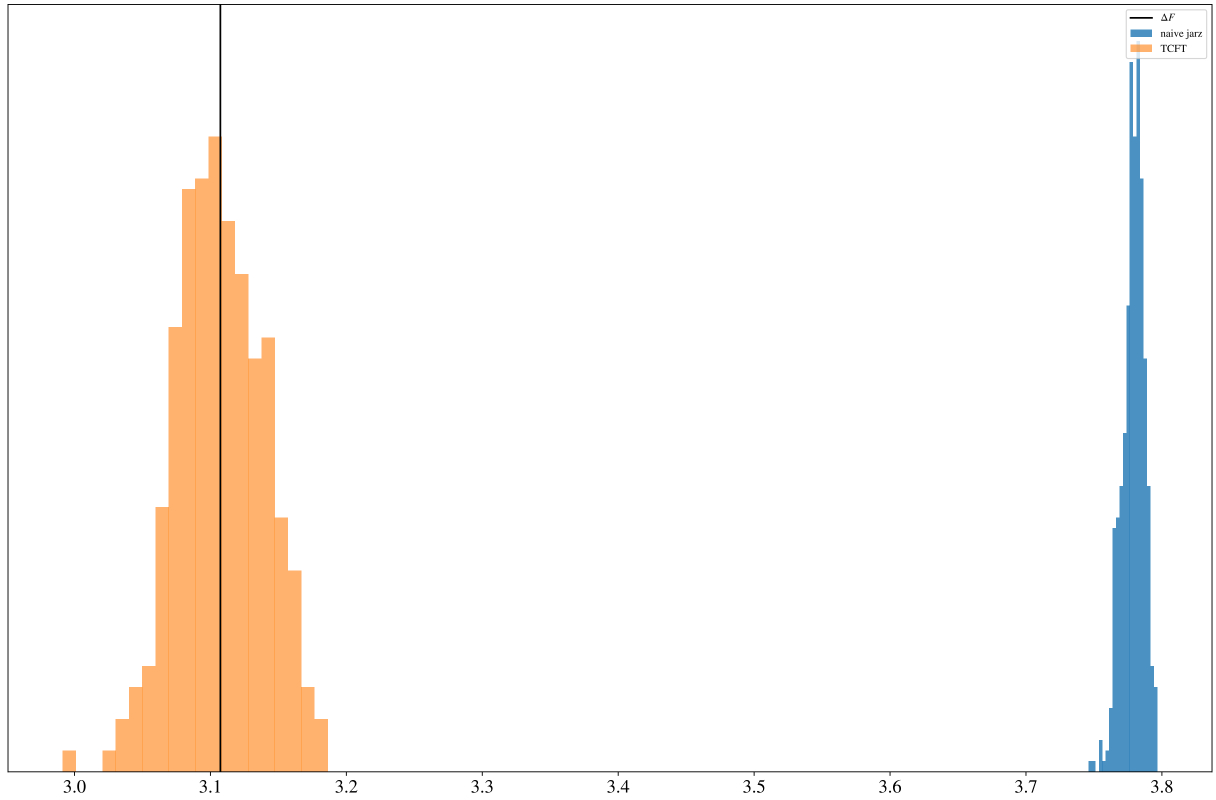


FIGURE 3.4.4. A histogram of taking 100 samples of 1000 trajectories each from a pool of 500,000 trajectories. In each case, a 'naive Jarzynski' estimate was performed by throwing away the trials that did not fit in the class 0, and then using the JE to estimate the free energy. The 'TCFT' estimate was made by simulating the reverse trajectory and estimating the probability of the 0 class.

the focus will be on the TCFT as a correction to JE in the absence of an unbiased measurement device.

A first example will be the following: suppose the measurement device ignored every trajectory in the forward process that ever occupied the 0 state. In this case, the experiment is still able to measure about 98% of all trials– and so it still might seem reasonable that it wont bias the free energy estimate too much. However, knowing that even rare events can dominate the free energy estimate, this might not be the case. Ignoring the TCFT, an experiment might use the faulty data– knowing that a bias was inherent in the machine but unable to account for it. However, in this case the class that is missed by the machine is symmetric with respect to $C^\dagger$ conjugation, so if the probability of failing to measure a trajectory in the reverse process is estimated– the TCFT can be used to completely correct the shortcomings of the machine. Figure 3.4.4 displays this by showing two different distributions. In one, the impoverished data is used to apply the JE to find the free energy and in the other and estimate of a measurement failure in the reserve process is estimated as well (by simulation) and then the TCFT is used. It is immediately obvious that the missing 2% of trajectories provides an enormous bias to the result, but that the TCFT is able to reverse the bias even with relatively few trials. The additional error in the TCFT corrected distribution comes from having to estimate both $P(C)$ and $R(C)$.

3.5. Thermodynamic Uncertainty Theorem

Fluctuation relation symmetries have proven themselves useful in estimating free energies, but they can also be used for other interesting applications. One common recent focus is on a class of results termed Thermodynamic Uncertainty Relations (TURs) [**56**]. The chapter to come will discuss these at length, but here we show how a type of fluctuation theorem symmetry can lead to a new theorem: the Thermodynamic Uncertainty Theorem.

Define current J to be observations of a system $\mathcal{S}$ that flip sign under time reversal. Formally, a system trajectory is the sequence $\vec{s} \equiv s_0 s_{dt} \cdots s_{\tau - dt} s_\tau$, where each $s_t \in \mathcal{S}$ is the system's state at time t. The current associated with a reversed trajectory $R(\vec{s}) \equiv s^\dagger_\tau s^\dagger_{\tau-dt} \cdots s^\dagger_{dt} s^\dagger_0$ is minus the current of the forward trajectory:

$$J(R(\vec{s})) = -J(\vec{s}). \tag{3.90}$$

If the system $\mathcal{S}$ is influenced by an external control parameter $\lambda^\dagger_{\tau-t} = \lambda_t$ and the protocol conjugates the distribution under the operation $p_\tau(s^\dagger) = p_0(s)$ then the probability of a reverse trajectory is exponentially damped by the entropy production [**66**]:

$$\Pr(R(\vec{s}), -\Sigma) = e^{-\Sigma} \Pr(\vec{s}, \Sigma). \tag{3.91}$$

This is the Detailed Fluctuation Theorem (DFT) for a Time-Symmetrically Controlled Computation (TSCC) [**67**, **68**, **69**]. It includes non equilibrium steady state (NESS) dynamics for which the control parameter is constant $\lambda_t = \lambda_{t'}$. It can also describe, as explored here, computations that begin in equilibrium and are then allowed to relax after the application of a time-symmetric control signal. These latter symmetries are ubiquitous in computing [**70**].

The symmetry imposed by the TSSC imbues J's statistics with special properties in stochastic nonequilibrium systems when compared with the entropy production Σ. Namely, the TURs related the entropy production to the scaled variance $\epsilon^2(J) = \mathrm{var(J)}/\langle J \rangle^2$. Given a TSCC operating over the time interval $[0, \tau]$, described by probability distribution $\Pr(\vec{s}, \Sigma)$ over state trajectories $\vec{s}$ and entropy productions Σ, our task is to find a current function $J(\vec{s}) = -J(R(\vec{s}))$ of the state trajectories $\vec{s}$ that minimizes this scaled variance.

3.5.1. Proof. Currents are functions of state trajectories, so they can be applied to the distribution that characterizes the TSCC $\Pr(\vec{s}, \Sigma)$ to derive a three-variable joint distribution over currents, state trajectories, and entropy productions:

$$\Pr(J, \vec{s}, \sigma) \equiv \delta_{J, J(\vec{s})} \Pr(\vec{s}, \Sigma).$$

From this, one can determine a variety of marginal distributions. Most relevant to us is the joint probability of currents and entropy productions $\Pr(J, \Sigma)$, and the probability of a current given an entropy production $\Pr(J|\Sigma)$. From the latter, we define the entropy-conditioned current $j(\Sigma)$ as the average current in the system given that the entropy Σ was dissipated in the NESS:

$$j(\Sigma) \equiv \sum_J J \Pr(J|\Sigma). \tag{3.92}$$

If the entropy production is a function of the state trajectory, as is the case for systems satisfying local detailed balance, then we can use the entropy conditioned current to define a new function of the trajectories

$$J'(\vec{s}) \equiv j(\Sigma(\vec{s})). \tag{3.93}$$

J' is a well-defined current within the system, because $J'(R(\vec{s})) = -J'(\vec{s})$, as shown in Appendix 3.A.

The newly defined entropy-conditioned current has the convenient property that it's average $\langle J' \rangle \equiv \langle j \rangle$ is the same as for the current that was used to define it

$$\begin{aligned}
\langle J' \rangle &= \sum_{\Sigma} \Pr(\Sigma) j(\Sigma) \\
&= \sum_{\Sigma} \Pr(\Sigma) \sum_{J} J \Pr(J|\Sigma) \\
&= \sum_{\Sigma, J} J \Pr(J, \Sigma) \\
&= \sum_{J} J \Pr(J) \\
&\equiv \langle J \rangle .
\end{aligned}$$

On the other hand, the variance of the entropy-conditioned current is not the same. When we evaluate the average square of the newly defined current

$$\begin{aligned}
\langle J'^2 \rangle &= \sum_{\sigma} \Pr(\sigma) j(\sigma)^2 \\
&= \sum_{\sigma} \Pr(\sigma) \left(\sum_{J} J \Pr(J|\sigma) \right)^2 ,
\end{aligned}$$

we can apply Jensen's inequality $\left(\sum_J J \Pr(J|\sigma)\right)^2 \leq \left(\sum_J J^2 \Pr(J|\sigma)\right)$ to show that

$$\begin{aligned}
\langle J'^2 \rangle &\leq \sum_{\sigma} \Pr(\sigma) \left(\sum_{J} J^2 \Pr(J|\sigma) \right) \\
&= \sum_{J, \sigma} \Pr(J, \sigma) J^2 \\
&= \langle J^2 \rangle .
\end{aligned}$$

As a result, the entropy-conditioned current produces scaled variance that is less than or equal to the current that was used to define it

$$\epsilon_J^2 \geq \epsilon_{J'}^2 . \tag{3.94}$$

Thus, we need only consider currents which are functions of the entropy production in order to find the minimum-variance current.

Given that any current's scaled variance can be reduced by finding its corresponding entropy-conditioned current, given some TSCC process $\Pr(\vec{s}, \Sigma)$, we need only find the function $j(\Sigma)$ that minimizes the scaled variance $\epsilon_j^2 = \frac{\langle j^2 \rangle}{\langle j \rangle^2} - 1$. There is a single important constraint that applies to these functions, which is that $j(-\Sigma) = -j(\Sigma)$, meaning that this is a constrained optimization.

However, we can ignore this constraint by using the TSCC DFT and summing over the state trajectories of Eq. 3.91 to produce:

$$\Pr(-\Sigma) = e^{-\Sigma} \Pr(\Sigma), \tag{3.95}$$

where we've set Boltzmann's constant to $k_B = 1$ for ease of notation. We use this to express the average j and j^2 can in terms of positive entropy productions

$$\begin{aligned} \langle j^2 \rangle &= \sum_{\Sigma > 0} \Pr(\Sigma)(1 + e^{-\Sigma}) j(\Sigma)^2 \\ \langle j \rangle &= \sum_{\Sigma > 0} \Pr(\Sigma)(1 - e^{-\Sigma}) j(\Sigma). \end{aligned} \tag{3.96}$$

This means that ϵ_j^2 can be expressed in terms of only positive entropy productions. $j(\Sigma)$ is *unconstrained* over positive Σ, so the minimum occurs when

$$\begin{aligned} \frac{\partial}{\partial j(\Sigma)} \epsilon_j^2 &= \frac{1}{\langle j \rangle^2} \left(\frac{\partial \langle j^2 \rangle}{\partial j(\Sigma)} - \frac{2 \langle j^2 \rangle}{\langle j \rangle} \frac{\langle j \rangle}{\partial j(\Sigma)} \right) \\ &= 0 \text{ for all } \Sigma > 0. \end{aligned} \tag{3.97}$$

Applying the derivative with respect to the positive-entropy current $j(\Sigma)$ to the averages shown in Eq. 3.97 yields

$$\begin{aligned} \frac{\partial \langle j \rangle}{\partial j(\Sigma)} &= (1 - \epsilon^{-\Sigma}) \Pr(\Sigma) \\ \frac{\partial \langle j^2 \rangle}{\partial j(\Sigma)} &= (1 + \epsilon^{-\Sigma}) \Pr(\Sigma) 2 j(\Sigma). \end{aligned}$$

Finally, plugging these into Eq. 3.97, we solve for the current with the minimum scaled variance

$$\boxed{j_{\min}(\Sigma) = \frac{\langle j^2_{\min} \rangle}{\langle j_{\min} \rangle} \frac{1 - e^{-\Sigma}}{1 + e^{-\Sigma}}}, \tag{3.98}$$

which applies as long as Σ is in the support of the entropy distribution. Note that, even though the expression for the minimal current was derived for $\Sigma > 0$, it applies to $\Sigma \leq 0$ as well, because of the condition that a current must satisfy $j(-\Sigma) = -j(\Sigma)$. Indeed, $j_{\min}(0) = 0$ and $j_{\min}(-\Sigma) = \frac{\langle j^2 \rangle}{\langle j \rangle} \frac{1-e^{\Sigma}}{1+e^{\Sigma}} = \frac{\langle j^2 \rangle}{\langle j \rangle} \frac{e^{-\Sigma}-1}{e^{\Sigma}+1} = -j_{\min}(\Sigma)$. Thus, we have found the form of the current that minimizes scaled variance, and it depends exclusively on the entropy production of the process

$$J_{\min}(\vec{s}) = j_{\min}(\Sigma(\vec{s})).$$

For simplicity, note that we can choose any real value for $k = \langle j^2_{\min} \rangle / \langle j_{\min} \rangle$, and the entropy-conditioned current $j_{\min}(\Sigma)$ will minimize the scaled variance. Moreover, $\frac{1-e^{-\Sigma}}{1+e^{-\Sigma}} = \tanh(\Sigma/2)$. Whatever the entropy production distribution $\Pr(\Sigma)$ may be, the current

$$J_{\min}(\vec{s}) = k \tanh(\Sigma(\vec{s})/2), \tag{3.99}$$

will set a lower bound on the all other currents of the system

$$\begin{aligned} \epsilon^2_J &\geq \epsilon^2_{J\min} \\ &= \frac{\langle \tanh(\Sigma/2)^2 \rangle}{\langle \tanh(\Sigma/2) \rangle^2} - 1, \end{aligned} \tag{3.100}$$

where the constant $k = \langle j^2 \rangle / \langle j \rangle$ has factored out. This bound on the scaled variance is *tight*, because it is realized by our newly defined $J_{\min}(\vec{s})$. For this reason, $\epsilon^2_{J_{\min}}$ represents the tightest possible bound on the scaled variance for the process $\Pr(\Sigma, \vec{s})$.

Once again, the TSCC DFT ($\Pr(\Sigma) = e^{\Sigma}\Pr(-\Sigma)$) simplifies:

$$\begin{aligned}\langle \tanh(\Sigma/2)^2 \rangle &= \sum_{\Sigma>0} \Pr(\Sigma)(1+e^{-\Sigma})\left(\frac{1-e^{-\Sigma}}{1+e^{-\Sigma}}\right)^2 \\ &= \sum_{\Sigma>0} \Pr(\Sigma)(1-e^{-\Sigma})\left(\frac{1-e^{-\Sigma}}{1+e^{-\Sigma}}\right) \\ &= \langle \tanh(\Sigma/2) \rangle. \end{aligned} \tag{3.101}$$

As a result, we have the simplified bound on the scaled variance in terms of the entropy production.

$$\boxed{\epsilon_J^2 \geq \epsilon_{J\min}^2 = \frac{1}{\langle \tanh(\Sigma/2) \rangle} - 1}. \tag{3.102}$$

3.5.2. Testing the theory. The TUT, equation 3.102, provides an intriguing and theoretically saturable bound that relates the precision of a current to the entropy generation that underlies the dynamics of a system. While the bound is indeed, saturable in theory– it is quite a bit less clear what kind of physical systems display entropy production distributions for which this bound is useful. It turns out that the answers to these questions are rather more subtle than might be expected; in discussing, they reveal the possibility of alternative computational frameworks that outcompete the status quo significantly. As such: in order to really answer the question, a more detailed chapter discussing the past, present, and future of TURs will be necessary.

Appendix

3.A. Well-defined current

Let us define function J' of state trajectories in terms of the entropy-conditioned current

$$J'(\vec{s}) \equiv j(\Sigma(\vec{s})),$$

where the entropy conditioned current is defined

$$j(\Sigma) \equiv \sum_J J \Pr(J|\Sigma).$$

Note that we can evaluate J' for the time-reversal of a trajectory

$$J'(R(\vec{s})) = j(\Sigma(R(\vec{s}))).$$

The TSCC fluctuation theorem

$$\Pr(R(\vec{s}), -\Sigma) = e^{-\Sigma} \Pr(\vec{s}, \Sigma),$$

implies both a marginalized version,

$$\Pr(-\Sigma) = e^{-\Sigma} \Pr(\Sigma),$$

as well as equality of the conditional probabilities

$$\Pr(R(\vec{s})| - \Sigma) = \Pr(\vec{s}|\Sigma).$$

Thus, the entropy conditioned current is an odd function of the entropy

$$\begin{aligned}
j(-\Sigma) &= \sum_J J \Pr(J| - \Sigma) \\
&= \sum_{J,\vec{s}} J \Pr(J, \vec{s}| - \Sigma) \\
&= \sum_{J,\vec{s}} J\delta_{J,J(\vec{s})} \Pr(\vec{s}| - \Sigma) \\
&= \sum_{\vec{s}} J(\vec{s}) \Pr(\vec{s}| - \Sigma) \\
&= \sum_{\vec{s}} J(R(\vec{s})) \Pr(R(\vec{s})| - \Sigma) \\
&= \sum_{\vec{s}} -J(\vec{s}) \Pr(\vec{s}|\Sigma) \\
&= -j(\Sigma).
\end{aligned}$$

Having assumed that Σ is a function state trajectory, we can re-express the TSSC fluctuation theorem

$$\Pr(R(\vec{s}))\delta_{\Sigma(R(\vec{s})),-\Sigma'} = e^{-\Sigma'} \Pr(\vec{s})\delta_{\Sigma',\Sigma(\vec{s})},$$

This can only be true if the entropy production is itself a current $\Sigma(R(\vec{s})) = -\Sigma(\vec{s})$. Thus, we see that our new function of the trajectories J' is indeed a current as well

$$\begin{aligned}
J'(R(\vec{s})) &= j(\Sigma(R(\vec{s}))) \\
&= j(-\Sigma(\vec{s})) \\
&= -j(\Sigma(\vec{s})) \\
&= -J'(\vec{s}).
\end{aligned}$$

CHAPTER 4

The Thermodynamic Uncertainty Theorem

Since Onsager's and Kubo's pioneering discovery of the fluctuation-dissipation theorem (FDT) [**71**, **72**, **73**], determining the universal properties of fluctuations in out-of-equilibrium processes, as well as their role in dissipation, has been a cornerstone of stochastic thermodynamics. In the 90s, Jarzynski and Crooks generalized the FDT through the *fluctuation relations* (FR) [**65**, **74**, **75**, **76**, **77**, **78**, **79**, **80**, **81**, **82**, **83**, **84**]. At the microscopic scale, the FR refine the famous Second Law of Thermodynamics $\langle \Sigma \rangle \geq 0$ by determining the distribution of thermodynamic fluctuations. That is, the FRs replaced the familiar Second law inequality with an equality from which the Second Law is easily derived through Jensen's inequality.

More recently, a third milestone was crossed by connecting thermodynamic fluctuations out of equilibrium to dissipation. These broad results, called *thermodynamic uncertainty relations* (TURs), were originally discovered in nonequilibrium steady-states of classical time-homogeneous Markov jump-processes satisfying local detailed balance [**56**, **85**]. Today, though, TURs have been generalized to finite-time processes [**86**, **87**, **88**], periodically-driven systems [**89**, **90**, **91**, **92**, **93**, **94**, **95**], Markovian quantum systems undergoing Lindblad dynamics [**96**, **97**, **98**], and autonomous classical [**87**, **99**] and quantum [**99**, **100**, **101**] systems in steady-states close to linear response.

In all these, TURs bound the fluctuations of any (time-reversal anti-symmetric stochastic) thermodynamic quantity J as a function of the *average* entropy production $\langle \Sigma \rangle$:

$$\epsilon_J^2 \equiv \frac{\mathrm{var}(J)}{\langle J \rangle^2} \geq f(\langle \Sigma \rangle) \ . \tag{4.1}$$

with $\langle J \rangle$ and $\mathrm{var}(J) = \langle J^2 \rangle - \langle J \rangle^2$ being the average and variance of J, respectively. In this way, the scaled variance ϵ_J^2 can be seen as the inverse of current J's signal-to-noise ratio. Since f is generally a monotonically-decreasing function, TURs express the trade-off that increased precision in J inevitably comes at the cost of more dissipation. In the previous section, a proof was provided by analyzing the impact of higher statistical moments of the entropy production on the signal-to-noise ratio of thermodynamic currents J. Specifically, it replaces the r.h.s. $f(\langle \Sigma \rangle)$ with $\langle f_{\min}(\Sigma) \rangle$, giving:

$$\epsilon_J^2 \geq \frac{1}{\langle \tanh(\Sigma/2) \rangle} - 1. \tag{4.2}$$

Now, the bound is a functional of the *stochastic* entropy production Σ distribution and so, critically, accounts for all higher moments. It is important to stress that $g(\langle f_{\min}(\Sigma) \rangle)$ appearing in Eq. (4.2) agrees and coincides with a recent result obtained in Ref. [**102**], albeit via a completely different derivation. Our approach, however, focuses on realizing the bound by also finding an explicit expression for the *minimum-variance current* $J_{\min}(\vec{s})$ that saturates Eq. (4.2):

$$\boxed{J_{\min}(\vec{s}) = \frac{\langle J_{\min}^2 \rangle}{\langle J_{\min} \rangle} \tanh\left(\Sigma(\vec{s})/2\right)}. \tag{4.3}$$

This minimum depends sensitively on the entropy production's higher-order fluctuations. In much the same way that fluctuation theorems [**68**, **103**, **104**] reframe the Second Law from an inequality to an equality, the TUT replaces the bounds set by TUR with a saturable equality. Applying the TUT to thermodynamic simulations of fundamental bit swap and reset computations, we now demonstrate that current fluctuations can depart substantially from previous bounds set by TURs.

4.1. Background

Before analyzing examples that saturate (or not) the new bound, it is worthwhile to do a short review on some of the most important TUR bounds in the literature to date. Barato and Seifert [**105**] derived the first thermodynamic uncertainty relation in terms of rates of entropy production. They considered the precision of currents J through the inverse of the signal to noise ratio: the scaled variance. We focus on integrated quantities over the time interval $(0, \tau)$. Ref. [**106**] proves such a "finite time" uncertainty relation. Horowitz and Gingrich set a lower bound on the scaled variance of the time-integrated current J. Remarkably, they found that the precision couldn't be maximized without a corresponding increase in the average entropy production

$$\epsilon_J^2 \geq \frac{2}{\langle \Sigma \rangle}. \tag{4.4}$$

Because Barato and Seifert discovered this form of the bound, we denote it the Barato-Seifert (BS) bound on scaled variance

$$\epsilon_{\text{BS}}^2 \equiv \frac{2}{\langle \Sigma \rangle}. \tag{4.5}$$

Further exploration found that detailed fluctuation theorems [**68**, **104**] can be used to prove modified thermodynamic uncertainty relations. Hasegawa and Van Vu [**107**] used Eq. 3.91 [1] to demonstrate that the scaled variance is bounded below by:

$$\epsilon_J^2 \geq \frac{2}{e^{\langle \Sigma \rangle} - 1} \equiv \epsilon_{\text{HVV}}^2, \tag{4.6}$$

where we have labeled their bound by ϵ_{HVV}^2. Also using the fluctuation theorem, Timpanaro, Guarnieri, Goold, and Landi [**108**] showed another bound using the average entropy production

$$\epsilon_J^2 \geq \text{csch}^2[g(\langle \Sigma \rangle)] \equiv \epsilon_{\text{TGGL}}^2 \tag{4.7}$$

where $g(x)$ is the inverse of $x \tanh(x)$, and we have again labeled the bound for the authors. This bound is tighter than ϵ_{HVV}^2. In fact, it is the tightest possible bound on scaled variance that can be determined from the average entropy [**108**].

All three bounds above, while far from complete list, are functions of average entropy production, so we can consider which set the tightest and most accurate bounds for different TSCCs. It is also worth noting that our results is reminiscent of the recently introduced notion of *hyper-accurate currents* [**109**, **110**], defined as those possessing the maximum signal-to-noise ratio. While in that case, however, the form of these current, whose precision therefore can be used to bound the precision of any other thermodynamic current (thus much alike, in spirit, to Eq. (3.102)) was found within classical and quantum thermoelectrics given a coherent transport modelization in the

[1] Ref. [**107**] makes a slightly different assumption, which is that the computation preserves the initial distribution $p_\tau(s) = p_0(s)$. Their assumption is often sufficient, when time-antisymmetric variables like momentum are thermalized, such as in overdamped Langevin dynamics. However, it is necessary to account for the conjugation of the system state to establish the most general conditions for the DFT.

Landauer-Büttiker formalism, in our work we derive them by imposing the TSCC symmetry on $\Pr(\vec{s}, \Sigma)$.

4.2. Comparison to Past Uncertainty Relations

It is natural to ask for direct comparisons between these different bounds. Comparing these bounds independent of the TUT, one can see in Figs. 4.3.2 and 4.2.1 that they are ordered:

$$\epsilon^2_{\mathrm{BS}}(\langle\Sigma\rangle) > \epsilon^2_{\mathrm{TGGL}}(\langle\Sigma\rangle) > \epsilon^2_{\mathrm{HVV}}(\langle\Sigma\rangle). \tag{4.8}$$

(See also App. 4.A for a proof.) Note that, since $\epsilon^2_{\mathrm{TGGL}}$ and $\epsilon^2_{\mathrm{HVV}}$ were derived from the TSCC DFT, which is our starting point as well, the minimum scaled variance is bounded below by these TURS but not necessarily ϵ^2_{BS}.

With $\epsilon^2_{J\mathrm{min}}$'s exact form determined, though, a natural next question is how close the previous bounds, which all depend only on the average entropy production $\langle\Sigma\rangle$, are to the actual minimum. Fortunately, Timpanaro et al. also showed that a particular bimodal distribution:

$$\Pr_{\min}(\Sigma) \propto \delta(\Sigma - a) + e^{-a}\delta(\Sigma + a)$$

achieves their lower bound $\epsilon^2_{\mathrm{TGGL}}$. This is the simplest distribution satisfying $\Pr(-\Sigma) = e^{-\Sigma}\Pr(\Sigma)$. That is, it consists of a delta function at entropy production $\Sigma = a$ and then contains a mirror of that entropy production at $\Sigma = -a$ reduced by the exponential factor e^{-a}. Any other NESS entropy distribution can be constructed from a superposition of such distributions. We investigate the TUT by exploring a variety of possible distributions. We take a similar strategy, breaking the entropy distribution into the piecewise function:

$$\Pr\big(\Sigma|\mu, \sigma^2\big) = n(\mu, \sigma^2)\begin{cases} e^{\frac{(\Sigma-\mu)^2}{2\sigma^2}} & \text{if } \Sigma \geq 0 \\ e^{\Sigma}e^{\frac{(-\Sigma-\mu)^2}{2\sigma^2}} & \text{if } \Sigma < 0 \end{cases}. \tag{4.9}$$

Here, $n(\mu, \sigma^2) = \int_0^\infty e^{\frac{(\Sigma-\mu)^2}{2\sigma^2}} d\Sigma + \int_{-\infty}^0 e^{\Sigma}e^{\frac{(-\Sigma-\mu)^2}{2\sigma^2}} d\Sigma$ is the normalization factor. In essence, our probability distribution is a normal distribution with average μ and variance σ^2 over the positive interval. And, the TSCC DFT defines the distribution to be $\Pr(\Sigma) = e^{\Sigma}\Pr(-\Sigma)$ on the negative interval.

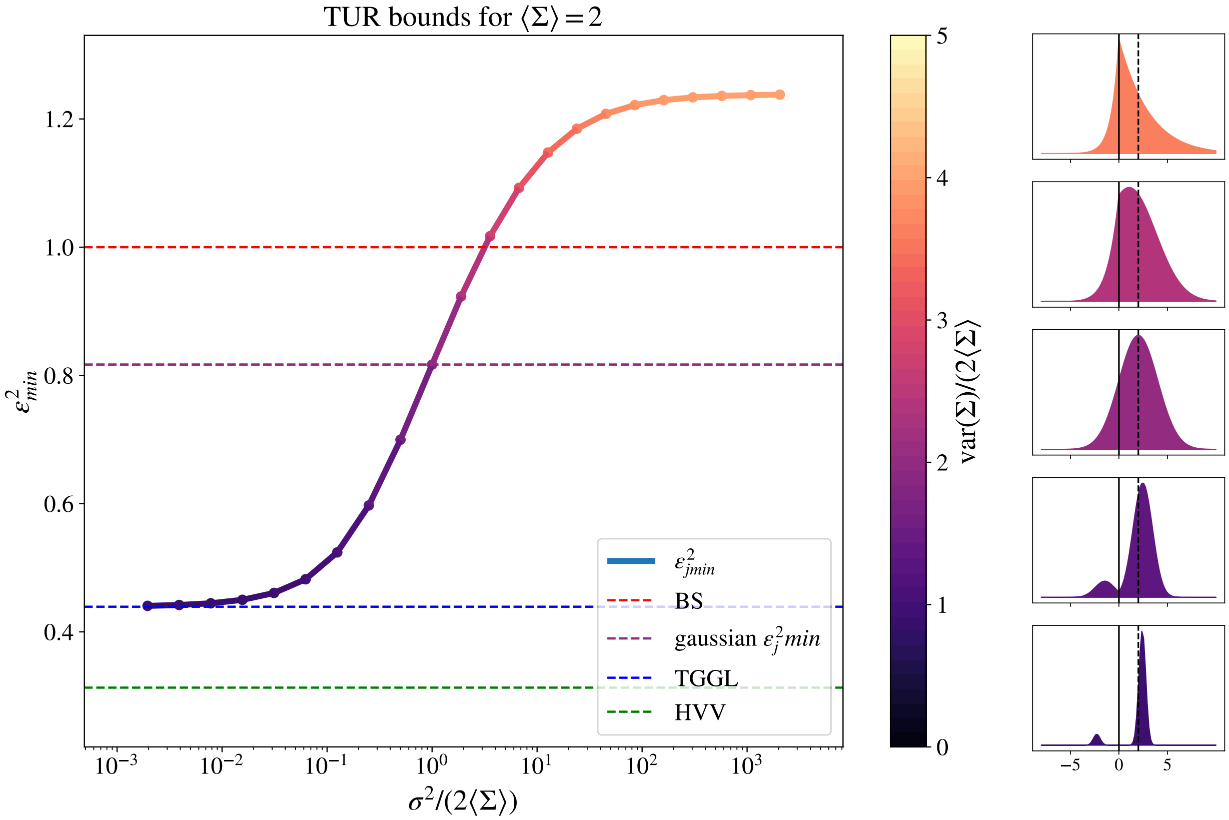


FIGURE 4.2.1. Three dashed lines show previous TURS—ϵ^2_{BS}, ϵ^2_{HVV}, and ϵ^2_{TGGL}—all functions of average entropy production $\langle\Sigma\rangle$. While they make nearly identical predictions for small entropy production, they diverge as entropy increases, setting very different bounds for average entropy production as low as $\langle\Sigma\rangle = 2k_B$. In contrast, the minimum scaled variance $\epsilon^2_{J\text{min}}$ is not strictly a function of average entropy. The entropy distribution $\Pr(\Sigma|\mu, \sigma^2)$ depends on the variance parameter σ^2 and is shown on a sliding scale from high to low variance (from dark to light). σ^2 ranges from $\approx 8 \times 10^{-3}$ to 8×10^3. While μ is adjusted to keep $\langle\Sigma\rangle$ fixed. This yields entropy distributions of the form shown on the right. The lowest values of σ^2 and var(Σ) closely match ϵ^2_{TGGL}. As σ^2 increases, the variance of the entropy production increases, and the curve become lighter, achieving and surpassing the dashed line for ϵ^2_{BS}. Between these two extremes, there is a purple dashed line that shows the minimum scaled variance of normal entropy distributions.

The variance σ^2 and average μ of the positive entropy portion of the distribution $\Pr(\Sigma|\mu, \sigma^2)$ define it. In the limit $\sigma^2 \ll k_B\mu$, the positive entropy distribution is nearly a delta function, and we recover roughly the distribution $\Pr_{\min}(\Sigma)$ proposed by Timpanaro et al [**108**]. We show this on the lefthand side of Fig. 4.2.1 where $\sigma^2 \approx 10^{-3}k_B\mu$, corresponding to a two-peaked distribution.

In this case, $\epsilon^2_{J\mathrm{min}}$ closely matches the bound $\epsilon^2_{\mathrm{TGGL}}$, as expected. However, Fig. 4.2.1 also shows that the average entropy production is not the sole determinant of the minimum scaled variance of the current.

As the variance of the positive normal distribution σ^2 increases, Fig. 4.2.1 shows that the minimum scaled variance increases. Amidst that progression is a special distribution, where $\sigma^2 = 2k_B\mu$, for which $\Pr\big(\Sigma|\mu, \sigma^2\big)$ is a normal distribution over the full range of entropy production. It can be quickly shown that the variance for a normal distribution that satisfies the TSCC DFT must be $\sigma^2 = 2k_B\mu$. We highlight this special case with a purple dashed line in Fig. 4.3.2. NESSs, the most frequently studied subclass of TSCC processes, approach a normal entropy production distribution in the long time limit. Interestingly, the minimal variance currents of the typical asymptotic behavior in NESSs clearly violate the original TUR given by ϵ^2_{BS} in the long-time limit.

If we continue beyond the normal distribution to higher variances labeled in lighter colors in Fig. 4.2.1, the minimum scaled variance $\epsilon^2_{J\mathrm{min}}$ continues to increase, until it surpasses the bound ϵ^2_{BS} [**105**]. Thus, by changing the parameter σ^2 of the NESS distribution $\Pr\big(\Sigma|\mu, \sigma^2\big)$ and thus its variance $\mathrm{var}(\Sigma)$ as well as other higher moments of the entropy distribution $\Pr(\Sigma)$, we can interpolate between the TUR set by Timpanaro et al. and that set by Barato and Seifert. Moreover, we can find entropy distributions that far exceed even Barato and Seifert's bound.

4.3. Thermodynamic Simulations

The entropy production distribution $\Pr\big(\Sigma|\mu, \sigma^2\big)$ is convenient to examine, but it is not obvious how it can be physically generated. We now describe two computational protocols that are able to show similar breadth of behavior, but are firmly rooted in dynamical models of physical processes. Both are TSCCs in that they are implemented with time symmetric control of a potential energy landscape, where the thermal influence of a bath is applied through Langevin dynamics.

4.3.1. Reset. First, consider a simple reset protocol. A system consisting of a single positional variable x in asymmetric double well potential U^{store} is initially set up in equilibrium with a thermal environment at temperature T. If the two metastable states are 0 and 1, then take 0 as the one with a deeper well and higher initial probability. Then, the energy landscape is tilted and the energy barrier is removed. This computational potential U^{comp} is held so that probability mass flows from A to B. This is a "reset" in that it re-initializes the system to the B state.

The system begins in equilibrium with a thermal reservoir, exposed to a storage potential U^{store}. At $t = 0$, a computational potential U^{comp} is applied until $t = \tau$ and then the system is re-exposed to U^{store}. Because the potential energy surface changes only at $t = 0$ and $t = \tau$, the work $W(x_0, x_\tau, U^{\text{comp}}, U^{\text{store}})$ done during any particular microscopic trajectory is given by the sum of the work invested at $t = 0$ and that invested at $t = \tau$. Once the system relaxes back to equilibrium, any energy added to the system in the form of work has been dissipated into the environment: generating entropy in the heat bath equal to βW. Thus, the entropy generated by a trajectory can be calculated simply as:

$$\begin{aligned}\beta^{-1}\Sigma(\vec{x}) &= \beta^{-1}\Sigma(x_0, x_\tau) = W_0 + W_\tau \\ &= U^{\text{comp}}(x_0) - U^{\text{store}}(x_0) + U^{\text{store}}(x_\tau) - U^{\text{comp}}(x_\tau) \qquad (4.10)\end{aligned}$$

A large enough ensemble of initial conditions allows for an estimate of the entropy production distribution and, through equation 3.102, an estimate of the minimum scaled variance that can be achieved by any current defined on the system. For both the reset, U^{comp} was chosen to be an asymmetric double square-well potential with wells of depths D_0, D_1, widths ℓ, and centered at $x = \pm L$ (See fig 4.3.1.)

The simulation of the reset used non-dimensionalized overdamped Langevin dynamics:

$$dx = -\Omega \partial_x U(x,t)dt + \xi\sqrt{2}\, r(t)\sqrt{dt} \ ,$$

Here, $r(t)$ is a memoryless Gaussian variable, and all parameters and variables have been scaled to be dimensionless by the scheme $q' = q \cdot q_c$. q' is some dimensional quantity with q_c a scaling factor and q the dimensionless variable. The dimensionless simulation parameters Ω and ξ are combinations of the scaling factors and the familiar dimensional Langevin parameters. For all overdamped simulations, $\Omega = \xi = 1$. This represents some relationship between the physical parameters of the system, but the exact relationship is not important for our purposes. Note also

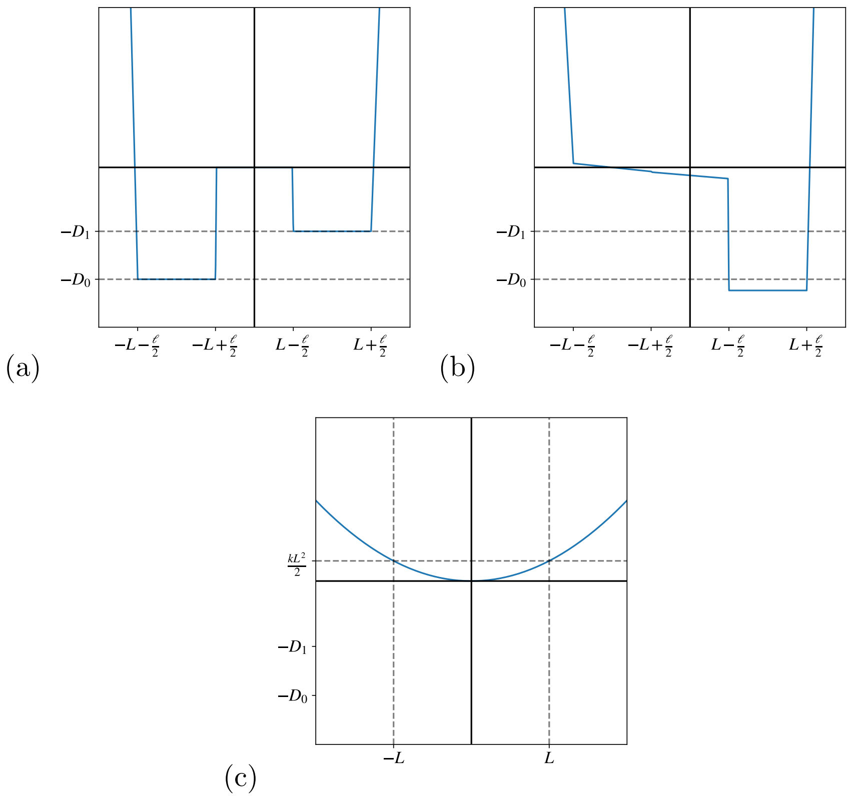


FIGURE 4.3.1. Potential energy landscapes during (a) storage, (b) the reset computation, and (c) the swap computation. The offset from D_0 in the right well during the reset computation creates the same energy barrier as the left well during the storage potential.

that we choose our scaling factor for energies to be $k_\mathrm{B}T$ so that the potential energy can be thought of as being in units of the thermal energy.

In order to see the reset could approach the TUT bound, a suite of 1366 simulations was performed using a Monte Carlo Markov Chain (MCMC) inspired approach to find parameters for which $\epsilon^2_{\mathcal{J}\mathrm{min}}$, as estimated by equation 3.102, is minimized. On each iteration of algorithm, a new value was chosen for 2 (chosen randomly, with replacement) of the 4 parameters L, ℓ, D_0, D_1 using a Gaussian distribution centered on its current value, checking to make sure that $\ell < L$ and $D_0 > D_1$. After performing the simulation and measuring $\epsilon^2_{\mathcal{J}\mathrm{min}}$, the proposed parameter change was accepted with certainty if the new $\epsilon^2_{\mathcal{J}\mathrm{min}}$ was less than the original and accepted with a probability $p \propto e^{-\Delta\epsilon^2_{\mathcal{J}\mathrm{min}}}$ if it was greater. Jumps for which the average entropy production did not satisfy $1.5 \leq \langle \beta\Sigma \rangle \leq 6$ we also rejected, to keep the algorithm from exploring an untenable range of parameter space. The end result is that for the simulations in figure 4.3.2 the parameters were sampled from the following ranges, though not uniformly or independently: $L \in (.2, 1.2)$, $\ell \in (0, 1.1)$, $D_0 \in (1, 6.2)$ and $D_1 \in (.2, 3.6)$. Here, L, ℓ are in units of the non-dimensional position and D_0, D_1 in units of $k_{\mathrm{B}}T$

The resulting entropy productions in minimum-variance currents are shown in the left panel of Fig. 4.3.2. Some computations are considerably less precise than specified by ϵ^2_{BS}, the original TUR, but many lie well below this bound. As expected, all computations are less precise than the bounds $\epsilon^2_{\mathrm{TGGL}}$ and $\epsilon^2_{\mathrm{HVV}}$, but none of them come very close to the tightest theoretical bound given by $\epsilon^2_{\mathrm{TGGL}}$. Instead, there is another curve that seems to bound all of these time-symmetric erasures, shown in dashed red. It might be expected that the resulting entropy production distribution would closely mirror the bimodal distribution case from figure 4.2.1 because the square shape of the well means there are only two dominant values that the work can take: most trajectories will go from the 0 to the 1 well, and experience a work value of $\approx D_0 - D_1$, a few will go the opposite direction and experience $D_1 - D_0$. However, the results are not in line with this assumption. This is because the 0 work trajectories that stay in their respective wells have a nontrivial effect.

4.3.1.1. *Discrete Bound.* In order to see why the reset operation as described cannot generate a truly bimodal distribution, consider a simplified version of the continuous state dynamics that implement the reset operation: a two-level system operating in the regime of rate equation dynamics. Here, the 'potential energy landscape' is defined simply by setting the energy levels of the two states $x \in \{A, B\}$. The U^{store} energy levels are $E_A = E, E_B = 0$ so that the equilibrium distribution over the two states is given by $\rho_0 = (Pr(X_0 = A), Pr(X_0 = B) = (p_E, 1 - p_E)$. Here, U^{comp} will swap the two energy levels so that $E_A = 0, E_B = E$ and τ will be long enough that the system has enough time to equilibrate to U^{comp} yielding $\rho_\tau = (1 - p_E, p_E)$. Using equation 4.10 reveals only three possible outcomes for $\Sigma(x_0, x_\tau)$:

$$\Sigma(A, A) = \Sigma(B, B) = 0$$

$$\Sigma(B, A) = -\Sigma(A, B) = 2\beta E.$$

Because the system has been given time to equilibrate the state at time $t = 0$ is not correlated with the state at time $t = \tau$ so the probabilities of these different events can be readily calculated,

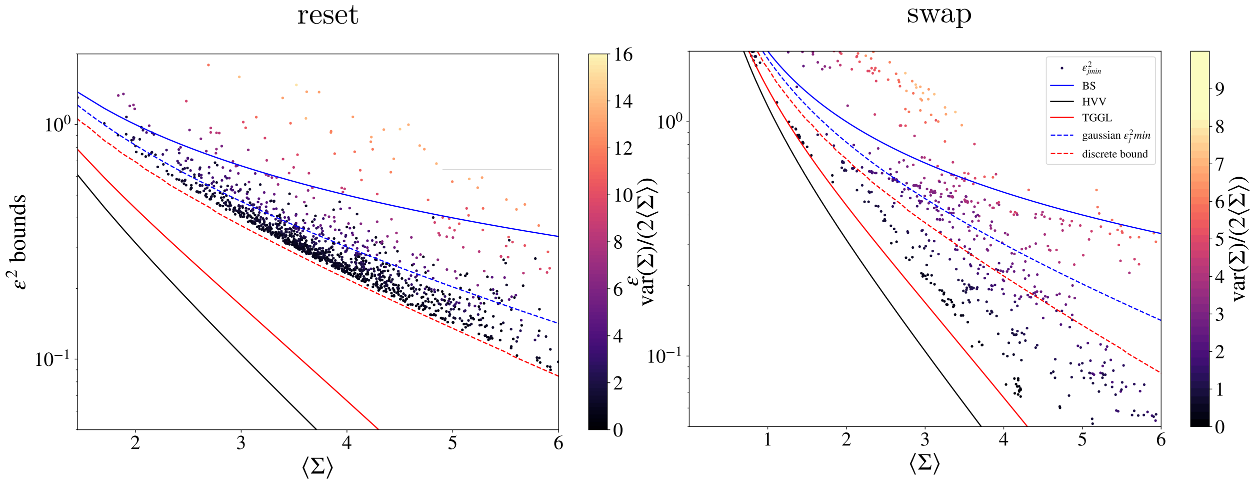


FIGURE 4.3.2. Bounds $\epsilon^2_{\text{BS}} > \epsilon^2_{\text{TGGL}} > \epsilon^2_{\text{HVV}}$ in solid lines (blue, red, and black respectively) and specific thermal processes with dashed lines: The blue dashed line is the minimum scaled variance $\epsilon^2_{\text{Gaussian}}$ of any process that generates a Gaussian entropy production distribution, which is achieved in the long-time limit of NESS processes. The red dashed line is the minimum scaled variance $\epsilon^2_{\text{Discrete}}$ of an ideal discrete erasure. We compare two computational classes to the these bounds. (Left) we plot 1366 different time-symmetric erasures. As expected (see App. 4.3.1.1) they are bounded by the scaled variance of the ideal discrete erasure $\epsilon^2_{\text{Discrete}}$, which lies well above the bounds ϵ^2_{TGGL} and ϵ^2_{HVV}. A number of erasure operations are well above the minimum ϵ^2_{TGGL} set by Barato and Seifert. (Right) we plot the result 1193 different bit-flips. As with the erasure protocol, many computations are above the Barato-Seifert bound. However, many computations achieve a minimum scaled variance well below the discrete erasure bound $\epsilon^2_{\text{Discrete}}$. Many computations are quite close to the strongest possible TUR ϵ^2_{TGGL}, indicating that this theoretical bound is indeed achievable with TSCCs.

yielding the full distribution of entropy production:

$$\begin{aligned} Pr(\Sigma(A,A)) = Pr(\Sigma(B,B)) = p_E * (1-p_E) \\ Pr(\Sigma(A,B)) = p_E^2 \\ Pr(\Sigma(B,A)) = (1-p_E)^2 \end{aligned}$$

For the two-level system, $p_E = \frac{e^{-\beta E}}{1+e^{-\beta E}}$. For any anti-symmetric function $J(\Sigma) = -J(-\Sigma)$, we have

$$\begin{aligned}
\langle J(\Sigma)\rangle(E) &= J(2\beta E)((1-p_E)^2 - p_E^2) \\
&= J(2\beta E)(1-2p_E) = J(2\beta E)\frac{1-e^{-\beta E}}{1+e^{-\beta E}} \\
&= J(2\beta E)\tanh(\beta E/2).
\end{aligned}$$

Where the zero entropy events do not appear directly in the first line because $J(0) = 0$ for any function J that is odd in Σ. We use this equation to readily calculate both the average entropy production

$$\langle \Sigma\rangle(E) = 2\beta E \tanh(\beta E/2), \tag{4.11}$$

and the minimum variance current for the distribution (through equation 3.102)

$$\begin{aligned}
\epsilon^2_{J\min}(E) &= \frac{1}{\langle \tanh(\Sigma/2)\rangle(E)} - 1 \\
&= \frac{1}{\tanh(\beta E)\tanh(\beta E/2)} - 1.
\end{aligned} \tag{4.12}$$

We can then use the parameter E to find the effective bound that equation 4.3 sets for a given average entropy production in the reset process. Fig. 4.3.2 shows that, while this bound lies below the one for a normally distributed entropy production– it lies far above previous bounds, $\epsilon^2_{\mathrm{TGGL}}$ and $\epsilon^2_{\mathrm{HVV}}$, that used only the TSCC DFT in 3.91. And, for the reset processes simulations the bound appears tight. This example showcases the flexibility of 4.3, as we see it can be used to set operationally useful regime and/or protocol specific bounds by including information about the system of interest.

4.3.2. Swap. In order to generating an entropy production distribution that allows for the most accurate current possible– we turn to a different kind of protocol with a different type of dynamic. Consider the same initial metastable states A and B, which are stored in local equilibrium. Then, instantaneously implement a harmonic potential and hold the energy landscape for half a period of the oscillation. If the coupling to the thermal environment is weak, then this implements a reliable swap between A and B, using momentum as memory to carry the distributions into their new states. Crucially, this simulation must use underdamped Langevin dynamics rather than overdamped:

$$\begin{aligned} dx &= v dt \\ dv &= -\lambda v dt - \Theta \partial_x U(x,t) dt + \eta \sqrt{2\lambda}\, r(t) \sqrt{dt}\ , \end{aligned}$$

A similar scaling strategy as that described in Appendix 4.3.1 leads to three dimensionless parameters: λ, Θ, η. $\Theta = \eta = 1$ for all simulations but λ, which parameterizes the system's coupling to its thermal environment, was allowed to change.

The computational potential for the swap is a harmonic potential with $k = m\pi^2$ (see Fig. 4.3.1.) If $\lambda = 0$, the system undergoes a harmonic oscillation with a period of 2 time units. Exactly halfway through the oscillation, the particles that were in the left(right) well of U^{store} should now be located where the right(left) well is. Thus, if U^{store} is turned back on at $t = 1$ we have implemented a 'swap' operation between asymmetric wells. Because the dynamics are underdamped, this type of protocol also persists in the case of nonzero λ, since the system will undergo the same oscillation, with some amount of dissipation. The work cost distribution to implement this protocol will approach a bimodal distribution, with particles starting in the left well costing an energy value near $D_0 - D_1$ and those starting in the right yielding an energy surplus near $D_1 - D_0$. This distribution is sharpened by narrower wells and lower values of λ.

Instead of only attempting to minimize $\epsilon^2_{J\text{min}}$, the point of this simulation is to showcase that $\epsilon^2_{J\text{min}}$ can faithfully capture all cases: from where ϵ^2_{TGGL} is tight, to where ϵ^2_{BS} is tight, to where neither is a good approximation. The protocol described above generates the bimodal distribution that saturates ϵ^2_{TGGL} under some cases, but can also produce entropy production distributions where the minimal scaled variance is well above ϵ^2_{BS} (see the right panel of figure 4.3.2.) To showcase this variety, a MCMC approach was again used. On each iteration of algorithm, a new value was chosen for 3 (chosen randomly, with replacement) of the 5 parameters $L, \ell, D_0, D_1, \lambda$ using a Gaussian distribution centered on its current value, checking to make sure that $\ell < L$, $D_0 > D_1$,and $\lambda > 0$. In this case, all jumps for which $\langle \Sigma \rangle$ fell between 2 and 5 were accepted, and those that did not were accepted with a probability that exponentially decayed in $|\langle \Sigma \rangle - 3.5|$. The plot in figure 4.3.2 shows a suite of 1193 simulations, that stem from 8 different starting points for $\lambda \in (0, .15)$, but all other parameters the same. As the algorithm evolved, all free parameters were allowed to shift with the result being that parameters were sampled from the following ranges: $L \in (.14, .68)$, $\ell \in (0, .39)$, $D_0 \in (1.2, 10.5)$ and $D_1 \in (.39, 3.6)$, $\lambda \in (0, .2)$.

The entropy and scaled variance of these swaps are shown on the right-hand side of Fig. 4.3.2. They span the same space of possibilities shown for $\Pr\left(\Sigma|\mu, \sigma^2\right)$ in Fig. 4.2.1: some lying above ϵ^2_{BS} and some TSCCs sitting just above the minimum set by $\epsilon^2_{\mathrm{TGGL}}$.

4.4. Momentum

Analogous to the transition from the Second Law inequality to the fluctuation relations of Crooks and Jarzynski, we have developed a treatment of the entropy production that recognizes it as a stochastic quantity with lawful fluctuations. This treatment yields a result—the Thermodynamic Uncertainty Theorem—that is (i) an equality rather than an inequality and (ii) depends on the stochastic entropy production's fluctuations rather than on only its average value. This generalization of previous TURs can be used to derive them, to better understand their domains of applicability, and to establish new system-specific bounds on the variance of currents. In this way, the Thermodynamic Uncertainty Theorem moves further to fully quantifying the relationship between fluctuations and dissipation in out-of-equilibrium processes—the unifying goal in stochastic thermodynamics. In the process of this discovery, we find that the underlying system's momentum coordinate was key in generating an entropy production distribution that permitted most accurate currents allowed by the TSCC DFT symmetry. The revelation opens the door for exploring a new class of computational protocols, named 'Momentum Computing', that leverage both the position-like and momentum like coordinates of a systems phase space.

Appendix

4.A. Proof that $\epsilon_J^2 \geq \epsilon_{J\mathbf{min}}^2 \geq \epsilon_{\mathbf{TGGL}}^2$

Let us start by considering the expressions for the two TUR bounds $\epsilon_{j_{min}}^2$ and ϵ_{TGGL}^2, which can be proven to bound the noise-to-signal ratio (or scaled variance) $\epsilon_J^2 \equiv \frac{\mathrm{Var}(J)}{\langle J \rangle^2}$ of any current J (assumed to be anti-symmetric under time-reversal) in steady-state regime under the constraint that the probability distribution for the entropy production Σ satisfies the so-called Evan-Searles (or exchange) fluctuation relation symmetry

$$p(-\Sigma) = p(\Sigma)e^{-\Sigma}. \tag{4.13}$$

Our main result is that, under no additional assumptions about the distribution over Σ, the following bound can be placed

$$\epsilon_J^2 \geq \epsilon_{j_{min}}^2 \equiv \frac{1}{\langle \tanh(\Sigma/2) \rangle} - 1. \tag{4.14}$$

The other bound, proven in [**108**] shows instead if on top of the above the averages of the entropy production and of a generic current, i.e. $\langle \Sigma \rangle$ and $\langle J \rangle$, are fixed and satisfy *a joint* fluctuation relation symmetry of the form

$$p(-\Sigma, J) = p(\Sigma, J)e^{-\Sigma}, \tag{4.15}$$

then the following TUR bound can be derived

$$\epsilon_J^2 \geq \epsilon_{TGGL}^2 \equiv \mathrm{csch}^2(g(\langle \Sigma \rangle/2)). \tag{4.16}$$

where $g(\langle \Sigma \rangle)$ is the function inverse of $\langle \Sigma \rangle \tanh(\langle \Sigma \rangle)$.

Is there a relationship between these two bounds? In these brief notes we provide a positive answer to this question.

The first step to show this is to re-express the right hand side of Eq. (4.16) as

$$\epsilon_J^2 \geq \epsilon_{TGGL}^2 \equiv \frac{1}{\tanh^2(g(\langle\Sigma\rangle/2))} - 1 \tag{4.17}$$

by using the identity $\operatorname{csch}^2(x)+1 = 1/\tanh^2(x)$. In order to proceed, let us then define the following quantities:

$$z(\Sigma) \equiv \tanh^2\left(\frac{\Sigma}{2}\right) \tag{4.18}$$

$$h(\Sigma) \equiv \Sigma \tanh\left(\frac{\Sigma}{2}\right). \tag{4.19}$$

Furthermore, let us introduce the inverse function of $h(\Sigma)$ and denote with by g, i.e. $g(h(\Sigma)) = \Sigma$. Notice that this is possible since $h(\Sigma)$ is a monotonically increasing function of Σ whenever $\Sigma \geq 0$. This is not a limitation, since, thanks to the fluctuation relation symmetry Eq. (4.13), it amounts to consider

$$\langle(\ldots)\rangle \equiv \sum_{\Sigma}(\ldots)p(\Sigma) = \sum_{\Sigma>0}(\ldots)p(\Sigma)\left(1+e^{-\Sigma}\right). \tag{4.20}$$

The way to make use of the above-introduced quantities is to realize that the composite function $w(h) \equiv f(g(h))$ is a concave function of h (when $h \geq 0$), since $w'(h) > 0$ $w''(h) < 0$. This allows us to exploit Jensen's inequality (with the "norm" being given by the average $\langle\ldots\rangle$ calculated with the probability distribution $p(\Sigma)(1+e^{-\Sigma})$, as discussed a few lines ago) and obtain the following inequality

$$\langle\tanh\left(\frac{\Sigma}{2}\right)^2\rangle = \langle z(\Sigma)\rangle = \langle z(g(h(\Sigma)))\rangle \tag{4.21}$$

$$\leq z(g(\langle h(\Sigma)\rangle)) = z(g(\langle\Sigma\rangle)), \tag{4.22}$$

where on the last passage we have used the fact that $\langle\Sigma\rangle = \langle h(\Sigma)\rangle$. Notice that the quantity that appears on the right hand side of this inequality, i.e. $z(g(\langle\Sigma\rangle))$, is nothing but the denominator in the first term of the r.h.s. of Eq. (4.16).

The final step to conclude the proof then consists in realizing that 3.101 means that the the l.h.s. of the inequality (4.21) is the denominator in the first term of the r.h.s. of Eq. (4.14).

In view of Eq. (4.21) then immediately follows that

$$\epsilon^2_{\jmath_{min}} \geq \epsilon^2_{TGGL}. \tag{4.23}$$

4.A.1. Simulation Details. Each dot in the simulation plot was calculated from an ensemble of 50,000 trajectories sampled from the equilibrium distribution, using a Monte Carlo method. In order to reduce the size of errorbars, averages shown in figure 4.3.2 were calculated using only the positive entropy production events according to equation 3.97, which assumes a system that obeys the TSCC DFT. This numerical trick does not change any of the qualitative results, but allows for plots in which the error bars are small enough to not be relevant in the plot. For example, figure 4.A.1 shows a plot of the same simulation data that generated figure 4.3.2 but with 3σ error bars calculated from the full simulation, rather than just the well-sampled positive events. The Langevin simulations of the dimensionless equations of motion for both the underdamped and overdamped cases employed a fourth-order Runge-Kutta method for the deterministic portion and Euler's method for the stochastic portion of the integration with dt set to 5×10^{-5}. Python NumPy's Gaussian number generator was used to generate the memoryless Gaussian variable r(t).

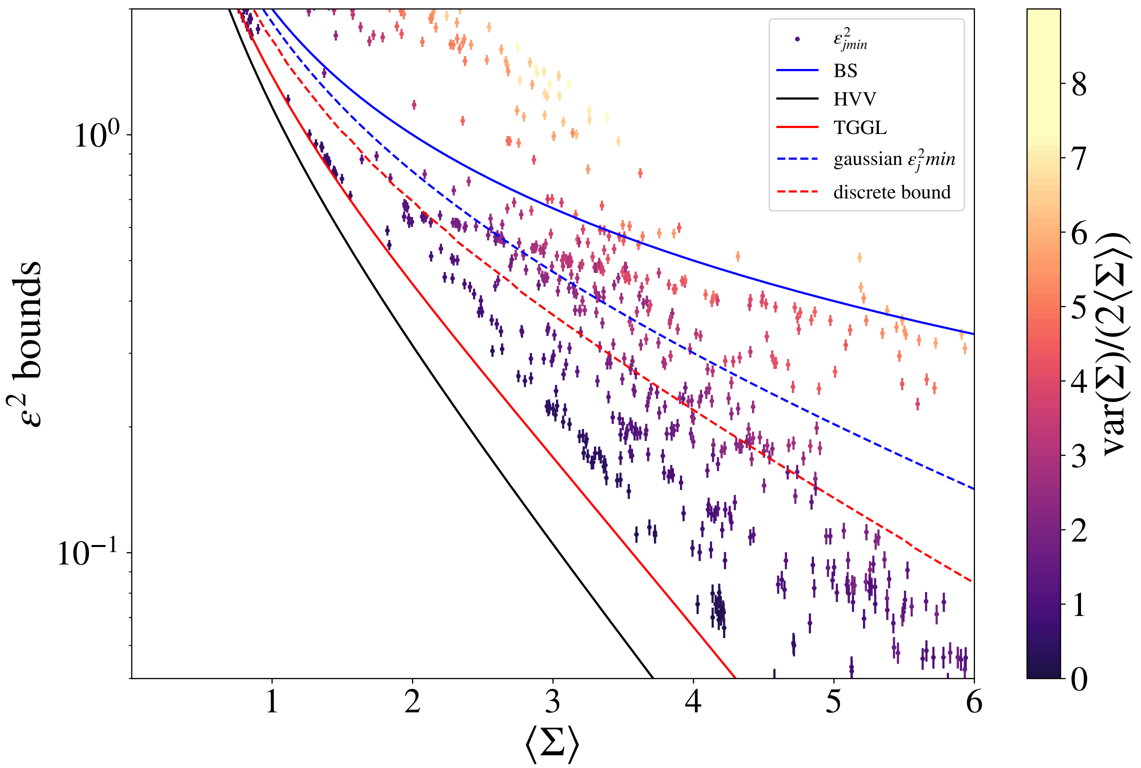


FIGURE 4.A.1. The same simulation data that generated figure 4.3.2 but with 3σ error bars calculated from the full simulation, rather than just the well-sampled positive events.

CHAPTER 5

Momentum Computing

Practical, useful computations are instantiated via physical processes. Information must be stored and updated within a system's configurations, whose energetics determine a computation's cost. To describe thermodynamic and biological information processing, a growing body of results embraces rate equations as the underlying mechanics of computation. Strictly applying these continuous-time stochastic Markov dynamics, however, precludes a universe of natural computing. Within this framework, operations as simple as a NOT gate (flipping a bit) and swapping bits are inaccessible. We show that expanding the toolset to continuous-time *hidden* Markov dynamics substantially removes the constraints, by allowing information to be stored in a system's latent states. We demonstrate this by simulating computations that are impossible to implement without hidden states. We design and analyze a thermodynamically-costless bit flip, providing a counterexample to rate-equation modeling. We generalize this to a costless Fredkin gate—a key operation in reversible computing that is Turing complete (computation universal). Going beyond rate-equation dynamics is not only possible, but necessary if stochastic thermodynamics is to become part of the paradigm for physical information processing.

The burgeoning field of thermodynamic computing leverages recent progress in nonequilibrium thermodynamics and information and computation theories [**23**, **111**, **112**, **113**, **114**] to establish a new paradigm for physical information processing. It promises to increase computational power and efficiency and to reduce energy dissipation in a next generation of computers [**45**]. Thermodynamic computing is distinguished from alternative paradigms by its focus on an information-processing device's physical embedding; specifically, by constructively working with $k_\mathrm{B}T$-scale fluctuations that a thermal environment generates. More broadly, a general framework rooted in thermodynamics, as thermodynamic computing is, will provide the tools to understand the physics of computation in all its many forms. The following illustrates its breadth by introducing non-Markovian, momentum-based computing—a paradigm that is both computation universal and thermodynamically efficient.

We describe physically-embedded computation as a stochastic mapping within a system's set $\mathcal{M}$ of memory states—Landauer's *information-bearing degrees of freedom* (IDoF) [**115**]. Carried out over time interval $t \in (0, \tau)$, the mapping is the conditional probability $\mathbf{p}$ of transitioning from an initial memory state $m(0) \in \mathcal{M}$ to a final state $m(\tau) \in \mathcal{M}$: $\mathbf{p}_{m(0) \to m(\tau)} = \Pr\left[m(\tau) | m(0)\right]$. The mapping $\mathbf{p}$ determines the probability of the final memory state given the initial memory state, and so updates the state distribution $\vec{p}(\tau) = \mathbf{p}\, \vec{p}(0)$.

This describes the physical dynamics underlying a computation, but what of its thermodynamic consequences? To address this, we must first identify the constraints on dynamics that can implement computations.

5.1. Computing with Continuous-Time Markov Chains

To date, proposed frameworks for the required mappings in thermodynamic computing assume that the memory state m obeys stochastic Markovian dynamics [**116**, **117**, and references therein]. Taking time to be continuous, the dynamics are continuous-time Markov chains (CTMCs), where the state distribution changes continuously as a function of itself $\dot{\vec{p}}(t) = f(\vec{p}, t)$. The resulting dynamics are necessarily represented by a master equation over the memory-state distribution $\dot{\vec{p}}(t) = \mathbf{A}(t)\vec{p}(t)$ [**116**, **117**]; that is, by *rate equations*. This is a powerful framework for stochastic thermodynamics [**111**, **118**, **119**] that yields insight into physical realizations of computations such as bit erasure and measurement [**120**].

The constraint that the computation $\mathbf{p}$ is generated by integrating continuous-time master equations comes at a substantial compromise, though, as it limits both the range of possible computations and the energetic consequences of the computations that are possible. One example is the 'discrete bound' derived in the previous chapter that prevented the overdamped 'reset' from generating entropy production distributions that could contain very precise currents. There are other limitations as well: for example, only input-output mappings whose determinants are positive are allowed when memory-state dynamics are restricted to obey CTMCs [**117**]. This eliminates many common and useful computations, including flipping a single bit of information. Reference [**116**] takes these restrictions as delineating the possible *physically realizable computations*. Given that any computation we can observe—a bit flip, to take one example—is necessarily physical, one must instead interpret the restrictions as a limitation of the CTMC framework, rather than of the physical world.

Understanding both the merits and limitations of the CTMC framework requires a look at the physical mechanisms that underpin it. It might seem natural to say that the memory states $\mathcal{M}$ are microstates of a physical memory system $\mathcal{S}$, evolving under Hamiltonian dynamics. However, a physical computation device is typically coupled to an environment—which suggests treating $\mathcal{S}$ as a stochastic subsystem of a deterministic universe. If the environment is a large weakly-coupled heat bath, with degrees of freedom that relax sufficiently quickly, the effective dynamics for $\mathcal{S}$ are also Markov, and therefore CTMC [**111**, **121**, **122**, **123**, **124**]. In essence, coarse-graining the thermal environment allows for accurate, probabilistic predictions about the memory system, while avoiding the task of tracking the full Hamiltonian dynamics of the joint system and bath.

This justification of the Markov evolution of a memory system recognizes an important fact: $\mathcal{S}$'s states are not themselves full descriptions of physical degrees of freedom. Instead, they are mesostates defined by a coarse-graining over the thermal environment's microstates. This coarse-graining is appropriate since the environment does not retain information about the past.

Computationally-useful memory states $\mathcal{M}$—Landauer's IDoF—are mesostates that also coarse-grain over $\mathcal{S}$'s CTMC-evolving states. It is possible, depending on the variables and timescales of interest, that this coarse-graining ignores only rapidly-relaxing subsystems of $\mathcal{S}$ and, then, the IDoF inherit the Markov property of the memory system [**125**]. The result is a powerful and widely-used framework for thermodynamic computing in which the IDoF also obey CTMC dynamics. This case is typified by IDoF that are positional degrees of freedom and where $\mathcal{S}$ is described by overdamped Langevin dynamics. It is from this perspective that a bit flip is forbidden: in order to cause realizations that fall in the region representing $m = 0 (m = 1)$ at $t = 0$ to move to the region of state space representing $m = 1 (m = 0)$ at $t = \tau$, the two must overlap at some intermediate time. If the dynamics of $\mathcal{M}$ are restricted to be Markovian (memoryless), the two disparate initial conditions cannot be distinguished from each other once the overlap occurs—rendering it impossible to selectively control them to end in separate memory states.

5.2. Computing with Continuous-Time Hidden Markov Chains

However, in the most general case, the IDoF coarse-grain over subsets of $\mathcal{S}$ that carry information relevant for predicting a computation's performance. That is, the states that $\mathcal{M}$ coarse-grains over are *hidden* in that they contain dynamically relevant information not determined from instantaneous realizations of the memory. The resulting memory dynamics are non-Markovian, since information is transmitted from past to future without ever appearing in the present memory state [**126**]. The sobering fact is that a general analytical treatment of partially-observed (and therefore non-Markovian) systems is highly nontrivial [**114**, **124**, **127**, **128**, **129**]. No matter, hidden states allow for more general forms of computation [**117**, **130**], since non-Markov dynamics relax the constraints imposed by CTMCs. Following this argument to its conclusion, the following demonstrates that the appropriate setting for thermodynamic computing is continuous-time *hidden* Markov chains (CTHMCs), in which hidden variables store computationally-relevant information.

Moreover, when memory is stored in positional degrees of freedom, the conjugate momentum variables are particularly useful hidden variables for flexibly designing computations. We demonstrate this first by implementing a thermodynamically-costless bit flip—a simple computation that is explicitly forbidden by CTMCs. We then generalize this to a costless Fredkin gate—a key component in reversible computing that is also impossible to implement with CTMCs. This operation is computation universal (Turing complete), meaning that combinations of the Fredkin gate can implement any logical operation [**131**]. The implementation of this universal and reversible-logic gate via CTHMCs demonstrates that non-Markovian dynamics are essential to thermodynamic computing and that a new class of momentum-based computation is within reach.

5.2.1. Momentum Computing Realization.

Consider a computation that happens faster than the equilibration timescale of the physical substrate and its thermal environment. In this regime, a particle's instantaneous momentum can be commandeered to carry useful information about its future behavior. Our protocol operates on this timescale, using the full phase space of the underlying system's degrees of freedom to transiently store information in their momenta. Due to this, the instantaneous microstate distribution is necessarily far from equilibrium during the computation. Moreover, the coarse-grained memory-state dynamics during the swap are not

Markovian; despite both the net transformation over the memory states and the microscopic phase space dynamics being Markovian. Nonetheless, the system operates orders of magnitude more efficiently than current CMOS but, competing with CMOS, the dynamics evolve non-adiabatically in finite time—on nanosecond timescales for our physical implementation below.

In this way, momentum computing offers up device designs and protocols that accomplish information processing that is at once fast, efficient, and low error. There is a trade-off—a loss of Markovianity in the memory-state dynamics. That noted, the dynamics of the memory states are faithfully described by continuous-time *hidden* Markov chains (CTHMCs) [**130**, **132**, **133**], rather than the continuous-time Markov chains (CTMCs) that are common in stochastic thermodynamics [**114**, **134**].

Formally, auxiliary systems can be added to the set of memory states. Done correctly this again permits using CTMCs in the augmented state space to accomplish the computation [**135**]. However, physically-embedded computations do not generally allow the required perfect control over the system Hamiltonian. Indeed, one need look no further than the present work to see how nontrivial it is to implement an operation as simple as a harmonic oscillation in a physically-realistic device.

Moreover, adding auxiliary subsystems increases state-space dimension and complicates control apparatus and control protocols. Due to the increased complication, in many settings, adding auxiliary dimensions is simply not physically feasible. On top of this, the timescale of these augmented computations must be longer than the equilibration time of the auxiliary systems and thermal environment. In this way, adding auxiliary systems imposes additional speed limits to computations. In short, adding auxiliary subsystems addresses the shortfalls of CTMCs, but does not sidestep their fundamental limitations.

We illustrate this by considering an efficient bit swap implemented via a Markovian embedding. First, it augments the system with an unoccupied auxiliary state A to serve as a transient memory. It then quasistatically translates memory state 0 to A, while memory state 1 is translated to 0. Finally, it quasistatically translates A to 1.

Quasistatic processes cost arbitrarily little work, but they take arbitrarily-long times. To compute faster ($\tau \to 0$), the work cost will diverge as $1/\tau$ [**46**, **52**, **53**, **54**]. Increasing fidelity requires raising the scale of the barrier separating the states. Doing so, though, increases the energetic cost at a given computational speed; maintaining the same work cost, then, requires slowing the operation. In short, the trade-offs in Markovian embedding complicate design and, more to the point, reduce performance.

5.3. Flipping a Bit

The ideal bit swap has no error, but in the thermodynamic setting one is also interested in an implementation's fidelity. And so, we write a swap with error rates ϵ_0 and ϵ_1 as a stochastic mapping between memory states $m \in \{0, 1\}$ from time 0 to time τ:

$$P_\epsilon(m_\tau | m_0) = \begin{bmatrix} \epsilon_0 & 1 - \epsilon_0 \\ 1 - \epsilon_1 & \epsilon_1 \end{bmatrix} .$$

To execute a single bit flip over a time interval $t \in [0, \tau]$, the first step is to store a bit of information. One candidate is a particle with a single position dimension $x \in \mathbb{R}$ and corresponding momentum $p \in \mathbb{R}$ in an even potential energy landscape $V^{\text{store}}(x)$ containing only two potential minima at $x = \pm x_0$ with an associated energy barrier between them equal to $\{V^{\text{store}}(0) - V^{\text{store}}(x_0)$. The particle's environment is a thermal bath at temperature T. As the height of the potential energy barrier rises relative to the bath energy scale $k_B T$, the probability that the particle transitions between left ($x < 0$) and right ($x \geq 0$) decreases exponentially. In this way, if we assign the left half of the position space to memory state 0 and the right half to memory state 1, the energy landscape is capable of metastably storing a bit $m \in \{0, 1\}$.

To execute a flip operation, we instantaneously reduce the coupling to the thermal reservoir to zero such that the memory system now follows dissipationless Hamiltonian dynamics. Simultaneously, the potential energy landscape changes to a positive quadratic well: $V^{\text{comp}}(x, t = 0^+) = kx^2/2$. The resulting particle motion is harmonic oscillation: $x(t) = x^* \cos\left(t\sqrt{k/\mu} + \phi\right)$, where μ is the particle mass, x^* is the maximum distance from the cycle's origin, and ϕ is the phase difference from maximum distance at the time $t = 0^+$. Maintaining the thermally decoupled system in the quadratic potential energy landscape for half the oscillation period $t \in (0, \pi\sqrt{\mu/k})$, the particle's new position becomes: $x(\pi\sqrt{\mu/k}) = x^* \cos(\pi + \phi) = -x^* \cos(\phi) = -x(0)$. Thus, over the computation interval $\tau = \pi\sqrt{\mu/k}$, the position flipped sign so that the memory state has flipped as well: $m(\tau) = 1 - m(0)$. Finally, we instantaneously return the potential energy landscape to $V^{\text{store}}(x)$ and recouple to the thermal bath.

The work W involved is the time-integrated rate of potential energy change due to the change in the protocol parameter [**136**]: $W = \int dt \partial U(x,t')/\partial t'|_{x(t),t}$. The work cost for a particular trajectory is the instantaneous change in potential energy at $t = 0$ and $t = \tau$:

$$W = V(x(0), 0^+) - V(x(0), 0) + V(x(\tau), \tau) - V(x(\tau), \tau^-), \tag{5.1}$$

where 0^+ and τ^- are times immediately after and before $t = 0$ and $t = \tau$, respectively. Recall that the potential is time-symmetric ($V(x,t) = V(x, \tau - t))$), that $x(\tau) = -x(0)$, and that the potential is even in x. These three qualities yield $-V(x(0), 0) + V(x(\tau), \tau) = V(x(0), 0^+) - V(x(\tau), \tau^-) = 0$. No net work is generated during the protocol.

5.4. The Fredkin Gate

The bit-flip implementation may seem obvious in its simplicity. However, sophisticated and functional computing can be built from a similar passive processes. Below we outline an implementation of the Fredkin gate, a reversible and computation universal logical gate [**131**], using the same strategy. This establishes that CTHMCs give straightforward access to complex and Turing-complete thermodynamic computing.

The *Fredkin gate* operates on three bits $\mathcal{M} = \{0,1\}^3$. That is, we encode the physical substrate as three particle-position variables (x, y, z) that are each separated into negative and positive memory-state regions, as above. This splits the memory states into eight respective octants: $(x < 0, y < 0, z < 0)$ corresponds memory state $m = 000$, $(x < 0, y \geq 0, z < 0)$ to $m = 010$, and so on. The information-storing Hamiltonian is a straightforward sum of bistable, even, one-dimensional storage potentials: $V^{\text{store}}(x,y,z) = V^{\text{store}}(x) + V^{\text{store}}(y) + V^{\text{store}}(z)$. This provides metastable regions corresponding to each memory state $m_x m_y m_z \in \{0,1\}^3$.

Given this construction, we design physical transformations that implement the Fredkin gate with zero cost in finite time. The Fredkin gate is also known as the *controlled swap gate*, as it exchanges inputs m_y and m_z only if the control m_x is set to 1. In other words, the gate maps all inputs to themselves, excluding 101 and 110 that swap with each other. The implementation uses the bit-flip strategy of decoupling from the thermal reservoir and applying a harmonic potential over the time interval $t \in (0, \tau)$, then recoupling and resetting the original information-storing Hamiltonian. The only difference is that the harmonic potential driving the computation is now embedded in the higher-dimensional space.

To execute the Fredkin gate, first note that the memory-state x-index must always be fixed: $m_x(\tau) = m_x(0)$. Moreover, behavior in the $y - z$ plane should only depend on x up to whether it is positive or negative. Thus, we first split the potential into two pieces: $V(x, y, z, t) = V^{\mathrm{store}}(x) + V^{yz}(x, y, z, t)$. If $m_x(0) = 0$ then m_y and m_z must also not change. This suggests using the information-storing potential for this region of state space: $V(x < 0, y, z, t) = V^{\mathrm{store}}(x, y, z)$ during the entire computation. For $m_x = 1$, however, we must nontrivially compute on m_y and m_z: $V^{yz}(x \geq 0, y, z, t \in (0, \tau)) = V^{\mathrm{comp}}(y, z)$. Here, V^{comp} determines that part of the Hamiltonian which implements the switch 101 → 110 and 110 → 101 and remains unchanging over $t \in (0, \tau)$. Due to decoupling from the x-axis, particle behavior in either the positive or negative x regions can be considered as being purely the result of two-dimensional dynamics.

To swap 101 and 110, while keeping 111 and 100 fixed, consider a new basis for the yz-space. Define new variables: $y' = (y - z)/\sqrt{2}$ and $z' = (y + z)/\sqrt{2}$, such that the local equilibrium distributions for states 110 and 101 are centered around $z' = 0$ and those for states 111 and 100 are centered around $y' = 0$. Thus, our goal is to swap the distributions in the y'-coordinate while preserving their z'-coordinate.

Given this, we split the computation Hamiltonian again into independent components: $V^{\mathrm{comp}}(y, z) = V(y') + V(z')$. Flipping in the y'-coordinate employs the same Hamiltonian as for the previous bit-flip protocol: $V(y') = ky'^2/2$. As a result, when waiting half a period $\tau = \pi\sqrt{\mu/k}$, the y' coordinate changes sign $y'(\tau) = -y'(0)$, as does its momentum. We choose the z' coordinate's potential to be quadratic as well, but with an induced period of oscillation that is half as long: $V(z') = 2kz'^2$. z' then undergoes a full cycle after the duration τ, returning to its original value $z'(\tau) = z(0)$, as does its momentum. Over the control interval $t \in (0, \tau)$ the Hamiltonian operates piecewise with

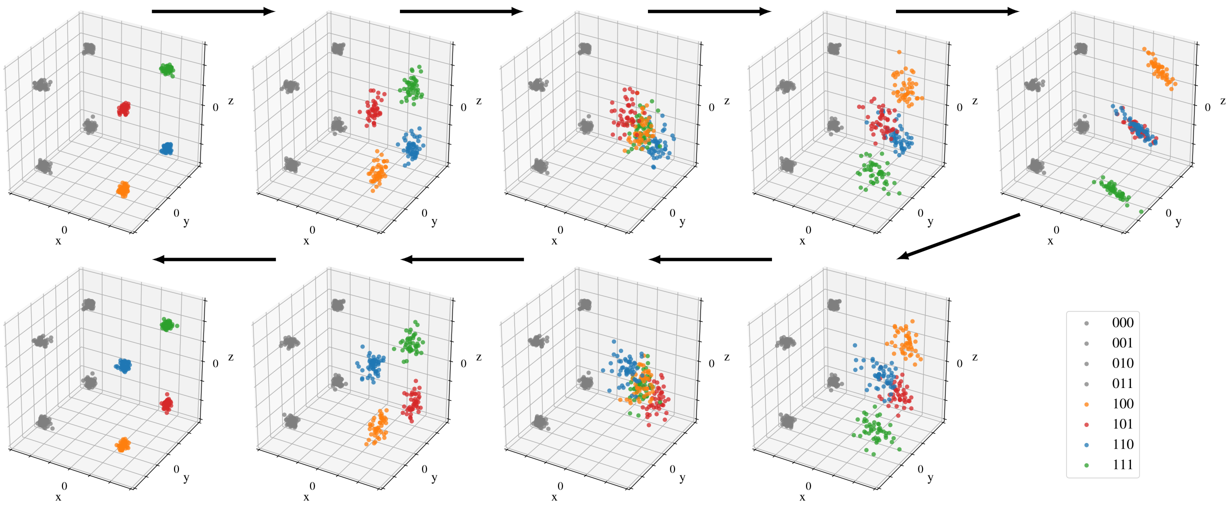


FIGURE 5.4.1. Particle ensemble initialized in equilibrium with $V^{\text{store}}(x, y, z)$ undergoing the Fredkin gate protocol with zero coupling to the thermal reservoir. Each snapshot of the state evolution is separated by a time interval of $\tau/8$, with the black arrows indicating forward time. Color encodes in which informational state each trial begins. The 101 (red) and 110 (blue) states only oscillate by a quarter period in the time ($\tau/2$) it takes the 100 (yellow) and 111 (green) states to oscillate by a half cycle. As the 100 and 111 trials return to their initial positions, the 110 and 101 states approach their final positions: a half cycle from where they started (right). The states have been swapped. Additional Animations are available online at `https://kylejray.github.io/fredkin/`.

$V(x, y', z', t) = V^{\text{store}}(x) + V^{yz}(x, y', z', t)$, where:

$$V(x, y', z', t) = \begin{cases} V^{\text{store}}(x, y, z) & \text{if } x < 0 \\ V^{\text{store}}(x) + \frac{ky'^2}{2} + 2kz'^2 & \text{if } x \geq 0 \end{cases}.$$

In our original coordinates, this passive Hamiltonian transforms the particle's state by swapping y and z, but only when when $x > 0$ ($m_x = 1$); thus, it implements the Fredkin gate.

For a particular trajectory $(x, y, z)(t)$, the work invested only comes from the initial and final instantaneous changes in the energy landscape, as noted above. Recall that $x(t)$ is exponentially unlikely to change sign, since the energy barrier between states is much higher than the vast majority of thermal fluctuations can access. Thus, we assume that paths maintain a single sign for $x(t)$. If $x(t)$ is negative, then there is no instantaneous change, as the system is held in the same double-well potential, so $W = 0$.

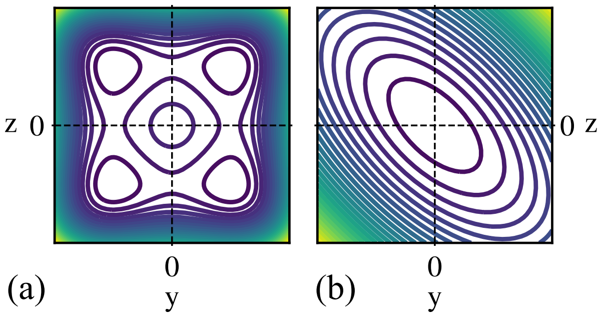


FIGURE 5.4.2. Slice of the potential energy landscape $V(x, y, z)$ in the $y-z$ plane: (a) information-storing domain ($x = -x_0$) and (b) controlled-swap domain ($x = x_0$) during the Fredkin gate operation.

That said, if $x(t)$ is positive, then the work invested also vanishes. For these trajectories, note that the $y-z$ subspace potential is symmetric with respect to exchange of the y and z coordinates and that the action of the map is to swap y and z. Thus, using the same arguments as for the bit flip, we see that the work production vanishes. The only work-producing trajectories are the exponentially suppressed barrier crossing events—so the average work production is nearly zero.

Figure 5.4.1 demonstrates the evolution of the phase space on an ensemble of initial conditions drawn from the equilibrium distribution of a quartic storage potential. As shown by the particle coloring, those that start in 110 and 101 swap while all others are fixed. Moreover, none of the particles' x-coordinates change informationally, confirming the effectiveness of the overall transformation.

5.5. Langevin Simulation

The preceding stipulated that the logical system be isolated from its thermal environment during the swap. It might not seem surprising then, that we are able to accomplish a work-free bit flip, given that other classical implementations of efficient reversible computing—such as ballistic computing with billiards [**131**]—necessarily operate in a dissipationless environment.

However, a key and somewhat surprising point is that the Fredkin gate implemented above tolerates imperfect isolation from its thermal environment. The gate's robustness to fluctuations separates it from other implementations that are dynamically unstable, such as billiard computing.

To demonstrate this, we investigated how robust the operation is to thermal agitation by using *underdamped* Langevin dynamics. A simulation was carried out by initializing particles in the equilibrium distribution with a thermal reservoir under a quartic information-storing potential $V^{\text{store}}(x, y, z)$. Next, as described above, we exert work on the system by turning on the computational potential V^{comp} in the region $x > 0$. However, rather than reducing the thermal coupling to $\lambda = 0$, we drop the coupling coefficient to a nonzero value in the weak coupling regime. This coupling value and potential are held fixed for time $\tau = \pi\sqrt{\mu/k}$. (The Appendix provides additional detail).

5.6. Thermodynamically Robust Fredkin Gate

The particles experience thermal fluctuations as the weak coupling to the bath perturbs their trajectories from the otherwise expected harmonic motion. The work gained from shutting off the potential will not generally be the same as the work invested to turn it on (as in the idealized case of zero thermal coupling). In fact, the Second Law guarantees that, generally, positive work is invested for such cyclical transformations, because the net change in equilibrium free energy is zero. Nevertheless, one expects the behavior to approximate the desired Fredkin-gate dynamics if the coupling is sufficiently weak. Figure 5.6.1 shows that the logical fidelity approaches unity. And, it does so with *zero slope*, revealing that this Fredkin gate implementation is robust even in the presence of thermal fluctuations.

This also gives evidence that the implementation should have practical use for reversible universal computing in a thermal environment. This regime is particularly well suited, for example, to superconducting flux qubits working in the classical regime [**137**] in which a tunable resistance can act as a control parameter for damping.

As expected and shown in Fig. 5.6.1, the work invested approaches zero with decreasing coupling. However, as the coupling to the thermal reservoir increases, the average work required to compute increases to multiples of $k_B T$. This cost scaling is evidence that the thermal agitation has become significant enough to take the particles appreciably far from their ideal (costless) trajectories; nevertheless the protocol maintains high fidelity, even in this regime. Note that the work cost

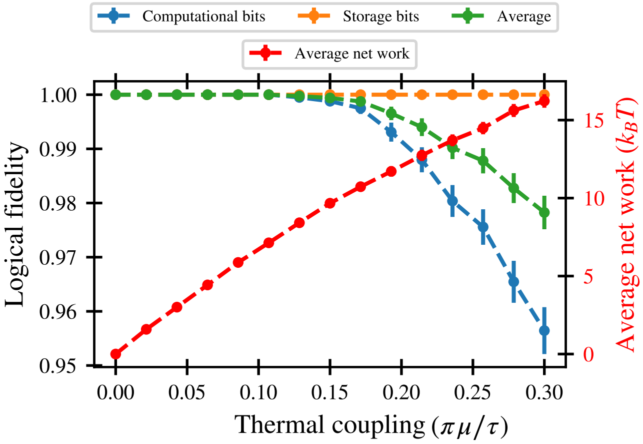


FIGURE 5.6.1. Logical fidelity (successful trials/total trials) in the low-coupling Fredkin gate and the average net work required to implement it for different values of the thermal coupling constant λ, measured in units of $\pi\mu/\tau$. *Computational bits* refers to states that fall in the region $x > 0$, where the computational potential is in effect.

is exponentially unlikely to come from trajectories that "jump" over the $x = 0$ boundary, since the storage bits maintain perfect fidelity.

5.7. The Future of Computation?

Ever since the first exorcism of Maxwell's demon [**4**], determining how much energetic input a particular computation requires has been a broadly-appreciated theoretical question. In the current century, however, the question has taken on a markedly practical bent; a familiar example is the evolution of Moore's Law from initially provocative speculations decades ago to now addressing material, thermodynamic, and fabrication restrictions [**138**, **139**, **140**, **141**, **142**]. Transistor-based microprocessing presents fundamental scaling challenges that strictly limit potential directions for future optimization, and these challenges are no longer speculative. Clock speed, to take one example, has been essentially capped for two decades due to energy dissipation at high rates [**143**, **144**]. By some measures, Moore's law is already dead—as integrated circuit manufacturers go vertical, rather than face the expense of creating smaller transistors for 2D circuits that yield only marginal gains [**145**, **146**, **147**].

In the preceding sections, we introduced a design framework and theory for an arbitrarily low-cost, high-speed bit swap, a logically-reversible gate (the only known logical framework with no nontrivial lower bound on its dissipation [**6**, **148**, **149**].) Additionally, we showed that a universal reversible gate—a Fredkin gate [**131**, **150**]—can be built by coupling three such devices together. However, any particular physically-instantiated implementation will come with its own restrictions and considerations that are likely to disallow performing the swap exactly as theorized. And so, an implementation linked to a particular substrate must be built and analyzed in its own right.

In order to test the physical feasibility of such protocols, we next present a physically-realizable device and control protocols that implement a bit swap gate that operates in the sub-$k_\mathrm{B}T$ energy regime using superconducting Josephson junctions (JJs)—a well-known and scalable microtechnology. This device was recently used to measure the thermodynamic performance of bit erasure [**151**,**152**]. That extensive experimental effort demonstrated in practical terms that the device proposed here is realizable with today's microfabrication technologies and allows for detailed studies of thermodynamic costs. We use extensive, physically-calibrated simulations to demonstrate that the device performance is robust and that momentum computing can support thermodynamically-efficient, high-speed, large-scale general-purpose computing that circumvents Landauer's bound. And so, the device's design and control protocol open up exploring the energy scales of highly energy-efficient, high-speed, general-purpose computing.

5.8. The Landauer

The first task is to come up with a reasonable benchmark to compare the efficiency of computational devices. While there are many different quantities one might wish to optimize, the perspective here sets the goal as minimizing the net work invested W when performing logical operations. It is well known that the most pressing physical limits on modern computation are power constraints [**153**], thus the measure is well suited to diagnose the problems with current devices as well as potential strengths of new ones.

For over half a century now *Landauer's Principle* has exerted a major impact on the contemporary approach to thermodynamic costs of information processing [**18**, **19**]. Its lower bound of

$k_B T \ln 2$ energy dissipated per bit erased has served as standard candle for energy use in physical information processing. To aid comparing other computing paradigms and protocols, we refer to this temperature-dependent information-processing energy scale as a *Landauer*: approximately a few zeptojoules at room temperature, and a few hundredths of a zeptojoule at liquid He temperatures.

To appreciate the potential benefits of momentum computing operating at sub-Landauer energies we ask where contemporary computing is on the energy scale. Consider recent stochastic thermodynamic analyses of single-electron transistor logic gates [**154**, **155**]—analogs to conventional CMOS technology. The upshot is that these technologies currently operate between 10^3 and 10^4 Landauers. More to the point, devices using CMOS-based technology will only ever be able to operate accurately above $\approx 10^2$ Landauers [**6**, **148**]. In short, momentum computing can promise substantial improvements in efficiency with no compromise in speed.

Environment	Temperature T	Thermodynamic Energy
	Kelvin(K)	Joules(J)
Microprocessor	373	5.2×10^{-21}
Room Temp	293	4.0×10^{-21}
Liquid N_2	77	1.1×10^{-21}
Liquid He	4.2	5.7×10^{-23}
1 K	1.0	1.4×10^{-23}
1 mK	0.001	1.4×10^{-26}

Table 5.8.1. Thermodynamic energy in environments at various temperatures.

Operation	Landauers (L)	Environment T	Energy
		Kelvin (K)	Joules (J)
CMOS gate [**154**]	7000	293	1.9×10^{-17}
CMOS gate [**155**]	3000	293	8.4×10^{-18}
CMOS bound [**6**, **148**]	100	293	2.8×10^{-19}
Bit Erase (Ideal) [**19**]	1	293	2.8×10^{-21}
Bit Erase (Ideal) [**19**]	1	1	9.6×10^{-24}
Bit Swap (JJ)	0.43	1	4.1×10^{-24}
Bit Swap (Ideal)	0	293	0
Bit Swap (Ideal)	0	1	0

Table 5.8.2. Landauers and work energies (Joules) for various information processing operations in environments and at temperatures where thermodynamic computers may operate.

Recently, the paradigm of *thermodynamic computing* emerged to frame probing the limits of efficient computation [**156**]. In this setting, Landauer's Principle says that $k_B T \ln 2$ energy units must be expended to erase a single bit of information. Beyond erasure, though, his principle also stands as a challenge—Can conventional computing paradigms operate at sub-Landauer scales? It seems not. Landauer's theory and follow-on results [**16**,**157**,**158**] and recent experiments [**49**,**159**] verified the lower bound.

To apply more broadly, *Landauer's Principle* generalizes to $W \geq k_{\mathrm{B}}T\Delta H$, where ΔH is the change in Shannon entropy between a computational system's initial and final information-bearing states [**16**, **17**, **160**]. Despite the Principle's generalization beyond bit erasure, the Landauer scale remains a familiar reference point for the energy costs of binary operations; its familiar use coming at the expense of ignoring specifics of any given logical operation [**51**].

An efficient bit-swap operation, for example, has zero generalized Landauer cost, as it is logically reversible. However, since many thermodynamic computing architectures do not have access to dynamics that can accomplish reversible computing efficiently, the Landauer scale provides a common reference to compare gate performance across physical substrates and design paradigms. It also facilitates comparing across substrates that operate at different temperatures. Table 5.8.1 lists thermodynamic energies for a range of physical environments. Table 5.8.2 gives Landauer work energies for various information processing operations in environments and at temperatures where thermodynamic computers operate.

5.9. The Energetic Scale of Momentum Computing

The Landauer cost stood as a reference for so long since bit erasure is the dominant source of unavoidable dissipation when implementing universal computing with transistor logic gates. It is the elementary binary computation that most changes the Shannon entropy of the distribution over memory states. In this way, one sees $k_{\mathrm{B}}T \ln 2$ not just as the cost of erasure, but as the cost of the maximally dissipative elementary operation on which conventional computing relies. And so, the Landauer naturally sets the energy scale for conventional computing.

Taking inspiration from Landauer's pioneering work, we investigate the cost of the most expensive operation necessary to physically implement universal momentum computing: a bit swap. The bit swap's dominance in the cost of universal momentum computing can be appreciated by considering the input-output mapping of the Fredkin gate—a 3-bit universal gate with memory

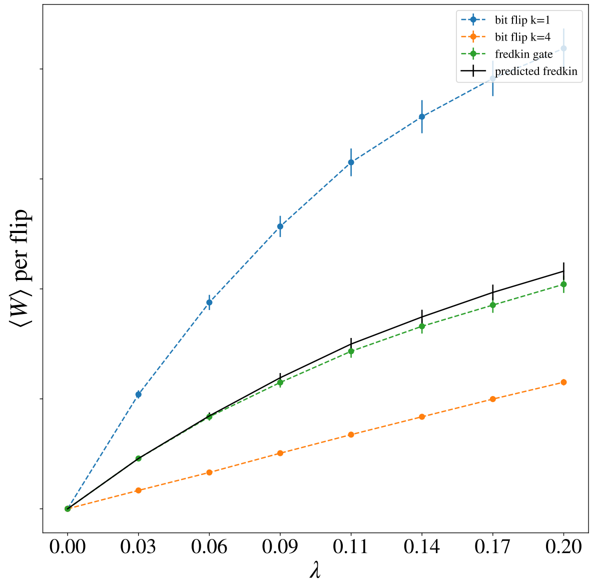


Figure 5.9.1. A display of the 'work per swap' in the Fredkin gate, and two 1D swaps. We see that the Fredkin gate cost lies very close to the line that we would expect if its cost was dominated by the costs of its component 1D swaps.

states $m_x m_y m_z$, $m_i \in \{0,1\}$. All inputs are preserved except for the exchange 101 ↔ 110. We can decompose the informational state space into two regions. If $m_x = 0$, the operation is simply an identity, which trivially is costless. If $m_x = 1$ and $m_y = m_z$, we once again have an identity. Thus, it is only the subspace of $m_x = 1$, where $m_y \neq m_z$ that a swap must take place.

Provided that the above holds true, the Fredkin gate from section 5.6 should have a work cost that is dominated by the different 1D swaps that make it up. The Fredkin gate essentially has three different swaps occurring at two different timescales: 110 ↔ 101 at a timescale of τ, and then two 111 ↔ 100 swaps at double the speed. The obvious assumption is that $W_{fred} = 2W_{fast} + W_{slow}$ where the work costs are the cost of the full Fredkin, the fast swap and the slow swap, respectively. In order to test this, simulations were preformed to supplement those of the thermally agitated Fredkin gate in figure 5.6.1: bit swaps at the slow speed $k = 1$ and the fast speed $k = 4$. Since the Fredkin gate has three swaps, we divide its cost by three to arrive at an average cost per swap. The results are shown in figured 5.9.1; remarkably, the naive assumption that the Fredkin gate is entirely dominated by the costs of the 1D swaps holds rather well. Thus, we view the cost of a bit swap as momnetum computing's analog to the cost of an erasure.

5.10. Physical Instantiation

Due to its conceptual simplicity the protocol from section 5.9 does not require any particular physical substrate. That said, the practical feasibility of performing such a computation must be addressed. One obvious point of practical concern is assuming the system can be isolated from its thermal environment during the computation. However, total isolation is not necessary. If $\tau \ll \tau_R$—the relaxation timescale associated with the energy flux rate between the system and its thermal bath—then the device performs close to the ideal case of zero coupling.

The simulations in section 5.6 showed that this class of protocol is robust: thermodynamic performance persists in the presence of imperfect isolation from the thermal environment, albeit at an energetic cost. Thus, a system that obeys significantly-underdamped Langevin dynamics is an ideal candidate as the physical substrate for bit swap.

We analyze in detail one physical instantiation—a *gradiometric flux logic cell* (Fig. 5.10.1), a mature technology for information processing. With suitable scale definitions, the effective degrees of freedom—Josephson phase sum φ and difference φ_{dc}—follow a dimensionless Langevin equation [**151**, **152**, **161**, **162**, **163**, **164**]:

$$dv' = -\lambda v' dt' - \theta \partial_{x'} U' + \eta r(t) \sqrt{2dt'} \ , \tag{5.2}$$

where $x' \equiv (\varphi, \varphi_{\text{dc}})$ and $v' \equiv (\dot{\varphi}, \dot{\varphi_{\text{dc}}})$ are vector representations of the dynamical coordinates. Enacting a control protocol on this system involves changing the parameters of the potential over time:

$$\begin{aligned} U'(t') &= U/U_0 \\ &= (\varphi - \varphi_x(t'))^2/2 + \gamma(\varphi_{\text{dc}} - \varphi_{\text{xdc}}(t'))^2/2 \\ &\quad + \beta \cos\varphi \cos(\varphi_{\text{dc}}/2) - \delta\beta \sin\varphi \sin(\varphi_{\text{dc}}/2) \ . \end{aligned} \tag{5.3}$$

The relationships between the circuit parameters and the parameters in the effective potential U' are as follows. $\varphi = (\varphi_1 + \varphi_2)/2 - \pi$ and $\varphi_{\text{dc}} = (\varphi_2 - \varphi_1)$, where φ_1 and φ_2 are the phases across the two Josephson elements; $\varphi_x = 2\pi\phi_x/\Phi_0 - \pi$ and $\varphi_{\text{xdc}} = 2\pi\phi_{xdc}/\Phi_0$, where Φ_0 is the magnetic flux quantum and (ϕ_x, ϕ_{xdc}) are external magnetic fluxes applied to the circuit; $U_0 = (\Phi_0/2\pi)^2/L$,

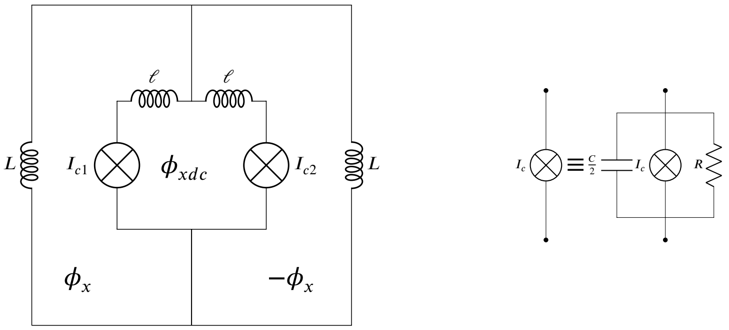


FIGURE 5.10.1. Gradiometric flux logic cell: The superconducting current has two important flow modes. One circulation around the inner loop—a DC SQUID. And, the other, a flow through the Josephson junctions in the inner loop and around the outer conductor pickup loops—an AC SQUID [**162**]. This is the origin of the variable subscripts to distinguish φ from φ_{dc} and φ_x from φ_{xdc}.

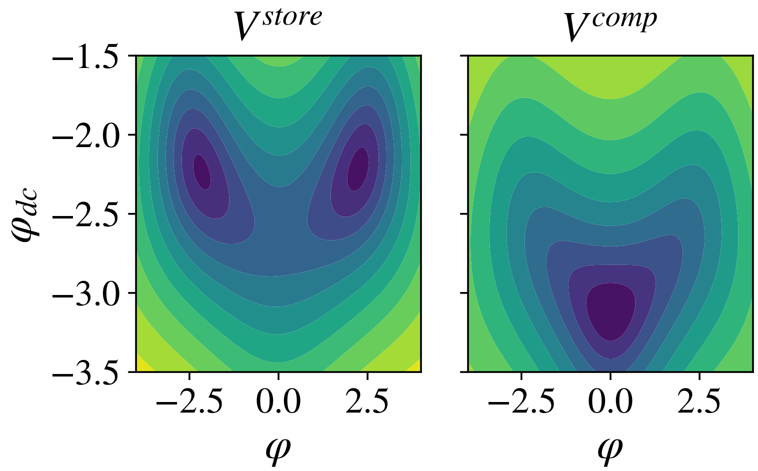


FIGURE 5.10.2. (Left) V^{store}, the bistable storage potential. (Right) V^{comp}, the "banana-harmonic" potential. These potential energy profiles serve as qualitative pictures to represent prototypical computational and storage potentials, and do not represent any particularly favorable parameter set.

$\gamma = L/2\ell$, $\beta = I_+ 2\pi L/\Phi_0$, and $\delta\beta = I_- 2\pi L/\Phi_0$, where L and 2ℓ are geometric inductances; and $I_\pm \equiv I_{c1} \pm I_{c2}$ are the sum and difference of the critical currents of the two Josephson junctions. All parameters are real and it is assumed that $\gamma > \beta > 1 \gg \delta\beta$.

Some particularly important parameters of U' are φ_x and φ_{xdc}, which control the potential's shape by where the the dynamical variables φ and φ_{dc} localize in equilibrium, and γ, which controls how quickly φ_{dc} localizes to the bottom of the quadratic well centered near $\varphi_{\mathrm{dc}} = \varphi_{\mathrm{xdc}}$. At certain control parameters $(\varphi_x, \varphi_{\mathrm{xdc}})$, the effective potential contains only two minima: one located at

$\varphi < 0$ and one at $\varphi > 0$. So, the device is capable of metastably storing a bit, as described above. In point of fact, the logic cell has been often used as a double well in φ with a controllable tilt and barrier height [**151**, **162**, **164**].

The Langevin equation's coupling constants, λ and η, determine the rate of energy flow between the system and its thermal environment and the. They depend on the parameters L, R, and C. In the regimes at which one typically finds L, C, and R and with temperatures around 1 K, the system is very underdamped; ring-down times are $\mathcal{O}(10^3)$ oscillations about the local minima. (Notably, the device thermalizes at a rate proportional to R^{-1}. A tunable R allows the device to transition from the underdamped to overdamped regime, allowing for rapid thermalization, if desired.) Finally, θ is a dimensionless factor that depends on the relative inertia of the two degrees of freedom, it depends on the circuit architecture. Appendix 5.C gives the equations of motion and thorough definitions of all parameters and variables in terms of dimensional quantities.

5.10.1. Realistic Protocol. With the device's physical substrate set, we now show how to design energy-efficient bit-swap control protocols. There are four parameters that depend primarily on device fabrication: I_{c1}, I_{c2}, R, and C. Two that depend on the circuit design: L and ℓ. And, four that allow external control: φ_x, φ_{xdc}, T (the environmental temperature), and τ (the computation time). Without additional circuit complexities to allow tunable L, R, and C, we assume that once a device is made, any given protocol can only manipulate φ_x, φ_{xdc}, T, and τ. A central assumption is that computation happens on a timescale over which the thermal environment has minimal effect on the dynamics, so the primary controls are φ_x, φ_{xdc}, and τ. φ_x is associated with asymmetry in the informational subspace, and will only take a nonzero value to help offset asymmetry from the $\delta\beta$ term in U'. Thus, φ_{xdc} primarily controls the difference between V^{comp} and V^{store}, while τ governs how long we subject the system to V^{comp}.

V^{store} must be chosen to operate the device in a parameter regime admitting two minima on either side of $\varphi = 0$ as in Fig. 5.10.2. They must also be sufficiently separated so that they are distinct memory states when immersed in an environment of temperature T.

In the ideal case, V^{comp} is a quadratic well with an oscillation period $\tau = \pi\sqrt{m/k}$. However, U will never give an exact quadratic well unless $\beta = \delta\beta = 0$. So, a suitable replacement is necessary. The closest approximate is at the relatively obvious choice $\varphi_{\text{xdc}} = -2\pi$. In this case, the minima of

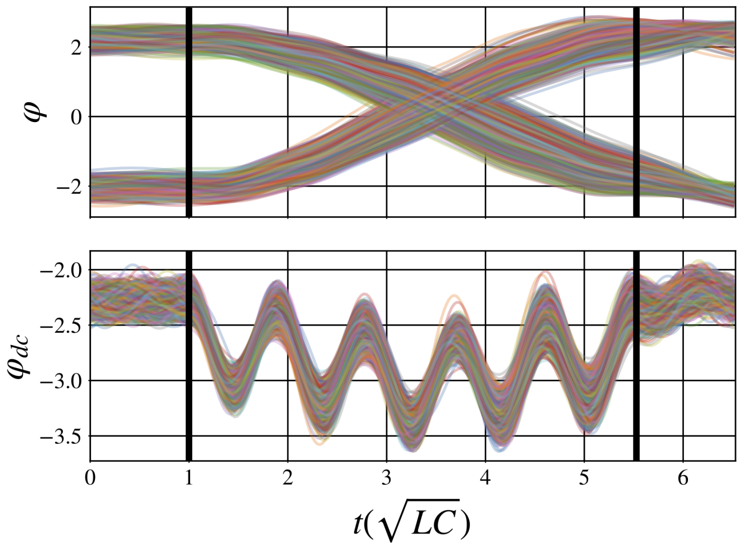


FIGURE 5.10.3. A dynamic computation: $1,500$ trajectories from V^{store}'s equilibrium distribution in the φ (top) and φ_{dc} (bottom) dimensions. V^{comp} is applied at $t \in (1, 1+\tau)$, denoted by heavy black lines. φ_{dc} oscillations are several times faster than the others, as expected when $\gamma \gg 1$. The work done on the system by the control apparatus, $W_0 = V^{\text{comp}}(t=1) - V^{\text{store}}(t=1)$, by its intervention at $t=1$ is largely offset by the work absorbed into the apparatus by its intervention at $t = 1+\tau$, $W_\tau = V^{\text{store}}(t=1+\tau) - V^{\text{comp}}(t=1+\tau)$, when V^{comp} re-engages. Visually, we can track this energy flux by the nonequilibrium oscillations induced at $t=1$ and the return to a near-equilibrium distribution at $t = 1+\tau$. Time is measured in units of $\sqrt{LC}$, which is $\approx$ 2ns for the JJ device. Animations of the protocol and a sample of simulated trajectories can be found at `https://kylejray.github.io/gslmc/`

both the quadratic and the periodic part of the potential lie on top of each other and the potential is well approximated by a quadratic function over most of the relevant position-domain.

However, due to restrictions on V^{store}, transitioning between V^{store} and V^{comp} may induce unnecessarily large dissipation since the oscillations in the φ_{dc} dimension have a large amplitude. (See Appendix 5.D for details.) Instead, to dissipate the minimum energy, the control parameters must balance placing the system as close as possible to the pitchfork bifurcation where the two wells merge, while still maintaining dynamics that induce the $\varphi < 0$ and $\varphi > 0$ informational states to swap places due to an approximately harmonic oscillation. Near this parameter value, one typically finds a "banana-harmonic" potential energy landscape. (See Fig. 5.10.2 for a comparison of the distinct potential profiles for storage and computation.)

5.10.2. Computation Time. The final design task determines the computation timescale τ. Under a perfect harmonic potential, the most energetically efficient τ is simply $\pi\sqrt{m/k}$. This ensures that $x(t=0) = -x(t=\tau)$. Since the design has an additional degree of freedom beyond that necessary—the φ_{dc} dimension—however, we must not only ensure our information-bearing degree of freedom switches signs, but also ensure that $\varphi_{\mathrm{dc}}(t=0) \approx \varphi_{\mathrm{dc}}(t=\tau)$. This means that during time τ, the φ variables must undergo $n+1/2$ oscillations and the φ_{dc} variables must undergo an integer number of complete oscillations. (See Fig. 5.10.3.) Hence, τ must satisfy matching conditions for the periods of the oscillations in both φ and φ_{dc} during the computation:

$$\omega\tau \approx (2n-1)\pi$$

$$\omega_{dc}\tau \approx 2n\pi \ .$$

Figure 5.10.4 showcases this by displaying the behavior observed during simulations near the ideal timescale. The local work minima coincide with local minima in the average kinetic energy, but not every kinetic energy minimum coincides with a work minimum. While there are kinetic energy minima every half-integer oscillation in φ_{dc}, only integer multiples of φ_{dc} oscillations yield minimum work.

The equations of motion governing the system are stochastic, dissipative, and nonlinear, so the frequencies of the different oscillations ω, ω_{dc} are nontrivial nonlinear stochastic mappings of device parameters, initial positions, and protocol parameters. They are not easily determined analytically. However, they change smoothly with small changes in the parameters they depends on. Thus, we were able to use an algorithmic approach to find the timescales that yield local minima and explore the regions surrounding them.

5.10.3. Physically-Calibrated Bit Swap. We are most interested in the effect of parameters that are least constrained by fabrication. And so, all simulations assume constant fabrication parameters with I_+, R, and C set to $2.0\,\mu\mathrm{A}$, $371\,\Omega$, and $4.0\,\mathrm{nF}$, respectively. To explore how the I_- asymmetry affects work cost, we simulated protocols with both a nearly-symmetric device ($I_- = 7\,\mathrm{nA}$) and a moderately-asymmetric device ($I_- = 35\,\mathrm{nA}$). Given devices with the parameters above, what values of the other parameters yield protocols with minimum work cost? This involves

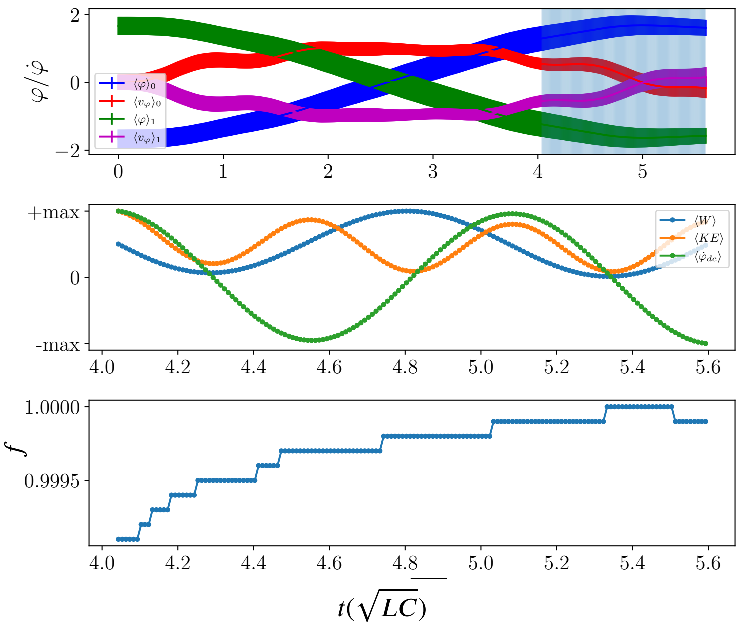


FIGURE 5.10.4. Performing a successful and low-cost bit swap: (Top) Ensemble averages, conditioned on initial memory state, of the fluxes and their conjugate momenta. Line width tracks the distribution's variance. The shaded region indicates timescales that are potentially successful swap operations. These are probed more closely in the bottom two plots. (Middle) Ensemble averaged work, kinetic energy, and conjugate momentum in the φ_{dc} coordinate. Note that work minima occur only at whole-integer oscillations of the momentum. Each dataset is scaled to its maximum value, so that it saturates at 1. This emphasizes the qualitative relationships rather than the quantitative. (Bottom) Computational fidelity f of the swap, approaching a perfect swap.

a twofold procedure. First, create a circuit architecture by setting L and γ, thus fully specifying the device; details in Appendix 5.E. Second, determine the ideal protocols for that combination of device parameters.

5.10.4. Computational Fidelity. To determine the best successful protocol, we must define what a successful bit swap is. First, we set a lower bound for the *fidelity* f: $f \geq 0.99$. We define f over an ensemble of N independent trials as: $f = 1 - N_e/N$, with N_e counting the number of failed trials, trials for which $\mathrm{sign}[\varphi(t=0)] = \mathrm{sign}[\varphi(t=\tau)]$. Second, the distribution over both $\varphi(t=\tau)$ and $\varphi(t=0)$ must be bimodal with clear and separate informational states. The criteria used for this second condition is:

$$\langle \varphi < 0 \rangle + 3\sigma_{\varphi<0} < \langle \varphi > 0 \rangle - 3\sigma_{\varphi>0} \ , \tag{5.4}$$

were σ_s and $\langle s \rangle$ are standard deviations and means of φ conditioned on statement s being true.

The final choice concerns the initial distribution from which to sample trial runs. For this, we used the equilibrium distribution associated with V^{store} with the environmental temperature set to satisfy $k_{\mathrm{B}}T = 0.05 U_0$. Here, we ensure fair comparisons between different parameter settings by fixing a relationship between the potential's energy scale and that of thermal fluctuations. This resulted in temperatures from $400 - 1400$ mK, though it is possible to create superconducting circuits at much higher temperatures [**165**, **166**, **167**, **168**] using alternative materials.

Sampling initial conditions from a thermal state assumes no special intervention created the system's initial distribution. We only need wait a suitably long time to reach it. Moreover, this choice is no more than an algorithmic way to select a starting distribution. It is not a limitation or restriction of the protocol. Indeed, if some intervention allowed sampling initial conditions from a lower-variance distribution, it could be leveraged into even higher performance.

5.11. Performance

Appendix 5.E lays out the computational strategy used to find minimal $\langle W \rangle$ implementations among the protocols that satisfy the conditions above. Since the potential is held constant between $t = 0$ and $t = \tau$, work is only done when turning V^{comp} on at $t = 0$ and turning it off at $t = \tau$. The ensemble average work done at $t = 0$ is $W_0 \equiv \langle V^{\mathrm{comp}}(\varphi(0), \varphi_{\mathrm{dc}}(0)) - V^{\mathrm{store}}(\varphi(0), \varphi_{\mathrm{dc}}(0)) \rangle$ and returning to V^{comp} at time τ costs $W_\tau \equiv \left\langle V^{\mathrm{store}}(\varphi(\tau), \varphi_{\mathrm{dc}}(\tau)) - V^{\mathrm{comp}}(\varphi(\tau), \varphi_{\mathrm{dc}}(\tau)) \right\rangle$. Thus, the mean net work cost is the sum $\langle W \rangle = W_0 + W_\tau$. As we detail shortly, this yielded large regions of parameter space that implement bit swaps at sub-Landauer work cost. This result and others demonstrate the notable and desirable aspects of momentum computing: accuracy, low thermodynamic cost, and high speed. Let's recount these one by one.

5.11.1. Accuracy. Tradeoffs between a computation's fidelity and its thermodynamic cost are now familiar—an increase in accuracy comes at the cost of increased W or computation time [**46**, **47**, **48**, **49**, **50**, **51**]. These analyses conclude that accuracy generally raises computation costs.

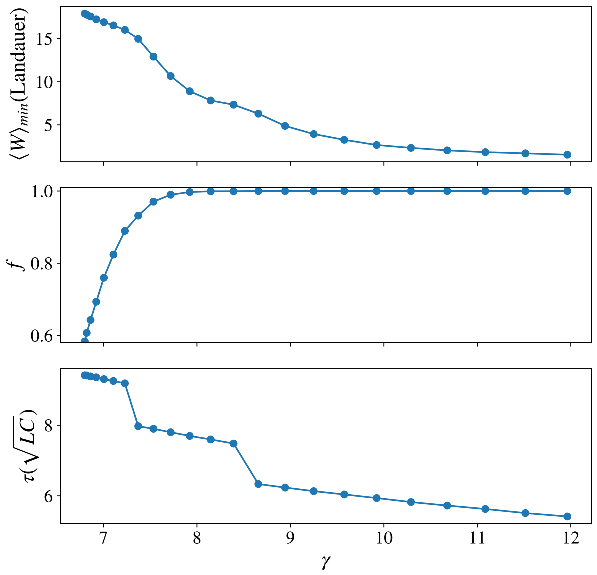


FIGURE 5.11.1. Performance of the minimum work protocol as γ, the ratio of device inductances, goes from a region where the computation fails ($f < 0.99$) to a region of perfect fidelity ($f = 1.0$). Note that in the parameter space region in which the computation becomes successful, the work costs decrease as the fidelity approaches unity. Finally, τ decreases as the work cost minimizes to ≈ 1 Landauer—showing that the work cost does not display $1/\tau$ adiabatic compute-time scaling. The parameter γ controls the starting parameters for the suite of simulations represented by each data point and should not be read as the primary independent variable responsible for the behavior. Rather, the plots show τ, f, and $\langle W \rangle_{min}$ evolving jointly to more preferable values.

Momentum computing does not work this way. In fact, it works in the opposite way. The low cost of a momentum computing protocol comes from controlling the distribution over the computing system's final state. Due to this, fidelity and low operation cost are not in opposition, but go hand in hand, as Figs. 5.10.4 and 5.11.1 demonstrate.

5.11.2. Low Thermodynamic Cost. Conventional computing, based on transistor-network steady-state currents, operates nowhere near the theoretical limit of efficiency for logical gates. Even gates in Application Specific Integrated Circuits (ASICs) designed for maximal efficiency operate on the scale of $10^4 - 10^6$ Landauers [**169**, **170**]. The physically-calibrated simulations described above achieved average costs well below a Landauer for a wide range of parameter values with an absolute minimum of $\langle W \rangle_{\text{min}} = 0.43$ Landauers, as shown in Figure 5.11.2 (left). For the less-ideal asymmetric critical-current device (right panel), the cost increases to only $\langle W \rangle_{\min} = 0.60$ Landauers. And, the bulk of the protocols we explored operated at < 10 Landauers. Altogether, the momentum computing devices operated many orders of magnitude lower than the status quo. Moreover, the wide basins reveal robustness in the device's performance: an important feature for practical optimization and implementation.

5.11.3. High Speed.

Paralleling accuracy, the now-conventional belief is that computational work generally scales inversely with the computation time: $W \sim 1/\tau$ [**46**, **52**, **53**, **54**]. Again, this is not the case for momentum computing, as Figs. 5.10.4 and 5.11.1 demonstrate. Instead, there are optimal times τ^* that give local work minima and around which the work cost increases.

Optimal τ^*s are upper bounded: the devices must operate *faster* than particular timescales—timescales determined by the substrate physics. The bit swap's low work cost requires operating on a timescale faster than the rates at which the system exchanges energy and information with the environment. Thus, momentum computing protocols have a *speed floor* rather than a speed limit.

However, even assuming perfect thermal isolation there is a second bound on τ^*. The computation must terminate before the initially localized ensemble—storing the memory—decoheres in position space due to dispersion. For our JJ device this is the more restrictive timescale. Due to local curvature differences in the potential, the initially compact state-space regions corresponding to peaks of the storage potential's equilibrium distribution begin to decohere after only one or two oscillations. Once they have spread to cover both memory states, the stored information is lost. This means it is most effective to limit the duration of the swap to just a half-oscillation of the φ coordinate. For our devices, this typically corresponds to operating on timescales < 15 ns.

5.12. Other Approaches to Reversible and Ballistic Computing

Reversible computing implementations of various operations have been proposed many times over many decades. Perhaps the most famous is the Fredkin billiards implementation [**131**]. While ingenious, it suffers from inherent dynamical instability (deterministic chaos) and cannot abide any interactions with the environment. At the other end of the spectrum is a family of superconducting adiabatic implementations [**171**, **172**, **173**, **174**, **175**, **176**, **177**, **178**]. These are low cost in terms of

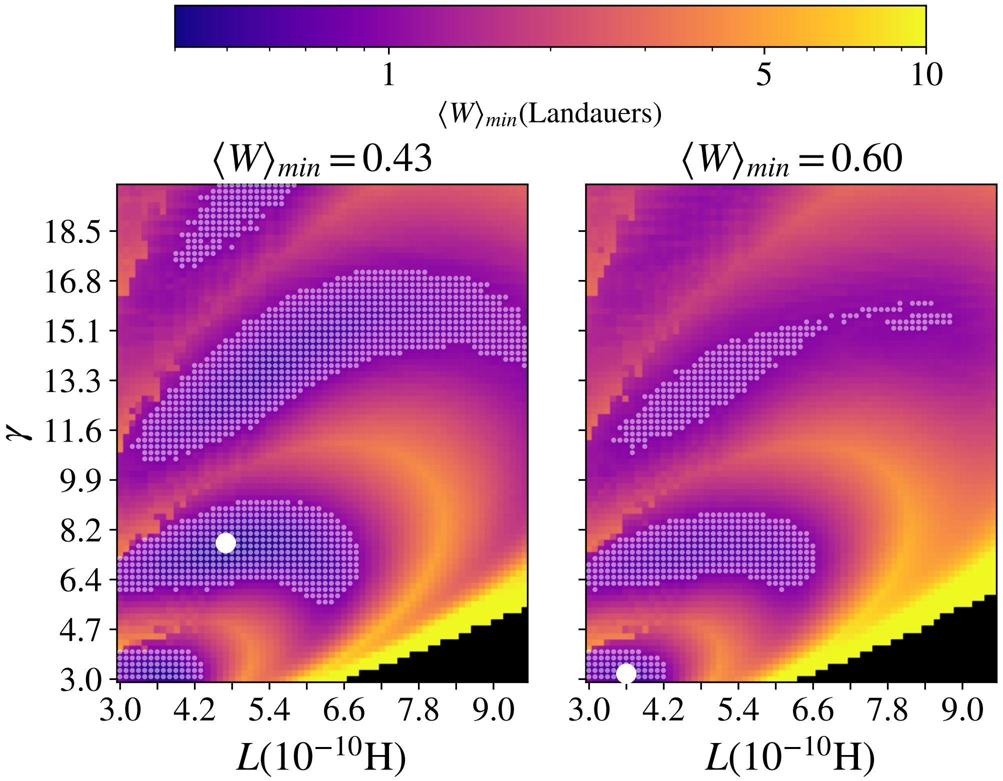


FIGURE 5.11.2. Thermodynamic energy cost $\langle W\rangle_{\text{min}}$ for momentum-computing bit-swap over 5,120 parameter combinations of L and γ. (Left) Slightly asymmetric device with $I_- = 7\,\text{nA}$ gives the overall minimum $\langle W\rangle_{\text{min}} = 0.43$ Landauers (large solid white circle). (Right) Substantially asymmetric device with $I_- = 35\,\text{nA}$ gives the overall minimum $\langle W\rangle_{\text{min}} = 0.60$ Landauers (large solid white circle). (Both) Small white circles indicate parameter values with protocols yielding $\langle W\rangle_{\text{min}} < 1$ Landauer. Black squares (lower right in each) represent parameter values where no successful swap was accomplished. Note that when the asymmetry is low, it can effectively be offset by the parameter φ_x, but for higher asymmetry, protocols that cost less than 1 Landauer are less common.

dissipation and are stable, but they suffer from fundamental speed limits due to the adiabaticity requirement: $\langle W\rangle \propto 1/\tau$.

Other recent implementations [**179**, **180**, **181**] of reversible logic using JJs are more akin to the proposal at hand, in that they require nearly-ballistic dynamics and attempt to recapture the energy used in a swap at the final step. While these implementations are markedly different, their motivation follows similar principles. Particularly, the framework for asynchronous reversible computing proposed in [**180**, **181**] might serve as a testbed for momentum computing elements.

Another distinguishing feature of the present design is that the phenomenon supporting the computing is inherently linked to microscopic degrees of freedom evolving in the device's phase space. This moves one closer to the ultimate goal of using reversible nanoscale phenomena as the primitives for reversible computing—a goal whose importance and difficulty were recognized by Ref. [**182**]. Working directly with the underlying phase space also allows incorporating the thermal environment. And, this facilitates characterizing the effect of (inevitable) imperfect isolation from the environment.

It is worth noting the similarity between the optimal timescales τ^* and the principal result in Ref. [**183**] in which a similar local minima emerges when comparing thermodynamic dissipation to computation time. These minima also come from certain matching conditions between the rate of thermalization and the system's response time to its control device. Another qualitatively similar result [**184**] found faster operation could lead to reduced errors in overdamped JJs under periodic driving. These similarities could point to a more general principle at play.

5.13. What Next?

Rate-equation dynamics is certainly a venerable and powerful framework, central to reaction kinetics in chemistry [**185**, **186**] and key to the master equations of applied statistical mechanics [**111**, **118**, **119**]. Due to the remarkable successes of continuous-time Markov chain predictions of many thermodynamic behaviors, it might seem natural to claim that to be "physically realizable", thermodynamic computing and biological information processing should *only* be described and analyzed as rate-equation dynamics [**116**].

The results here demonstrated that this does not hold generally. And so, it cannot form a complete basis for thermodynamic computing. Moreover, it levies a heavy penalty, precluding engineering and analyzing Maxwellian information ratchets, which are the physical equivalent of Turing machines [**157**, **187**, **188**, **189**]. The limits are especially draconian, since they preclude efficient time-symmetrically controlled computations consisting of involutions [**70**]— such operations are composed of bit swaps and identity maps in positional memory.

As a constructive alternative, we proposed employing continuous-time hidden Markov chains to realize *non-Markovian momentum computing.* We demonstrated it provides a more complete framework, using two explicit examples that are forbidden if one is restricted to rate equations to describe the evolution between memory states [**116**]. Additionally, we introduced explicit mechanisms for implementing both in finite time with zero work, proving them "physically realizable". However, we did fully acknowledge the increased analytical complexity posed by CTHMC dynamics. Fortunately, requisite tools have been developed that render the behaviors analytically tractable and in closed form [**190**, **191**].

Our detailed, thermodynamically-calibrated simulation of microscopic trajectories demonstrated that momentum computing can reliably (i) implement a bit swap at sub-Landauer work costs at (ii) nanosecond timescales in (iii) a well-characterized superconducting circuit.

These simulations serve two main purposes. The first highlights momentum computing's advantages. The proposed framework uses the continuum of momentum states to serve as the auxiliary system that allows a swap. In doing so, it eliminates the associated tradeoffs between energetic, temporal, or accuracy costs that are commonly emphasized in thermodynamic control analyses [**46**, **47**, **48**, **49**]. Momentum computing protocols are holistic in that low energy cost, high fidelity, and fast operation times all come from matching parallel constraints rather than competing ones.

The second purpose points out key aspects of the proposed JJ circuit's physics. The simulations reveal several guiding principles—those that contribute most to decreasing work costs for the proposed protocols. The system is so underdamped that thermal agitation is not the primary cause of inefficiency. The two main contributors are (i) the appearance of dispersive behavior in the dynamics of an initially-coherent region of state space and (ii) asymmetries inherent to the device that arise from differing critical currents in the component superconducting JJ elements. Notably, if the elements are very close to each other in I_c, then symmetry can be effectively restored by setting the control parameter φ_x to counteract the difference. However, the more asymmetry, the harder it is to find ultra low-cost protocols; cf. Fig. 5.11.2 left and right panels. Note, too, that initial-state dispersion can be ameliorated by using a V^{comp} that is as harmonic (quadratic) as possible. However, this typically requires lower inductance L, possibly complicating circuit fabrication. Additionally, the potential-well separation parameter β's linear dependence on L hinders the system's ability to create two distinct states during information storage. Though these tradeoffs are complicated, our simulations suggest that dispersion can be controlled, yielding swap protocols with even lower work costs.

Since the protocol search space is quite high-dimensional and contains many local-minima, we offer no proof that the protocols found give the global work minimum. Very likely, the thermodynamic costs and operation speed of our proposed JJ momentum computing device can be substantially improved using more sophisticated parameter optimization and alternative materials. Even with the work cost as it stands, though, sub-Landauer operation represents a radical change from transistor-based architectures. One calibration for this is given in the recent stochastic thermodynamic analysis of a NOT gate composed of single-electron-state transistors [**154**] that found work costs 10^4 times larger.

Note, too, that running at low temperatures requires significant off-board cooling costs, as required in superconducting quantum computing. Our current flux qubit implementation requires operating at liquid He temperatures [**151**, **152**]. However, there are also JJs that operate at N_2 temperatures, promising system cooling costs that would be 2 to 3 orders of magnitude lower [**165**, **166**, **167**, **168**].

Additionally, the physics necessary to build a momentum computing swap—underdamped behavior and controllable multiwell dynamics—is far from unique to superconducting circuits. As an example, nanoelectromechanical systems (NEMS) are another well-known technology that is scalable with modern microfabrication techniques. NEMS provide the needed nonlinearity for multiple-well potentials, are extremely energy efficient, and have high Q factors even while operating at room temperature [**192**, **193**, **194**]. Momentum computing implemented with NEMS rather than superconductors completely obviates the cooling infrastructure and so may be better suited for large-scale implementations.

That said, the JJ implementation at low temperatures augmented with appropriate calorimetry will provide a key experimental platform for careful, controlled, and detailed study of the physical limits of the thermodynamic costs of information processing. Thus, these devices are necessary to fully understand the physics of thermodynamic efficiency. And so, beyond technology impacts, the proposed device and protocols provide a fascinating experimental opportunity to measure energy flows that fluctuate at GHz timescales and at energy scales below thermal fluctuations. Success in these will open the way to theoretical investigations of the fundamental physics of information storage and manipulation, time symmetries, and fluctuation theorems [**50**, **195**].

Appendix

5.A. Langevin Dynamics for the Fredkin Gate

To explore the performance of the proposed Fredkin gate protocol when there is nonzero coupling to a thermal bath, we modeled the system as obeying the Langevin equations of motion:

$$\begin{aligned} dq &= v_q dt \\ \mu\, dv_q &= -\lambda v_q dt - \partial_q V(q,t) dt + \sqrt{2 k_{\mathrm{B}} T \lambda}\, r(t) \sqrt{dt}\ , \end{aligned}$$

over three position coordinates $q = x$, y, or z. Here, v_q is the corresponding velocity, m is the mass, λ is the damping coefficient, and $r(t)$ is a memoryless Gaussian random variable with zero mean and unit variance. We used a quartic storage potential of the form $V^{\text{store}}(q) = \alpha q^4 - \beta q^2$ (see Fig. 5.4.2) with coefficients α and β.

To simulate the above system, we nondimensionalized the equations of motion. First, we defined nondimensional quantities via the following equalities ($\widetilde{\cdot}$ denotes a nondimensional quantity):

$$t = \widetilde{t}\sqrt{\frac{\mu}{k}} \qquad q = \widetilde{q}\sqrt{\frac{k_B T}{k}} \qquad v_q = \widetilde{v}_q \frac{q/\widetilde{q}}{t/\widetilde{t}} = \widetilde{v}_q \sqrt{\frac{k_B T}{\mu}}$$

$$\alpha = \widetilde{\alpha}\frac{k^2}{k_B T} \qquad \beta = \widetilde{\beta} k \qquad V = \widetilde{V} k_B T.$$

Inserting these scales into the equation of motion, yields the nondimensional potentials:

$$\begin{aligned} \widetilde{V}^{\text{store}}(\widetilde{q}) &= \widetilde{\alpha}\widetilde{q}^4 - \widetilde{\beta}\widetilde{q}^2 \\ \widetilde{V}^{\text{comp}}(\widetilde{y}', \widetilde{z}') &= \frac{1}{2}\widetilde{y'}^2 + 2\widetilde{z'}^2\ , \end{aligned}$$

and the following equation of motion for the nondimensional variables,

$$\begin{aligned} d\widetilde{q} &= \widetilde{v}_q d\widetilde{t} \\ d\widetilde{v}_q &= -\gamma\widetilde{v}\, d\widetilde{t} + \partial_{\widetilde{q}}\widetilde{V} d\widetilde{t} + \sqrt{2}\eta\, r(\widetilde{t})\sqrt{d\widetilde{t}}\ , \end{aligned}$$

where and γ and η are two additional nondimensional parameters implicitly defined as being equal to whatever is left over after the substitution. The first of these two is $\gamma = \lambda/\sqrt{\mu k}$, suggesting that γ is a nondimensional version of the thermal coupling parameter $\lambda = \sqrt{\mu k}\gamma$. Plugging this definition of λ into the expression for η yields $\eta = \sqrt{\gamma}$. For all simulations, the following nondimensional parameters were fixed: $\widetilde{\tau} = \pi$, $\widetilde{\alpha} = 2$, $\widetilde{\beta} = 16$, where $\widetilde{\tau}$ is the nondimensional duration of the computation interval.

These choices are equivalent to relationships between the dimensional parameters. The following three equalities held for all simulations: (i) $\tau = \pi\sqrt{\mu/k}$, (ii) $w = 2\sqrt{k_{\mathrm{B}}T/k}$, and (iii) $h = 32k_{\mathrm{B}}T$, where τ is the dimensional duration of the computation interval, $w = \sqrt{\beta/2\alpha}$ is the positional distance from the central maximum to the minima in the one-dimensional storage potential V^{store}, and $h = \beta^2/4\alpha$ is the energy difference between those points.

As a final note, we can use the expression for τ to write the relationship between λ and γ as $\lambda = \pi\mu\gamma/\tau$. Thus, we see that setting $\gamma = 1$, for example, corresponds to setting $\lambda = \pi\mu/\tau$. All simulations were carried out by simulating the nondimensionalized equations above and then converting to dimensional relationships using the relevant scales. For clarity, the next section discusses the simulation exclusively in terms of dimensional variables and parameters.

5.B. Simulation and Figure Generation for the Fredkin Gate

Figure 5.6.1 was generated from simulation using the following procedure. First, an ensemble of 20,000 initial values were chosen from an approximate equilibrium distribution of $V^{\mathrm{store}}(x, y, z)$ using the Monte Carlo algorithm. Second, this ensemble was thermalized while coupled to a bath ($\lambda = \pi\mu/\tau$) until the ensemble energy changed by no more than 1 part in 1,000 over a time interval of $\sqrt{\mu/k}$. This ensemble was then used as the start state for the Fredkin gate operation. Third, for each value of thermal coupling tested, λ dropped down to a low coupling value $\lambda \in (0, \frac{3}{10}\frac{\pi\mu}{\tau})$

and exposed the particles to the computational potential:

$$
\begin{aligned}
V(x, y', z', t) &= V^{\text{store}}(x) + V^{yz}(x, y', z', t) \\
&= \begin{cases} V^{\text{store}}(x, y, z) & \text{if } x < 0 \\ V^{\text{store}}(x) + \frac{ky'^2}{2} + 2kz'^2 & \text{if } x \geq 0 \end{cases} .
\end{aligned}
\tag{S5}
$$

Fourth, we measured the work required to change the potential across our ensemble. Fifth, the potential was then held fixed for the computation duration τ using an integration step $dt \approx 0.0005\tau/\pi$. Finally, immediately following the computation interval, we measured the second work contribution—the work that would be harvested by dropping the potential back to V^{store}. The average net work is the ensemble average difference between the work invested when raising the potential and the work harvested when lowering it. The plot displays 3σ error bars. The errors, though, are sufficiently small that they do not show up appreciably. Statistical errors were estimated using standard procedures for sample means and proportions.

Figure 5.4.1 was generated by starting the particles in the approximate equilibrium distribution described above and running the simulation above with $\lambda = 0$, to simulate dissipationless oscillatory dynamics. For clarity, the plot shows a sample of 200 trials, rather than the full $20,000$. Figure 5.4.1 gives a more complete picture, with snapshots every $\tau/8$.

All the simulations of the nondimensional equations of motion above employed a fourth-order Runge-Kutta method for the deterministic portion and Euler's method for the stochastic portion of the integration. (Python NumPy's Gaussian number generator was used to generate the memoryless Gaussian variable r(t).)

5.C. Flux Qubit Dimensionless Equations of Motion

In terms of the dimensional degrees of freedom, the flux qubit equations of motion are [**151**, **162**, **164**]:

$$\ddot{\widehat{\varphi}} = -\frac{2}{RC}\dot{\widehat{\varphi}} - \frac{1}{C}\partial_{\widehat{\varphi}} U(\widehat{\varphi}, \widehat{\varphi}_{dc}) \tag{S6}$$

$$\ddot{\widehat{\varphi}}_{dc} = -\frac{2}{RC}\dot{\widehat{\varphi}}_{dc} - \frac{4}{C}\partial_{\widehat{\varphi}_{dc}} U(\widehat{\varphi}, \widehat{\varphi}_{dc}) \ , \tag{S7}$$

where the dimensional $\widehat{\varphi}$s are related to the main text's dimensionless fluxes and phases by the magnetic flux quantum $2\pi/\Phi_0$. With the addition of thermal noise, the Langevin equation is:

$$dv_i = -\frac{\nu_i}{m_i} v_i dt - \frac{1}{m_i}\partial_{x_i} U(x) dt + \frac{1}{m_i} r(t)\sqrt{2\nu_i \kappa dt} \ , \tag{S8}$$

where $\kappa \equiv k_B T$. Matching these variables to the equations of motion yields:

$$x = (\widehat{\varphi}, \widehat{\varphi}_{dc}) \tag{S9}$$

$$v = \left(\dot{\widehat{\varphi}}, \dot{\widehat{\varphi}}_{dc}\right) \tag{S10}$$

$$m = \left(C, \frac{C}{4}\right), \text{ and} \tag{S11}$$

$$\nu = \left(\frac{2}{R}, \frac{1}{2R}\right) \ , \tag{S12}$$

where subscript i has been dropped in favor of a vector representation.

The task is to write each physical quantity z in terms of a dimensional constant and dimensionless variable by defining scaling factors according to the following prescription: $z \equiv z' z_c$, where z_c is a dimensional constant.

Setting $m_c = C$ and $\nu_c = 1/R$ are obvious choices. Additionally, since the potential factors into $U = U_0 \times U'(\frac{2\pi}{\Phi_0} \cdot x)$, a good choice for positional scaling is $x_c = \Phi_0/2\pi$.

It is advantageous to write nondimensional kinetic energies as $\frac{1}{2}m'v'^2$ without additional scaling factors. This means setting the energy scaling as:

$$E_c = m_c \frac{x_c^2}{t_c^2} \ . \tag{S13}$$

This does not uniquely determine the energetic scale, since t_c is still free. The two obvious choices are to scale to the temperature—$KE' = 1$ corresponds to $k_B T$ units of dimensional energy—or to the potential energy scale—$KE' = 1$ corresponds to U_0 units of dimensional energy. Choosing the latter yields:

$$E_c = U_0 = m_c \frac{x_c^2}{t_c^2} \text{ and} \tag{S14}$$

$$\frac{x_c^2}{L} = m_c \frac{x_c^2}{t_c^2} \ . \tag{S15}$$

Evidently, the timescale is $t_c = \sqrt{LC}$, which is a workable timescale for our purposes given that the dynamics of interest happen on the scale of $\tau \approx \omega_{LC}$. Setting the timescale to the potential energy rather than the thermal energy may well become common practice in simulating momentum computation, since protocols must be timed precisely with respect to the dynamics of the potential energy surface.

The Langevin equation, in terms of the nondimensional quantities defined above, becomes:

$$\begin{aligned} dv'\frac{x_c}{t_c} = &-\frac{\nu'\nu_c}{m'm_c}v'x_c dt' - \frac{1}{m'm_c}\left(\frac{U_0}{x_c}\partial_{x'}U'(x')\right)t_c dt' \\ &+ \frac{1}{m'm_c}r(t)\sqrt{2\nu'\nu_c E_c \kappa' t_c dt'}\ . \end{aligned} \tag{S16}$$

Simplifying algebra then yields:

$$\begin{aligned} dv' = &-\frac{\sqrt{LC}}{RC}\frac{\nu'}{m'}v'dt' - \frac{1}{m'}\partial_{x'}U'(x')dt' \\ &+ \left(\frac{L}{R^2C}\right)^{1/4}\frac{\sqrt{\nu'\kappa'}}{m'}r(t)\sqrt{2dt'}\ . \end{aligned} \tag{S17}$$

Finally, we define λ, θ, and η as nondimensional parameters that serve as our dimensionless Langevin coefficients. This yields the Langevin equation for the simulations detailed in Appendix 5.E:

$$dv' = -\lambda v'dt' - \theta\partial_{x'}U' + \eta r(t)\sqrt{2dt'}\ , \tag{S18}$$

with:

$$\lambda = \frac{\sqrt{LC}}{RC}\frac{\nu'}{m'}, \tag{S19}$$

$$\theta = \frac{1}{m'},\ \text{and} \tag{S20}$$

$$\eta = \sqrt{\frac{\lambda\kappa'}{m'}}\ , \tag{S21}$$

where:

(S22) $$x' = (\varphi, \varphi_{dc}),$$

(S23) $$v' = \frac{d}{dt'}x',$$

(S24) $$\nu' = (2, 1/2),$$

(S25) $$m' = (1, 1/4), \text{ and}$$

(S26) $$\kappa' = \frac{k_B T}{U_0} \ .$$

5.D. Effective Potential and Simulation Details for the Flux Qubit

We consider two cases: critical-current symmetric and asymmetric JJ pairs.

5.D.1. Symmetric Approximation. We can obtain reasonable estimates for good $\varphi_{\rm xdc}$ values by assuming a perfectly symmetric device $\delta\beta = 0$. Furthermore, we also set $\varphi_x = 0$ for all cases. This allows two symmetric wells on either side of $\varphi = 0$. In practice, since $\delta\beta \neq 0$ in a real device, φ_x would be calibrated to compensate for the asymmetry; see Sec. 5.D.2.

In the symmetric case, the potential splits into two components—periodic and quadratic:

(S27) $$\beta \cos\varphi \cos\frac{\varphi_{\rm dc}}{2} + \frac{1}{2}\varphi^2 + \frac{\gamma}{2}(\varphi_{\rm dc} - \varphi_{\rm xdc})^2 \ .$$

The periodic term allows for multiple minima, while the quadratic terms force the dynamical variables to stay close to their respective parameters. This localization means we focus only on the the area near $\varphi = \varphi_x$ and $\varphi_{\rm dc} = \varphi_{\rm xdc}$.

To employ the potential most flexibly, we must characterize the relevant fixed points that occur in this region. Following Refs. [**151, 163**], we choose to search in the domain $-\pi < \varphi < \pi$ and $-2\pi < \varphi_{\rm dc} < 0$. Fixed points occur when all components of the gradient vanish:

(S28) $$\partial_\varphi U' = -\beta \sin\varphi \cos\frac{\varphi_{\rm dc}}{2} + \varphi = 0$$

(S29) $$\partial_{\varphi_{\rm dc}} U' = -\frac{\beta}{2}\sin\frac{\varphi_{\rm dc}}{2}\cos\varphi + \gamma(\varphi_{\rm dc} - \varphi_{\rm xdc}) = 0$$

The first condition is met whenever $\varphi = 0$ and, also, when $\frac{\varphi}{\beta \sin \varphi} = \cos \frac{1}{2}\varphi_{\mathrm{dc}}$. Consider the case where $\varphi = 0$—the "central" fixed point. To find the φ_{dc} location of the fixed point φ^0_{dc}, we look to the gradient's second term. This yields the condition:

$$\begin{aligned} \varphi^0_{\mathrm{dc}} - \frac{\beta}{2\gamma} \sin \frac{\varphi^0_{\mathrm{dc}}}{2} &= \varphi_{\mathrm{xdc}} \\ F^0(\varphi_{\mathrm{dc}} = \varphi^0_{\mathrm{dc}}, \beta, \gamma) &= \varphi_{\mathrm{xdc}} \ . \end{aligned} \tag{S30}$$

The central fixed point occurs close to the parameter φ_{xdc}, but is offset by a value $\leq \beta/2\gamma$.

The equation above can be solved numerically with ease to find the location of the central fixed point. To classify the fixed point, we look at the Hessian. While the general expression for the eigenvalues is rather verbose, the case where $\varphi = 0$ simplifies to:

$$\lambda_1 = -\beta \cos \frac{\varphi^0_{\mathrm{dc}}}{2} + 1 \tag{S31}$$

$$\lambda_2 = \gamma - \frac{\beta}{4} \cos \frac{\varphi^0_{\mathrm{dc}}}{2} \ . \tag{S32}$$

$\lambda_2 > 0$ as long as $\gamma > \beta/4$. And, since we assume $\gamma > \beta$, this condition is always met. Thus, this fixed point is either a saddle point or a minimum based on whether φ^0_{dc} is greater or less than $\varphi^c_{\mathrm{dc}} \equiv -2\cos^{-1} \frac{1}{\beta}$, respectively. (We only use the negative branch of $\cos^{-1}$ due to the domain of φ_{dc}.) See Fig. S1 for an example of the behavior of the central fixed point for typical parameters.

We can also find an expression for $\varphi^c_{\mathrm{xdc}}(\beta, \gamma) \equiv F^0(\varphi_{\mathrm{dc}} = \varphi^c_{\mathrm{dc}})$, the critical value of the control parameter at which the central fixed point transitions between a saddle point and a minimum:

$$\begin{aligned} \varphi^c_{\mathrm{xdc}}(\beta, \gamma) &= \varphi^c_{\mathrm{dc}} - \frac{\beta}{2\gamma} \sin \frac{\varphi^c_{\mathrm{dc}}}{2} \\ &= -2\cos^{-1} \frac{1}{\beta} + \frac{\beta}{2\gamma} \sqrt{1 - \frac{1}{\beta^2}} \ . \end{aligned} \tag{S33}$$

Naively, the best strategy to form a low cost protocol is to take values of φ_{xdc} just above and below φ^c_{xdc}. However, there are several factors that introduce complications. For one, the energy scale separating the two wells when $\varphi_{\mathrm{xdc}} \approx \varphi^c_{\mathrm{xdc}}$ is very small and it will typically be overwhelmed by thermal energy at the temperatures of interest ($400 - 1400$ mK). A second is that the approximation of $\delta\beta = 0$ actually has a most pernicious effect near φ^x_{xdc}. (This is discussed in Sec. 5.D.2.)

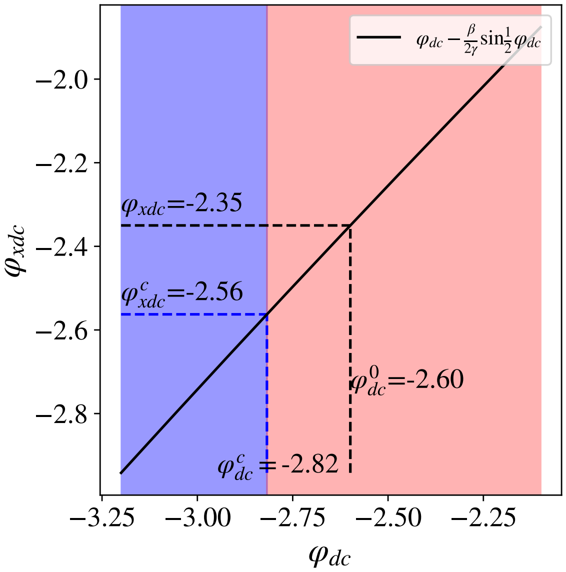


FIGURE S1. Fixed point at $\varphi = 0$ in an ideal device with $\beta = 6.2$ and $\gamma = 12.0$: Red (Blue) background indicates regions where the fixed point is a saddle point (local minimum). For example, if $\varphi_{\text{xdc}} = -2.35$, the central fixed point is a saddle point at $\varphi_{\text{dc}} = -2.6$. To find a stable fixed point at $\varphi = 0$, a control parameter less than φ^c_{xdc} is necessary, which falls at -2.56 in the example above.

Finally, we have yet to consider the other fixed points at $\varphi \neq 0$. Doing so reveals that sometimes φ^c_{xdc} corresponds to a subcritical pitchfork bifurcation—yielding a potential with a third (undesirable) minimum rather than a single one.

When $\varphi \neq 0$ we can rewrite Eqs. (S28) and (S29):

$$\frac{\varphi}{\beta \sin \varphi} = \cos \frac{1}{2}\varphi_{\text{dc}} \tag{S34}$$

$$\frac{\beta}{4\gamma} \sin \frac{\varphi_{\text{dc}}}{2} \cos \varphi - \frac{1}{2}\varphi_{\text{xdc}} = \frac{1}{2}\varphi_{\text{dc}} \ . \tag{S35}$$

The potential is symmetric, so these fixed points come in pairs $\varphi^{\pm}$. Substituting $\varphi_{\text{dc}}/2 = -\cos^{-1}(\varphi^{\pm}/\beta \sin \varphi^{\pm})$ into the second equation yields the following for $\varphi^{\pm}$:

$$\varphi_{\text{xdc}} = \frac{\beta}{2\gamma}\sqrt{1 - \left(\frac{\varphi^{\pm}}{\beta \sin \varphi^{\pm}}\right)^2} \cos \varphi^{\pm} - 2\cos^{-1}\frac{\varphi^{\pm}}{\beta \sin \varphi^{\pm}} \tag{S36}$$

$$\varphi_{\text{xdc}} = F^{\pm}(\varphi = \varphi^{\pm}, \beta, \gamma) \ . \tag{S37}$$

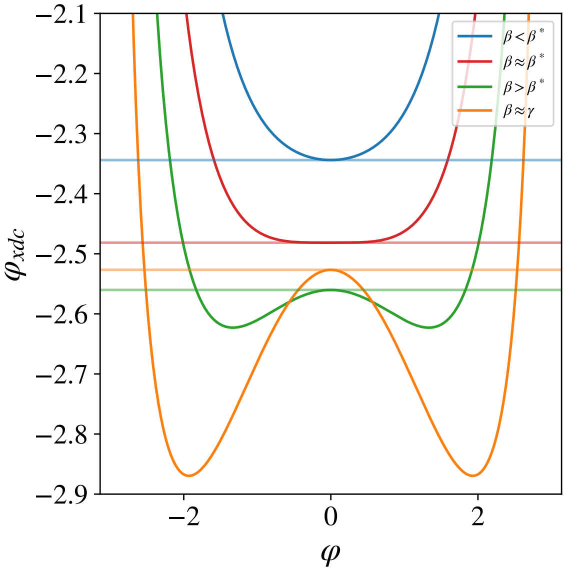


FIGURE S2. $\varphi \neq 0$ fixed points appear when the value of the function plotted equals the external φ_{xdc} parameter. Note that for some β and γ combinations, there is a qualitatively different behavior. Especially for larger β, there is a coexistence region of three potential minima. For the $\beta \approx \gamma$ example, one would want to set $\Delta C > 0.5$ to make sure V^{comp} falls well outside of the three minima range. Horizontal lines show the values of φ^c_{xdc}. (See Appendix 5.E.)

Note that the sign changes due to the domain restriction of φ_{dc}. Figure S2 shows how these fixed points behave as β, γ, and φ_{xdc} change. The value of φ_{xdc} tangent to the curve when $\varphi = 0$ corresponds to the critical control parameter value φ^c_{xdc}, which can be seen by verifying $\lim_{\varphi \to 0} F^{\pm}(\varphi) = \varphi^c_{\text{xdc}}$.

As a last note, different values of β and γ have qualitatively different fixed point profiles depending on whether the central fixed point undergoes a supercritical or subcritical pitchfork bifurcation when $\varphi_{\text{xdc}} = \varphi^c_{\text{xdc}}$. The critical value β^* where the bifurcation of the central fixed point transitions between being supercritical and subcritical is given by:

$$\lim_{\varphi \to 0} \partial^2_{\varphi} F^{\pm}(\varphi, \beta^*, \gamma) = 0 \ . \tag{S38}$$

Once again, the full derivative is quite verbose. However, taking the limit $\varphi \to 0$ gives:

$$\text{(S39)} \qquad \frac{\sqrt{\beta^{*^2}-1}}{6\beta^{*^2}}\left(-3\beta^{*^2}+4\gamma+2\right)=0$$

$$\text{(S40)} \qquad \beta^* = \sqrt{\frac{4\gamma+2}{3}}\ .$$

Interestingly, when $\beta > \beta^*$, there is always a parameter space region with three distinct minima. This might be useful, in fact, for single-bit computations that require more states. For bit swap, though, the goal is for the system to jump between a V^{store} with 2 minima and a V^{comp} with a single minimum (see Figure S4). And so, care must be taken to avoid the three-minima regions when $\beta > \beta^*$.

5.D.2. $\delta\beta \neq 0$**.** The device just considered is ideal. In reality $\delta\beta \neq 0$, and exact analytic work is much less fruitful. Introducing the asymmetric terms augments the potential:

$$\text{(S41)} \qquad U_{\text{asym}}(\varphi, \varphi_x, \delta\beta, \varphi_{\text{dc}}) = \frac{1}{2}\varphi_x^2 - \varphi\varphi_x - \delta\beta \sin\varphi \cos\frac{\varphi_{\text{dc}}}{2}\ .$$

In short, one must vary φ_x to offset the effect of $\delta\beta$, provided a symmetric potential is preferred.

There are two obvious strategies to minimize the effects of asymmetry. Either a strategy that minimizes the effect of U_{asym} at the central fixed point—the "min of mid" strategy—or at the fixed points at $\varphi^{\pm}$—the "min of max" strategy. It stands to reason that one uses the former to set φ_x for V^{comp} and the latter for V^{store}.

The "min of mid" strategy is easy to implement. Simply set the derivative of $\partial_\varphi U_{asym}|_{\varphi=0} = 0$, with the intent of having the asymmetrical part of the potential be as flat as possible near $\varphi = 0$. Simple algebra yields: $\varphi_x = -\delta\beta \sin\varphi_{\text{dc}}/2$.

The "min of max" strategy requires numerical solution. First, note that the maximum value of U_{asym} occurs when $\varphi = \varphi_{max} = \arccos\left(\frac{\varphi_x}{\delta\beta \sin .5\varphi_{\text{dc}}}\right)$. Then, use a symbolic solver (e.g., SymPy's *nsolve* function) to find the value of φ_x that minimizes $U_{asym}(\varphi_{max}, \varphi_x, \delta\beta, \varphi_{\text{dc}})$.

Figure S3 shows that the effect of $\delta\beta \neq 0$ is, unsurprisingly, the most noticeable near the bifurcation of the central fixed point. For the bit swap, as described in Sec. 5.3, we need only two different profiles for the potential: one in which we have two symmetric wells and one in which we have a single well placed midway between them. Thus, we must keep the φ_{xdc} parameter

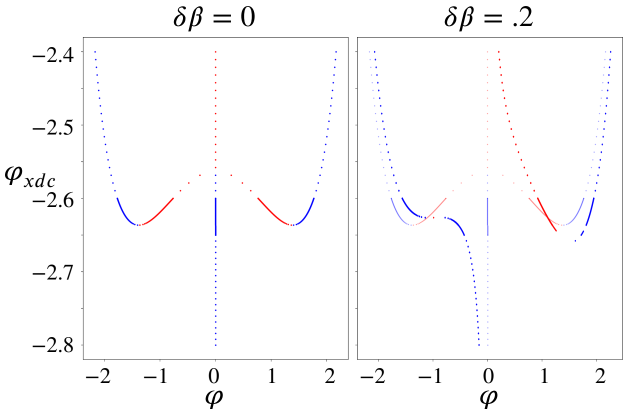


FIGURE S3. Fixed point bifurcation diagram for the (left) idealized $\delta\beta = 0$ device and (right) a device with $\delta\beta = 0.2$. Blue indicates stable minima and red saddle points. On the right plot, the $\delta\beta = 0$ fixed points are plotted as well, with low opacity to help see the difference. The naive "minimum of maximum" strategy has been used to minimize the effect of U_{asym}. And, we can see that the symmetric approximation works fairly well as long as $|\varphi_{\mathrm{xdc}} - \varphi^c_{\mathrm{xdc}}| > .2$. It is likely that more evolved solution strategies will improve results.

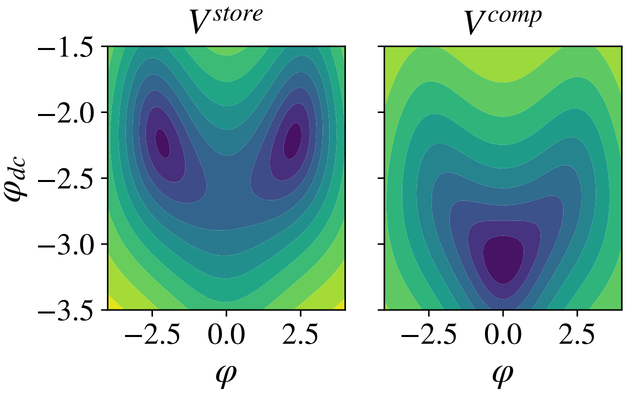


FIGURE S4. (Left) V^{store}, the bistable storage potential. (Right) V^{comp}, the "banana-harmonic" potential. These potential energy profiles serve as qualitative pictures to represent prototypical computational and storage potentials, and do not represent any particularly favorable parameter set.

sufficiently far away from φ^c_{xdc}. The strategy employed in the simulations described below always involves setting a minimum distance that φ_{xdc} must be from φ^c_{xdc}, in order to avoid falling into the pitfalls described here.

5.E. Searching for Minimal-Work Bit Swaps

The following lays out the computational strategy to find low work-cost implementations.

We are most interested in the effect of parameters that are the most removed from fabrication, so all simulations assume JJ elements with I_+, R, and C set to $2.0\,\mu\text{A}$, $371\,\Omega$, and $4.0\,\text{nF}$, respectively. To explore how asymmetry affects work cost, we simulated protocols with a nearly-symmetric device with $I_- = 7\,\text{nA}$, a moderately-symmetric device with $I_- = 35\,\text{nA}$, and an asymmetric device with $I_- = 60\,\text{nA}$. Additionally, $k_\text{B}T$ is always scaled to U_0, so that $\kappa' \equiv k_\text{B}T/U_0 = 0.05$.

Given devices with the parameters above, what values of the remaining parameters yield protocols with minimum work cost? This involves a twofold procedure. First, create the circuit architecture by setting L and γ by hand; thus, fully specifying the device. Second, determine the ideal protocols for that combination of device parameters through simulation.

L's order of magnitude was chosen from previous results [**151**, **152**, **161**, **162**, **163**, **164**] to be $10^{-9}H$. Noting that a lower L results in a more harmonic potential during computation, we set a minimum L to be $0.3nH$. This is in order to stay within the parameter range for which $\beta > 1$ and we can still use the analytic expressions derived above. To assure $\gamma > \beta$, γ values were tested in the range $[3.0, 20.0]$.

After choosing a pair of circuit parameters L and γ, we turn to simulation. First, V^store must be chosen by setting φ_x^store and $\varphi_\text{xdc}^\text{store}$. This is done by calculating $\varphi_\text{xdc}^\text{store} \equiv \varphi_\text{xdc}^c + \Delta S$, where $\varphi_\text{xdc}^c(\gamma, \beta)$ is from Eq. (S33). The parameter ΔS is initialized manually to a value ΔS^* when starting a new round of simulations. ($\Delta S^* = 0.16$ was used in the heatmaps shown in Fig. 5.11.2.) Then, using the "min of max" method (Sec. 5.D.2), we set φ_x^store.

Finally, V^store is tested by sampling 50,000 states from V^store's equilibrium distribution using a Monte Carlo algorithm. The resulting ensemble is verified by determining that it contains two well-separated informational states by asserting that:

$$\langle \varphi < 0 \rangle + 3\sigma_{\varphi<0} < \langle \varphi > 0 \rangle - 3\sigma_{\varphi>0}\ , \tag{S42}$$

where $\langle s \rangle$ and σ_s are means and standard deviations of φ conditioned on s being true. If the ensemble fails the test, ΔS is incremented and the process is repeated. If the ensemble succeeds, we have found a viable V^store.

Then, we move on to establish V^comp by choosing φ_x^comp and $\varphi_\text{xdc}^\text{comp}$. Similar to $\varphi_\text{xdc}^\text{store}$, $\varphi_\text{xdc}^\text{comp} \equiv \varphi_\text{xdc}^c - \Delta C$ with ΔC manually set. The value of ΔC does effect the eventual work cost, but the

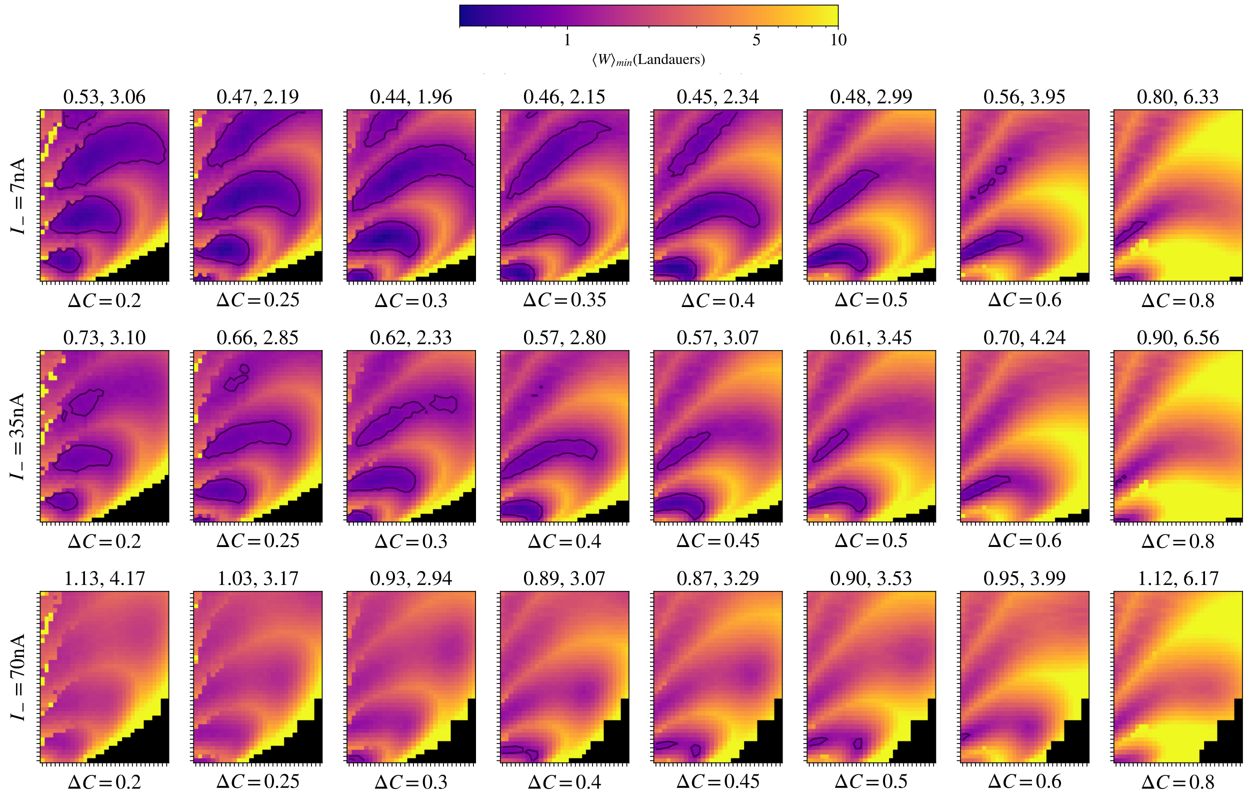


FIGURE S1. Thermodynamic performance under changing ΔC for devices with three different symmetry parameters: In each case, the x axis variable is $L \in (0.3, 1)$nH and the y axis $\gamma \in (3, 20)$. The numerical figures at the top of each panel are the minimum and average values of $\langle W \rangle_{\text{min}}$. The outlined (black line) regions represent pieces of parameter space where the minimal work protocols cost less than one Landauer. The simulations represented by each point in the heatmaps used $10,000$ samples from the equilibrium distribution. And, $1,200$ parameter sets were tested in each map.

work costs vary smoothly, and a single value of ΔC tends to work well over a large parameter range. Manually setting a single value for ΔC, rather than allowing it to adjust itself to fall into a local minimum, substantially reduces simulation run time. However, we expect that given more compute resources a wider range of sub-Landauer protocols will be discovered. Figure S1 shows the effect of changing ΔC for three different devices. Once ΔC is chosen, we use the "min of mid" (Sec. 5.D.2) method to set φ_x^{comp} and fully determine V^{comp}.

Next, a preliminary simulation is run to identify an approximate value of the computation time τ. To make the simulation run quickly, the ensemble above is coarse-grained into two partitions based on whether $\varphi > 0$ or $\varphi < 0$. Then, each partition is coarse-grained again into ≈ 250

representative points through histogramming. A Langevin simulation is run over the histogram data, exposing it to V^{comp} for a time $\mathcal{O}(10)\sqrt{LC}$. This ensures capturing the time with the best bit swap. Next, weighting the simulation results by histogram counts within each partition, we obtain conditional averages for an approximation of the behavior over the entire ensemble. These averages are parsed for a set of times at which there are indications of a successful and low-cost bit swap: $\langle \varphi(t=0) < 0 \rangle > 0$, $\langle \varphi(t=0) > 0 \rangle < 0$, and values of $\langle \dot{\varphi} \rangle$ and $\langle \dot{\varphi_{\mathrm{dc}}} \rangle$ that are close to zero. See, for example, the blue highlighted portion on the top panel of Fig. 5.10.4. In this way, a range $(\tau_{\mathrm{min}}, \tau_{max})$ is determined for τ.

Now, a larger simulation is completed to determine τ that give the lowest work value. Another 40,000 samples are generated from V^{store}'s equilibrium distribution, and a Langevin simulation is run on the full ensemble by exposing it to V^{comp} for τ_{max} time units. Since the potential is held constant between $t = 0$ and $t = \tau$, work is only done when turning V^{comp} on at $t = 0$ and turning it off at $t = \tau$. The average work done at $t = 0$ is $W_0 \equiv \langle V^{\mathrm{comp}}(\varphi(0), \varphi_{\mathrm{dc}}(0)) - V^{\mathrm{store}}(\varphi(0), \varphi_{\mathrm{dc}}(0)) \rangle$ and returning to V^{comp} at time t costs $W_t \equiv \left\langle V^{\mathrm{store}}(\varphi(t), \varphi_{\mathrm{dc}}(t)) - V^{\mathrm{comp}}(\varphi(t), \varphi_{\mathrm{dc}}(t)) \right\rangle$. Thus, the mean net work cost at time t is the sum $W(t) = W_0 + W_t$.

Additionally, for each $t \in (\tau_{\mathrm{min}}, \tau_{max})$ we calculate the fidelity $f(t)$ and whether the final states are well-separated informational states, $s(t)$:

$$\begin{aligned} f(t) &= 1 - \frac{1}{N} \sum_{i=1}^{N} \mathrm{bool}\left[\mathrm{sign}\varphi_i(t=0) = \mathrm{sign}\varphi_i(t=t)\right] \\ s(t) &= \mathrm{bool}\left[\langle \varphi < 0 \rangle + 3\sigma_{\varphi<0} < \langle \varphi > 0 \rangle - 3\sigma_{\varphi>0}\right] \ . \end{aligned} \tag{S43}$$

Finally, we choose the minimum work protocol via $\inf\left(W(t) : f(t) \geq 0.99, s(t) = \mathrm{True}\right)$.

After this, we move on to the next pair of L and γ. Typically, these are chosen to be individually close to the last pair. And, and instead of re-initializing ΔS to its initial value by hand, we decrement ΔS from its current value by a small amount if $\Delta S > \Delta S^{*}$, using this value as the starting point for the next L and γ pair. This allows the value of ΔS to drift from its starting point towards more favorable values as the parameters change, while still preferring to be close to the known well-behaved parameter value ΔS^{*}. Setting a new initial value for ΔS goes full circle, to find the next minimum work protocol by repeating the procedure.

This procedure yielded rather large ranges of parameter space over which we found very low work-cost bit swap protocols. Here, we offer no proof that the protocols found achieve the global minimum work, since the protocol space is high dimensional and contains many local minima. That said, improved algorithms and a larger parameter-range search should result in even lower work costs.

Langevin simulations of the dimensionless equations of motion employed a fourth-order Runge-Kutta method for the deterministic portion and Euler's method for the stochastic portion of the integration with dt set to $0.005\sqrt{LC}$. (Python NumPy's Gaussian number generator was used to generate the memoryless Gaussian variable r(t).)

Bibliography

[1] J. C. Maxwell. *Theory of Heat*. Longmans, Green and Co., London, United Kingdom, 1871.

[2] H. Leff and A. Rex. *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*. Taylor and Francis, New York, 2002.

[3] W. Thomson. The sorting demon of Maxwell. In *Roy. Soc. Proc.*, volume 9, pages 113–114, 1879.

[4] L. Szilard. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.*, 53:840–856, 1929.

[5] W. Lanouette and B. Szilard. *Genius in the Shadows: A Biography of Leo Szilard, The Man Behind The Bomb*. Skyhorse Publishing, New York, New York, 2013.

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

[7] R. Feynman. Simulating physics with computers. *Intl. J. Theo. Phys.*, 21(6/7):467–488, 1982.

[8] Werner Ehrenberg. Maxwell's demon. *Scientific American*, 217(5):103–111, 1967.

[9] Karl Popper. The philosophy of karl popper. 1974.

[10] Oliver Penrose. Foundations of statistical mechanics. *Reports on Progress in Physics*, 42(12):1937, 1979.

[11] 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.

[12] R. K. Pathria and P. D. Beale. *Statistical Mechanics*. Butterwork-Heinemann, Oxford, United Kingdom, second edition, 1996.

[13] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[15] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports Prog. Phys.*, 75(12):126001, 2012.

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

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

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

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

[20] K. Shizume. Heat generation required by information erasure. *Phys. Rev. E*, 52(4):3495–3499, 1995.

[21] F. N. Fahn. Maxwell's demon and the entropy cost of information. *Found. Physics*, 26:71–93, 1996.

[22] M. M. Barkeshli. Dissipationless information erasure and Landauer's principle. *arXiv:0504323*.

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

[24] A. Lasota and M. C. Mackey. *Probabilistic Properties of Deterministic Systems.* Cambridge University press, Cambridge, United Kingdom, 1985.

[25] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012.

[26] S. H. Strogatz. *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry, and engineering.* Addison-Wesley, Reading, Massachusetts, 1994.

[27] N. Barnett and J. P. Crutchfield. Computational mechanics of input-output processes: Structured transformations and the ϵ-transducer. *J. Stat. Phys.*, 161(2):404–451, 2015.

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

[29] W. Ross Ashby. *An Introduction to Cybernetics.* John Wiley and Sons, New York, second edition, 1960.

[30] 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.

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

[32] J. Bengtsson, M. N. Tengstrand, A. Wacker, P. Samuelsson, M. Ueda, H. Linke, and S. M. Reimann. Quantum Szilard engine with attractively interacting bosons. *Phys. Rev. Let.*, 120(10):100601, 2018.

[33] M. H. Mohammady and J. Anders. A quantum Szilard engine without heat from a thermal reservoir. *New J. Physics*, 19(11):113026, 2017.

[34] S. Vaikuntanathan and C. Christopher. Modeling Maxwell's demon with a microcanonical Szilard engine. *Phys. Rev. E*, 83(6):061120, 2011.

[35] L. B. Kish and C. G. Granqvist. Energy requirement of control: Comments on Szilard's engine and Maxwell's demon. *Europhys. Lett.*, 98(6):68001, 2012.

[36] W. H. Zurek. Eliminating ensembles from equilibrium statistical physics: Maxwell's demon, Szilard's engine, and thermodynamics via entanglement. *Phys. Rep.*, 755:1–21, 2018.

[37] G. M. Zaslavsky. From Hamiltonian chaos to Maxwell's demon. *Chaos*, 5:653–661, 1995.

[38] T. Admon, S. Rahav, and Y. Roichman. Experimental realization of an information machine with tunable temporal correlations. *Phys. Rev. Let.*, 121(18):180601, 2018.

[39] J. V. Koski, V. F. Maisi, J. P. Pekola, and D. V. Averin. Experimental realization of a Szilard engine with a single electron. *Proc. Natl. Acad. Sci. USA*, 111(38):13786–13789, 2014.

[40] K. V. Koski, V. F. Maisi, T. Sagawa, and J. P. Pekola. Experimental observation of the role of mutual information in the nonequilibrium dynamics of a Maxwell demon. *Phys. Rev. Let.*, 113(3):030601, 2014.

[41] J. V. Koski and J.P. Pekola. Maxwell's demons realized in electronic circuits. *Compt. Rend. Phys.*, 17(10):1130–1138, 2016.

[42] L. B. Kish and C.-G. Granqvist. Electrical Maxwell demon and Szilard engine utilizing Johnson noise, measurement, logic and control. *PloS One*, 7(10), 2012.

[43] K. Choi, A. Droudian, R. W. Wyss, K.-P. Schlichting, and H. G. Park. Multifunctional wafer-scale graphene membranes for fast ultrafiltration and high permeation gas separation. *Sci. Adv.*, 4(11):eaau0476, 2018.

[44] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.*, 136(6):1005–1034, 2009.

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

[46] A. B. Boyd, A. Patra, C. Jarzynski, and J. P. Crutchfield. Shortcuts to thermodynamic computing: The cost of fast and faithful information processing. *J. Stat. Physics*, in press, 2021.

[47] S. Lahiri, J. Sohl-Dickstein, and S. Ganguli. A universal tradeoff between power, precision and speed in physical communication. *arXiv preprint arXiv:1603.07758*, 2016.

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

[49] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz. Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, 483(7388):187–189, 2012.

[50] P. M. Riechers, A. B. Boyd, G. W. Wimsatt, and J. P. Crutchfield. Balancing error and dissipation in computing. *Phys. Rev. Res.*, 2(3):033524, 2020.

[51] L. Gammaitoni. Beating the Landauer's limit by trading energy with uncertainty. *arXiv preprint arXiv:1111.2937*, 2011.

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

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

[54] D. Reeb and M. M. Wolf. An improved Landauer principle with finite-size corrections. *New J. Physics*, 16(10):103011, 2014.

[55] Gregory Wimsatt, Alexander B. Boyd, and James P. Crutchfield. Trajectory class fluctuation theorem, 2022.

[56] Andre C Barato and Udo Seifert. Thermodynamic uncertainty relation for biomolecular processes. *Physical Review Letters*, 114(15):158101, 2015.

[57] Albert Einstein et al. On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der physik*, 17(549-560):208, 1905.

[58] GN Bochkov and Yu E Kuzovlev. General theory of thermal fluctuations in nonlinear systems. *Zh. Eksp. Teor. Fiz*, 72:238–243, 1977.

[59] Denis J Evans and Debra J Searles. Equilibrium microstates which generate second law violating steady states. *Physical Review E*, 50(2):1645, 1994.

[60] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.

[61] Gavin E Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721, 1999.

[62] Chris Jarzynski. Nonequilibrium work theorem for a system strongly coupled to a thermal environment. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(09):P09005, 2004.

[63] Jan Liphardt, Sophie Dumont, Steven B Smith, Ignacio Tinoco Jr, and Carlos Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of jarzynski's equality. *Science*, 296(5574):1832–1835, 2002.

[64] Christopher Jarzynski. Rare events and the convergence of exponentially averaged work values. *Physical Review E*, 73(4):046105, 2006.

[65] Gavin E Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *Journal of Statistical Physics*, 90(5):1481–1487, 1998.

[66] Yoshihiko Hasegawa and Tan Van Vu. Generalized Thermodynamic Uncertainty Relation via Fluctuation Theorem. (0), 2019.

[67] Denis J Evans and Debra J Searles. The Fluctuation Theorem. *Advances in Physics*, 51(7):1529–1585, nov 2002.

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

[69] Christopher Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. *Journal of Statistical Physics*, 98:77–102, 2000.

[70] Paul M Riechers, Alexander B Boyd, Gregory W Wimsatt, and James P Crutchfield. Balancing error and dissipation in highlyreliable computing. *arXiv preprint arXiv:1909.06650*, 2019.

[71] Ryogo Kubo. Brownian motion and nonequilibrium statistical mechanics. *Science*, 233(4761):330–334, 1986.

[72] L Onsager. Reciprocal relations in irreversible processes. II. *Physical Review*, 38:2265, 1931.

[73] Rep Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

[74] G. Gallavotti and E. G. D. Cohen. Dynamical ensembles in stationary states. *Journal of Statistical Physics*, 80(5-6):931–970, sep 1995.

[75] Christopher Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018–5035, nov 1997.

[76] Hal Tasaki. Jarzynski Relations for Quantum Systems and Some Applications. 2000.

[77] Jorge Kurchan. A Quantum Fluctuation Theorem. 2000.

[78] Christopher Jarzynski and D. K. Wójcik. Classical and Quantum Fluctuation Theorems for Heat Exchange. *Physical Review Letters*, 92(23):230602, jun 2004.

[79] D. Andrieux, P. Gaspard, T. Monnai, and S. Tasaki. The fluctuation theorem for currents in open quantum systems. *New Journal of Physics*, 11:043014, 2009.

[80] Keiji Saito and Yasuhiro Utsumi. Symmetry in full counting statistics, fluctuation theorem, and relations among nonlinear transport coefficients in the presence of a magnetic field. *Physical Review B - Condensed Matter and Materials Physics*, 78(11):115429, 2008.

[81] Massimiliano Esposito, U. Harbola, and S. Mukamel. Nonequilibrium fluctuations, fluctuation theorems, and counting statistics in quantum systems. *Reviews of Modern Physics*, 81(4):1665–1702, dec 2009.

[82] Michele Campisi, Peter Hänggi, and Peter Talkner. Colloquium: Quantum fluctuation relations: Foundations and applications. *Reviews of Modern Physics*, 83(3):771–791, jul 2011.

[83] Christopher Jarzynski. Equalities and Inequalities: Irreversibility and the Second Law of Thermodynamics at the Nanoscale. *Annual Review of Condensed Matter Physics*, 2(1):329–351, mar 2011.

[84] Peter Hänggi and Peter Talkner. The other QFT. *Nature Physics*, 11(2):108–110, 2015.

[85] Todd R. Gingrich, Jordan M. Horowitz, Nikolay Perunov, and Jeremy L. England. Dissipation bounds all steady-state current fluctuations. *Physical Review letters*, 116(12):120601, 2016.

[86] Patrick Pietzonka, Felix Ritort, and Udo Seifert. Finite-time generalization of the thermodynamic uncertainty relation. *Physical Review E*, 96(1):012101, 2017.

[87] Andreas Dechant. Multidimensional thermodynamic uncertainty relations. *Journal of Physics A: Mathematical and General*, 52:035001, 2018.

[88] Jordan M. Horowitz and Todd R. Gingrich. Proof of the finite-time thermodynamic uncertainty relation for steady-state currents. *Physical Review E*, 96(2):020103, 2017.

[89] Andre C. Barato, Raphael Chetrite, Alessandra Faggionato, and Davide Gabrielli. Bounds on current fluctuations in periodically driven systems. *New Journal of Physics*, 20(10):103023, 2018.

[90] Viktor Holubec and Artem Ryabov. Cycling tames power fluctuations near optimum efficiency. *Physical Review Letters*, 121(12):120601, 2018.

[91] Karel Proesmans and Christian Van den Broeck. Discrete-time thermodynamic uncertainty relation. *Europhysics Letters*, 119(2):20001, 2017.

[92] Tan Van Vu and Yoshihiko Hasegawa. Thermodynamic uncertainty relations under arbitrary control protocols. *Physical Review Research*, 2(1):013060, 2020.

[93] Harry JD Miller, M Hamed Mohammady, Martí Perarnau-Llobet, and Giacomo Guarnieri. Thermodynamic uncertainty relation in slowly driven quantum heat engines. *Physical Review Letters*, 126(21):210603, 2021.

[94] Harry J.D. Miller, M. Hamed Mohammady, Martí Perarnau-Llobet, and Giacomo Guarnieri. Joint statistics of work and entropy production along quantum trajectories. *Physical Review E*, 103(5):052138, 2021.

[95] Gianmaria Falasco and Massimiliano Esposito. Dissipation-time uncertainty relation. *Physical Review Letters*, 125(12):120604, 2020.

[96] Yoshihiko Hasegawa. Quantum thermodynamic uncertainty relation for continuous measurement. *Physical Review Letters*, 125(5):050601, 2020.

[97] Yoshihiko Hasegawa. Thermodynamic uncertainty relation for general open quantum systems. *Physical Review Letters*, 126(1):010602, 2021.

[98] Tan Van Vu and Keiji Saito. Thermodynamics of precision in markovian open quantum dynamics. *Physical Review Letters*, 128(14):140602, 2022.

[99] Andreas Dechant and Shin-ichi Sasa. Fluctuation–response inequality out of equilibrium. *Proceedings of the National Academy of Sciences*, 117(12):6430–6436, 2020.

[100] Katarzyna MacIeszczak, Kay Brandner, and Juan P. Garrahan. Unified Thermodynamic Uncertainty Relations in Linear Response. *Physical Review Letters*, 121:130601, 2018.

[101] Giacomo Guarnieri, Gabriel T. Landi, Stephen R Clark, and John Goold. Thermodynamics of precision in quantum non equilibrium steady states. 2019.

[102] Gianmaria Falasco, Massimiliano Esposito, and Jean-Charles Delvenne. Unifying thermodynamic uncertainty relations. *New Journal of Physics*, 22(5):053046, 2020.

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

[104] Chirstopher Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. *Journal of Statistical Physics*, 98(1):77–102, 2000.

[105] B. Barany. On iterated function systems with place-dependent probabilities. *Proc. Am. Math. Soc.*, 143(1), 2015.

[106] Jordan M. Horowitz and Todd R. Gingrich. Proof of the finite-time thermodynamic uncertainty relation for steady-state currents. *Physical Review E*, 96(020103), 2017.

[107] Yoshihiko Hasegawa and Tan Van Vu. Fluctuation theorem uncertainty relation. *Physical Review Letters*, 123(110602), 2019.

[108] Andre M. Timpanaro, Giacomo Guarnieri, John Goold, and Gabriel T. Landi. Thermodynamic uncertainty relations from exchange fluctuation theorems. *Physical Review Letters*, 123(090604), 2019.

[109] Daniel Maria Busiello and Simone Pigolotti. Hyperaccurate currents in stochastic thermodynamics. *Physical Review E*, 100(6):060102, 2019.

[110] M Timpanaro, Giacomo Guarnieri, Gabriel T Landi, et al. Hyperaccurate thermoelectric currents. *arXiv preprint arXiv:2108.05325*, 2021.

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

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

[113] Y. Hasegawa and T. Van Vu. Fluctuation theorem uncertainty relation. *Phys. Rev. Lett.*, 123(11):110602, 2019.

[114] U. Seifert. From stochastic thermodynamics to thermodynamic inference. *Ann. Rev. Cond. Mat. Physics*, 10:171–192, 2019.

[115] R. Landauer. Information is physical. *Physics Today*, pages 23–29, May 1991.

[116] E. Stopnitzky, S. Still, T. E. Ouldridge, and L. Altenberg. Physical limitations of work extraction from temporal correlations. *Phys. Rev. E*, 99:042115, 2019.

[117] J. A. Owen, A. Kolchinsky, and D. H. Wolpert. Number of hidden states needed to physically implement a given conditional distribution. *New J. Physics*, 21(1):013022, 2019.

[118] C. van den Broeck. Stochastic thermodynamics: A brief introduction. *Phys. Complex Colloids*, 2013.

[119] C. van den Broeck and M. Esposito. Ensemble and trajectory thermodynamics: A brief introduction. *Phyisca A*, 418:6–16, 2015.

[120] T. E. Ouldridge, C. C. Govern, and P. Rein ten Wolde. Thermodynamics of computational copying in biochemical systems. *Phys. Rev. X*, 7(2):021004, 2017.

[121] R. Alicki. The quantum open system as a model of the heat engine. *J. Phys. A*, 12(5), 1979.

[122] C. Jarzynski. Stochastic and macroscopic thermodynamics of strongly coupled systems. *Phys. Rev. X*, 7(011008), 2017.

[123] S. Deffner and E. Lutz. Nonequilibrium entropy production for open quantum systems. *Phys. Rev. Lett.*, 107(140404), 2011.

[124] P. Strasberg, G. Schaller, N. Lambert, and T. Brandes. Nonequilibrium thermodynamics in the strong coupling and non-Markovian regime based on a reaction coordinate mapping. *New J. Physics*, 18(073007), 2016.

[125] M. Esposito. Stochastic thermodynamics under coarse graining. *Phys. Rev. E*, 85(4):041125, 2012.

[126] P. M. Ara, R. G. James, and J. P. Crutchfield. The elusive present: Hidden past and future dependence and why we build models. *Phys. Rev. E*, 93(2):022143, 2016. SFI Working Paper 15-07-024; arxiv.org:1507.00672 [cond-mat.stat-mech].

[127] T. Koyuk and U. Seifert. Operationally accessible bounds on fluctuations and entropy production in periodically driven systems. *Phys. Rev. Lett.*, 122(23):230601, 2019.

[128] C. Maes. Frenetic bounds on the entropy production. *Phys. Rev. Lett.*, 119(16):160601, 2017.

[129] P. Strasberg and M. Esposito. Non-Markovianity and negative entropy production. *Phys. Rev. E*, 99(1)(012120), 2019.

[130] J. Bechhoefer. Hidden Markov models for stochastic thermodynamics. *New J. Physics*, 17(7):075003, 2015.

[131] E. Fredkin, R. Landauer, and T. Toffoli. Physics of computation. *Intl. J. Theo. Phys.*, 21(12):903, 1982.

[132] P. Strasberg, G. Schaller, N. Lambert, and T. Brandes. Nonequilibrium thermodynamics in the strong coupling and non-Markovian regime based on a reaction coordinate mapping. *New J. Physics*, 18(7):073007, 2016.

[133] P. M. Ara, R. G. James, and J. P. Crutchfield. Elusive present: Hidden past and future dependency and why we build models. *Phys. Rev. E*, 93(2):022143, 2016.

[134] M. Esposito. Stochastic thermodynamics under coarse graining. *Phys. Rev. E*, 85(4):041125, 2012.

[135] J. A. Owen, A. Kolchinsky, and D. H. Wolpert. Number of hidden states needed to physically implement a given conditional distribution. *New J. Physics*, 21(1):013022, 2019.

[136] S. Deffner and C. Jarzynski. Information processing and the second law of thermodynamics: An inclusive, hamiltonian approach. *Phys. Rev. X*, 3(4):041003, 2013.

[137] O.-P. Saira, M. H. Matheny, R. Katti, W. Fon, G. Wimsatt, J. P. Crutchfield, S. Han, and M. L. Roukes. Nonequilibrium thermodynamics of erasure with superconducting flux logic. *Phys. Rev. Res.*, 2(1):013249, 2020.

[138] G. E. Moore. The future of integrated electronics. *Fairchild Semiconductor internal publication*, 2, 1964.

[139] G. E. Moore. Cramming more components onto integrated circuits. *Proc. IEEE*, 86(1):82–85, 1998.

[140] G. E. Moore. Lithography and the future of moore's law. *IEEE Solid-State Circuits Society Newsletter*, 11(3):37–42, 2006.

[141] J. D. Hutcheson and G. D. Hutcheson. Is semiconductor manufacturing equipment still affordable? In *IEEE 1993 Interl. Symp. Semiconductor Manufacturing*, pages 54–62. VLSI Research Inc., 1993.

[142] G. D. Hutcheson and J. D. Hutcheson. Technology and economics in the semiconductor industry. *Scientific American*, 274(1):54–62, 1996.

[143] P. P. Gelsinger, P. A. Gargini, G. H. Parker, and A. Y. C. Yu. Microprocessors circa 2000. *IEEE Spectrum*, 26(10):43–47, 1989.

[144] M. M. Waldrop. More than Moore. *Nature*, 530(7589):144–148, 2016.

[145] P. Ball. Semiconductor technology looks up. *Nature Materials*, 21(2):132–132, 2022.

[146] M. Vinet, P. Batude, C. Tabone, B. Previtali, C. LeRoyer, A. Pouydebasque, L. Clavelier, A. Valentian, O. Thomas, S. Michaud, et al. 3D monolithic integration: Technological challenges and electrical results. *Microelectronic Engineering*, 88(4):331–335, 2011.

[147] R. Courtland. Transistors could stop shrinking in 2021. *IEEE Spectrum*, 53(9):9–11, 2016.

[148] M. P. Frank. Approaching the physical limits of computing. In *35th International Symposium on Multiple-Valued Logic (ISMVL'05)*, pages 168–185. IEEE, 2005.

[149] S. Bhattacharya and A. Sen. A review on reversible computing and it's applications on combinational circuits. *International Journal*, 9(6), 2021.

[150] T. Toffoli. Reversible computing. In *Intl. Colloquium on Automata, Languages, and Programming*, pages 632–644. Springer, 1980.

[151] O.-P. Saira, M. H. Matheny, R. Katti, W. Fon, G. Wimsatt, J. P. Crutchfield, S. Han, and M. L. Roukes. Nonequilibrium thermodynamics of erasure with superconducting flux logic. *Phys. Rev. Res.*, 2(1):013249, 2020.

[152] 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.

[153] D. J. Frank. Power-constrained CMOS scaling limits. *IBM J. Res. Dev.*, 46(2.3):235–244, 2002.

[154] C. Y. Gao and D. T. Limmer. Principles of low dissipation computing from a stochastic circuit model. *Phys. Rev. Res.*, 3(3):033169, 2021.

[155] N. Freitas, J.-C. Delvenne, and M. Esposito. Stochastic thermodynamics of nonlinear electronic circuits: A realistic framework for computing around kT. *Phys. Rev. X*, 11:031064, Sep 2021.

[156] T. Conte et al. Thermodynamic computing. *arxiv:1911.01968*, 2019.

[157] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18:023049, 2016. SFI Working Paper 15-07-025; arxiv.org:1507.01537 [cond-mat.stat-mech].

[158] G. W. Wimsatt, A. B. Boyd, P. M. Riechers, and J. P. Crutchfield. Refining Landuaer's stack: Balancing error and dissipation when erasing information. *J. Stat. Physics*, 183(16):1–23, 2021.

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

[160] S. Deffner and C. Jarzynski. Information processing and the second law of thermodynamics: An inclusive, Hamiltonian approach. *Phys. Rev. X*, 3(4):041003, 2013.

[161] A. Barone and G. Paterno. *Physics and applications of the Josephson effect*, volume 1. Wiley Online Library, 1982.

[162] S. Han. Variable β RF SQUID. In *Single-electron Tunneling and Mesoscopic Devices: Proceedings of the 4th International Conference, SQUID'91 (sessions on SET and Mesoscopic Devices), Berlin, Fed. Rep. of Germany, June 18-21, 1991*, volume 31, page 219. Springer Verlag, 1992.

[163] 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(10):6338, 1992.

[164] R. Rouse, S. Han, and J. E. Lukens. Observation of resonant tunneling between macroscopically distinct quantum levels. *Phys. Rev. Let.*, 75(8):1614, 1995.

[165] A. A. Yurgens. Intrinsic Josephson junctions: recent developments. *Supercond. Sci. Technol.*, 13:R85–R100, 2000.

[166] L. Longobardi, D. Massarotti, D. Stornaiuolo, L. Galletti, G. Rotoli, F. Lombardi, and F. Tafuri. Direct transition from quantum escape to a phase diffusion regime in YBaCuO biepitaxial Josephson junctions. *Phys. Rev. Lett.*, 109:050601, 2012.

[167] S. A. Cybart, E. Y. Cho, T. J. Wong, B. H. Wehlin, M. K. Ma, C. Huynh, and R. C. Dynes. Nano Josephson superconducting tunnel junctions in $YBa_2Cu_3O_7-\delta$ directly patterned with a focused helium ion beam. *Nature Nanotech*, 10(7):598–602, 2015.

[168] L. S. Revin, D. V. Masterov, A. E. Parafin, S. A. Pavlov, and A. L. Pankratov. Nonmonotonous temperature dependence of shapiro steps in YBCO grain boundary junctions. *Beilstein J. Nanotechnol.*, 12:1279–1285, 2021.

[169] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. *ACM SIGARCH Computer Architecture News*, 42(1):269–284, 2014.

[170] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X*, 9(2):021032, 2019.

[171] K. K. Likharev. Classical and quantum limitations on energy consumption in computation. *Intl. J. Theo. Physics*, 21(3):311–326, 1982.

[172] K. K. Likharev and A. N. Korotkov. Single-electron parametron: Reversible computation in a discrete-state system. *Science*, 273(5276):763–765, 1996.

[173] N. Takeuchi, D. Ozawa, Y. Yamanashi1, and N. Yoshikawa. An adiabatic quantum flux parametron as an ultra-low-power logic device. *Supercond. Sci. Technol.*, 26(3):035010, 2013.

[174] N. Takeuchi, Y. Yamanashi, and N. Yoshikawa. Simulation of sub-k_bt bit-energy operation of adiabatic quantum-flux-parametron logic with low bit-error-rate. *App. Physics Lett.*, 103:062602, 2013.

[175] N. Takeuchi, Y. Yamanashi, and N. Yoshikawa. Reversible logic gate using adiabatic superconducting devices. *Scientific Reports*, 4:6354, 2014.

[176] I. I. Soloviev, N. V. Klenov, S. V. Bakurskiy, M. Yu. Kupriyanov, A. L. Gudkov, and A. S. Sidorenko. Beyond Moore's technologies: operation principles of a superconductor alternative. *Beilstein J. Nanotech.*, 8:2689–2710, 2017.

[177] I. I. Soloviev, A. E. Schegolev, N. V. Klenov, S. V. Bakurskiy, M. Y. Kupriyanov, M. V. Tereshonok, A. V. Shadrin, V. S. Stolyarov, and A. A. Golubov. Adiabatic superconducting artificial neural network: Basic cells. *J. Appl. Physics*, 124(15):152113, 2018.

[178] A. E. Schegolev, N. V. Klenov, I. I. Soloviev, and M. V. Tereshonok. Adiabatic superconducting cells for ultra-low-power artificial neural networks. *Beilstein J. Nanotech.*, 7:1397–1403, 2016.

[179] K. D. Osborn and W. Wustmann. Reversible fluxon logic for future computing. In *2019 IEEE International Superconductive Electronics Conference (ISEC)*, pages 1–5. IEEE, 2019.

[180] M. P. Frank. Asynchronous ballistic reversible computing. In *2017 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8. IEEE, 2017.

[181] M. P. Frank, R. M. Lewis, N. A. Missert, M. A. Wolak, and M. D. Henry. Asynchronous ballistic reversible fluxon logic. *IEEE Trans. Appl. Superconductivity*, 29(5):1–7, 2019.

[182] K. Morita. Reversible computing. In R. A. Meyers, editor, *Encyclo. Complexity Sys. Sci.*, pages 7695–7712. Springer, 2009.

[183] S. S. Pidaparthi and C. S. Lent. Energy dissipation during two-state switching for quantum-dot cellular automata. *J. Appl. Physics*, 129(2):024304, 2021.

[184] A. L. Pankratov and B. Spagnolo. Suppression of timing errors in short overdamped josephson junctions. *Phys. Rev. Lett.*, 93:177001, Oct 2004.

[185] K. J. Laidler. The development of the Arrhenius equation. *J. Chem. Edu.*, 61.6(494), 1984.

[186] J. I. Steinfeld, J. S. Francisco, and W. L. Hase. *Chemical kinetics and dynamics*. Prentice Hall, 1999.

[187] 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. arXiv.org:1612.08616 [cond-mat.stat-mech].

[188] A. Jurgens and J. P. Crutchfield. Functional thermodynamics of Maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors. *Phys. Rev. Res.*, 2(3):033334, 2020.

[189] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Above and beyond the landauer bound: Thermodynamics of modularity. 2017. SFI Working Paper 2017-08-030; arxiv.org:1708.03030 [cond-mat.stat-mech].

[190] P. M. Riechers and J. P. Crutchfield. Fluctuations when driving between nonequilibrium steady states. *J. Stat. Phys.*, 168(4):873–918, 2017. Santa Fe Institute Working Paper 16-10-023; arxiv.org:1610.09444 [cond-mat.stat-mech].

[191] P. M. Riechers and J. P. Crutchfield. Beyond the spectral theorem: Decomposing arbitrary functions of nondiagonalizable operators. 2016. Santa Fe Institute Working Paper 16-07-015; arxiv.org:1607.06526 [math-ph].

[192] R. Lifshitz and M. C. Cross. Nonlinear dynamics of nanomechanical and micromechanical resonators. In *Reviews of Nonlinear Dynamics and Complexity*, volume 1. Wiley-VCH Verlag GmbH and Co. KGaA, 2008.

[193] M. H. Matheny, M. Grau, L. G. Villanueva, R. B. Karabalin, M. C. Cross, , and M. L. Roukes. Phase synchronization of two anharmonic nanomechanical oscillators. *Phys. Rev. Lett.*, 112:014101, 2014.

[194] J. W. Ryu, A. Lazarescu, R. Marathe, and J. Thingna. Stochastic thermodynamics of inertial-like Stuart-Landau dimer. *New J. Physics*, 23:105005, 2021.

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