Thermodynamics of Correlations and Structure in Information Engines

By

Alexander Blades Boyd

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

Rajiv Singh

John Rundle

Committee in Charge

2018

To my loving family.

Contents

List of Figures

Abstract

Thermodynamics of Correlations and Structure in Information Engines

Understanding structured information and computation in thermodynamics systems is crucial to progress in diverse fields – from biology at a molecular level to designed nano-scale information processors. Landauer's principle puts a bound on the energy cost of erasing a bit of information. This suggests that devices which exchange energy and information with the environment, which we call information engines, can use information as a thermodynamic fuel to extract work from a heat reservoir, or dissipate work to erase information. However, Landauer's Principle on its own neglects the detailed dynamics of physical information processing – the mechanics and structure between the start and end of a computation. Our work deepens our understanding of these nonequilibrium dynamics, leading to new principles of efficient thermodynamic control. We explore a particular type of information engine called an information ratchet, which processes a symbol string sequentially, transducing its input string to an output string. We derive a general energetic framework for these ratchets as they operate out of equilibrium, allowing us to exactly calculate work and heat production. We show that this very general form of computation must obey a Landauer-like bound, the Information Processing Second Law (IPSL), which shows that any form of temporal correlations are a potential thermodynamic fuel. We show that in order to leverage that fuel, the autonomous information ratchet must have internal states which match the predictive states of the information reservoir. This leads to a thermodynamic principle of requisite complexity, much like Ashby's law of requisite variety in cybernetics. This is a result of the modularity of information transducers. We derive the *modularity dissipation*, which is an energetic cost beyond Landauer's bound that predicts the structural energy costs of different implementations of the same computation. Applying the modularity dissipation to information ratchets establishes design principles for thermodynamically efficient autonomous information processors. They prescribe the ratchet's structure such that the computation saturates the bound set by the IPSL and, thus, achieves maximum thermodynamic efficiency.

Acknowledgments

I am grateful for the chance to work with my Ph.D. advisor, Professor James P. Crutchfield, because of his incredible support and vision. Without his mentorship and guidance, and the financial support of the Army Research Office, through the Multidisciplinary Research Initiative on Information Engines, this work would not be possible. The University of California in Davis also provided a wonderful collaborative learning environment where I could pursue my interests and work with amazing people. Professor Crutchfield fosters a brilliant community of scientists at the University of California, Davis, and beyond, because of his deep insight into the structure of the natural world, as well as his bold and creative attitude towards science. He gave me, and many others, the space and means to pursue scientific passions, all while insisting on incisive analytic precision, and guiding the work with his impressive intuition. Studying the physics of information processing with Professor Crutchfield has been one of the greatest adventures of my life.

The work presented in this dissertation started with a small seed, when I first met Dr. Dibyendu Mandal at the winter Mini Stat Mech meeting at UC Berkeley in 2013. I found his proposed model of an autonomous Maxwell's demon fascinating but perplexing. Dr. Mandal graciously answered my questions, helping me to better understand the nonequilibrium thermodynamics that underly the model, but I left feeling that I still had much more to learn. Fortunately for me, I got that opportunity when Professor Crutchfield organized joint meetings between the three of us. Since then, we've taken a deep dive into the physics of transducers presented in this dissertation.

Dr. Mandal's thoughtful and grounded approach to physics has been an invaluable asset to our work. He showed me how to be a diligent student of the field's to which we contribute. Without his deep understanding nonequilibrium statistical mechanics, this work could have never taken off. His voice is interwoven in every part of it. I am grateful for his integrity and all that he taught me about physics and being a scientist.

Dr. Paul Riechers, in addition to making invaluable contributions to the thermodynamics of synchronization as a colleague, is a good friend. His tenacious work ethic, sparkling intellect, and infectious optimism serve as inspiration for me. He challenged me

to consider beyond what *is*, and imagine what *can be*. I've taken this lesson to heart, both in my work and in my life. Through example, he shows that with hopeful confidence and commitment, you can achieve incredible things and surprise even yourself.

Professor Crutchfield, Dr. Mandal, and Dr. Riechers are the principal collaborators in the work presented in this dissertation, but many more people have contributed to my six year journey at Davis.

Dr. Korana Burke is a champion, pulling me into the scientific community by inviting me to share my findings and connect with other scientists. I am lucky to have her as a friend and mentor. Her helpful critiques and comments refined my public voice so that I could effectively communicate my work to my peers. Watching her commitment to building our community of nonequilibrium dynamicists, I learned that an essential part of being a good scientist is being a good person. I strive to pay forward the invaluable generosity she has shown me through her mentorship.

While I don't present the results of our efforts here, working with Greg Wimsatt is a joy and has been an essential part of my growth. His desire to understand on the deepest level is a constant reminder of why I choose to pursue physics. He grounds me and challenges me to consider fundamental questions. As a friend, he helps me see the world from new angles, which can be a struggle. However, I am always grateful for how he expands my perspective.

My mother, Joan, showed me the value of patience and persistence. I strive to follow the example she sets through her kindness and shining empathy. Her commitment to the idea that everyone deserves compassion helped me see the importance of understanding others' contributions and cultivating humility.

My father, Wes, nurtured my curiosity and fostered my love of physics from a young age. He wouldn't answer many of my questions, but he encouraged them nonetheless. He showed me that the process of inquiry is a reward in itself, even if the answers are inaccessible. He helped me see the joy of a good question.

My sister, Robin, and cousin, Willie, are incredible friends. I am grateful that, while we pursued our separate adventures, we stayed close. Growing with both of them has

been one of the great joys of my life.

My family's unwavering love anchored me through the highs and lows of my graduate degree. Failure, success, confusion, inspiration; my family was there for all of it. They gave me the courage to pursue a dream that I've nurtured since before I knew what physics was: to be a discoverer, searching for the fundamental laws of nature.

Chapter 1

Nonequilibrium Thermal Physics of Information Processing

1.1 Introduction

In 1867 James Clerk Maxwell famously proposed that an intelligent and "neat fingered" being, later referred to as Maxwell's demon [3], would be able to observe and subsequently control its environment in such a way that it could turn thermal fluctuations into useful work. This suggested that there was a thermodynamic benefit to processing information in the environment, but also posed a challenge to the Second Law of thermodynamics. The demon seemed to violate fundamental tenant of thermodynamics, that the entropy of the universe never decreases, as the demon could steadily transform disordered thermal energy into ordered work energy, decreasing the disorder/entropy of the universe. It wasn't until physicists began building physical information processing devices in the past century that they resolved this apparent and paradoxical violation of the Second Law of thermodynamics.

Physically processing information has a thermodynamic cost. Landauer showed that, because of state space compression, erasing a bit at temperature T requires a minimum work investment of

$$\langle W^{\text{erase}} \rangle_{\text{min}} = k_B T \ln 2, \tag{1.1}$$

known as Landauer's Principle [4]. Thus, Maxwell's demon is limited by the thermody-

namic cost of its own information processing or intelligence. It may measure its environment and control it to extract work, as described in Maxwell's experiment, but Landauer's Principle guarantees that the cost of this information processing outweighs any benefit it could gain from that measurement. There are costs and benefits information processing, but the benefits are always outweighed by the costs *if* the demon had finite memory.

With finite memory, the demon eventually reaches the limits of its storage and must begin erasing what it observed. But, what if, rather than erasing what it measured, it is able to write and store that information in a memory bank. The demon would, in theory, function as originally intended and turn thermal energy into useful work. However, in this new picture, the demon no longer violates the Second Law of thermodynamics, because the increase in Shannon entropy of the memory bank corresponds to an increase in thermodynamic entropy which balances the work generation. This memory bank is a type of thermodynamic resource, called an information reservoir, which contains the information bearing degrees of freedom [5]. An information reservoir can be used in conjunction with work and heat reservoirs to perform useful operations.

A device that interacts with an information reservoir, heat reservoir, and work reservoir is called an *information engine*. An information engine functions much like a heat engine, which uses the temperature difference between a hot and cold heat reservoir to move energy into a work reservoir, as shown in Fig. 1.1. In the same way, an information engine can use the order in an information reservoir as a thermal "fuel" [6] to produce work.

The first example of a device that can reliably use an information reservoir as a resource was an exactly solvable Maxwell's demon [7]. This demon interacted with each bit in a semi-infinite string sequentially, making thermally activated transitions and so facilitating work generation. The work produced by such a device was bounded by $k_B T \ln 2$ times the Shannon entropy produced in the string, preserving the Second Law of thermodynamics. This particular type of information engine, which autonomously processes the information in a symbol string unidirectionally, transducing inputs to outputs, is known as an *information ratchet*. It is the focus of this work.

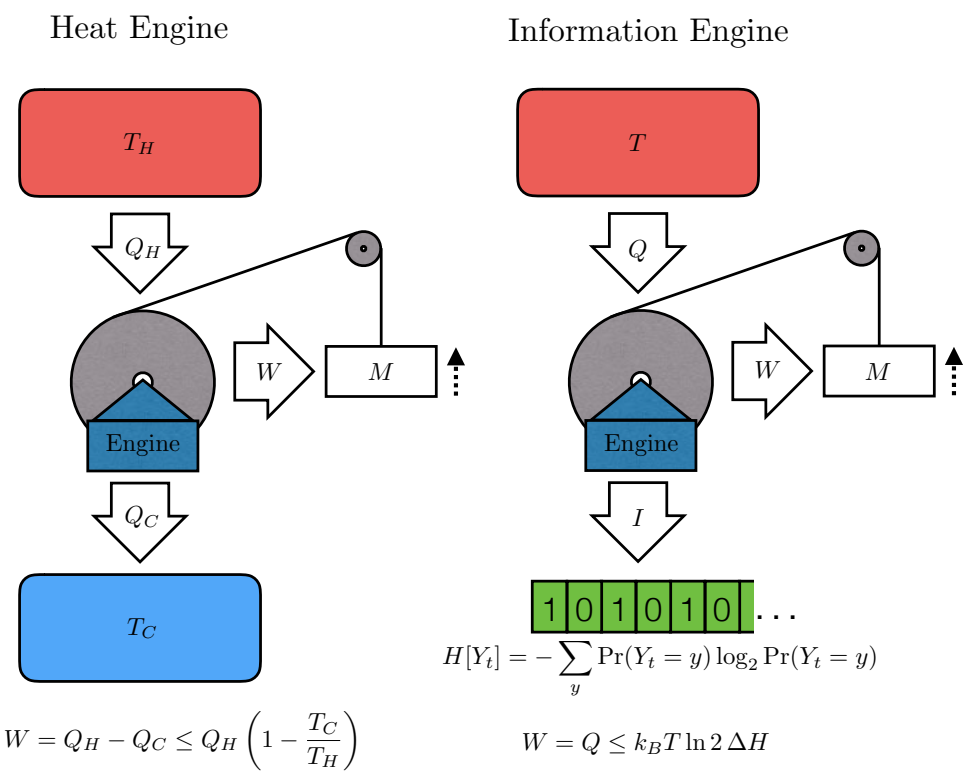


Figure 1.1. (Left) A heat engine uses the natural tendency of energy to from a hot to cold reservoir, and syphons off some of the energy to generate useful work. The work is bounded by the net heat dissipated from a hot thermal reservoir times the Carnot efficiency. (Right) An information engine uses the natural tendency of an information reservoir (shown in the bottom) to randomize in order to move energy directly from a thermal reservoir to a work reservoir. The work production is bounded by the change in uncertainty in the information reservoir ΔH times the temperature of the heat reservoir T.

These devices express a wide variety of behavior—creating, destroying, and transforming structured (temporally correlated) signals. They present a general thermodynamic framework for computational devices that use internal states to map inputs to outputs. We show that there are thermodynamic limits on the energetics of their behavior, much like Landauer's bound. Moreover, we show how achievable these bounds are, quantifying the thermodynamic efficiency of different implementations of the same computation. These measures lead to a general theory of thermodynamically efficient computing.

1.2 Identifying Functional Thermodynamics

In the original work on information ratchets, it was shown that they can autonomously use ordered inputs to turn thermal energy into work, acting as an engine. Or, they can create order in the output using work, acting as an "eraser" [7]. These two functionalities illustrate the interplay between information and energy in the equation for average global entropy production $\langle \Sigma \rangle$ when manipulating information bearing degrees of freedom [8, 9]

$$\langle \Sigma \rangle = -\langle W \rangle - k_B T \ln 2 \Delta H, \tag{1.2}$$

where $\langle W \rangle$ is the average work produced by the system and ΔH is the change in Shannon entropy of the information bearing degrees of freedom. Because the entropy production must be non-negative, the work investment in a thermodynamic process must exceed the change in uncertainty of the information bearing degrees of freedom.

However, initial investigations of information ratchets only considered the order in a single bit of the input, independent of other inputs. It was shown [7, 10] that the change in uncertainty of a single bit, from input to output, bounded the work production in special cases

$$\langle W \rangle \leq k_B T \ln 2 (H[Y'_N] - H[Y_N]). \tag{1.3}$$

Here, $H[Z]$ denotes the Shannon entropy of the random variable Z [11], which is expressed

$$H[Z] = -\sum_{z \in \mathcal{Z}} \Pr(Z = z) \log_2 \Pr(Z = z), \tag{1.4}$$

where z are taken from the alphabet of possible realizations $\mathcal{Z}$. However, all order in an input is not necessarily detectible through single-symbol entropies. The true change in accessible free energy in the input comes from considering the Shannon entropy of the whole string $Y_{0:\infty} = Y_0 Y_1 Y_2 \cdots$. Temporal correlations can exist within the string, such that the uncertainty per symbol, known as the entropy rate [12]

$$h_\mu = \lim_{\ell \to \infty} \frac{H[Y_{0:\ell}]}{\ell}, \tag{1.5}$$

diverges from the single symbol entropy. The difference $H[Y_N] - h_\mu$ is the length-1 redundancy [12]—one useful measure of temporal correlations. This redundancy reflects

the contribution of global correlations to the available nonequilibrium free energy in the symbol string. We show that if you take into account the global correlations in the information reservoir, which amount to temporal correlations in the ratchet's input, the work investment is asymptotically bounded by the difference in entropy rates of input h_μ and output h'_μ

$$\langle W \rangle \leq k_B T \ln 2 (h'_\mu - h_\mu), \tag{1.6}$$

rather than by the change in single-symbol entropy $H[Y'_N] - H[Y_N]$, as shown in Eq. 1.3. This is the Information Processing Second Law (IPSL) [13]: any form of order is potentially a thermodynamic resource. The more ordered the input is, the smaller the entropy rate h_μ, and the more work one can potentially extract from that input. Similarly, the more ordered and predictable the output is, the smaller the entropy rate h'_μ, and the more work you must invest to create the desired order.

Well known since the invention of communication theory, entropy rates are generally a challenge to calculate, since this requires accounting for a process' temporal correlations at all time scales. Due to this, we turn to the tools of computational mechanics to evaluate accurate thermodynamic bounds on information ratchets. The temporal correlations that we must account for come from memory in one form or another. For the input string, its ϵ-machine is the hidden Markov model (HMM) which is the minimal predictive machine [12]. Its hidden states store the temporal correlations in the string, in that they are the minimal amount of memory required to predict the string.

The information ratchet's internal states also serve as a form of memory, storing information about the past inputs and outputs. The ratchet behaves as information transducer [13], using its internal memory to react to the structured temporal order in the input or to create it in the output [1]. The form of this transducer is given by the Markov process over the joint space of the ratchet states and interaction bit, as described in Fig. 2.1, which in turn exactly specifies the form of the HMM that generates the output. Thus, from this transducer formalism, we can exactly predict the entropy rates in order to set better thermodynamic bounds on ratchet work generation.

Additionally, detailed balance constrains the information engine's energetics. Specif-

ically, the energy differences are proportional to the log ratio of forward and reverse thermal transition probabilities

$$\Delta E_{a\to b} = k_B T \ln \frac{M_{b\to a}}{M_{a\to b}}. \tag{1.7}$$

In this way we can exactly calculate average heat and work flows for the ratchets. And this allows us to compare to the bounds set by the IPSL.

We start our exploration with systems driven by temporally uncorrelated inputs, where the ratchet reads the realization of a biased coin with every time step. We consider a memoryful ratchet, which is able to take such an input, and create temporal correlations at a thermodynamic cost. We identify the functionality of the ratchet for a variety of parameter values, showing that it can either behave as (i) an engine, randomizing the input string to generate work; (ii) an eraser, using work to create order in the output; or (iii) a dud, which also uses work, but does not create any useful order in the outputs with it.

The benefit of the computational mechanics framework, as applied to autonomous thermodynamic information processors, is that it allows us to detect any type of order created by the ratchet. With this, we can detect new erasure functionalities. For example, we identify an entirely new eraser region of ratchet parameter space. This novel functionality is missed by past single-symbol entropy analyses, since the single-symbol uncertainty grows. Instead, the ratchet creates order purely by generating temporal correlations in the output string.

In this specific case, where the ratchet has memory, but the input is IID (independent identically distributed) and thus memoryless, the IPSL is a stricter and more accurate bound on the work production than the single-symbol bound. Thus, for memoryless input and memoryful ratchets, the single symbol bound is still valid but a less informative bound. This, however, changes when the inputs are temporally correlated.

1.3 Correlation-powered Information Engines

Our theory establishes that temporal correlations are a thermodynamic resource—a resource that can be leveraged to produce work. However, there has been disagreement

about whether the IPSL (Eq. (1.6)) or single-symbol entropy bounds (Eq. (1.3)) are valid for information ratchets [10, 14]. The two bounds make different predictions for an input where all order comes in the form of temporal correlations, such as the period-2 process of alternating 0s and 1s, which has two equally likely phases

$$\Pr(Y_{0:\infty} = 0101...) = \Pr(Y_{0:\infty} = 1010...) = 1/2. \tag{1.8}$$

The IPSL suggests that this input could be used to run an engine, which is impossible according to the single-symbol bounds. We investigate how to make use of this resource.

To do this, we first show that there are ratchets that provably cannot make use of temporal correlations. We show that memoryless ratchets are only sensitive to the single-symbol bias of their input [6]. Thus, to memoryless ratchets, a temporally correlated memoryful input is functionally equivalent to a memoryless input that has the same single-symbol bias but no correlations. And so, if the ratchet obeys the IPSL for the temporally uncorrelated approximation of the input, it must obey single symbol bounds, confirming previous results for memoryless ratchets [10]. This means that we must use ratchets with internal states if we hope to create an information engine that runs on temporal correlations, guiding our design of information ratchets.

For concreteness, we introduce a three state engine, shown in Fig. 3.3, which runs off the period-2 input to produce work. Thus, with memoryful ratchets, it is possible to not only impose stricter bounds on work production with the IPSL, but also violate the single-symbol bounds, which state that such inputs cannot produce any work. Asymptotically we see that we're able to produce

$$\langle W \rangle = k_B T \frac{1-\delta}{e}, \tag{1.9}$$

while randomizing the output, where δ is a parameter of the ratchet which determines how quickly it transitions into synchronization states. Synchronization appears to be an essential part of leveraging temporal correlations. For this particular ratchet, we find that the states of the ratchet approach synchrony with the causal states of the input process, where the ratchet is in a stationary work-producing dynamical phase.

However, the ratchet may start out of phase with the input, in which case it will initially dissipate energy in a work-dissipating dynamical phase. The parameter δ is the rate at which the ratchet transitions out of this phase into the aforementioned synchronization states, which facilitate the transition to the stationary work-producing dynamical phase. Thus, it appears that synchronization is required to produce work from temporally correlated sources. This is unlike how engines use single-symbol statistical biases, which take no synchronization to leverage.

There is, however, a cost to synchronization. Another parameter of the ratchet, γ gives the rate of synchronization from the synchronization state

$$R_{\text{sync}}(\delta, \gamma) = \gamma, \tag{1.10}$$

but because of detailed balance, we find that this parameter also contributes to an energy cost to synchronizing

$$Q_{\text{sync}}(\delta, \gamma) = k_B T \ln \frac{\delta}{\gamma}. \tag{1.11}$$

As a result, we see a three-way tradeoff between asymptotic work production, synchronization rate, and synchronization cost

$$Q_{\text{sync}} + k_B T \ln R_{\text{sync}} - k_B T \ln \left(1 - \frac{e\langle W \rangle}{k_B T}\right) = 0. \tag{1.12}$$

This is much like the tradeoff between the fidelity, rate, and energy cost of information processing in finite time [15, 16].

Synchronization also allows the ratchet to cope with noisy inputs, such as phase slips in the input which push the ratchet out of the work-producing dynamical phase. Functionally, this kind of synchronization is tantamount to error correction, where the ratchet autonomously corrects its estimate of the causal state of the input. The faster the ratchet transitions to the synchronization state, the less time the ratchet spends in a dissipative phase. However, because of the tradeoff with synchronization cost and asymptotic work production, there is a sweet spot in synchronization rate, where the ratchet is able, on average, to leverage the most work from a noisy source. We analytically calculate this

value for different rates of phase slips, and show there is a critical frequency of phase slips past which it is impossible for this ratchet effectively autocorrect and extract work.

Thus, we see that temporal correlations are indeed a thermal resource, but that the ratchet must synchronize its internal states to the causal structure of its environment in order leverage that resource. That is, an effective ratchet requires memory. This adds credence to the IPSL, because it predicted that period-2 inputs could be used to generate work. The emerging challenge is to find the principles of ratchet design that allow us to take advantage of the environment.

1.4 Thermodynamics of Requisite Complexity

Up to this point, we saw that memoryless ratchets can leverage the order of temporally uncorrelated inputs, but cannot leverage structure in the form of temporal correlations. Instead, we require memoryful ratchets to leverage those correlations. This is in line with the information theoretic bounds, and, because temporal correlations require memory to generate, these bounds suggest that memoryless ratchets are best for leveraging memoryless inputs and memoryful ratchets are best for leveraging memoryful inputs. We refer to this matching between the structure of the input and the ratchet, where memory requires memory, as the thermodynamic principle of requisite variety [17], much like Ashby's law of Requisite Variety in the field of cybernetics [18].

While the information theoretic bounds naturally suggest the matching between ratchet and input, information ratchets fundamentally operate out of equilibrium. The result is that they may not satisfy these bounds. For instance, we show that even when designed optimally for memoryful inputs, memoryless ratchets fall short of the thermodynamic bounds, dissipating energy and creating entropy. This is because they interact with their input out of equilibrium. Therefore, we must consider the explicit energetic dynamics of these ratchets to predict their behavior.

In the previous section, we already made dynamical arguments that memory is required to take advantage of temporal correlations, and thus memoryful ratchets are better suited than memoryless for extracting work from those information sources. However, it may

well be that, since memoryless ratchets cannot achieve the IPSL, there are memoryful ratchets which can do better. That is, the law of requisite variety/memory may not hold.

Fortunately, our analysis of ratchet energetic dynamics resolves the issue as one might expect. We show that, for a ratchet driven by memoryless inputs, the dynamically-evaluated average work production increases if the ratchet states are coarse-grained. This gives us the exact form of a memoryless ratchet which can at least match work production of any memoryful ratchet, confirming that memory matches memory.

However, this result is limited to finite ratchets. If we consider ratchets with infinitely many states, then the IPSL no long holds, because the ratchet's internal states may not ever asymptote to a steady state distribution. Thus, the ratchet states can act as a reservoir of free energy, much like the bit string. We show how this works with a particular model that, through continual expansion into new ratchet states, is able to produce more than $k_B T \ln 2$ work per cycle from a sequence of incoming bits. The ratchet is able to extract more nonequilibrium free energy than is available within each bit, but when we account for the expansion of the ratchet's memory, Landauer's principle is preserved, as is the second law of thermodynamics.

In our exploration of ratchet memory, we see that it is of critical importance to ratchet functionality. Memory is necessary to create temporal correlations in the output, or leverage temporal correlations in the input. We confirm our guess at the principle of requisite variety with finite memory dynamical models of information ratchets, which operate out of equilibrium, and thus are not maximally thermodynamically efficient, leaving some energy on the table. However, while these example cases illustrate principles that allow us to distinguish between memoryless, memoryful, and infinite memory thermodynamics. We want general principles of ratchet design that tell us explicitly how to construct thermodynamic agents that optimally leverage their environment. We want a thermodynamics of structure.

1.5 Thermodynamics of Synchronization

As we are beginning to see, different implementations of the same computation (input to output) can have radically different thermodynamic consequences. While we may be able to randomize a temporally correlated input string, we will not be able to leverage that structure without memory. The example ratchet we used to leverage the inputs was able to produce work by synchronizing to the predictive states of its input. However, when leveraging a memoryful input process, the presciption that one must have a memoryful ratchet leaves much unspecified about ratchet design. The period-2 driven ratchet was constructed with trial, error, and intuition. Fully rational design is a goal for the future.

As a first push towards information theoretic principles of design, we look more closely at the entropic bounds on the whole state space: tape and ratchet together. The IPSL, which predicts that temporal correlations are a thermodynamic resource, was derived as an asymptotic approximation of such global entropy change bounds [13]. While the asymptotic IPSL allows us to predict asymptotic energetics, memoryful ratchets take time to synchronize to their asymptotic behavior. For instance, for the period-2 driven ratchet, synchronizing to the input's causal states required both time and energy. The difference between the asymptotic and global changes in entropy gives a transient dissipation, which accumulates as the ratchet synchronizes. This, in essence, is a bound on the synchronization dissipation, which can be applied beyond engines, which synchronize to their input, to generators and other information processors.

For a particular computational input-output process, we can isolate the component of the transient cost that depends on the ratchet states, yielding the implementation cost

$$\langle Q^{\text{impl}} \rangle_{\text{min}} = k_B T \ln 2 I[X_0; \overleftarrow{Y}', \overrightarrow{Y}]. \tag{1.13}$$

We have shifted to the frame of the ratchet after many operations, so that X_0 is the current state of the ratchet, $\overrightarrow{Y}$ is the input future, and $\overleftarrow{Y}'$ is the output past [19].

The implementation cost applies to any form of transduction, but we consider the particular case of pattern generators, because a previous analysis [20] argued that simpler ratchets are thermodynamically more efficient for generating patterns. We see that, if you restrict to predictive pattern generators, then, indeed, more memory does correspond to

more transient dissipation and a greater cost to synchronization, because the implementation dissipation is proportional to the uncertainty in the ratchet. Thus, for predictive ratchets, this dissipation is bounded by the statistical complexity of the output. However, generators are not required to predict their outputs, in that the ratchet states need not be a function of the output past [21]. They can, instead, neglect information about their past, such that they store less information about the past than the statistical complexity. It turns out that retrodictive ratchets, which minimize the information stored about the output past, are most efficient, with zero transient dissipation. Moreover, these ratchets need not be simple, as long as additional states do not store information about the past and so do not induce transient dissipation. Thus, we see one of our first information theoretic structural measures of thermodynamic efficiency. The implementation dependent transient dissipation gives principles of thermodynamic design that can be used to eliminate unnecessary energy loss as an agent adapts to its environment. However, there are also costs that apply beyond this transient regime—costs that depend explicitly on the modular structure of information processing.

1.6 Thermodynamics of Modularity

Delving further into the thermodynamics of structure, we find that localized modular information processing has an unavoidable thermodynamic cost that comes from dissipated global correlations. Though the principle of thermodynamic requisite complexity/variety, as discussed in Chapter 4, requires matching between ratchet memory and input memory, it leaves much undetermined about the structure of thermodynamically efficient ratchets. By recognizing that ratchets operate locally, on a single bit of the symbol sequence at time, we see that it is a fundamentally modular information processor. Thus, the ratchet's controller is potentially unaware of the various global correlations between bits that extend throughout the tape. As such, while the next bit may be predictable from past inputs, unless the autonomous ratchet states are able to store those correlations within, the bit is effectively random to ratchet controller. And the latter leverages energies as if the input symbol is more random that it really is. As such, it dissipates global correlations, which

are a source of global free energy. The net result is that the global entropy production is positive:

$$\langle \Sigma \rangle = -\langle W \rangle - \Delta F^{\text{neq}} > 0. \tag{1.14}$$

This phenomenon generalizes to any computation that decomposes into a sequence localized computations over a subset of the system variables. Such computations are modular, in that each localized computation can be performed independently of all others. Modularity is a common property of many designed computing systems. For example, elementary universal logic gates can be composed to implement any computation. Designing the physical implementation one of these logic gates, then connecting to a collection of others is simpler than designing the global computation from scratch. However, because these gates are designed without considering the global correlations that may exist in the calculation, they necessarily dissipate energy and produce entropy. For a particular step in a modular computation, where only the subset of the information bearing degrees of freedom $\mathcal{Z}^i$ is changing and the rest of the system $\mathcal{Z}^s$ is held fixed, we quantify a bound on the global entropy production. We call this the modularity dissipation [22]. If the particular step of the computation happens over the time interval $(t, t+\tau)$, then the modularity dissipation is proportional to the information shared between the initial state of the interacting subsystem Z_t^i and the initial state of the stationary subsystem Z_t^s, conditioned on the final state of the interacting modular subsystem $Z_{t+\tau}^i$

$$\langle \Sigma^{\text{mod}}_{t \to t+\tau} \rangle_{\min} = k_B \ln 2 I[Z_t^i; Z_t^s | Z_{t+\tau}^i]. \tag{1.15}$$

This is the amount of information the modularly interacting component forgets about the non-interacting subsystem. Thus, it corresponds to the global correlations between the two systems that are lost over the course of the operation.

Modularity dissipation applies to many different systems, from physical logical circuits to biological organisms. Here, we only apply it to thermodynamic information ratchets. Plugging in the ratchet and interacting bit as the interacting subsystem, we find immediate results which quantify the thermodynamic efficiency of information transduction.

Maximizing the thermodynamic efficiency, and thus minimizing the modularity dissipation, leads to interesting results for two subclasses of transducers.

For information extractors, which transform a structured input process into sequence of uncorrelated IID outputs, we find that thermodynamic efficiency means that the ratchet must be predictive of its input. This leads directly to the principle of requisite complexity, because in order to be predictive, the asymptotic state uncertainty of the ratchet must at least match the statistical complexity C_μ of the input [22]

$$\lim_{N\to\infty} H[X_N] \geq C_\mu. \tag{1.16}$$

We find similar results for information generators, which transform unstructured IID inputs into structured outputs. Rather than being predictors of their inputs, thermodynamically efficient generative ratchets are *retrodictive* of their output, meaning that they contain as little information about past outputs, while simultaneously containing all the information shared between the past and future. In this way, we show how to construct thermodynamically efficient pattern extractors and generators, such that they store all the relevant global correlations in the ratchet states, rather than dissipating free energy in the form of lost global correlations.

Despite this progress, there is a breadth of open questions about the thermodynamics of transduction, such as how to implement these information processors as efficiently as possible in finite time. We propose a method for quasistatically implementing the ratchet, but such implementations require long times, because they are quasistatic. Also, there are many computations/ratchets which fall outside the purview of pattern generation or extraction. These ratchets take patterns from one form to another, rather than creating them purely from a physical resource or completely destroying them for an energetic benefit. More work is required to understand how to minimize the modularity dissipation of arbitrary information transduction.

Chapter 2

Identifying Functional Thermodynamics

2.1 Introduction

The Second Law of Thermodynamics is only statistically true: while the entropy production in any process is nonnegative on the average, $\langle \Sigma \rangle \geq 0$, if we wait long enough, we shall see individual events for which the entropy production is negative. This is nicely summarized in the recent fluctuation theorem for the probability of entropy production ΔS [23, 24, 25, 26, 27, 28, 29]:

$$\frac{\Pr(\Sigma)}{\Pr(-\Sigma)} = e^{\Sigma} \ , \tag{2.1}$$

implying that negative entropy production events are exponentially rare but not impossible. Negative entropy fluctuations were known much before this modern formulation. In fact, in 1867 J. C. Maxwell used the negative entropy fluctuations in a clever thought experiment, involving an imaginary intelligent being—later called Maxwell's Demon—that exploits fluctuations to violate the Second Law [3, 30]. The Demon controls a small frictionless trapdoor on a partition inside a box of gas molecules to sort, without any expenditure of work, faster molecules to one side and slower ones to the other. This gives rise to a temperature gradient from an initially uniform system—a violation of the Second Law. Note that the "very observant and neat fingered" Demon's "intelligence" is necessary; a frictionless trapdoor connected to a spring acting as a valve, for example,

cannot achieve the same feat [31].

Maxwell's Demon posed a fundamental challenge. Either such a Demon could not exist, even in principle, or the Second Law itself needed modification. A glimmer of a resolution came with L. Szilard's reformulation of Maxwell's Demon in terms of measurement and feedback-control of a single-molecule engine. Critically, Szilard emphasized hitherto-neglected information-theoretic aspects of the Demon's operations [32]. Later, through the works of R. Landauer, O. Penrose, and C. Bennett, it was recognized that the Demon's operation necessarily accumulated information and, for a repeating thermodynamic cycle, erasing this information has an entropic cost that ultimately compensates for the total amount of negative entropy production leveraged by the Demon to extract work [4, 33, 34]. In other words, with intelligence and information-processing capabilities, the Demon merely shifts the entropy burden temporarily to an information reservoir, such as its memory. The cost is repaid whenever the information reservoir becomes full and needs to be reset. This resolution is concisely summarized in Landauer's Principle [4]: the Demon's erasure of one bit of information at temperature T requires at least $k_{\mathrm{B}}T \ln 2$ amount of heat dissipation, where k_{B} is Boltzmann's constant. (While it does not affect the following directly, it has been known for some time that this principle is only a special case [35].)

Building on this, a modified Second Law was recently proposed that explicitly addresses information processing in a thermodynamic system [5, 36]:

$$\langle \Sigma \rangle + k_{\mathrm{B}} \ln 2 \, \Delta \mathrm{H} \geq 0 \ , \tag{2.2}$$

where $\Delta \mathrm{H}$ is the change in the information reservoir's configurational entropy over a thermodynamic cycle. This is the change in the reservoir's "information-bearing degrees of freedom" as measured using Shannon information H [11]. These degrees of freedom are coarse-grained states of the reservoir's microstates—the mesoscopic states that store information needed for the Demon's thermodynamic control. Importantly for the following, this Second Law assumes *explicitly observed* Markov system dynamics [5] and quantifies this relevant information only in terms of the distribution of *instantaneous* system microstates; not, to emphasize, microstate path entropies. In short, while the system's

instantaneous distributions relax and change over time, the information reservoir itself is not allowed to build up and store memory or correlations.

Note that this framework differs from alternative approaches to the thermodynamics of information processing, including: (i) active feedback control by external means, where the thermodynamic account of the Demon's activities tracks the mutual information between measurement outcomes and system state [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]; (ii) the multipartite framework where, for a set of interacting, stochastic subsystems, the Second Law is expressed via their intrinsic entropy production, correlations among them, and transfer entropy [51, 52, 53, 54]; and (iii) steady-state models that invoke time-scale separation to identify a portion of the overall entropy production as an information current [55, 56]. A unified approach to these perspectives was attempted in Refs. [57, 58, 59].

Recently, Maxwellian Demons have been proposed to explore plausible automated mechanisms that appeal to Eq. (3.10)'s modified Second Law to do useful work, by deceasing the physical entropy, at the expense of positive change in reservoir Shannon information [7, 60, 56, 61, 62, 63, 64]. Paralleling the modified Second Law's development and the analyses of the alternatives above, they too neglect correlations in the information-bearing components and, in particular, the mechanisms by which those correlations develop over time. In effect, they account for Demon information-processing by replacing the Shannon information of the components as a whole by the sum of the components' *individual* Shannon informations. Since the latter is larger than the former [11], using it can lead to either stricter or looser bounds than the true bound which is derived from differences in total configurational entropies. More troubling, though, bounds that ignore correlations can simply be violated. Finally, and just as critically, they refer to configurational entropies, not the intrinsic dynamical entropy over system trajectories.

This Letter proposes a new Demon for which, for the first time, all correlations among system components can be explicitly accounted. This gives an exact, analytical treatment of the thermodynamically relevant Shannon information change—one that, in addition, accounts for system trajectories not just information in instantaneous state distributions.

The result is that, under minimal assumptions, we derive a Information Processing Second Law (IPSL) that refines Eq. (3.10) by properly accounting for intrinsic information processing reflected in temporal correlations via the overall dynamic's Kolmogorov-Sinai entropy [65].

Notably, our Demon is highly functional: Depending on model parameters, it acts both as an engine, by extracting energy from a single reservoir and converting it into work, and as an information eraser, erasing Shannon information at the cost of the external input of work. Moreover, it supports a new and counterintuitive thermodynamic functionality. In contrast with previously reported erasure operations that only decreased single-bit uncertainty, we find a new kind of erasure functionality during which multiple-bit uncertainties are removed by adding correlation (i.e., by adding temporal order), while single-bit uncertainties are actually increased. This new thermodynamic function provocatively suggests why real-world ratchets support memory: The very functioning of memoryful Demons relies on leveraging temporally correlated fluctuations in their environment.

2.2 Information Ratchets

Our model consists of four components, see Fig. 2.1: (1) an ensemble of bits that acts as an information reservoir; (2) a weight that acts as a reservoir for storing work; (3) a thermal reservoir at temperature T; and (4) a finite-state ratchet that mediates interactions between the three reservoirs. The bits interact with the ratchet sequentially and, depending on the incoming bit statistics and Demon parameters, the weight is either raised or lowered against gravity.

As a device that reads and processes a tape of bits, this class of ratchet model has a number of parallels that we mention now, partly to indicate possible future applications. First, one imagines a sophisticated, stateful biomolecule that scans a segment of DNA, say as a DNA polymerase does, leaving behind a modified sequence of nucleotide base-pairs [66] or that acts as an enzyme sequentially catalyzing otherwise unfavorable reactions [67]. Second, there is a rough similarity to a Turing machine sequentially recognizing tape symbols, updating its internal state, and taking an action by modifying the tape cell

and moving its read-write head [68]. When the control logic is stochastic, this sometimes is referred to as "Brownian computing" [69, and references therein]. Finally, we are reminded of the deterministic finite-state tape processor of Ref. [70] that, despite its simplicity, indicates how undecidability can be imminent in dynamical processes. Surely there are other intriguing parallels, but these give a sense of a range of applications in which sequential information processing embedded in a thermodynamic system has relevance.

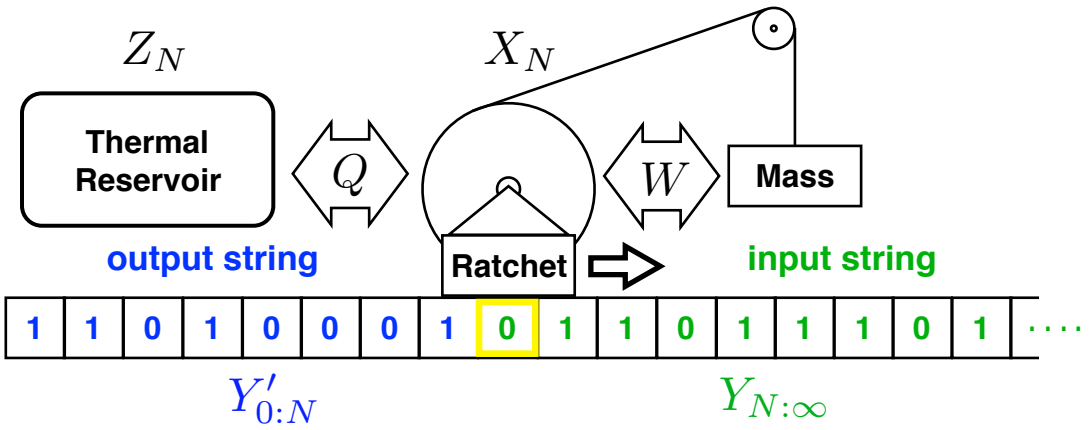


Figure 2.1. Information ratchet sequentially processing a bit string: At time step N, X_N is the random variable for the ratchet state and Z_N that for the thermal reservoir. $Y_{N:\infty}$ is the block random variable for the input bit string and $Y'_{0:N}$ that for the output bit string. The last bit Y_N of the input string, highlighted in yellow, interacts with the ratchet and is called the interaction bit. The arrow on the right of the ratchet indicates the direction the ratchet moves along the tape as it sequentially interacts with each input bit in turn.

The bit ensemble is a semi-infinite sequence, broken into incoming and outgoing pieces. The ratchet runs along the sequence, interacting with each bit of the input string step by step. During each interaction at step N, the ratchet state X_N and interacting bit Y_N fluctuate between different internal joint states within $\mathcal{X} \otimes \mathcal{Y}$, exchanging energy with the thermal reservoir and work reservoir, and potentially changing Y_N's state. At the end of step N, after input bit Y_N interacts with the ratchet, it becomes the last bit Y'_N of the output string. By interacting with the ensemble of bits, transducing the input string into the output string, the ratchet can convert thermal energy from the heat reservoir into work energy stored in the weight's height.

The ratchet interacts with each incoming bit for a time interval τ, starting at the 0th bit Y_0 of the input string. After N time intervals, input bit Y_{N-1} finishes interacting with the ratchet and, with the coupling removed, it is effectively "written" to the output string, becoming Y'_{N-1}. The ratchet then begins interacting with input bit Y_N. As Fig. 2.1 illustrates, the state of the overall system is described by the realizations of four random variables: X_N for the ratchet state, $Y_{N:\infty}$ for the input string, $Y'_{0:N}$ for the output string, and Z_N for the thermal reservoir. A random variable like X_N realizes elements x_N of its physical state space, denoted by alphabet $\mathcal{X}$, with probability $\Pr(X_N = x_N)$. Random variable blocks are denoted $Y_{a:b} = Y_a Y_{a+1} \ldots Y_{b-1}$, with the last index being exclusive. In the following, we take binary alphabets for $\mathcal{Y}$ and $\mathcal{Y}'$: $y_N, y'_N \in \{0, 1\}$. The bit ensemble is considered two joint variables $Y'_{0:N} = Y'_0 Y'_1 \ldots Y'_{N-1}$ and $Y_{N:\infty} = Y_N Y_{N+1} \ldots$ rather than one $Y_{0:\infty}$, so that the probability of realizing a *word* $w \in \{0, 1\}^{b-a}$ in the output string is not the same as in the input string. That is, during ratchet operation typically $\Pr(Y_{a:b} = w) \neq \Pr(Y'_{a:b} = w)$.

The ratchet steadily transduces the input bit sequence, described by the input word distribution $\Pr(Y_{0:\infty}) \equiv \{\Pr(Y_{0:\infty} = w)\}_{w \in \{0,1\}^\infty}$—the probability for every semi-infinite input word—into the output string, described by the word distribution $\Pr(Y'_{0:\infty}) \equiv \{\Pr(Y'_{0:\infty} = v)\}_{v \in \{0,1\}^\ell}$. A useful assumption for computational mechanics is that the word distributions are stationary, meaning that $\Pr(Y_{a:a+b}) = \Pr(Y_{0:b})$ for all nonnegative integers a and b.

A key question in working with a sequence such as $Y_{0:\infty}$ is how random it is. One commonly turns to information theory to provide quantitative measures: the more informative a sequence is, the more random it is. For words at a given length ℓ the average amount of information in the $Y_{0:\infty}$ sequence is given by the *Shannon block entropy* [12]:

$$\mathrm{H}[Y_{0:\ell}] \equiv -\sum_{w \in \{0,1\}^\ell} \Pr(Y_{0:\ell} = w) \log_2 \Pr(Y_{0:\ell} = w). \tag{2.3}$$

Due to correlations in typical process sequences, the irreducible randomness per symbol is not the *single-symbol entropy* $\mathrm{H}[Y_0]$. Rather, it is given by the *Shannon entropy rate*

[12]:

$$h_\mu \equiv \lim_{\ell \to \infty} \frac{\mathrm{H}[Y_{0:\ell}]}{\ell} \ . \tag{2.4}$$

When applied to a physical system described by a suitable symbolic dynamics, as done here, this quantity is the *Kolmogorov-Sinai dynamical entropy* of the underlying physical generator of the process.

Note that these ways of monitoring information are quantitatively quite different. For large ℓ, $h_\mu \ell \ll \mathrm{H}[Y_{0:\ell}]$ and, in particular, anticipating later use, $h_\mu \leq \mathrm{H}[Y_0]$, typically much less. Equality between the single-symbol entropy and entropy rate is only achieved when the generating process is memoryless. Calculating the single-symbol entropy is typically quite easy, while calculating h_μ for general processes has been known for quite some time to be difficult [71] and it remains a technical challenge [72]. The entropy rates of the output sequence and input sequence are $h'_\mu = \lim_{\ell \to \infty} \mathrm{H}[Y'_{0:\ell}]/\ell$ and $h_\mu = \lim_{\ell \to \infty} \mathrm{H}[Y_{0:\ell}]/\ell$, respectively.

The informational properties of the input and output word distributions set bounds on energy flows in the system. Appendix 2.7.1 establishes one of our main results: The average work done by the ratchet is bounded above by the difference in Kolmogorov-Sinai entropy of the input and output processes [1]:

$$\begin{aligned} \langle W \rangle &\leq k_\mathrm{B} T \ln 2 \, (h'_\mu - h_\mu) \\ &= k_\mathrm{B} T \ln 2 \, \Delta h_\mu \ . \end{aligned} \tag{2.5}$$

We refer to this as the Information Processing Second Law (IPSL), because it bounds the transformation of structured information. In light of the preceding remarks on the basic difference between $\mathrm{H}[Y_0]$ and h_μ, we can now consider more directly the differences between Eqs. (3.10) and (2.5). Most importantly, the ΔH in the former refers to the instantaneous configurational entropy H before and after a thermodynamic transformation. In the ratchet's steady state operation, ΔH vanishes since the configuration distribution

[1] Reference [7]'s appendix suggests Eq. (2.5) without any detailed proof. An integrated version appeared also in Ref. [10] for the special case of memoryless demons. Our App. 2.7.1 gives a more general proof of Eq. (2.5) that, in addition, accounts for memory.

is time invariant, even when the overall system's information production is positive. The entropies h'_μ and h_μ in Eq. (2.5), in contrast, are dynamical: rates of active information generation in the input and output giving, in addition, the correct minimum rates since they take all temporal correlations into account. Together they bound the overall system's information production in steady state away from zero. In short, though often conflated, configurational entropy and dynamical entropy capture two very different kinds of information and they, per force, are associated with different physical properties supporting different kinds of information processing. They are comparable only in special cases.

For example, if one puts aside this basic difference to facilitate comparison and considers the Shannon entropy change $\Delta\mathrm{H}$ in the joint state space of all bits, the two equations are analogous in the current setup. However, often enough, a weaker version of Eq. (3.10) is considered in the discussions on Maxwell's Demon [7, 60, 61, 58, 10] and information reservoirs [36], wherein the statistical correlations between the bits are neglected, and one simply interprets $\Delta\,\mathrm{H}$ to be the change in the marginal Shannon entropies $\mathrm{H}[Y_0]$ of the individual bits. This implies the following relation in the current context:

$$\langle W \rangle \leq k_\mathrm{B} \ln 2 \, \Delta \mathrm{H}[Y_0] \ , \tag{2.6}$$

where $\Delta\,\mathrm{H}[Y_0] = \mathrm{H}[Y'_0] - \mathrm{H}[Y_0]$. While Eq. (2.6) is valid for the studies in Refs. [7, 60, 61, 58, 36, 10], it can be violated under certain scenarios [73]. In comparison, Eq. (2.5) is generally valid.

As an example, consider the case where the ratchet has memory and, for simplicity of exposition, is driven by an uncorrelated input process, meaning the input process entropy rate is the same as the single-symbol entropy: $h_\mu = \mathrm{H}[Y_0]$. However, the ratchet's memory can create correlations in the output bit string, so:

$$\begin{aligned} \Delta h_\mu &= h'_\mu - \mathrm{H}[Y_0] \\ &\leq \mathrm{H}[Y'_0] - \mathrm{H}[Y_0] \\ &= \Delta \mathrm{H}[Y_0] \ . \end{aligned} \tag{2.7}$$

In this case, Eq. (2.5) is a tighter bound on the work done by the ratchet—a bound that explicitly accounts for correlations within the output bit string generated by the

ratchet during its operation. For example, for the particular ratchet we consider with the parameter combination $\{p = 0.5, q = 0.1, b = 0.9\}$, the block entropies $H[Y'_{0:L}]$ of the output process do not converge to the entropy rate even when looking at block lengths up to $L = 13$. This means that there are correlations within the output that are not captured even when looking at long blocks of symbols, resulting in an over-estimate of randomness. In short, generally the entropy rate is necessary in order to properly account for the effects of all correlations in the output [12].

Previously, the effect of these correlations has not been calculated, but they have important consequences. Due to correlations, it is possible to have an increase in the single-symbol entropy difference $\Delta H[Y_0]$ but a decrease in the Kolmogorov-Sinai entropy rate Δh_μ. In this situation, it is erroneous to assume that there is an increase in the information content in the bits. There is, in fact, a decrease in information due to correlations; cf. Sec. 2.5. As a result, we get an unexpected eraser regime in the phase diagram of the system (Fig. 2.7). A similar regime may be present also in the model of Ref. [7] where the outgoing bits were observed to have small but finite correlations.

Note that a somewhat different situation was considered in Ref. [10], a memoryless channel (ratchet) driven by a correlated process. In this special case—ratchets unable to leverage or create temporal correlations—either Eq. (2.6) or Eq. (2.5) can be a tighter quantitative bound on work. When a memoryless ratchet is driven by uncorrelated input, though, the bounds are equivalent. Critically, for memoryful ratchets driven by correlated input Eq. (2.6) can be violated. In all settings, Eq. (2.5) holds.

While we defer it's development to a later chapter, Eq. (2.5) also has implications for ratchet functioning when the input bits are correlated as well. Specifically, correlations in the input bits can be leveraged by the ratchet to do additional work—work that cannot be accounted for if one only considers single-symbol configurational entropy of the input bits [74].

2.3 Energetics and Dynamics

To predict how the ratchet interacts with the bit string and weight, we need to specify the string and ratchet energies. While we developed general tools for energetic analysis in later chapters with equivalent asymptotic results, we begin our exploration of ratchets with a detailed description of the ratchet's energetic interaction with the input bit. When not interacting with the ratchet the energies, E_0 and E_1, of both bit states, $Y = 0$ and $Y = 1$, are taken to be zero for symmetry and simplicity: $E_0 = E_1 = 0$. For simplicity, too, we say the ratchet mechanism has just two internal states A and B. When the ratchet is not interacting with bits, the two states can have different energies. We take $E_A = 0$ and $E_B = -\alpha k_{\mathrm{B}} T$, without loss of generality. Since the bits interact with the ratchet one at a time, we only need to specify the interaction energy of the ratchet and an individual bit. The interaction energy is zero if the bit is in the state $Y = 0$, regardless of the ratchet state, and it is $-\beta k_{\mathrm{B}} T$ (or $+\beta k_{\mathrm{B}} T$) if the bit is in state $Y = 1$ and the ratchet is in state A (or B). See Fig. 2.2 for a graphical depiction of the energy scheme under "Ratchet $\otimes$ Bit".

The scheme is further modified by the interaction of the weight with the ratchet and bit string. We attach the weight to the ratchet-bit system such that when the latter transitions from the $B \otimes 0$ state to the $A \otimes 1$ state it lifts the weight, doing a constant amount $w k_{\mathrm{B}} T$ of work. As a result, the energy of the composite system—Demon, interacting bit, and weight—increases by $w k_{\mathrm{B}} T$ whenever the transition $B \otimes 0 \to A \otimes 1$ takes place, the required energy being extracted from the heat reservoir Z_N. The rightmost part of Fig. 2.2 indicates this by raising the energy level of $A \otimes 1$ by $w k_{\mathrm{B}} T$ compared to its previous value. Since the transitions between $A \otimes 1$ and $B \otimes 1$ do not involve the weight, their relative energy difference remains unaffected. An increase in the energy of $A \otimes 1$ by $w k_{\mathrm{B}} T$ therefore implies the same increase in the energy of $B \otimes 1$. Again, see Fig. 2.2 for the energy scheme under "Ratchet $\otimes$ Bit $\otimes$ Weight".

The time evolution over the joint state space of the ratchet, last bit of the input string, and weight is governed by a Markov dynamic, specified by state-transition matrix M. If, at the beginning of the Nth interaction interval at time $t = \tau(N-1) + 0^+$, the ratchet

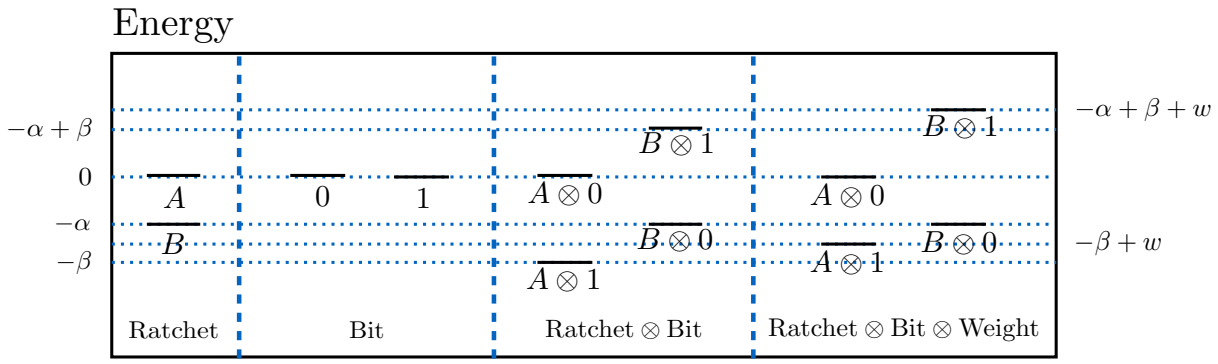


Figure 2.2. Energy levels of the Demon states, interacting bits, their joint system, and their joint system with a weight in units of $[k_{\mathrm{B}}T]$.

is in state $X_N = x_N$ and the input bit is in state $Y_N = y_N$, then let $M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N}$ be the probability $\Pr(X_{N+1} = x_{N+1}, Y'_N = y'_N | X_N = x_N, Y_N = y_N)$ that the ratchet is in state $X_N = x_{N+1}$ and the bit is in state $Y_N = y'_N$ at the end of the interaction interval $t = \tau(N-1)+\tau^-$. X_N and Y_N at the end of the Nth interaction interval become X_{N+1} and Y'_N respectively at the beginning of the $N+1$th interaction interval. Since we assume the system is thermalized with a bath at temperature T, the ratchet dynamics obey detailed balance. And so, transition rates are governed by the energy differences between joint states:

$$\frac{M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N}}{M_{x_{N+1} \otimes y'_N \to x_N \otimes y_N}} = e^{(E_{x_{N+1} \otimes y'_N} - E_{x_N \otimes y_N})/k_{\mathrm{B}}T} \quad . \tag{2.8}$$

There is substantial flexibility in constructing a detailed-balanced Markov dynamic for the ratchet, interaction bit, and weight. Consistent with our theme of simplicity, we choose one that has only six allowed transitions: $A \otimes 0 \leftrightarrow B \otimes 0$, $A \otimes 1 \leftrightarrow B \otimes 1$, and $A \otimes 1 \leftrightarrow B \otimes 0$. Such a model is convenient to consider, since it can be described by just two transition probabilities $0 \leq p \leq 1$ and $0 \leq q \leq 1$, as shown in Fig. 2.3.

The Markov transition matrix for this system is given by:

$$M = \begin{bmatrix} 0 & 1-p & 0 & 0 \\ 1 & 0 & q & 0 \\ 0 & p & 0 & 1 \\ 0 & 0 & 1-q & 0 \end{bmatrix} . \tag{2.9}$$

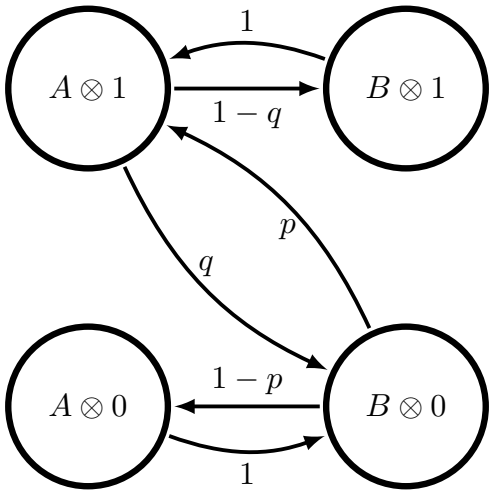


Figure 2.3. The Markovian, detailed-balance dynamic over the joint states of the ratchet and interacting bit.

This allows allows us to calculate the state distribution $\mathbf{p}((N-1)\tau + \tau^-)$ at the end of the Nth interaction interval from the state distribution $\mathbf{p}((N-1)\tau + 0^+)$ at the interval's beginning via:

$$\mathbf{p}((N-1)\tau + \tau^-) = M\mathbf{p}((N-1)\tau + 0^+) \ , \tag{2.10}$$

where the probability vector is indexed $\mathbf{p} = (\Pr(A \otimes 0), \Pr(B \otimes 0), \Pr(A \otimes 1), \Pr(B \otimes 1))^\top$. To satisfy detailed balance, we find that α, β, and w should be:

$$\alpha = -\ln(1-p) \ , \tag{2.11}$$

$$\beta = -\frac{1}{2}\ln\left[(1-p)(1-q)\right] \ , \text{ and} \tag{2.12}$$

$$w = \ln\left(\frac{q\sqrt{1-p}}{p\sqrt{1-q}}\right) \ . \tag{2.13}$$

(Appendix 2.7.2 details the relationships between the transitions probabilities and energy levels.)

This simple model is particularly useful since, as we show shortly, it captures the full range of thermodynamic functionality familiar from previous models and, more importantly, it makes it possible to exactly calculate informational properties of the output string analytically.

Now that we know how the ratchet interacts with the bit string and weight, we need to characterize the input string to predict the energy flow through the ratchet. As in

the ratchet models of Refs. [7, 63], we consider an input generated by a biased coin—$\Pr(Y_N = 0) = b$ at each N—which has no correlations between successive bits. For this input, the steady state distributions at the beginning and end of the interaction interval τ are:

$$\mathbf{p}^{\mathrm{s}}(0^+) = \frac{1}{2}\begin{bmatrix} b \\ b \\ 1-b \\ 1-b \end{bmatrix} \text{ and}$$

$$\mathbf{p}^{\mathrm{s}}(\tau^-) = \frac{1}{2}\begin{bmatrix} b(1-p) \\ b+q-bq \\ bp+1-b \\ (1-b)(1-q) \end{bmatrix} . \tag{2.14}$$

These distributions are needed to calculate the work done by the ratchet.

To calculate net extracted work by the ratchet we need to consider three work-exchange steps for each interaction interval: (1) when the ratchet gets attached to a new bit, to account for their interaction energy; (2) when the joint transitions $B \otimes 0 \leftrightarrow A \otimes 1$ take place, to account for the raising or lowering of the weight; and (3) when the ratchet detaches itself from the old bit, again, to account for their nonzero interaction energy. We refer to these incremental works as W_1, W_2, and W_3, respectively.

Consider the work W_1. If the new bit is in state 0, from Fig. 2.2 we see that there is no change in the energy of the joint system of the ratchet and the bit. However, if the new bit is 1 and the initial state of the ratchet is A, energy of the ratchet-bit joint system decreases from 0 to $-\beta$. The corresponding energy is gained as work by the mechanism that makes the ratchet move past the tape of bits. Similarly, if the new bit is 1 and the initial state of the ratchet is B, there is an increase in the joint state energy by β; this amount of energy is now taken away from the driving mechanism of the ratchet. In the steady state, the average work gain $\langle W_1 \rangle$ is then obtained from the average decrease in

energy of the joint (ratchet-bit) system:

$$\begin{aligned}\langle W_1\rangle &= -\sum_{\substack{x\in\{A,B\}\\ y\in\{0,1\}}} p^{\rm s}_{x\otimes y}(0^+)\,(E_{x\otimes y}-E_x-E_y)\\ &= 0\ ,\end{aligned} \tag{2.15}$$

where we used the probabilities in Eq. (2.14) and Fig. 2.2's energies.

By a similar argument, the average work $\langle W_3\rangle$ is equal to the average decrease in the energy of the joint system on the departure of the ratchet, given by:

$$\langle W_3\rangle = -\frac{k_{\rm B}T}{2}\beta[q+b(p-q)]\ . \tag{2.16}$$

Note that the cost of moving the Demon on the bit string (or moving the string past a stationary Demon) is accounted for in works W_1 and W_3.

Work W_2 is associated with raising and lowering of the weight depicted in Fig. 2.1. Since transitions $B\otimes 0 \to A\otimes 1$ raise the weight to give work $k_{\rm B}Tw$ and reverse transitions $B\otimes 0 \leftarrow A\otimes 1$ lower the weight consuming equal amount of work, the average work gain $\langle W_2\rangle$ must be $k_{\rm B}Tw$ times the net probability transition along the former direction, which is $[T_{B\otimes 0\to A\otimes 1}p^{\rm s}_{B\otimes 0}(0^+) - T_{A\otimes 1\to A\otimes 1}p^{\rm s}_{A\otimes 1}(0^+)]$. This leads to the following expression:

$$\langle W_2\rangle = \frac{k_{\rm B}Tw}{2}[-q+b(p+q)]\ , \tag{2.17}$$

where we used the probabilities in Eq. (2.14).

The total work supplied by the ratchet and a bit is their sum:

$$\begin{aligned}\langle W\rangle &= \langle W_1\rangle + \langle W_2\rangle + \langle W_3\rangle \\ &= \frac{k_{\rm B}T}{2}[(pb-q+qb)\ln\left(\frac{q}{p}\right)\\ &\qquad + (1-b)q\ln(1-q) + pb\ln(1-p)]\ .\end{aligned} \tag{2.18}$$

Note that we considered the total amount amount of work that can be gained by the system, not just that obtained by raising the weight. Why? As we shall see in Sec. 2.5, the former is the thermodynamically more relevant quantity. A similar energetic scheme that incorporates the effects of interaction has also been discussed in Ref. [64].

In this way, we exactly calculated the work term in Eq. (2.5). We still need to calculate the entropy rate of the output and input strings to validate the proposed Information Processing Second Law. For this, we introduce an information-theoretic formalism to monitor processing of the bit strings by the ratchet.

2.4 Information

To analytically calculate the input and output entropy rates, we consider how the strings are generated. A natural way to incorporate temporal correlations in the input string is to model its generator by a finite-state hidden Markov model (HMM), since HMMs are strictly more powerful than Markov chains in the sense that finite-state HMMs can generate all processes produced by Markov chains, but the reverse is not true. For example, there are processes generated by finite HMMs that cannot be by any finite-state Markov chain. In short, HMMs give a compact representations for a wider range of memoryful processes which are generated by physical systems with states that are hidden to the observer.

Consider possible input strings to the ratchet. With or without correlations between bits, they can be described by an HMM generator with a finite set of, say, K states and a set of two symbol-labeled transition matrices $T^{(0)}$ and $T^{(1)}$, where:

$$T^{(y_N)}_{s_N \to s_{N+1}} = \Pr(Y_N = y_N, S_{N+1} = s_{N+1} | S_N = s_N) \tag{2.19}$$

is the probability of outputting y_N for the Nth bit of the input string and transitioning to the hidden internal state s_{N+1} given that the HMM was in state s_N.

When it comes to the output string, in contrast, we have no choice. We are forced to use HMMs. Since the current input bit state Y_N and ratchet state X_N are not explicitly captured in the current output bit state Y'_N, Y_N and X_N are hidden variables. As we noted before, calculating HMM entropy rates is a known challenging problem [71, 72]. Much of the difficulty stems from the fact that in HMM-generated processes the effects of internal states are only indirectly observed and, even then, appear only over long output sequences.

We can circumvent this difficulty by using *unifilar* HMMs, in which the current state

and generated symbol uniquely determine the next state. This is a key technical contribution here since for unifilar HMMs the entropy rate is exactly calculable, as we now explain. Unifilar HMMs internal states are a causal partitioning of the past, meaning that every past w maps to a particular state through some function f and so:

$$\Pr(Y_N = y_N | Y_{0:N} = w) = \Pr(Y_N = y_N | S_N = f(w)) \ . \tag{2.20}$$

As a consequence, the entropy rate h_μ in its block-entropy form (Eq. (2.4)) can be re-expressed in terms of the transition matrices. First, recall the alternative, equivalent form for entropy rate: $h_\mu = \lim_{N\to\infty} \mathrm{H}[Y_N|Y_{0:N}]$. Second, since S_N captures all the dependence of Y_N on the past, $h_\mu = \lim_{N\to\infty} \mathrm{H}[Y_N|S_N]$. This finally leads to a closed-form for the entropy rate [12]:

$$\begin{aligned} h_\mu &= \lim_{N\to\infty} \mathrm{H}[Y_N|S_N] \\ &= -\sum_{y_N, s_N, s_{N+1}} \pi_{s_N} T^{(y_N)}_{s_N \to s_{N+1}} \log_2 T^{(y_N)}_{s_N \to s_{N+1}} \ , \end{aligned} \tag{2.21}$$

where π is the stationary distribution over the unifilar HMM's states.

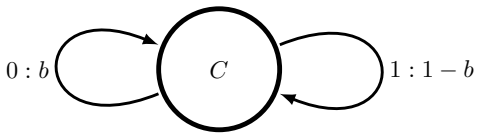


Figure 2.4. Biased coin input string as a unifilar hidden Markov model with bias $\Pr(Y = 0) = b$.

Let's now put these observations to work. Here, we assume the ratchet's input string was generated by a memoryless biased coin. Figure 2.4 shows its (minimal-size) unifilar HMM. The single internal state C implies that the process is memoryless and the bits are uncorrelated. The HMM's symbol-labeled (1×1) transition matrices are $T^{(0)} = [b]$ and $T^{(1)} = [1 - b]$. The transition from state C to itself labeled $0 : b$ means that if the system is in state C, then it transitions to state C and outputs $Y = 0$ with probability b. Since this model is unifilar, we can calculate the input-string entropy rate from Eq. (2.21) and

see that it is the single-symbol entropy of bias b:

$$\begin{aligned} h_\mu &= \mathrm{H}(b) \\ &\equiv -b \log_2 b - (1-b) \log_2 (1-b) \ , \end{aligned} \tag{2.22}$$

where $\mathrm{H}(b)$ is the (base 2) binary entropy function [11].

The more challenging part of our overall analysis is to determine the entropy rate of the output string. Even if the input is uncorrelated, it's possible that the ratchet creates temporal correlations in the output string. (Indeed, these correlations reflect the ratchet's operation and so its thermodynamic behavior, as we shall see below.) To calculate the effect of these correlations, we need a generating unifilar HMM for the output process—a process produced by the ratchet being driven by the input.

When discussing the ratchet energetics, there was a Markov dynamic M over the ratchet-bit joint state space. Here, it is now controlled by bits from the input string and writes the result of the thermal interaction with the ratchet to the output string. In this way, M becomes an input-output machine or *transducer* [1]. In fact, this transducer is a communication channel in the sense of Shannon [75] that communicates the input bit sequence to the output bit sequence. However, it is a channel with memory. Its internal states correspond to the ratchet's states. To work with M, we rewrite it componentwise as:

$$M^{(y'_N|y_N)}_{x_N \to x_{N+1}} = M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} \tag{2.23}$$

to evoke its re-tooled operation. The probability of generating bit y'_N and transitioning to ratchet state x_{N+1}, given that the input bit is y_N and the ratchet is in state x_N, is:

$$\begin{aligned} M^{(y'_N|y_N)}_{x_N \to x_{N+1}} = & \\ & \Pr(Y'_N = y'_N, X_{N+1} = x_{N+1} | Y_N = y_N, X_N = x_N) \ . \end{aligned} \tag{2.24}$$

This allows us to exactly calculate the symbol-labeled transition matrices, $T'^{(0)}$ and $T'^{(1)}$, of the HMM that generates the output string:

$$T'^{(y'_N)}_{s_N \otimes x_N \to s_{N+1} \otimes x_{N+1}} = \sum_{y_N} M^{(y'_N|y_N)}_{x_N \to x_{N+1}} T^{(y_N)}_{s_N \to s_{N+1}} \ . \tag{2.25}$$

The joint states of the ratchet and the internal states of the input process are the internal states of the output HMM, with $x_N, x_{N+1} \in \{A, B\}$ and $s_N, s_{N+1} \in \{C\}$ in the present case. This approach is a powerful tool for directly analyzing informational properties of the output process.

By adopting the transducer perspective, it is possible to find HMMs for the output processes of previous ratchet models, such as in Refs. [7, 63]. However, their generating HMMs are highly nonunifilar, meaning that knowing the current internal state and output allows for many alternative internal-state paths. And, this precludes writing down closed-form expressions for informational quantities, as we do here. Said simply, the essential problem is that those models build in too many transitions. Ameliorating this constraint led to the Markov dynamic shown in Fig. 2.3 with two ratchet states and sparse transitions. Although this ratchet's behavior cannot be produced by a rate equation, due to the limited transitions, it respects detailed balance.

Figure 2.5 shows our two-state ratchet's transducer. As noted above, it's internal states are the ratchet states. Each transition is labeled $y'|y : p$, where y' is the output, conditioned on an input y, with probability p.

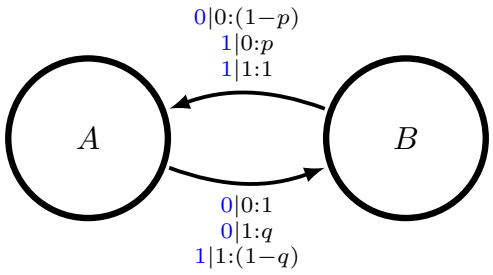


Figure 2.5. The Maxwellian ratchet's transducer.

We can drive this ratchet (transducer) with any input, but for comparison with previous work, we drive it with the memoryless biased coin process just introduced and shown in Fig. 2.4. The resulting unifilar HMM for the output string is shown in Fig. 2.6. The

corresponding symbol-labeled transition matrices are:

$$T'^{(0)} = \begin{bmatrix} 0 & (1-p)b \\ b + q(1-b) & 0 \end{bmatrix} \text{ , and} \tag{2.26}$$

$$T'^{(1)} = \begin{bmatrix} 0 & 1-(1-p)b \\ (1-q)(1-b) & 0 \end{bmatrix} . \tag{2.27}$$

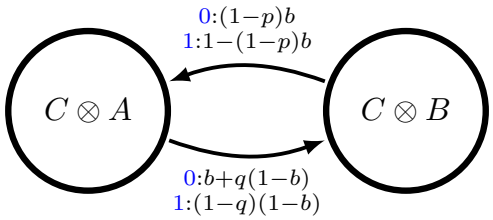


Figure 2.6. Unifilar HMM for the output string generated by the ratchet driven by a coin with bias b.

Using these we can complete our validation of the proposed Second Law, by exactly calculating the entropy rate of the output string. We find:

$$\begin{aligned} h'_\mu &= \lim_{N\to\infty} \mathrm{H}[Y'_N | Y'_{0:N}] \\ &= \lim_{N\to\infty} \mathrm{H}[Y'_N | S_N] \\ &= \frac{\mathrm{H}(b(1-p))}{2} + \frac{\mathrm{H}((1-b)(1-q))}{2} . \end{aligned} \tag{2.28}$$

We note that this is less than or equal to the (unconditioned) single-symbol entropy for the output process:

$$\begin{aligned} h'_\mu &\leq \mathrm{H}[Y'_0] \\ &= \mathrm{H}\left((b(1-p) + (1-b)(1-q))/2\right) . \end{aligned} \tag{2.29}$$

Any difference between h'_μ and single-symbol entropy $\mathrm{H}[Y_0]$ indicates correlations that the ratchet created in the output from the uncorrelated input string. In short, the entropy rate gives a more accurate picture of how information is flowing between bit strings and the heat bath. And, as we now demonstrate, the entropy rate leads to correctly identifying important classes of ratchet thermodynamic functioning—functionality the single-symbol entropy misses.

2.5 Thermodynamic Functionality

Let's step back to review and set context for exploring the ratchet's thermodynamic functionality as we vary its parameters. Our main results are analytical, provided in closed-form. First, we derived a modified version of the Second Law of Thermodynamics for information ratchets in terms of the difference between the Kolmogorov-Sinai entropy of the input and output strings:

$$\langle W \rangle \leq k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu} \ , \tag{2.30}$$

where $\Delta h_{\mu} = h'_{\mu} - h_{\mu}$. The improvement here takes into account correlations within the input string and those in the output string actively generated by the ratchet during its operation. From basic information-theoretic identities we know this bound is stricter for memoryless inputs than previous relations [76] that ignored correlations. However, by how much? And, this brings us to our second main result. We gave analytic expressions for both the input and output entropy rates and the work done by the Demon. Now, we are ready to test that the bound is satisfied and to see how much stricter it is than earlier approximations.

We find diverse thermodynamic behaviors as shown in Figure 2.7, which describes ratchet thermodynamic function at input bias $b = 0.9$. We note that there are analogous behaviors for all values of input bias. We identified three possible behaviors for the ratchet: *Engine*, *Dud*, and *Eraser*. Nowhere does the ratchet violate the rule $\langle W \rangle \leq k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu}$. The engine regime is defined by (p, q) for which $k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu} \geq \langle W \rangle > 0$ since work is positive. This is the only condition for which the ratchet extracts work. The eraser regime is defined by $0 > k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu} \geq \langle W \rangle$, meaning that work is extracted from the work reservoir while the uncertainty in the bit string decreases. In the dud regime, those (p, q) for which $k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu} \geq 0 \geq \langle W \rangle$, the ratchet is neither able to erase information nor is it able to do useful work.

At first blush, these are the same behavior types reported by Ref. [7], except that we have stronger bounds on the work now with $k_{\mathrm{B}} T \ln 2 \, \Delta h_{\mu}$, compared to the single-symbol entropy approximation. The stricter bound gives deeper insight into ratchet functionality. To give a concrete comparison, Fig. 2.8 plots the single-symbol entropy difference $\Delta \, \mathrm{H}[Y_0]$

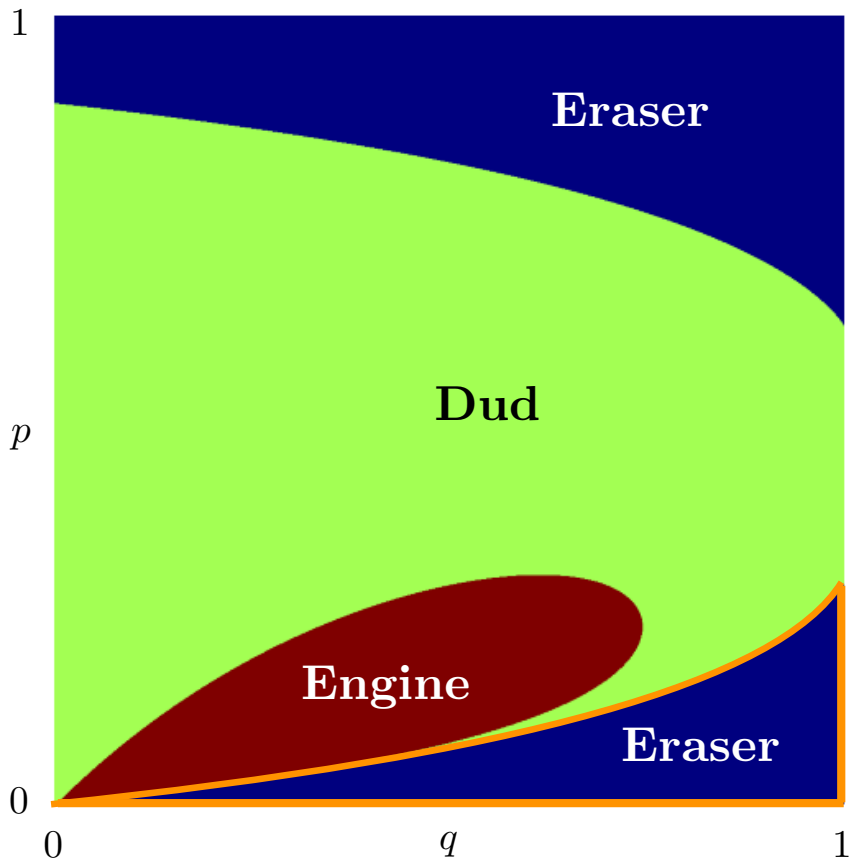


Figure 2.7. Information ratchet thermodynamic functionality at input bias $b = 0.9$: Engine: (p, q) such that $0 < \langle W \rangle \leq k_{\mathrm{B}} T \ln 2 \, \Delta h_\mu$. Eraser: (p, q) such that $\langle W \rangle \leq k_{\mathrm{B}} T \ln 2 \, \Delta h_\mu < 0$. Dud: (p, q) such that $\langle W \rangle \leq 0 \leq k_{\mathrm{B}} T \ln 2 \, \Delta h_\mu$.

and the entropy rate difference Δh_μ, with a flat surface identifying zero entropy change, for all p and q and at $b = 0.9$.

In the present setting where input symbols are uncorrelated, the blue $\Delta \mathrm{H}[Y_0]$ surface lies above the red Δh_μ surface for all parameters, confirming that the single-symbol entropy difference is always greater than the entropy rate difference. It should also be noted for this choice of input bias b and for larger p, $\Delta \mathrm{H}[Y_0]$ and Δh_μ are close, but they diverge for smaller p. They diverge so much, however, that looking only at single-symbol entropy approximation misses an entire low-p region, highlighted in orange in Fig. 2.8 and 2.7, where Δh_μ dips below zero and the ratchet functions as eraser.

The orange-outlined low-p erasure region is particularly interesting, as it hosts a new functionality not previously identified: The ratchet removes multiple-bit uncertainty, effectively erasing incoming bits by adding temporal order, all the while increasing the

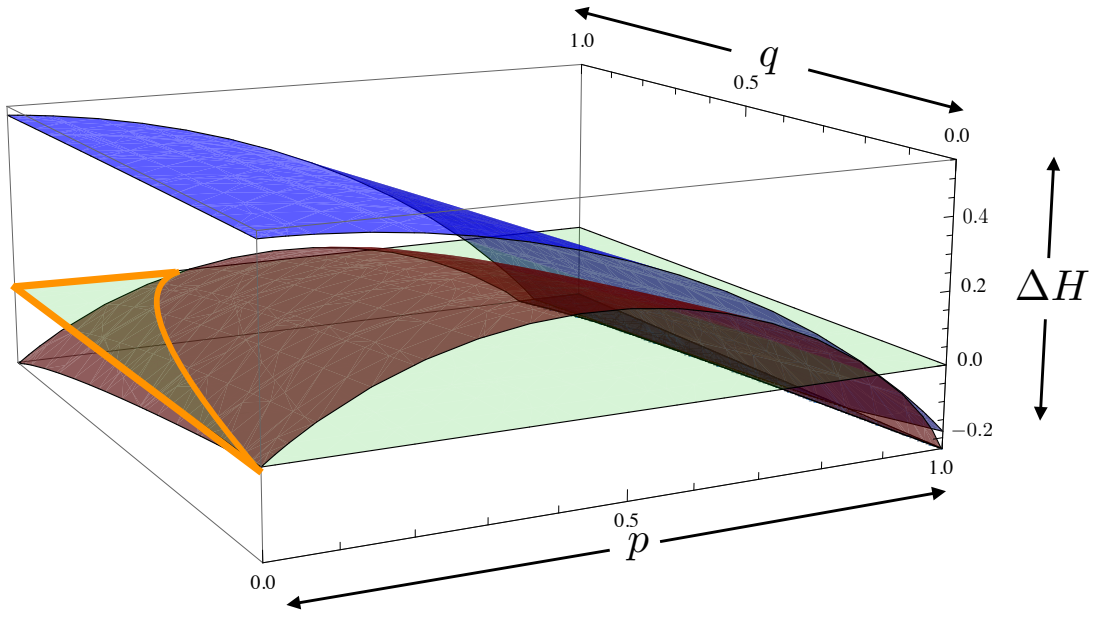


Figure 2.8. Exact entropy rate difference Δh_μ (red) is a much stricter bound on work than the difference in single-symbol entropy $\Delta\,\mathrm{H}[Y_0]$ (blue). The zero surface (light green) highlights where both entropies are greater than zero and so is an aid to identifying functionalities.

uncertainty in individual incoming bits. The existence of this mode of erasure is highly counterintuitive in light of the fact the Demon interacts with only one bit at a time. In contrast, operation in the erasure region at high p, like that in previous Demons, simply reduces single-bit uncertainty. Moreover, the low-p erasure region lies very close to the region where ratchet functions as an engine, as shown in Fig. 2.7. As one approaches $(p, q) = (0, 0)$ the eraser and engine regions become arbitrarily close in parameter space. This is a functionally meaningful region, since the device can be easily and efficiently switched between distinct modalities—an eraser or an engine.

In contrast, without knowing the exact entropy rate, it appears that the engine region of the ratchet's parameter space is isolated from the eraser region by a large dud region and that the ratchet is not tunable. Thus, knowing the correlations between bits in the output string allows one to predict additional functionality that otherwise is obscured when one only considers the single-symbol entropy of the output string.

As alluded to above, we can also consider structured input strings generated by memoryful processes, unlike the memoryless biased coin. While correlations in the output

string are relevant to the energetic behavior of this ratchet, it turns out that input string correlations are not. The work done by the ratchet depends only on the input's single-symbol bias b. That said, in the next chapter we will explore more intelligent ratchets that take advantage of input string correlations to do additional work.

2.6 Conclusion

Thermodynamic systems that include information reservoirs as well as thermal and work reservoirs are an area of growing interest, driven in many cases by biomolecular chemistry or nanoscale physics and engineering. With the ability to manipulate thermal systems on energy scales closer and closer to the level of thermal fluctuations $k_{\mathrm{B}}T$, information becomes critical to the flow of energy. Our model of a ratchet and a bit string as the information reservoir is very flexible and our methods showed how to analyze a broad class of such controlled thermodynamic systems. Central to identifying thermodynamic functionality was our deriving Eq. (2.5), based on the control system's Kolmogorov-Sinai entropy, that holds in all situations of memoryful or memoryless ratchets and correlated or uncorrelated input processes and that provides the tightest quantitative bound on work for memoryless inputs. This improvement comes directly from tracking Demon information production over system trajectories, not from time-local, configurational entropies.

Though its perspective and methods were not explicitly highlighted, *computational mechanics* [77] played a critical role in the foregoing analyses, from its focus on structure and calculating all system component correlations to the technical emphasis on unifilarity in Demon models. Its full impact was not fully explained here, but is further explored in later chapters. Two complementary computational mechanics analyses of information engines come to mind, in this light. The first is Ref. [78]'s demonstration that the chaotic instability in Szilard's Engine, reconceived as a deterministic dynamical system, is key to its ability to extract heat from a reservoir. This, too, highlights the role of Kolmogorov-Sinai dynamical entropy. Another is the thorough-going extension of fluctuation relations to show how intelligent agents can harvest energy when synchronizing to the fluctuations from a structured environment [74].

This is to say, in effect, the foregoing showed that computational mechanics is a natural framework for analyzing a ratchet interacting with an information reservoir to extract work from a thermal bath. The input and output strings that compose the information reservoir are best described by unifilar HMM generators, since they allow for exact calculation of any informational property of the strings, most importantly the entropy rate. In fact, the control system components are the ϵ-machines and ϵ-transducers of computational mechanics [77, 1].

By allowing one to exactly calculate the asymptotic entropy rate, we identified more functionality in the effective thermodynamic ϵ-transducers than previous methods can reveal. Two immediate consequences were that we identified a new kind of thermodynamic eraser and found that our ratchet is easily tunable between an eraser and an engine—functionalities suggesting that real-world ratchets exhibit memory to take advantage of correlated environmental fluctuations, as well as hinting at useful future engineering applications.

2.7 Appendices

2.7.1 Derivation of Eq. (2.5)

Here, we reframe the Second Law of Thermodynamics, deriving an expression of it that makes only one assumption about the information ratchet operating along the bit string: the ratchet accesses only a finite number of internal states. This constraint is rather mild and, thus, the bounds on thermodynamic functioning derived from the new Second Law apply quite broadly.

The original Second Law of Thermodynamics states that the total change in entropy of an isolated system must be nonnegative over any time interval. By considering a system composed of a thermal reservoir, information reservoir, and ratchet, in the following we derive an analog in terms of rates, rather than total configurational entropy changes.

Due to the Second Law, we insist that the change in thermodynamic entropy of the closed system is positive for any number N of time steps. If X denotes the ratchet, Y the

bit string, and Z the heat bath, this assumption translates to:

$$\triangle S[X, Y, Z] \geq 0 \ . \tag{2.31}$$

Note that we do not include a term for the weight (a mechanical energy reservoir), since it does not contribute to the thermodynamic entropy. Expressing the thermodynamic entropy S in terms the Shannon entropy of the random variables $S[X, Y, Z] = k_{\mathrm{B}} \ln 2 \, \mathrm{H}[X, Y, Z]$, we have the condition:

$$\triangle \mathrm{H}[X, Y, Z] \geq 0 \ . \tag{2.32}$$

To be more precise, this is true over any number of time steps N. If we have our system X, we denote the random variable for its state at time step N by X_N. The information reservoir Y is a semi-infinite string. At time zero, the string is composed entirely of the bits of the input process, for which the random variable is denoted $Y_{0:\infty}$. The ratchet transduces these inputs, starting with Y_0 and generating the output bit string, the entirety of which is expressed by the random variable $Y'_{0:\infty}$. At the Nth time step, the first N bits of the input Y have been converted into the first N bits of the output Y', so the random variable for the input-output bit string is $Y_{N:\infty} \otimes Y'_{0:N}$. Thus, the change in entropy from the initial time to the Nth time step is:

$$\begin{aligned} \triangle \, \mathrm{H}_N[X, Y, Z] &= \mathrm{H}[X_N, Y_{N:\infty}, Y'_{0:N}, Z_N] \\ &\quad - \mathrm{H}[X_0, Y_{0:\infty}, Z_0] \qquad (2.33) \\ &= \mathrm{H}[X_N, Y_{N:\infty}, Y'_{0:N}] + \mathrm{H}[Z_N] \\ &\quad - \mathrm{I}[X_N, Y_{N:\infty}, Y'_{0:N}; Z_N] \\ &\quad - \mathrm{H}[X_0, Y_{0:\infty}] - \mathrm{H}[Z_0] \\ &\quad + \mathrm{I}[X_0, Y_{0:\infty}; Z_0] \ . \qquad (2.34) \end{aligned}$$

Note that the internal states of an infinite heat bath do not correlate with the environment, since they have no memory of the environment. This means the mutual informations $\mathrm{I}[X_N, Y_{N:\infty}, Y'_{0:N}; Z_N]$ and $\mathrm{I}[X_0, Y_{0:\infty}; Z_0]$ of the thermal reservoir Z with the bit string Y and ratchet X vanish. Also, note that the change in thermal bath entropy can be

expressed in terms of the heat dissipated Q_N over the N time steps:

$$\begin{aligned} \triangle \mathrm{H}[Z] &= \mathrm{H}[Z_N] - \mathrm{H}[Z_0] \\ &= Q_N / k_\mathrm{B} T \ln 2 \ . \end{aligned} \tag{2.35}$$

Thus, the Second Law naturally separates into energetic terms describing the change in the heat bath and information terms describing the ratchet and bit strings:

$$\begin{aligned} \triangle \, \mathrm{H}_N[X, Y, Z] = \frac{Q_N}{k_B T \, \ln 2} \\ + \mathrm{H}[X_N, Y_{N:\infty}, Y'_{0:N}] - \mathrm{H}[X_0, Y_{0:\infty}] \ . \end{aligned} \tag{2.36}$$

Since $\triangle \, \mathrm{H} \geq 0$, we can rewrite this as an entirely general lower bound on the dissipated heat over a length $N\tau$ time interval, recalling that τ is the ratchet-bit interaction time:

$$Q_N \geq k_B T \, \ln 2 \left(\mathrm{H}[X_0, Y_{0:\infty}] - \mathrm{H}[X_N, Y_{N:\infty}, Y'_{0:N}] \right) \ . \tag{2.37}$$

This bound is superficially similar to Eq. (2.6), but it's true in all cases, as we have not yet made any assumptions about the ratchet. However, its informational quantities are difficult to calculate for large N and, in their current form, do not give much insight. Thus, we look at the infinite-time limit in order tease out hidden properties.

Over a time interval $N\tau$, the average heat dissipated per ratchet cycle is Q_N/N. When we classify an engine's operation, we usually quantify energy flows that neglect transient dynamics. These are just the heat dissipated per cycle over infinite time $\langle Q \rangle = \lim_{N \to \infty} Q_N/N$, which has the lower bound:

$$\langle Q \rangle \geq \lim_{N \to \infty} k_B T \, \ln 2 \frac{H[X_0, Y_{0:\infty}] - H[X_N, Y_{N:\infty}, Y'_{0:N}]}{N}. \tag{2.38}$$

Assuming the ratchet has a finite number of internal states, each with finite energy, then the bound can be simplified and written in terms of work. In this case, the average work produced is the opposite of the average dissipated heat: $\langle W \rangle = -\langle Q \rangle$. And so, it

has the upper bound:

$$\begin{aligned}\langle W \rangle \leq k_B T \, \ln 2 \lim_{N \to \infty} \Bigg(& \frac{H[Y_{N:\infty}, Y'_{0:N}] - H[Y_{0:\infty}]}{N} \\ & + \frac{H[X_N] - H[X_0]}{N} \\ & + \frac{I[X_0; Y_{0:\infty}] - I[X_N; Y_{N:\infty}, Y'_{0:N}]}{N} \Bigg),\end{aligned} \tag{2.39}$$

where the joint entropies are expanded in terms of their single-variable entropies and mutual informations.

The entropies over the initial X_0 and final X_N ratchet state distributions monitor the change in ratchet memory—time-dependent versions of its statistical complexity $C_\mu(N) = \mathrm{H}[X_N]$ [77]. This time dependence can be used to monitor how and when the ratchet synchronizes to the incoming sequence, recognizing a sequence's temporal correlations. However, since we assumed that the ratchet has finite states, the ratchet state-entropy and also mutual information terms involving it are bounded above by the logarithm of the number states. And so, they go to zero as $N \to \infty$, leaving the expression:

$$\langle W \rangle \leq k_B T \, \ln 2 \lim_{N \to \infty} \left(\frac{H[Y_{N:\infty}, Y'_{0:N}] - H[Y_{0:\infty}]}{N} \right) . \tag{2.40}$$

With this, we have a very general upper bound for the work done by the ratchet in terms of just the input and output string variables.

Once again, we split the joint entropy term into it's components:

$$\begin{aligned}\langle W \rangle \leq k_B T \, \ln 2 \lim_{N \to \infty} \Bigg(& \frac{H[Y_{N:\infty}] - H[Y_{0:\infty}]}{N} \\ & + \frac{H[Y'_{0:N}]}{N} - \frac{I[Y_{N:\infty}; Y'_{0:N}]}{N} \Bigg) .\end{aligned} \tag{2.41}$$

In this we identify the output process's entropy rate $h'_\mu = \lim_{N \to \infty} \mathrm{H}[Y'_{0:N}]/N$. While $\lim_{N \to \infty} \big(\mathrm{H}[Y_{N:\infty}] - \mathrm{H}[Y_{0:\infty}] \big)/N$ looks unfamiliar, it is actually the negative entropy rate h_μ of the input process, so we find that:

$$\langle W \rangle \leq k_B T \, \ln 2 \Bigg(h'_\mu - h_\mu - \lim_{N \to \infty} \frac{I[Y_{N:\infty}; Y'_{0:N}]}{N} \Bigg). \tag{2.42}$$

To understand the mutual information term, note that $Y'_{0:N}$ is generated from $Y_{0:N}$, so it is independent of $Y_{N:\infty}$ conditioned on $Y_{0:N}$. Essentially, $Y_{0:N}$ causally shields $Y'_{0:N}$

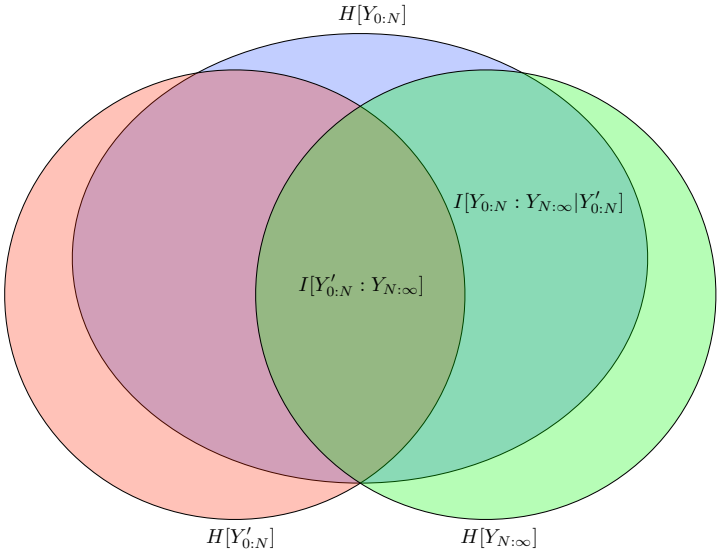


Figure 2.9. The N most recent variables of the input process shield the N variables of output from the rest of the input variables.

from $Y_{N:\infty}$, as shown in information diagram [79] of Fig 2.9. This means:

$$\mathrm{I}[Y_{N:\infty}; Y'_{0:N}] = \mathrm{I}[Y_{N:\infty}; Y_{0:N}] - \mathrm{I}[Y_{N:\infty}; Y_{0:N}|Y'_{0:N}] \ . \tag{2.43}$$

This, in turn, gives: $\mathrm{I}[Y_{N:\infty}; Y_{0:N}] \geq \mathrm{I}[Y_{N:\infty}; Y'_{0:N}] \geq 0$. Thus, we find the input process's excess entropy $\mathbf{E}$ [12]:

$$\begin{aligned} \lim_{N\to\infty} \mathrm{I}[Y_{N:\infty}; Y'_{0:N}] &\leq \lim_{N\to\infty} \mathrm{I}[Y_{N:\infty}; Y_{0:N}] \\ &= \mathbf{E} \ . \end{aligned} \tag{2.44}$$

However, dividing by N it's contribution vanishes:

$$\begin{aligned} \lim_{N\to\infty} \frac{I[Y_{N:\infty}; Y_{0:N}]}{N} &= \lim_{N\to\infty} \left(\frac{H[Y_{0:N}]}{N} - \frac{H[Y_{0:N}|Y_{N:\infty}]}{N} \right) \\ &= h_\mu - h_\mu \\ &= 0 \ . \end{aligned} \tag{2.45}$$

Thus, we are left with the inequality of Eq. (2.5):

$$\langle W \rangle \leq k_B T \, \ln 2 \left(h'_\mu - h_\mu \right) \ ; \tag{2.46}$$

derived with minimal assumptions. Also, the appearance of the statistical complexity and excess entropy, whose contributions this particular derivation shows are asymptotically

small, does indicate the potential role of correlations in the input for finite time—times during which the ratchet synchronizes to the incoming information [80].

One key difference between Eq. (2.46) (equivalently, Eq. (2.5)) and the more commonly used bound in Eq. (2.6), with the change in single-variable configurational entropy $\mathrm{H}[Y_0'] - \mathrm{H}[Y_0]$, is that the former bound is true for all finite ratchets and takes into account the production of information over time via the Kolmogorov-Sinai entropies h_μ and h'_μ.

There are several special cases where the single-variable bound of Eq. (2.6) applies. In the case where the input is uncorrelated, it holds, but it is a weaker bound than Eq. (2.5) using entropy rates. Also, in the case when the ratchet has no internal states and so is memoryless, Eq. (2.6) is satisfied. Interestingly, either it or Eq. (2.46) can be quantitatively stricter in this special case. However, in the most general case where the inputs are correlated and the ratchet has memory, the bound using single-variable entropy is incorrect, since there are cases where it is violated [14]. Finally, when the input-bit-ratchet interaction time τ grows the ratchet spends much time thermalizing. The result is that the output string becomes uncorrelated with the input and so the ratchet is effectively memoryless. Whether by assumption or if it arises as the effective behavior, whenever the ratchet is memoryless, it is ignorant of temporal correlations and so it and the single-symbol entropy bounds are of limited physical import. These issues will be discussed in detail in sequential sections, but as a preview see Ref. [14].

2.7.2 Designing Ratchet Energetics

Figure 2.3 is one of the simplest information transducers for which the outcomes are unifilar for uncorrelated inputs, resulting in the fact that the correlations in the outgoing bits can be explicitly calculated. As this calculation was a primary motivation in our work, we introduced the model in Fig. 2.3 first and, only then, introduced the associated energetic and thermodynamic quantities, as in Fig. 2.2. The introduction of energetic and thermodynamic quantities for an abstract transducer (as in Fig. 2.3), however, is not trivial. Given a transducer topology (such as the reverse "Z" shape of the current model), there are multiple possible energy schemes of which only a fraction are consistent with all possible values of the associated transition probabilities. However, more than one scheme

is generally possible.

To show that only a fraction of all possible energetic schemes are consistent with all possible parameter values, consider the case where the interaction energy between the ratchet and a bit is zero, as in Ref. [7]. In our model, this implies $\beta = 0$, or equivalently, $p = q = 0$ (from Eq. (2.12)). In other words, we cannot describe our model, valid for all values $0 < p, q < 1$, by the energy scheme in Fig. 2.2 with $\beta = 0$. This is despite the fact that we have two other independent parameters α and w.

To show that, nonetheless, more than one scheme is possible, imagine the case with $\alpha = \beta = 0$. Instead of just one mass, consider three masses such that, whenever the transitions $A \otimes 0 \to B \otimes 0$, $B \otimes 0 \to A \otimes 1$, and $A \otimes 1 \to B \otimes 1$ take place, we get works $k_{\mathrm{B}}T\widetilde{W}_1$, $k_{\mathrm{B}}T\widetilde{W}_2$, and $k_{\mathrm{B}}T\widetilde{W}_3$, respectively. We lose the corresponding amounts of work for the reverse transitions. This picture is consistent with the abstract model of Fig. 2.3 if the following requirements of detailed balance are satisfied:

$$\frac{1}{1-p} = \frac{M_{A\otimes 0 \to B\otimes 0}}{M_{B\otimes 0 \to A\otimes 0}} = e^{-\widetilde{W}_1} \ , \tag{2.47}$$

$$\frac{p}{q} = \frac{M_{B\otimes 0 \to A\otimes 1}}{M_{A\otimes 1 \to B\otimes 0}} = e^{-\widetilde{W}_2} \ , \text{ and} \tag{2.48}$$

$$1-q = \frac{M_{A\otimes 1 \to B\otimes 1}}{M_{B\otimes 1 \to A\otimes 1}} = e^{-\widetilde{W}_3} \ . \tag{2.49}$$

Existence of such an alternative scheme illustrates the fact that given the abstract model of Fig. 2.3, there is more than one possible consistent energy scheme. We suggest that this will allow for future engineering flexibility.

Chapter 3

Correlation-powered Information Engines

3.1 Introduction

Intriguing connections between statistical mechanics and information theory have emerged repeatedly since the latter's introduction in the 1940s. Thermodynamic entropy in the canonical ensemble is the Shannon information of the Boltzmann probability distribution [81]. Average entropy production during a nonequilibrium process is given by the relative entropy [82, 83], an information-theoretic quantity, of the forward trajectories with respect to the time-reversed trajectories [27]. Perhaps the most dramatic connection, though, appears in the phenomenon of Maxwell's demon, a thought experiment introduced by James C. Maxwell [84]. This is a hypothetical, intelligent creature that can reverse the spontaneous relaxation of a thermodynamic system, as mandated by the Second Law of thermodynamics, by gathering information about the system's microscopic fluctuations and accordingly modifying its constraints, *without expending any net work*. A consistent physical explanation can be obtained only if we postulate, following Szilard [32], a thermodynamic equivalent of information processing: Writing information has thermodynamic benefits whereas erasing information has a minimum thermodynamic cost, $k_{\mathrm{B}}T \ln 2$ for the erasure of one bit of information. This latter is Landauer's celebrated principle [4, 34].

The thermodynamic equivalent of information processing has the surprising impli-

cation that we can treat the carrying capacity of an information storage device as a thermodynamic fuel. This observation has led to a rapidly growing literature exploring the potential design principles of nanoscale, autonomous machines that are fueled by information. References [60, 7], for example, introduced a pair of stochastic models that can act as an engine without heat dissipation and a refrigerator without work expenditure, respectively. These strange thermal devices are achieved by writing information on a tape of "bits"—that is, on a tape of two-state, classical systems. A more realistic model was suggested in Ref. [63]. These designs have been extended to enzymatic dynamics [85], stochastic feedback control [86], and quantum information processing [87, 73].

The information tape in the above designs can be visualized as a sequence of symbols where each symbol is chosen from a fixed alphabet, as shown in Fig. 2.1 for binary tape symbols. There is less raw information in the tape if the symbols in the sequence are statistically correlated with each other. For example, the sequence $\ldots 101010 \ldots$, consisting of alternating 0s and 1s, encodes only a single bit of information on the whole since there are only two such sequences (differing by a phase shift). Whereas, a sequence of N random binary symbols encodes N bits of information. The thermodynamic equivalent of information processing, therefore, says that we can treat the former (ordered) sequence as a thermodynamic fuel. This holds even though it contains equal numbers of 0s and 1s on average as in the fully random sequence, which provides no such fuel.

The design principles of *information engines* [78] explored so far, however, are not generally geared towards temporally correlated information tapes [7, 60, 56, 61, 62, 63, 64, 10] since, by and large, only a tape's single-letter frequencies have been considered. However, the existence of statistical correlations among the symbols—that is, between environmental stimuli—is the rule, not an exception in Nature. Even technologically, producing a completely correlation-free (random) sequence of letters is a significant challenge [88, 89, 90]. The thermodynamic value of statistical correlations [8, 49] and quantum entanglement [91, 92, 93, 94, 95, 96, 97, 98, 99] have been discussed widely in the literature. Our goal here is to extend the design of tape-driven information engines to accommodate this more realistic scenario—information engines that leverage temporally correlated

environments to convert thermal energy to useful work.

Other studies have taken a somewhat different approach to the description and utilization of the thermodynamic equivalent of information processing. References [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 64, 17] explored active feedback control of a stochastic system by external means, involving measurement and feedback or measurement, control, and erasure. While Refs. [51, 52, 53, 54] explored a multipartite framework involving a set of interacting, stochastic subsystems and Refs. [55, 56] studied steady-state models of Maxwell's demon involving multiple reservoirs. And, finally, Refs. [57, 58, 59] indicated how several of these approaches can be combined into single framework.

Here, we use computational mechanics [77] for thermal information ratchets [13] to derive a general expression for work production that takes into account temporal correlations in the environment as well as correlations created in the output by the information engine's operation. The functional form of the work expression establishes that memoryless information ratchets cannot leverage anything more than single-symbol frequencies in their input and are, therefore, insensitive to temporal correlations. Thus, to the extent that it is possible to leverage temporally correlated environments, memoryful information engines are the only candidates. This indicates, without proof, that the memory of an information engine must reflect the memory of its environment to most efficiently leverage structure in its input.

Adding credence to this hypothesis, we introduce an *ergodic* information engine that is driven *solely* by temporal correlations in the input symbols to produce work. The states of the engine wind up reflecting the memory states of the generator of the input process. This makes good on the conjecture [13] as to why one observes thermodynamically functional ratchets in the real world that support memory [13]: Only Demons with memory can leverage temporally correlated fluctuations in their environment.

Similar behavior was demonstrated by Maxwell's refrigerator [60], when Ref. [73] showed it to be a nonergodic refrigerator when driven by a nonergodic process that is statistically unbiased over all realizations. However, we focus on our ergodic engine, since ergodicity leads to robust and reliable work production. This contrast is notable. Without

ergodicity, an engine does not function during many realizations, from trial to trial. In this sense, a "nonergodic engine" is unreliable in performing its intended task, such as being an engine (converting thermal energy to work), generating locomotion, and the like. During one trial it functions; on another it does not.

If one is willing to broaden what one means by "engine", then one can imagine constructing an "ensemble engine" composed of a large collection of nonergodic engines and then only reporting ensemble-averaged performance. Observed over many trials, the large trial-by-trial variations in work production are masked and so the ensemble-average work production seems a fair measure of its functionality. However, as noted, this is far from the conventional notion of an engine but, perhaps, in a biological setting with many molecular "motors" it may be usefully considered functional.

Our design of an ergodic engine that can operate solely on temporal correlations should also be contrasted with a recent proposal [100] that utilizes mutual information between two tapes, i.e., *spatial* correlations, as a thermodynamic fuel.

The overarching thermodynamic constraints on functioning at all are analyzed in a companion work [17]. The following, in contrast, focuses on the particular functionality of self-correcting Demons in the presence of temporally correlated environments and on analyzing the thermodynamic regimes that support them. First, we review the information engine used and give a synopsis of our main results so that they are not lost in the more detailed development. Second, the technical development begins as we introduce the necessary tools from computational mechanics and stochastic thermodynamics. Third, using them, we analyze the engine's behavior and functioning in the presence of a correlated input, calling out the how the Demon recognizes (or not) correlations in the input and either (i) responds constructively by using them to convert thermal energy to work or (ii) dissipates energy as it attempts to re-synchronize and regain engine functioning. Fourth, we note how these two dynamical modes represent a type of dynamical nonergodicity over the ratchet's state space when the ratchet cannot re-synchronize, which leads to temporary nonergodicity in the work production. However, with re-synchronization, these two dynamical modes become accessible from each other, which leads to ergodicity of the

engine and its work production. And, finally, we derive the physical consequences for the costs of self-correction and its operational limits.

3.2 A Self-Correcting Information Engine: Synopsis

Figure 2.1 shows our model [7, 13] of an information engine implemented as a thermal ratchet consisting of four elements: a thermal reservoir, a work reservoir (mass in a gravitational field), an information tape (or reservoir), and a ratchet controlled by the values in the input tape cells. The ratchet acts as the communication medium between the three reservoirs as it moves along the tape and transforms the input information content. In the process, it mediates energy exchange between the heat and work reservoirs.

To precisely specify the kinds of temporal correlation in the ratchet's environment, we represent the generator of the sequences on the information tape via a hidden Markov model (HMM), a technique introduced in the last chapter and Ref. [13]. This has several advantages. One is that the full distribution over infinite sequences of the input tape $\Pr(\overleftrightarrow{Y})$ is represented in a compact way. The most extreme case of this comes in recalling that finite-state HMMs can finitely represent infinite-order Markov processes [101]. And so, HMMs give a desirable flexibility in the kinds of environments we can analyze, from memoryless to finite- and infinite-order Markovian. Another is that many statistical and informational properties can be directly calculated, as we discuss shortly. In this setup, the ratchet is a transducer in the sense of computational mechanics [1]. And this, in turn, allows exact analysis of informational bounds on work production [13, 10]. Here, though, in Sec. 3.3.3 we go further, expanding the toolset of the HMM-transducer formalism by deriving a general work production expression for any finite-state input HMM driving a finite-state thermal ratchet.

With this powerful new expression for work production, Sec. 3.4.1 then considers the case of a perfectly correlated information tape. Though nominally simple, this case is of particular interest since previous single-symbol entropy bounds erroneously suggest this class of input should generate nonpositive work. Our entropy rate bounds, in contrast, suggest it is possible to generate net positive work. And, indeed, we see that the single-

symbol bounds are violated, as our ratchet produces positive work. In examining this concrete model, moreover, we realize that the ratchet's synchronizing to the correlations in its input is an essential part of work production: Synchronization is how the ratchet comes to leverage the thermodynamic "fuel" in a memoryful input process.

This result emphasizes a key feature of our ratchet design: Useful thermodynamic functioning is driven purely by the temporal correlations in the input tape. That is, if the symbols are perfectly correlated—a sequence with temporal memory, e.g., with 1s always following 0s and vice versa—the ratchet acts as an engine, writing new information on the output tape and transferring energy from the heat to the work reservoir. However, if the correlation is not perfect, depending on engine parameters, the ratchet can act as an information-eraser or dud, converting work into heat. Thus, there exists a critical level of corrupted input correlation beyond which engine functionality is no longer possible. Our tools allow us to give explicit expressions for work in all these cases, including the parameter limits of thermodynamic functioning.

Perhaps most importantly, the analysis reveals a novel mechanism underlying the functioning and its disappearance. This can be explained along the following lines. An exclusive feature of the ratchet design is the presence of a synchronizing state, denoted C in the (state $\otimes$ bit)-transition diagram of Fig. 3.3. Absent C and for perfectly correlated input, the ratchet is equally likely to be in two stable dynamical modes: "clockwise" in which heat is converted into work and "counterclockwise" in which work is converted into heat. (See Fig. 3.5.) Since the counterclockwise mode dissipates more per cycle than can be compensated by the clockwise mode, without C the ratchet cannot function as an engine. With C, though, the counterclockwise mode becomes a transient and the clockwise mode an attractor, making possible the net conversion of heat into work (engine mode). The phenomenon of an observer (ratchet) coming to know the state of its environment (phase of the memoryful input tape) is referred to as *synchronization* [12]. (For a rather different notion of synchronization and its thermodynamic interpretation see Ref. [102].)

In contrast, when the input symbols are not perfectly correlated due to phase slips, say, the ratchet is randomly thrown into the dissipative counterclockwise mode. Nonetheless,

repeated re-synchronization may compensate, allowing the engine mode, if the transition probabilities into C are enhanced, up to a level. This is a form of *dynamical error correction*. Beyond a certain level of corruption in the input correlations, however, dynamical error correction is not adequate to resynchronize to the input phase. The Demon cannot act as an engine, no matter how large the transition probabilities into C. This critical corruption level is shown in the thermodynamic-function diagram of Fig. 3.11 by the vertical dotted line, where the horizontal axis denotes level of corruption as the frequency of phase slips.

The current situation must be contrasted with the usual error correction schemes in communication theory and biological copying. In the former context, redundancy is built into the data to be transmitted so that errors introduced during transmission can be corrected by comparing to redundant copies, up to a certain capacity. In the biological context of copying, as in DNA replication [103], error correction corresponds to the phenomenon of active reduction of errors by thermodynamic means [104, 105, 106]. In the current context, we use the term *self-correction* to refer to the fact the proposed information engine can predict and synchronize itself with the state of the information source to produce positive work even when the engine is initiated in or driven by fluctuations to a dissipative mode. Section 3.5 discusses this self-correcting behavior of the engine in detail.

To analyze how dynamical error correction operates quantitatively, the following shows how the presence of state C renders the counterclockwise phase transient. This reveals a novel three-way tradeoff between synchronization rate (transition probability from C to the clockwise phase), work produced during synchronization, and average extracted work per cycle. Section 3.5 then turns to analyze re-synchronization, considering the case of imperfectly correlated information tape with phase slips. It demonstrates how the ratchet dynamically corrects itself and converts heat into work over certain parameter ranges. The section closes by giving the expression for maximum work and the parameter combinations corresponding to achieving optimum conversion.

Throughout the exploration, several lessons stand out. First, to effectively predict

bounds on a input-driven ratchet's work production, one must consider *Shannon entropy rates* of the input and output strings; and not single-variable entropies. Second, the expression for the work production shows that correlations coming from memoryful environments can only be leveraged by memoryful thermodynamic transformations (Demons). While it remains an open question how to design ratchets to best leverage memoryful inputs, the particular ratchet presented here demonstrates how important it is for the ratchet's structure to "match" that of the input correlations. In short, the ratchet only produces work when its internal states are synchronized to the internal states of the input sequence generator. Otherwise, it is highly dissipative. And last, synchronization has energetic consequences that determine the effectiveness of dynamical error correction and the tradeoffs between average work production, work to synchronize, and synchronization rate.

3.3 Thermal Ratchet Principles

As described in Chapter 2 and shown in Fig. 2.1, our ratchet moves along the information tape unidirectionally, interacting with each symbol sequentially. The ratchet interacts with each symbol for time τ and possibly switches the symbol value contained in the cell. We refer to time period τ as the *interaction interval* and the transitions that happen in the joint state space as *interaction transitions*. Through this process, the ratchet transduces a semi-infinite input string, expressed by random variable $Y_{0:\infty} = Y_0 Y_1 \ldots$, into an output string $Y'_{0:\infty} = Y'_0 Y'_1 \ldots$. Here, the symbols Y_N and Y'_N realize the elements y_N and y'_N, respectively, over the same information alphabet $\mathcal{Y}$.

For example, as in Fig. 2.1, the alphabet consists of just 0 and 1. Consider the case in which the ratchet was initiated at the leftmost end at time $t = 0$. At time $t = N\tau$ the entire tape is described by the random variables $Y'_{0:N} Y_{N\infty} = Y'_0 Y'_2 \ldots Y'_{N-2} Y'_{N-1} Y_N Y_{N+1} \ldots$, because in N time-steps N input symbols have been transduced into N output symbols. The state of the ratchet at time $t = N\tau$ is denoted by the random variable X_N, which realizes an element $x_N \in \mathcal{X}$, where $\mathcal{X}$ is the ratchet's state space.

Since we chose the input alphabet to consist of just two symbols 0 and 1, we refer

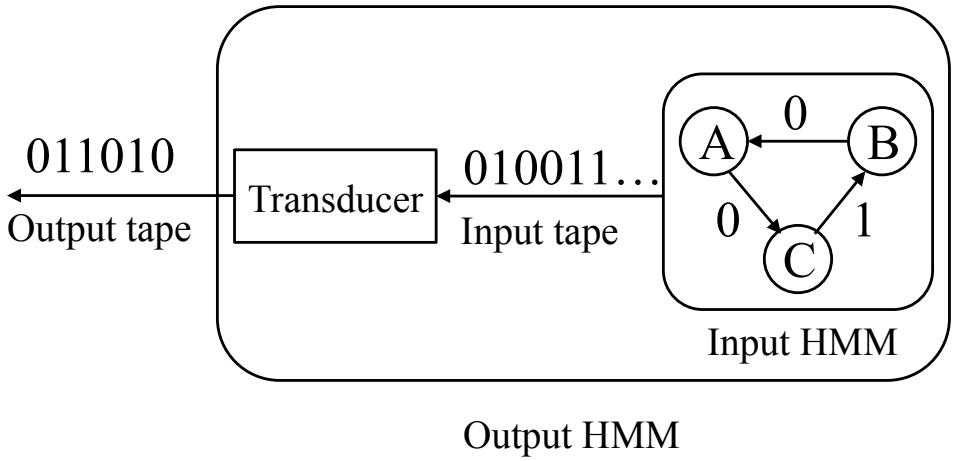


Figure 3.1. Computational mechanics of information engines: The input tape values are generated by a hidden Markov model (HMM) with, say, three hidden states—A, B, and C. Specifically, transitions among the hidden states produce 0s and 1s that form the input tape random variables $Y_{0:\infty}$. The ratchet acts as an informational transducer that converts the input HMM into an output process, that is also represented as an HMM. That is, the output tape $Y'_{0:\infty}$ can be considered as having been generated by an effective HMM that is the composition of the input HMM and the ratchet's transducer [1].

to the values in the tape cells as *bits*. That this differs from the information unit "bit" should be clear from context. *Tape* generally refers to the linear chain of cells and *string* to the stored sequence of symbols or cell values.

Finally, the ratchet is connected with two, more familiar reservoirs—a thermal reservoir and a work reservoir. The state of the thermal reservoir at time $t = N\tau$ is denoted by Z_N. We assume that the thermal reservoir is at absolute temperature T K. The work reservoir consists of a mass being pulled down by gravity, but kept suspended by a pulley. Certain, specified ratchet transitions lower and raise the mass, exchanging work.

To set up the analysis, we must first review how to measure information, structure, and energy as they arise during the ratchet's operation.

3.3.1 Ratchet Informatics: Computational Mechanics

To monitor information generation and storage, computational mechanics views the sequence of symbols from the left of the input tape $Y_{0:\infty}$ as the temporal output of a kind of HMM, called an ϵ-machine [77]. The latter provides the most compact way to represent the statistical distribution of symbol sequences. In particular, many types of long-range

correlation among the symbols are encoded in the ϵ-machine's finite-state hidden dynamics. The correlations appear as the *memory*, characterized by its internal-state entropy or *statistical complexity* C_μ. Specifically, if the input can be produced by an HMM with a single hidden state, the input generator is memoryless and there cannot be any correlation among the symbols [1].

The ratchet functions as a memoryful communication channel that sequentially converts the input symbols into values in $Y'_{0:\infty}$, the output tape. Naturally, the output tape itself can be considered in terms of another HMM, as emphasized by the schematic in Fig. 3.1. There, the ratchet acts as an information transducer between two information sources represented by respective input and output HMMs [1].

These choices make it rather straightforward to measure ratchet *memory*. If the size of its state space is unity ($|\mathcal{X}| = 1$), then we say it is memoryless. Otherwise ($|\mathcal{X}| > 1$), we say it is memoryful. With memory, the ratchet at time $t = N\tau$ can store information about the past input symbols $y_{0:N}$ with which it has interacted, as well as past outputs $y'_{0:N}$. Similarly, the output HMM can have memory (its own positive statistical complexity $C_\mu > 0$) even when the input HMM does not. This was the case, for example in Refs. [7, 60, 63, 13]. Critically, the transducer formalism has the benefit that we can exactly calculate the distribution $\Pr(Y'_{0:\infty})$ of output tapes for any finite-memory ratchet with a finite-memory input process. Shortly, we add to this set of tools, introducing a method to calculate the work production by any finite-memory ratchet operating on a finite-memory input.

3.3.2 Ratchet Energetics: The First Law of Thermodynamics

Interactions between ratchet states and input symbols have energetic consequences. The internal states and symbols interact with a thermal reservoir at temperature T, whose configuration at time step N is denoted by the random variable Z_N, and with a work reservoir, that holds no information and so need not have an associated random variable. Through its operation, the current input symbol facilitates or inhibits energy flows

[1] Thus, in our use of the descriptor "correlated", the all 0s sequence and the all 1s sequence have no temporal correlation. Since their internal memory $C_\mu = 0$, they have no information to correlate. This is analogous to *autocorrelation* in which the zero frequency offset is subtracted.

between the work and thermal reservoirs.

The joint dynamics of the ratchet and incoming symbol occur over two alternating steps: a *switching* transition and an *interaction* transition. At time $t = N\tau$, the ratchet switches the tape cell with which it interacts from the $(N-1)$th output symbol y'_{N-1} to the Nth input symbol y_N. This is followed by the interaction transition between the ratchet, which is in the x_N state, and the symbol y_N. Together, they make a stochastic transition in their joint state space according to the Markov chain:

$$M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} = \Pr(X_{N+1} = x_{N+1}, Y'_N = y'_N | X_N = x_N, Y_N = y_N) \ .$$

M has detailed balance, since transitions are activated by the thermal reservoir. Energy changes due to these thermal interaction transitions are given by the Markov chain:

$$\Delta E_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} = k_B T \ln \frac{M_{x_{N+1} \otimes y'_N \to x_N \otimes y_N}}{M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N}} \ .$$

These energies underlie the heat and work flows during the ratchet's operation. Through interaction, the input symbol y_N is converted into the output symbol y'_N and written to the output tape cell as the ratchet switches to the next input bit y_{N+1} to start the next interaction at time $t = (N+1)\tau$.

Notably, previous treatments [7, 60, 13] of information engines associated the energy change during an interaction transition with work production by coupling the interaction transitions to work reservoirs. While it is possible to construct devices that have this work generation scheme, it appears to be a difficult mechanism to implement in practice. We avoid this difficulty, designing the energetics in a less autonomous way, not attaching the work reservoir to the ratchet directly.

So, instead of the ratchet effortlessly stepping along the tape unidirectionally on its own, it is driven. (And, an energetic cost can be included for advancing the ratchet without loss of generality.) In this way, heat flow happens during the interaction transitions and work flow happens during the switching transitions. Appendix 3.7.1 shows how this

strategy gives an exact asymptotic average work production per time step:

$$\langle W \rangle = \sum_{\substack{x,x' \in \mathcal{X} \\ y,y' \in \mathcal{Y}}} \pi_{x \otimes y} M_{x \otimes y \to x' \otimes y'} \Delta E_{x \otimes y \to x' \otimes y'} \ , \tag{3.1}$$

where $\pi_{x \otimes y}$ is the asymptotic distribution over the joint state of the Demon and interaction cell at the beginning of any interaction transition:

$$\pi_{x \otimes y} = \lim_{N \to \infty} \Pr(X_N = x, Y_N = y) \ . \tag{3.2}$$

It is important to note that π is not M's stationary distribution and, moreover, it is highly dependent on the input HMM. Despite calculating work production for a different mechanism, the asymptotic power calculated here is the same as in previous examinations [7, 13, 10].

From the expression of work given in Eq. (4.13), we see that memoryless ratchets have severe limitations in their ability to extract work from the heat reservoir. In this case, the ratchet state space $\mathcal{X}$ consists of a single state and π in Eq. (3.2) is just the single symbol distribution of the input string:

$$\pi_{x \otimes y} = \Pr(Y_0 = y) \ .$$

As a result, the calculation of work depends only on the single-symbol statistics of the input string, producing work from the string as if the input were independent and identically distributed (IID). Regardless of whether there are correlations among the input symbols, the work production of a memoryless ratchet is therefore the same for all inputs having the same single-symbol statistics. For example, a memoryless ratchet cannot distinguish between input strings 01010101 . . . and 00110011 . . . as far as work is concerned. Thus, for the ratchet to use correlations in the input string to generate work, it must have nonzero memory. This is in line with previous examinations of autonomous information engines [10, 13]. In any case, the general form for the work production here allows one to calculate it for any finite memoryful channel operating on any input tape generated by a finite HMM.

3.3.3 Ratchet Entropy Production: The Second Law of Thermodynamics

Paralleling Landauer's Principle [4, 34] on the thermodynamic cost of information erasure, several extensions of the Second Law of thermodynamics have been proposed for information processing. We refer to them collectively as the *thermodynamic equivalents of information processing*. For ratchets, these bounds on the thermodynamic costs of information transformation can be stated either in terms of the input and the output HMMs' *single-symbol entropy* (less generally applicable) or *entropy rate* (most broadly applicable). Let's review their definitions for the sake of comparison.

Consider the probability distribution of the symbols $\{0, 1\}$ in the output sequence of an HMM. If the single-symbol probabilities are $\{p, 1-p\}$, respectively, the single-symbol entropy H_1 of the HMM is given by the *binary entropy function* $\mathrm{H}(p)$ [11]:

$$\begin{aligned} \mathrm{H}_1 &= \mathrm{H}(p) \\ &\equiv -p \log_2 p - (1-p) \log_2 (1-p) \ . \end{aligned} \tag{3.3}$$

By definition, single-symbol entropy ignores sequential symbol-symbol correlations.

The entropy rate, as discussed in Chapter 2, is the asymptotic per-symbol uncertainty. To define it, we need to first introduce the concept of a word in the output sequence generated by an HMM. A word w is a subsequence of symbols of length ℓ over the space $\mathcal{Y}^\ell$. For example, a binary word of length $\ell = 2$ consists of a pair of consecutive symbols; an event in the space $\mathcal{Y}^2 = \{00, 01, 10, 11\}$. Thus, there are 2^ℓ possible length-ℓ words or elements in $\mathcal{Y}^\ell$. The *Shannon entropy rate* of the process generated by an HMM is then given by [11]:

$$h_\mu = -\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{w \in \mathcal{Y}^\ell} \Pr(w) \log_2 \Pr(w) \ , \tag{3.4}$$

where $\Pr(w)$ denotes the probability of $w \in \mathcal{Y}^\ell$. Entropy rate h_μ captures the effects of correlations in the symbols at all lengths.

For memoryless processes, $\mathrm{H}_1 = h_\mu$. Otherwise, $\mathrm{H}_1 > h_\mu$, with h_μ being the correct measure of information per symbol and H_1 being an overestimate. One relevant extreme

case arises with exactly periodic processes with period greater than 1: $h_\mu = 0$; whereas $\mathrm{H}_1 > 0$, it's magnitude being determined by the single-symbol frequencies.

Again we're presented with the two specific forms of the thermodynamic equivalent of information processing for information engines:

$$\langle W \rangle \leq k_B T \, \ln 2 \, \Delta \, \mathrm{H}_1 \tag{3.5}$$

$$\langle W \rangle \leq k_B T \, \ln 2 \, \Delta h_\mu \ , \tag{3.6}$$

where $\Delta \, \mathrm{H}_1$ and Δh_μ denote, respectively, the change in single-symbol entropy and in entropy rate from the input HMM to the output HMM [7, 60, 61, 5, 58, 36, 10, 13].

Let's compare them. Equation (3.5) says that correlations in the input string beyond single symbols cannot be used to produce work, while Eq. (3.6) suggests that it is possible. This follows since, if we keep the single-symbol probabilities constant while increasing the temporal correlations in the input, all while keeping the output fixed, $\Delta \, \mathrm{H}_1$ remains constant, but Δh_μ increases.

To resolve this seeming ambiguity, we appeal to the general expression of Eq. (3.1) for calculating work production. The expression says that work production depends on the memory of both the ratchet and the input HMM; see App. 3.7.1. In this way, temporal correlations in the input string can influence the ratchet's thermodynamic behavior. Only when the ratchet is memoryless is there no relevance of the correlations, so far as the average work is concerned. In the memoryless case, Eq. (3.5) as well as Eq. (3.6) are valid.

This observation suggests that, in contrast, for a ratchet to use correlations in the input string to generate work, it must have more than one internal state [17]. In addition, to generate correlations in the input string, its generating HMM must have memory. This leads to the intuitive hypothesis that to leverage work from the temporal order in the input string (correlations created by the input HMM's memory), the ratchet must also have memory.

We test this hypothesis by analyzing the specific example of a perfectly correlated environment—a periodic input process. As we do, keep in mind that, on the one hand, Eq. (3.5) says that no work production is possible, regardless of the binary output process

statistics. On the other hand, Eq. (3.6) suggests the opposite. As long as the output process has some uncertainty in sequential symbols, then $\Delta h_\mu > 0$. We also introduce a ratchet with three memory states that produces positive work and even appears to be nearly optimal for certain parameter ranges [17]. In short, a memoryful ratchet with a memoryful input process violates Eq. (3.5), demonstrating that bound's limited range of application.

3.4 Functional Ratchets in Perfectly Correlated Environments

Let's consider the case of a correlated environment and then design a thermal ratchet adapted to it.

3.4.1 The Period-2 Environment

Take the specific case of a period-2 input process. The state transition diagram for its HMM is given in Fig. 3.2. There are three internal states. D is a transient state from which the process starts. From D, the process transitions to either E or F with equal probabilities. If the system transitions to E, a 0 is emitted, and if the system transitions to F, a 1 is. Afterwards, the process switches between E and F with $E \to F$ transitions emitting 1 and $F \to E$ transitions emitting 0. As a result, the input HMM generates two possible sequences that drive the ratchet: $y_{0:\infty} = 010101\ldots$ or $y_{0:\infty} = 101010\ldots$. Note that these two sequences differ by a single phase shift.

The period-2 process is an ideal base case for analyzing how ratchets extract work out of temporal correlations. First, its sequences have no bias in the frequencies of 0's and 1's, as they come in equal proportions; thereby removing any potential gain from an initial statistical bias. And, second, the symbols in the sequence are perfectly correlated—a 0 is followed by 1 and a 1 by 0.

More to the point, previous information engines cannot extract work out of such periodic sequences since those engines were designed to obtain their thermodynamic advantage purely from statistical biases in the inputs [7, 60, 61, 13, 10]. By way of contrast, we now introduce and analyze the performance of a ratchet design that extracts work out of such

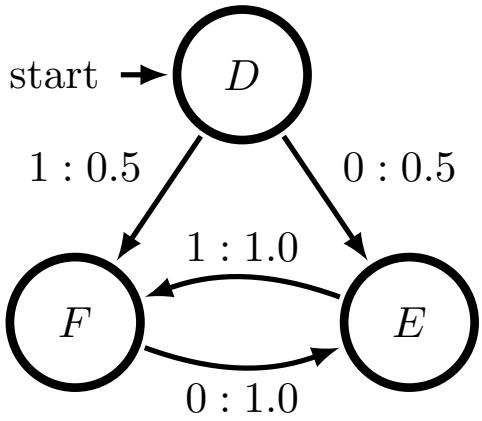


Figure 3.2. Period-2 process hidden Markov model with a transient start state D and two recurrent causal states E and F. Starting from D, the process makes a transition either to E or to F with equal probabilities while emitting $y = 0$ or $y = 1$, respectively. This is indicated by the transition labels from D: $y : p$ says generate symbol y when taking the transition with probability p. On arriving at states E or F, the process alternates between two states, emitting $y = 0$ for transitions $E \to F$ and $y = 1$ for transitions $F \to E$. In effect, we get either of two infinite sequences, $y_{0:\infty} = 0101\ldots$ and $y_{0:\infty} = 1010\ldots$, with equal probabilities.

perfectly correlated, unbiased input sequences. The following section then considers the more general case in which input correlations are corrupted by environmental fluctuations.

Let's explain the information-theoretic reasoning that motivates this. For a period-2 process, the single-symbol entropy H_1 is maximal: $\mathrm{H}[Y_N] = 1$. However, its entropy rate $h_\mu = 0$ due to its perfect predictability as soon as any symbol is known. This, on the one hand, implies $\Delta\,\mathrm{H}_1 \equiv \mathrm{H}[Y'_N] - \mathrm{H}[Y_N] \leq 0$. Equation (3.5), in turn, says that work cannot be extracted regardless of the realizations of the output string; no matter the design of the information engine. For the period-2 input, though, $\Delta h_\mu = h'_\mu \geq 0$. And, Eq. (3.6) indicates that work can be extracted as long as the output string has nonzero entropy rate h'_μ. This is achievable with appropriate thermal ratchet design. In other words, Eq. (3.5) suggests that it is impossible to extract work from input correlations beyond single-symbol bias, while Eq. (3.6) suggests it is possible. We resolve this disagreement in favor of Eq. (3.6) by explicit construction and exact analysis.

3.4.2 Memoryful Ratchet Design

Figure 3.3 gives a ratchet design that can extract work out of a period-2 process. As explained above in Sec. 3.3, the ratchet interacts with one incoming symbol at a time. As

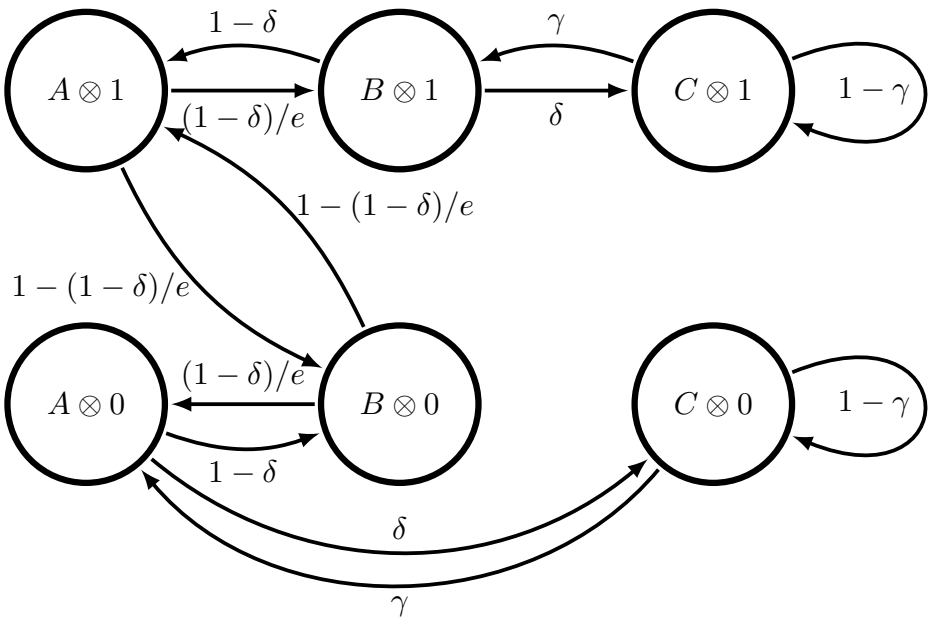


Figure 3.3. State transition diagram of a ratchet that extracts work out of environment correlations: A, B, and C denote the ratchet's internal states and 0 and 1 denote the values of the interacting cell. The joint dynamics of the Demon and interacting cell take place over the space of six internal joint states: $\{\mathrm{A}\otimes 0, \ldots, \mathrm{C}\otimes 1\}$. Arrows indicate the allowed transitions and their probabilities in terms of the ratchet control parameters δ and γ. Note that e here refers to the base of natural logarithm, not a variable.

a result, the ratchet's transducer specifies both ratchet internal states and the states of the input tape cell being read. In the figure, A, B, and C denote the ratchet's internal states and $x \otimes y$ denotes the joint transducer state of the ratchet state and interacting cell value, with the ratchet being in state $x \in \{A, B, C\}$ and the interacting cell with value $y \in \{0, 1\}$. Arrows denote the allowed transitions and their labels the transition probabilities in terms of ratchet control parameters, that we now introduce. For example, if the Demon is in state A and the input symbol has value 0, they make a transition to the joint state $B \otimes 0$ with probability $(1 - \delta)$ or to the joint state $C \otimes 0$ with probability δ. Due to conservation of probability, the sum of transition probabilities out of any joint state is unity. After the transition, the old symbol value in the tape cell is replaced by a new value. If the joint state made a transition to $B \otimes 0$ and the incoming symbol had value 1, the joint state is switched to $B \otimes 1$. Then, a transition from joint state $B \otimes 1$ takes place according to the rule described above.

The parameters δ and γ satisfy the following constraint: $0 \leq \delta, \gamma \leq 1$. The Markov chain matrix M corresponding to the transition dynamics depicted in Fig. 3.3 is given in App. 3.7.2. Due to the repetitive nature of the dynamics, the transducer reaches an asymptotic state (App. 3.7.1) such that its probability distribution does not change from one interaction interval to another.

Now, consider the transducer's response when driven by a period-2 input process. Appendix 3.7.2 calculates the work and entropy changes in the asymptotic limit, finding:

$$\langle W \rangle = \frac{1-\delta}{e} k_B T \ln 2 \ , \tag{3.7}$$

$$\Delta \, \mathrm{H}_1 = 0 \ , \text{ and} \tag{3.8}$$

$$\Delta h_\mu = \mathrm{H}\left(\frac{1-\delta}{e}\right) \ . \tag{3.9}$$

The work expression follows from the definition in Eq. (4.13). The single-symbol entropy difference $\Delta \, \mathrm{H}_1$ vanishes since the output tape consists of random, but still equal, mixtures of 0's and 1's, as did the input tape. The entropy rate change Δh_μ, though, is generally positive since, although the input entropy rate vanishes, the ratchet adds some randomness to the output.

From Eq. (3.9), we have a clear violation of Eq. (3.5). Whereas, Eq. (3.6) still holds:

$$0 = \Delta \, \mathrm{H}_1 < \frac{\langle W \rangle}{k_B T \ln 2} \leq \Delta h_\mu \ . \tag{3.10}$$

Since Ref. [13] established Eq. (3.6) for all finite ratchets, this difference in the bounds is expected. Nonetheless, it is worth calling out in light of recent discussions in the literature [14]. In any case, these results confirm the conclusion that to properly bound all finite information ratchets, including memoryful ratchets driven by memoryful inputs, we must use Eq. (3.6) rather than Eq. (3.5).

3.4.3 Dynamical Ergodicity and Synchronization

To provide intuition behind the work expression of Eq. (3.7), let's now analyze the ratchet's operation. This reveals a novel synchronization mechanism that's responsible for nonzero work production. First, consider the case in which the engine parameters δ and γ are zero; that is, the state C is disconnected from A and B. This effectively deletes

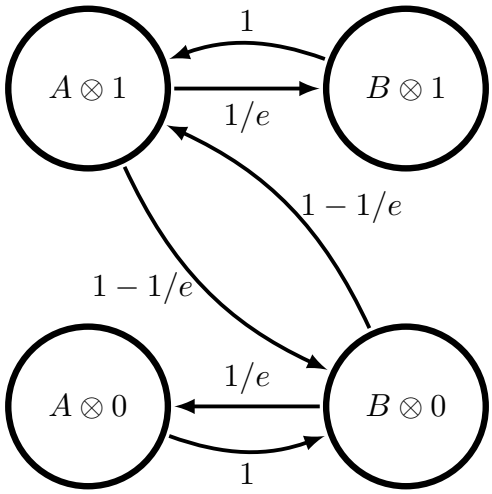


Figure 3.4. Ratchet dynamics absent the synchronizing state: Assuming system parameters δ and γ are set to zero, state C becomes inaccessible and the ratchet's joint state-symbol dynamics become restricted to that shown here—a truncated form of the dynamics of Fig. 3.3.

C from the joint dynamic, as shown in Fig. 3.4. This restricted model has the topology considered in our previous work [13].

It turns out that the ratchet has two equally likely dynamical modes, let's call them *clockwise* and *counterclockwise*. When in each mode, the ratchet behavior is periodic in time. The modes are depicted in Fig. 3.5, with the counterclockwise mode on the left and the clockwise mode on the right. The dashed (red) arrows show the paths taken through the joint state space due to an interaction transition followed by a switching transition when the switching transition is driven by input 0. And, the solid (blue) arrows show the paths taken when the switching transition is driven by input 1. The labels on the arrows indicate the amount of work done in the associated transitions. The clockwise mode extracts $k_B T/e$ amount of work per bit, while the counterclockwise mode expends $k_B T$ amount of work per bit.

There is a simple way to understand the existence and work performance of the two modes. Consider the counterclockwise mode first. The left state-transition diagram in Fig. 3.5 shows this mode arises when $A \otimes 0$ or $B \otimes 1$ happens to be the initial joint state. First, there is a horizontal interaction transition to a lower energy state. The energy difference $k_B T$ is fully dissipated in the thermal reservoir with no exchange of energy with

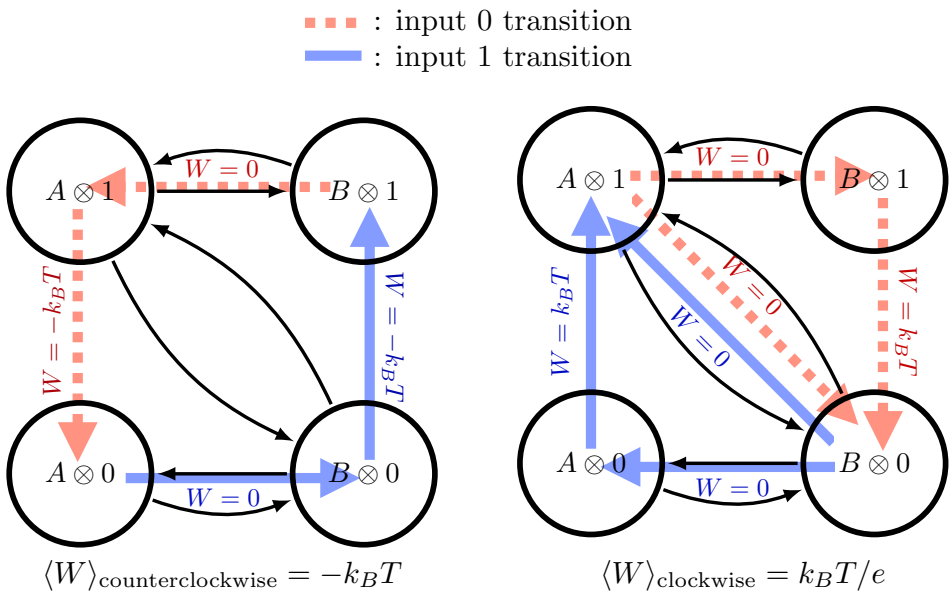


Figure 3.5. Two dynamical modes of the ratchet while driven by a period-2 input process: (a) *Counterclockwise* (left panel): ratchet is out of synchronization with the input tape and makes a steady counterclockwise rotation in the composite space of the Demon and the interacting cell. Work is steadily dissipated at the rate $-k_B T$ per pair of input symbols and no information is exchanged between the ratchet and the information reservoir. (b) *Clockwise* (right panel): ratchet is synchronized with the input correlated symbols on the tape, information exchange is nonzero, and work is continually accumulated at the rate $k_B T/e$ per pair of input symbols.

the work reservoir. Then, there is a vertical switching transition to a higher-energy state. The required energy $k_B T$ is taken from the work reservoir with no exchange of energy with the thermal reservoir. This energy is then dissipated as heat in the thermal reservoir at the next horizontal transition. The net amount of work produced per symbol—the net amount of energy supplied to the work reservoir—is $\langle W \rangle = -k_B T$.

Similarly, consider the clockwise mode. The righthand state-transition diagram in Fig. 3.5 shows that this mode arises when $A \otimes 1$ or $B \otimes 0$ is the initial joint state. First, there is an interaction transition along either the horizontal or diagonal paths of the Markov chain. (The horizontal transitions are opposite to those of the counterclockwise mode.) From microscopic reversibility, the horizontal interaction transitions lead to $k_B T$ energy taken from the thermal reservoir in order to move into higher energy states. No energy is exchanged with the work reservoir. On the diagonal transitions, on the other hand, no energy is exchanged with either reservoir. Then, there is a switching transition, which

corresponds to a vertical transition to a lower energy state if the horizontal interaction transition was made just before. The energy difference $k_B T$ is given to the work reservoir. However, if the diagonal transition was made, then the switching transition does not change the state and there is no work done. As shown in the figure, there are two possible paths the system can take between $A \otimes 1$ and $B \otimes 0$ in one operation cycle of the ratchet: $\{A \otimes 1 \to B \otimes 1 \to B \otimes 0\}$ and $\{A \otimes 1 \to B \otimes 0 \to B \otimes 0\}$. The same is true of transitions from $B \otimes 0$ to $A \otimes 1$: $\{B \otimes 0 \to A \otimes 0 \to A \otimes 1\}$ and $\{B \otimes 0 \to A \otimes 1 \to A \otimes 1\}$. Averaging over the probabilities of the two fundamental paths, the net average work produced is $\langle W \rangle = k_B T/e$. (See App. 3.7.2 for details.)

If the initial ratchet state is uncorrelated with the input HMM state, the clockwise and the counterclockwise modes occur with equal probability. Once in a particular mode, the ratchet cannot switch over to the other mode. In this sense, the two modes act as two different attractors for the Demon's joint state-symbol dynamics. In other words, the system is dynamically nonergodic, leading to nonergodic work production: either time averaged $-k_B T$ or $k_B T/e$. In this case, the ratchet dissipates on average $k_B T(1 - 1/e)/2$ units of energy from the work reservoir into the thermal reservoir as heat.

Comparing this ergodic ratchet, in which nonergodicity plays a dynamic and transient role, to the nonergodic engine discussed earlier is in order. Nonergodic engines (those driven by nonergodic input processes) can exhibit functional behavior when averaged over an ensemble of input realizations. As shown in Ref. [73], Maxwell's refrigerator [60] can refrigerate when driven by the nonergodic process consisting of two infinitely long realizations, one of all 0s and the other of all 1s. Similar to our ratchet driven by (ergodic) period-2 sequences, the refrigerator has two principal modes: the ratchet is driven by all 0s and refrigerates versus the ratchet is driven by all 1s and dissipates. However, one of these two modes is chosen at random in the beginning of a ratchet trial and remains fixed. This yields refrigeration that differs from the ensemble average (over the nonergodic input realizations). However, we can achieve robust and functional work production in our period-2 ratchet, by coupling modalities dynamically via the C state. Then, on every trial, the engine functions.

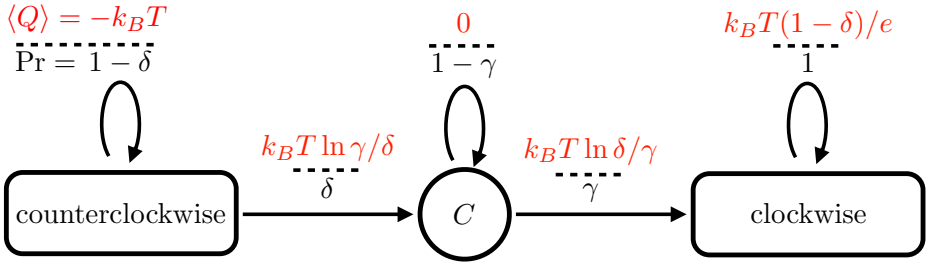


Figure 3.6. Crossover from the dissipative, counterclockwise mode to the generative, clockwise mode via synchronizing state C: Even though the microscopic dynamics satisfy time-reversal symmetry, a crossover is possible only from the counterclockwise mode to the clockwise mode because of the topology of the joint state space. With transitions between C and the other two modes, the engine becomes ergodic among its dynamic modes. The heats and transition probabilities are shown above each arrow.

Let's explain how its emergent nonergodicity makes this function robust. For $\delta \neq 0$ and $\gamma \neq 0$, state C becomes accessible to the ratchet, changing the stability of the counterclockwise attractor. And, this allows positive work production. (From here on we consider the original, full ratchet in Fig. 3.3.) We make a heuristic argument as to why the ratchet can generate positive net work using state C.

C's addition creates a "path" for the ratchet to shift from the dissipative, counterclockwise mode to the generative, clockwise mode. And, the latter becomes the only attractor in the system. In other words, the counterclockwise dynamical mode becomes a purely transient mode and the system becomes dynamically ergodic. The situation is schematically shown in Fig. 3.6, where the arrows denote allowed transitions in the dynamical sense. Heat and probability values of the transitions are shown there along each arrow. Recall that in the counterclockwise mode, the joint state is either $A \otimes 0$ or $B \otimes 1$ at the beginning of each interaction interval. According to the Markov model, both these states have probability δ of transitioning to a C state during interaction transitions. Thus, as depicted, δ is the probability of transitioning from the counterclockwise mode to C.

Once in state C, the ratchet cannot return to the counterclockwise mode, despite the fact there is probability γ of transitioning back to either $A \otimes 0$ or $B \otimes 1$ in an interaction transition. This is because the following switching transition immediately changes $A \otimes 0$ to $A \otimes 1$ and $B \otimes 1$ to $B \otimes 0$. That is, the system is in the clockwise mode at the beginning

of the next interaction interval. Thus, with probability γ the system makes a transition to the clockwise mode. After this transition, the system is necessarily synchronized, and it is impossible to transition out of the synchronized dynamic. In this way, the ratchet asymptotically extracts a positive amount of average heat from the environment, $\langle Q \rangle = k_B T(1-\delta)/e$ per symbol. Asymptotic heat extraction is the same as the work production for finite ratchets, confirming Eq. (3.7). Since the ratchet must move through C to arrive at the recurrent, clockwise, work-producing dynamic, we decide to start the ratchet in C. C serves as a synchronization state in that it is necessary for the ratchet state to synchronize to the input tape: once the ratchet transitions out of the C state, its internal states are synchronized with the input HMM states such that it produces work.

3.4.4 Trading-off Work Production Against Synchronization Rate and Work

With Fig. 3.6 in mind, we can define and calculate several quantities that are central to understanding the ratchet's thermodynamic functionality as functions of its parameters δ and γ: the *synchronization rate* R_{sync} and the *synchronization heat* Q_{sync} absorbed during synchronization. R_{sync} is the inverse of the average number of time steps until transitioning into the clockwise mode. It simplifies to the probability γ of transitioning into the clockwise mode:

$$\begin{aligned} R_{\text{sync}}(\delta, \gamma) &= \frac{1}{\langle t/\tau \rangle} \\ &= \frac{1}{\gamma \sum_{i=0}^{\infty} (i+1)(1-\gamma)^i} \\ &= \gamma \ . \end{aligned}$$

The heat Q_{sync} absorbed when synchronizing is the change in energy of the joint state as the ratchet goes from the synchronizing states ($C \otimes 0$ or $C \otimes 1$) into the recurrent synchronized states ($A \otimes 0$ or $B \otimes 1$):

$$Q_{\text{sync}}(\delta, \gamma) = k_B T \ln \frac{\delta}{\gamma} \ .$$

This is minus the energy dissipation required for synchronization.

Much like the speed, energy cost, and fidelity of a computation [107, 108, 15, 16], these two quantities and the average extracted work per symbol obey a three-way tradeoff in which each pair is inversely related, when holding the third constant. This is expressed most directly by combining the expressions above into a single relation that is independent of δ and γ:

$$Q_{\text{sync}} + k_B T \ln R_{\text{sync}} - k_B T \ln \left(1 - \frac{e\langle W \rangle}{k_B T}\right) = 0 \ . \tag{3.11}$$

Figure 3.7 illustrates this trade-off. Analytically, the same interdependence appears when taking the partial derivatives of the quantities with respect to each other:

$$\begin{aligned}
\frac{\partial Q_{\text{sync}}}{\partial \langle W \rangle} &= -\frac{-k_B T e}{k_B T - e\langle W \rangle} \ , \\
\frac{\partial Q_{\text{sync}}}{\partial R_{\text{sync}}} &= -\frac{-k_B T}{R_{\text{sync}}} \ , \text{ and} \\
\frac{\partial \langle W \rangle}{\partial R_{\text{sync}}} &= -\frac{k_B T - e\langle W \rangle}{e R_{\text{sync}}} \ .
\end{aligned}$$

These all turn out to be negative over the physical range of parameters: $\langle W \rangle \in (-\infty, 1/e]$, $R_{\text{sync}} \in [0, 1]$, and $Q_{\text{sync}} \in (-\infty, \infty)$.

The ratchet's successful functioning derives from the fact that it exhibits a dynamical mode that "resonates" with the input process correlation in terms of work production and that this mode can be made the only dynamical attractor. In other words, an essential element in constructing our ratchet its ability to synchronize its internal states with the effective states of the input process. This appears to be a basic principle for leveraging memoryful input processes and, more generally, correlated environments.

3.5 Fluctuating Correlated Environments

The preceding development considered a perfectly correlated environment that generates an input to the ratchet in which a 0 is always followed by 1 and a 1 by 0. Of course, this is an artificial and constrained input. It's purpose, though, was to isolate the role of structured, correlated environment signals and how a thermodynamic ratchet can leverage that order to function as an engine. Practically, though, it is hard to come by such perfectly correlated sequences in Nature. One expects sequences to involve errors, say

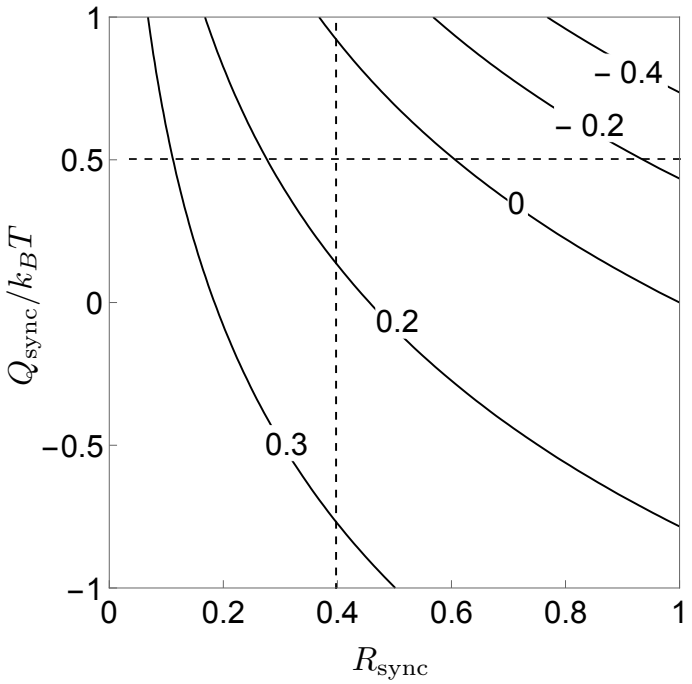


Figure 3.7. Trade-off between average work production, synchronization rate, and synchronization heat: Contour plot of average extracted work per symbol $\langle W \rangle$ as a function of rate of synchronization R_{sync} and synchronization heat Q_{sync} using Eq. (3.11). Work values are in the unit of k_BT. Numbers labeling contours denote the average extracted work $\langle W \rangle$. If we focus on any particular contour, increasing R_{sync} leads to a decrease in Q_{sync} and vice versa. Similarly, restricting to a fixed value of R_{sync}, say the vertical $R_{\text{sync}} = 0.4$ line, increasing Q_{sync} decreases values of $\langle W \rangle$. Restricting to a fixed vale of Q_{sync}, say the horizontal $Q_{\text{sync}} = 0.5$ line, increasing R_{sync} going to the right also decreases $\langle W \rangle$.

where a 0 is sometimes followed by a 0 and a 1 by 1. Such *phase slips* are one kind of error with which a thermodynamically functioning ratchet must contend.

In particular, whenever a phase slip occurs the ratchet is thrown out of its synchronization with the input, possibly into the dissipative, counterclockwise dynamical mode. Due to the presence of the synchronizing mechanism, shown in Fig. 3.6, the ratchet can recover via transiting through the synchronizing state C. If the frequency of phase slips is sufficiently low, then, the ratchet can still produce work, only at a lower rate. If the phase slip frequency is high enough, however, the ratchet does not have sufficient time in the clockwise mode to recover the work lost in the counterclockwise mode before it relaxed to the clockwise mode. At this error level the ratchet stops producing work; it

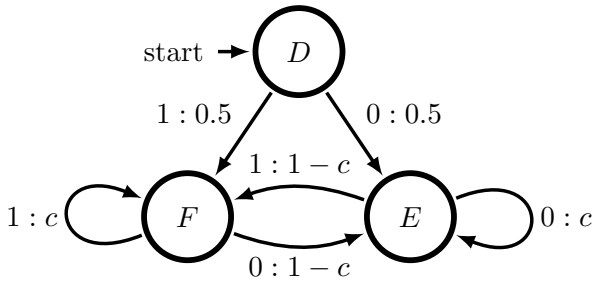


Figure 3.8. Noisy phase-slip period-2 (NPSP2) process: As with the exact period-2 process of Fig. 3.2, its HMM has a transient start state D and two recurrent causal states E and F. Starting from D, the process makes a transition either to E or F with equal probabilities while outputting 0 or 1, respectively. Once in state E, the process either stays with probability c and outputs a 0 or makes a transition to state F with probability $1-c$ and outputs a 1. If in state F, the process either stays with probability c and outputs a 1 or makes a transition to state E with probability $1-c$ and outputs a 0. For small nonzero c, the output is no longer a pure alternating sequence of 0s and 1s, but instead randomly breaks the period-2 phase. For $c = 1/2$, the generated sequences are flips of a fair coin. The process reduces to that in Fig. 3.2, if $c = 0$.

dissipates work even on average. This suggests there is a critical level of input errors where a transition from a functional to nonfunctional ratchet occurs. This section analyzes the transition, giving an exact expression for the critical phase-slip frequency at which the ratchet stops producing work.

To explore the ratchet's response to such errors, we introduce phase slips into the original period-2 input process. They occur with a probability c, meaning that after every transition, there is a probability c of emitting the same symbol again and remaining in the same hidden state rather than emitting the opposite symbol and transitioning to the next hidden state. An HMM corresponding this period-2 phase-slip dynamics is shown in Fig. 3.8—the *noisy phase-slip period-2* (NPSP2) process. It reduces to the original, exactly periodic process generated by the HMM in Fig. 3.2 when $c = 0$.

It is now straightforward to drive the ratchet (Fig. 3.3) inputs with the NPSP2 process (Fig. 3.8) and calculate exactly the average work production per symbol using Eq. (3.1). Appendix 3.7.2 does this for all values of δ, γ, and c. Here, let's first consider the special case of $\gamma = 1$. This is the regime in which the ratchet is most functional as an engine since, if the ratchet produces positive work, then $\gamma = 1$ maximizes that work production.

With $\gamma = 1$, once the ratchet is in state C, it immediately synchronizes in the next

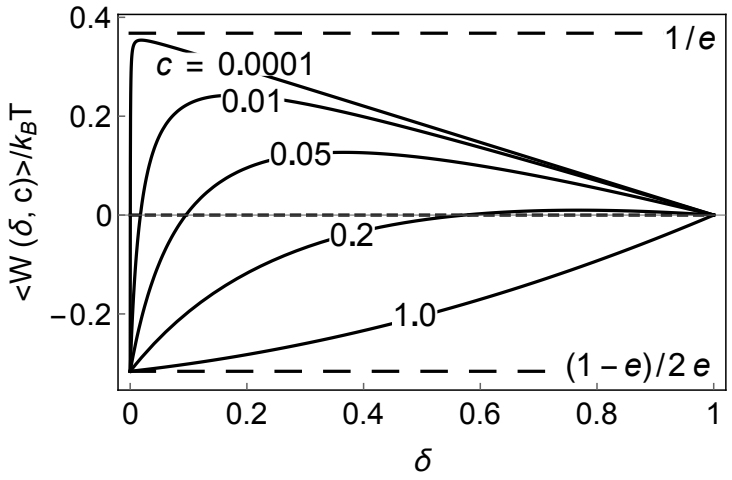


Figure 3.9. Average work production per symbol versus parameter δ and phase slip rate c at fixed $\gamma = 1$. Labels on the curves give c values. Long-dashed lines give the upper and lower bounds on work production: It is bounded above by k_BT/e and below by $k_BT(1-e)/(2e)$. (See text.)

interaction interval. In this case, δ parametrizes the relationship between the average work done when synchronized and the rate of synchronization. The higher δ is, the less work the ratchet extracts while synchronized, but the more often it transitions to the synchronizing state—recall Fig. 3.6—allowing it to recover from phase slips. The calculation for $\gamma = 1$ yields an average work rate (App. 3.7.2):

$$\langle W \rangle(\delta, c) = \frac{(1-\delta)[\delta + c - c(2\delta + e)]}{2ec + \delta e(1-c)} \ . \tag{3.12}$$

Thus, over the whole parameter space $c, \delta \in [0, 1]$, the average work varies over the range:

$$\langle W \rangle(\delta, c) \in \frac{k_BT}{e} \left[-\frac{e-1}{2}, 1 \right] \ .$$

Figure 3.9 shows how the work production varies with δ for different values of c. No matter the value of c, at $\delta = 0$ the average work attains its lower limit of $-k_BT(e-1)/2e$, which is the average work produced when both clockwise and counterclockwise modes have equal probability. As δ increases, there is an increase in the the average work until it reaches 0 at a particular value $\delta^*(c)$. Below $\delta^*(c)$—i.e., within the range $0 \leq \delta \leq \delta^*(c)$—the system consumes work; whereas above $\delta^*(c)$, the system acts as an engine, producing net positive work. Figure 3.9 shows that $\delta^*(c)$ is an increasing function of c, starting with 0^+ as c tends to 0 and ending up at 1 as c tends to unity.

The dependence is nonlinear, with sharp changes near $c = 0$ and saturating near $c = 1$. Since the average work vanishes as δ tends to 1 independent of c, there is a value of $\delta_{\max}(c)$ where the engine's work production is maximum. This maximum work $W_{\max}(c)$ is closer to its upper limit $k_B T/e$ for smaller values of c. As we increase c, there is a decrease in $W_{\max}(c)$ until it vanishes at $\delta = 1$.

Figure 3.10 shows the dependence of $W_{\max}(c)$ as a function of error rate c, revealing a critical value $c^* = 1/1 + e$ beyond which $W_{\max}$ vanishes. Thus, if the phase-slip frequency is too high, the ratchet cannot produce net positive work regardless of how quickly it synchronizes. This special value c^* actually partitions the expressions for $\delta^*(c)$, $\delta_{\max}(c)$, and $W_{\max}(c)$ into piecewise functions:

$$\delta^*(c) = \begin{cases} \frac{(1-e)c}{2c-1} & \text{if } c \leq \frac{1}{1+e} \\ 1 & \text{if } c > \frac{1}{1+e} \end{cases} \tag{3.13}$$

$$\delta_{\max}(c) = \begin{cases} \frac{4c^2+\alpha-2c}{2c^2-3c+1} & \text{if } c \leq \frac{1}{1+e} \\ 1 & \text{if } c > \frac{1}{1+e} \end{cases} \tag{3.14}$$

$$W_{\max}(c) = \begin{cases} k_B T \frac{-2\alpha+c(e-(5+e)c)+1}{e(c-1)^2} & \text{if } c \leq \frac{1}{1+e} \\ 0 & \text{if } c > \frac{1}{1+e} \end{cases}, \tag{3.15}$$

where $\alpha = \sqrt{c\,(2c^2 + c - 1)\,((3 + e)c - e - 1)}$.

The results in Fig. 3.10 should not be applied too broadly. They do not imply that positive net work cannot be extracted for the case $c > c^*$ for *any* information ratchet. On the contrary, there exist alternatively designed ratchets that can extract positive work even at $c = 1$. However, the design of such ratchets differs substantially from the current one. Sequels will take up the task of designing and analyzing this broader class of information engines.

Figure 3.11 combines the results in Figs. 3.9 and 3.10 into phase diagram summarizing the ratchet's thermodynamic functionality. It illustrates how the δ-c parameter space splits into two regions: the leftmost (red) region where the ratchet produces work, behaving as an engine, and the lower right (gray) region where the ratchet consumes work, behaving as either an information eraser (using work to erase information in the bit

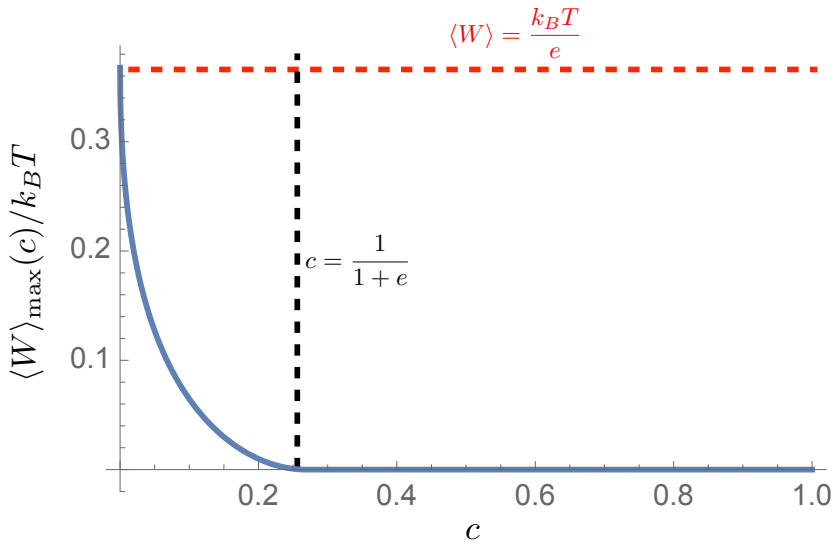


Figure 3.10. Maximum work production versus phase-slip rate c: Maximum work production decreases with c from $k_B T/e$ at $c = 0$ to 0 when $c \geq c^* = 1/(1+e)$.

string) or a dud (dissipating work without any erasure of information). It also shows that $c^* = 1/(1+e)$ corresponds to both the point at which $\delta_{\max}(c)$ reaches 1 and the point at which it is no longer possible to extract work from the input, independent δ. This is the point where phase slips happen so often that the ratchet finds it impossible to synchronize for long enough to extract any work.

3.6 Conclusion

We extended the functionality of autonomous Maxwellian Demons by introducing a new design for information engines that is capable of extracting work purely out of temporal correlations in an information source, characterized by an input HMM. This is in marked contrast with previous designs that can only leverage a statistically biased information source or the mutual, instantaneous correlation between a pair of information sources [7, 60, 61, 13, 10]. Our new design is especially appropriate for actual physical construction of information engines since physical, chemical, and biological environments (information sources) almost always produce temporally correlated signals.

The new design was inspired by trying to resolve conflicting bounds on the work production for information engines. On the one hand, Eq. (3.5) for monitoring information content only of isolated symbols suggests that no work can be produced from temporal

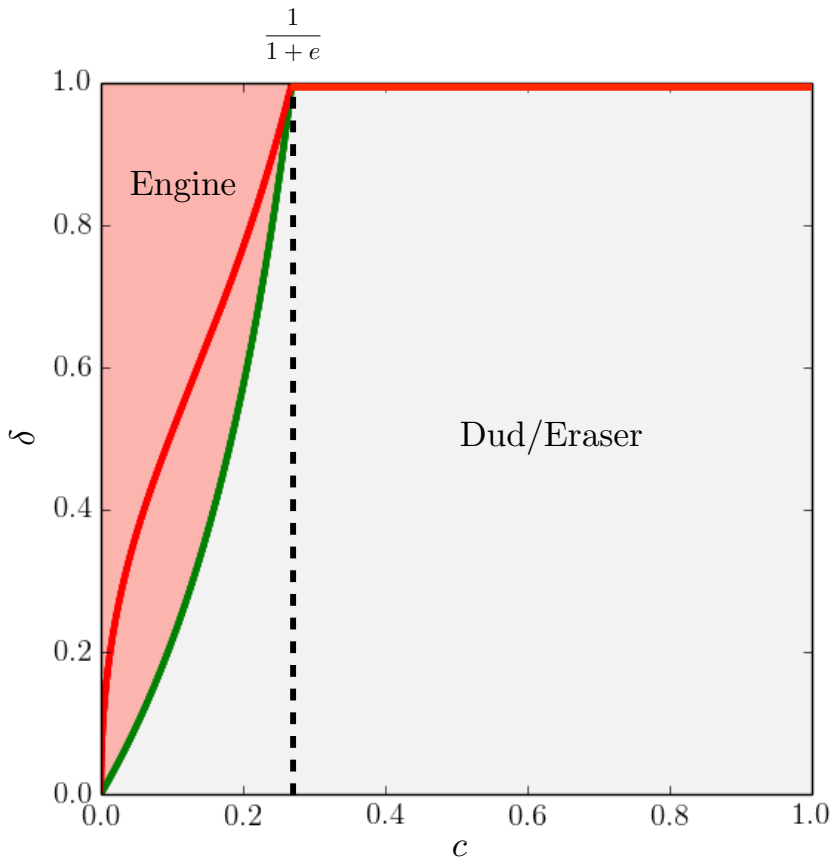


Figure 3.11. Ratchet thermodynamic-function phase diagram: In the leftmost (red) region, the ratchet behaves as an engine, producing positive work. In the lower right (gray) region, the ratchet behaves as either an eraser (dissipating work to erase information) or a dud (dissipating energy without erasing information). The solid (red) line indicates the parameter values that maximize the work output in the engine mode. The dashed (black) line indicates the critical value of c above which the ratchet cannot act as an engine.

correlations in input string; whereas, on the other, using entropy rates Eq. (3.6) indicates these correlations are an excellent resource. We showed, in effect, that this latter kind of correlational information is a thermodynamic fuel.

To disambiguate the two bounds, we described the exact analytical procedure to calculate the average work production for an arbitrary nonequilibrium memoryful channel and a HMM input process. The result is that it is now abundantly clear which bounds hold for correlated input processes.

We considered the specific example of a period-2 process for the input tape (Fig. 3.2), since it has structure in its temporal correlations, but no usable single-symbol information content. The ratchet we introduced to leverage this input process requires three memory states (Fig. 3.3) to produce positive work. This memoryful ratchet with a memoryful input process violates Eq. (3.5), establishing Eq. (3.6) as the proper information processing Second Law of thermodynamics.

It is intuitively appealing to think that ratchet memory must be in consonance with the input process' memory to generate positive work. In other words, the ratchet must be memoryful and be able to synchronize itself to the structured memory of the input HMM to be functional. We confirmed that this is indeed the case in general with our expression for work. If the ratchet has no memory, the only "structure" of consequence in the input process is simply, provably, the isolated-symbol statistical bias.

We see this nascent principle more concretely in the operation of the ratchet as it responds to the period-2 process. Critical to its behaving as an engine is the presence of state C (Fig. 3.3) through which the ratchet synchronizes itself to the input. As shown in Fig. 3.6, the synchronizing state C allows the system to make an irreversible transition from the counterclockwise, dissipative mode into the generative, clockwise mode. It demonstrates how key it is that the ratchet's effective memory match that of the input process generator.

We also discovered an intriguing three-way tradeoff (Fig. 3.7) between synchronization rate, synchronization heat (that absorbed during synchronization), and asymptotic average work production. For example, if the Demon keeps the synchronization rate fixed and

increases the synchronization heat, there is a decrease in the average work production. In other words, if the Demon becomes greedy and tries to extract energy from the thermal reservoir even during synchronization, on the one hand, it is left with less work in the end. If, on the other hand, the Demon actually supplies heat during the synchronization step, it gains more work in the end! Similarly, if it keeps the synchronization heat fixed, a slower rate of synchronization is actually better for the average work production. If the Demon waits longer for the ratchet to synchronize with its environment, it is rewarded more in terms of the work production. Thus, the Demon is better off in terms of work, by being patient and actually supplying more energy during synchronization. This three-way tradeoff reminds one of a recently reported tradeoff between the rate, energy production, and fidelity of a computation [15].

We then considered the robustness of our design in a setting in which the input process is not perfectly periodic, but has random phase slips (Fig. 3.8). As a result, the dissipative regime is no longer strictly transient. Every so often, the ratchet is thrown into the dissipative regime induced by the phase slips, after which the ratchet attempts to resynchronize to the generative mode. Thus, the ratchet seems remarkably robust with respect to the phase-slip errors, being able to dynamically correct its estimation of the input's hidden state due to the synchronization mechanism. This is true, however, only up to a certain probability of phase slips, beyond which the dissipative regime is simply too frequent for the ratchet to generate any work. For the region in which the ratchet is capable of generating work, we found the parametric combination for its optimal functionality for a given probability of phase slips (Fig. 3.9). We also determined the maximum net work that the ratchet can produce (Fig. 3.10). Finally, we gave a phase diagram of the ratchet's thermodynamic functionality over the control parameter space formed by δ and c for $\gamma = 1$ (Fig. 3.11).

In this way, we extended the design of information engines to include memoryful input processes and memoryful ratchets. The study suggests, via synchronization and dynamical self-correction, there are general principles that determine how autonomous devices and organisms can leverage arbitrary structure in their environments to extract

thermodynamic benefits.

Physical systems that demonstrate the thermodynamic equivalent of information processing are by now numerous. Most, in contrast to the present design, restrict themselves to single-step information processing. Moreover, many only consider information processing comprising the erasure of a single bit, staying within the setting of Landauer's Principle. The information-processing equivalence principle strongly suggests a much wider set of computational possibilities that use the capacity of stored information as a thermodynamic resource.

Practically implementing an information engine on the nanoscale, say, will require delicate control over system and materials properties. To achieve this in a convincing way will demand an unprecedented ability to measure heat and work. This has become possible only recently using single-electron devices [109], nanoelectronic mechanical systems (NEMS) [110, 111], and Bose-Einstein Condensates (BECs) [112, 113, 114]. The results and methods outlined here go some distance to realizing these possibilities by pointing to designs that are functionally robust and resilient, by identifying efficient information engines and diagnosing their operation, and by giving exact analytical methods for the quantitative predictions necessary for implementation.

3.7 Appendices

3.7.1 Ratchet Energetics: General Treatment

Here, we lay out the detailed calculations of the thermodynamic contributions made by the ratchet's transducer and the environmental input process.

3.7.1.1 Transducer Thermodynamic Contributions

We consider the case where the ratchet exchanges energy only with the work reservoir during the switching transitions and only with the heat reservoir during the interaction transitions. During the N-th switching transition, the ratchet "exhausts" the N-th input bit Y_N as the N-th output bit Y'_N and couples with the input bit Y_{N+1}. The joint state of the ratchet and the interacting bit changes from $X_{N+1} \otimes Y'_N$ to $X_{N+1} \otimes Y_{N+1}$. The corresponding decrease in energy is supplied to the work reservoir. So, the work output

at the N-th switching transition W_N is given by:

$$W_N = E_{x_{N+1} \otimes y'_N} - E_{x_{N+1} \otimes y_{N+1}} \ , \tag{3.16}$$

where $E_{x \otimes y}$ denotes the energy of the joint state $x \otimes y$. Via a similar argument, we write the heat absorbed by the ratchet during the N-th interaction transition Q_N:

$$Q_N = E_{x_{N+1} \otimes y'_N} - E_{x_N \otimes y_N} \ .$$

The main interest is in determining the asymptotic rate of work production:

$$\begin{aligned}
\langle W \rangle &= \lim_{N \to \infty} W_N \Pr(W_N) \\
&= \lim_{N \to \infty} \sum_{\substack{x_{N+1}, \\ y_{N+1}, y'_N}} (E_{x_{N+1} \otimes y'_N} - E_{x_{N+1} \otimes y_{N+1}}) \\
&\qquad \times \Pr(X_{N+1} = x_{N+1}, Y_{N+1} = y_{N+1}, Y'_N = y'_N) \\
&= \sum_{x', y'} E_{x' \otimes y'} \lim_{N \to \infty} \Pr(X_{N+1} = x', Y'_N = y') \\
&\qquad - \sum_{x, y} E_{x \otimes y} \lim_{N \to \infty} \Pr(X_{N+1} = x, Y_{N+1} = y) \ ,
\end{aligned} \tag{3.17}$$

where the second line uses Eq. (3.16) and the third relabels the realizations in the sum x and x', since these are dummy variables in separate sums.

Assuming the stationary distribution over the input variable and ratchet variable exists, the asymptotic probability $\lim_{N \to \infty} \Pr(X_{N+1} = x, Y_{N+1} = y)$ is the same as the asymptotic probability $\lim_{N \to \infty} \Pr(X_N = x, Y_N = y)$, which was defined as $\pi_{x \otimes y}$. In addition, note that the Markov matrix M controlling the joint ratchet-bit dynamic is stochastic, requiring $\sum_{x', y'} M_{x \otimes y \to x' \otimes y'} = 1$ from probability conservation. As a result,

the second summation in Eq. (3.17) is equal to:

$$\begin{aligned}
&-\sum_{x,y} E_{x\otimes y} \lim_{N\to\infty} \Pr(X_{N+1}=x, Y_{N+1}=y) \\
&\qquad = -\sum_{x,y} E_{x\otimes y}\pi_{x\otimes y} \\
&\qquad = -\sum_{x,y} E_{x\otimes y}\pi_{x\otimes y} \sum_{x',y'} M_{x\otimes y\to x'\otimes y'} \\
&\qquad = -\sum_{\substack{x,x',\\ y,y'}} E_{x\otimes y}\pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ .
\end{aligned}$$

To compute the first term in Eq. (3.17), we do a similar decomposition. Note that X_{N+1} and Y'_N are determined from X_N and Y_N by iterating with the joint Markov dynamic M, and so:

$$\begin{aligned}
&\Pr(X_{N+1}=x', Y'_N=y') \\
&\qquad = \sum_{x,y} \Pr(X_N=x, Y_N=y) M_{x\otimes y\to x'\otimes y'} \ . \qquad (3.18)
\end{aligned}$$

Using Eq. (3.18) we rewrite the first summation in Eq. (3.17) as:

$$\begin{aligned}
&\sum_{x',y'} E_{x'\otimes y'} \lim_{N\to\infty} \Pr(X_{N+1}=x', Y'_N=y') \\
&\quad = \sum_{x,y,x',y'} E_{x'\otimes y'} \lim_{N\to\infty} \Pr(X_N=x, Y_N=y) M_{x\otimes y\to x'\otimes y'} \\
&\quad = \sum_{x,y,x',y'} E_{x'\otimes y'}\pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ .
\end{aligned}$$

Combining the above, the resulting work production rate is:

$$\langle W\rangle = \sum_{x,x',y,y'} (E_{x'\otimes y'} - E_{x\otimes y})\pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ .$$

The same logic leads to the average heat absorption, which turns out to the same as the work production:

$$\langle Q\rangle = \langle W\rangle \ .$$

The intuition for this is that these equalities depend on the existence of the stationary distribution $\pi_{x\otimes y}$ over the ratchet and bit. This is guaranteed for a finite ratchet with

mixing dynamics. Only a finite amount of energy can be stored in a finite ratchet, so the heat energy flowing in must be the same as the work flowing out, on the average, to conserve energy. This, however, breaks down with infinite-state ratchets—an important and intriguing case which is addressed in the next chapter.

3.7.1.2 Input Process Contributions

The results above are expressed in terms of the ratchet, except for the stationary joint distribution over the input variable and ratchet state:

$$\pi_{x\otimes y} = \lim_{N\to\infty} \Pr(X_N = x, Y_N = y) \ .$$

This quantity is dependent on the input process, as we now describe. We describe the process generating the input string by an HMM with transition probabilities:

$$T^{(y_N)}_{s_N \to s_{N+1}} = \Pr(Y_N = y_N, S_{N+1} = s_{N+1} | S_N = s_N) \ , \tag{3.19}$$

where $s_i \in \mathcal{S}$ are the input process' hidden states [13]. Given that the input HMM is in internal state s_N, $T^{(y_N)}_{s_N \to s_{N+1}}$ gives the probability to make a transition to the internal; state s_{N+1} and produce the symbol y_N. The dependence between X_N and Y_N is determined by hidden state S_N. So, we rewrite:

$$\begin{aligned}
\Pr(X_N = x, Y_N = y) &= \sum_s \Pr(X_N = x, Y_N = y, S_N = s) \\
&= \sum_s \Pr(Y_N = y | S_N = s) \Pr(X_N = x, S_N = s) \\
&= \sum_{s,s'} \Pr(Y_N = y, S_{N+1} = s' | S_N = s) \Pr(X_N = x, S_N = s) \\
&= \sum_{s,s'} T^{(y)}_{s \to s'} \Pr(X_N = x, S_N = s) \ ,
\end{aligned}$$

The second line used the fact that Y_N depends on only S_N, as illustrated in Fig. 3.12. The last line used Eq. (3.19).

Combining the above equations gives:

$$\begin{aligned}
\pi_{x\otimes y} &= \lim_{N\to\infty} \Pr(X_N = x, Y_N = y) \\
&= \lim_{N\to\infty} \sum_{s,s'} T^{(y)}_{s\to s'} \Pr(X_N = x, S_N = s) \\
&= \sum_{s,s'} T^{(y)}_{s\to s'} \pi'_{x\otimes s} \ , \text{ and} \\
\pi'_{x\otimes s} &= \lim_{N\to\infty} \Pr(X_N = x, S_N = s) \ .
\end{aligned}$$

Thus, evaluating $\pi_{x\otimes y}$ requires knowing the input process $T^{(y)}_{s\to s'}$, which is given, and the stationary joint distribution $\pi_{x\otimes s}$ over the hidden states and the ratchet states.

To calculate $\pi_{x\otimes s}$, we must consider how X_{N+1} and S_{N+1} are generated from past variables. We notice that the output process is specified by an HMM whose hidden variables are composed of the hidden variable of the input HMM and the states of the transducer. In other words, the output HMM's hidden states belong to the product space $\mathcal{X} \otimes \mathcal{S}$. As a result, the transition probability of the output HMM is:

$$\begin{aligned}
T^{\prime(y')}_{x\otimes s\to x'\otimes s'} &= \Pr(Y'_N = y', X_{N+1} = x', S_{N+1} = s' | X_N = x, S_N = s) \\
&= \sum_y \Pr(Y'_N = y', X_{N+1} = x', Y_N = y, S_{N+1} = s' | X_N = x, S_N = s) \\
&= \sum_y \Pr(Y'_N = y', X_{N+1} = x | Y_N = y, S_{N+1} = s', X_N = x, S_N = s) \\
&\times \Pr(Y_N = y, S_{N+1} = s' | X_N = x, S_N = s) \\
&= \sum_y \Pr(Y'_N = y', X_{N+1} = x | Y_N = y, X_N = x) \Pr(Y_N = y, S_{N+1} = s' | S_N = s) \\
&= \sum_y M_{x\otimes y\to x'\otimes y'} T^{(y)}_{s\to s'} \\
&= \sum_y M^{(y'|y)}_{x\to x'} T^{(y)}_{s\to s'} \ ,
\end{aligned}$$

where the fourth used the facts that Y'_N and X_{N+1} are independent of S_N and S_{N+1}, if Y_N and X_N are known, and Y_N and S_{N+1} are independent of X_N, if S_N is known [1, 13].

Thus, summing over the output variable Y' yields a Markov dynamic over $\mathcal{X} \otimes \mathcal{S}$:

$$\begin{aligned} T'_{x\otimes s \to x'\otimes s'} &= \sum_{y'} T'^{(y')}_{x\otimes s \to x'\otimes s'} \\ &= \sum_{y,y'} M_{x\otimes y \to x'\otimes y'} T^{(y)}_{s\to s'} \ . \end{aligned} \tag{3.20}$$

The stationary distribution $\pi'_{x\otimes s}$ is this dynamics' asymptotic distribution:

$$\sum_{x,s} \pi'_{x\otimes s} T'_{x\otimes s \to x'\otimes s'} = \pi'_{x'\otimes s'} \ . \tag{3.21}$$

π' existence—that is, for a finite state Markov process like T'—is guaranteed by the Perron-Frobenius theorem and it is unique when T' is ergodic [115]. In short, we see that $\pi_{x\otimes y}$ is computable given the ratchet $M_{x\otimes y\to x'\otimes y'}$ and the input process generator $T^{(y)}_{s\to s'}$.

In this way, we derived an expression for the asymptotic work production of an arbitrary memoryful ratchet with an arbitrary memoryful input process in terms of HMM generator of the input and the Markovian dynamic over the input bit and ratchet state. Only a single assumption was made: there is an asymptotic distribution over the the input bit and ratchet state $\pi_{x\otimes y}$. In summary, there are three steps to calculate the average work production:

1. Calculate the stationary distribution $\pi'_{x\otimes s}$ over the hidden states of the output process $T'^{(y')}_{x\otimes s\to x'\otimes s'}$. The latter which is calculated from the operation of $M_{x\otimes y\to x'\otimes y'}$ on $T^{(y)}_{s\to s'}$;

2. Use π' and $T^{(y)}_{s\to s'}$ to calculate the stationary distribution over the ratchet and input bit at the beginning of the interaction interval $\pi_{x\otimes y}$;

3. Using this and the transducer's Markov dynamic, calculate the work production:

$$\langle W \rangle = k_B T \sum_{\substack{x,x',\\ y,y'}} \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ln \frac{M_{x'\otimes y'\to x\otimes y}}{M_{x\otimes y\to x'\otimes y'}}. \tag{3.22}$$

The following Appendix shows how to use this method to calculate average work production for the specific cases of the period-2 environment with and without phase-slips.

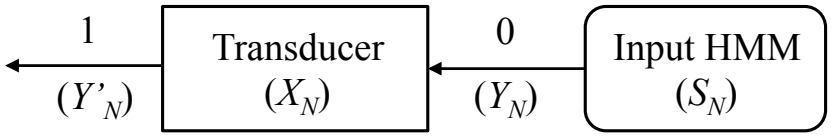


Figure 3.12. State variable interdependence: Input HMM has an autonomous dynamics with transitions $S_N \to S_{N+1}$ leading to input bits Y_N. That is, Y_N depends only on S_N. The joint dynamics of the transducer in state X_N and the input bit Y_N leads to the output bit Y'_N. In other words, Y'_N depend on X_N and Y_N or, equivalently, on X_N and S_N. Knowing the joint stationary distribution of X_N and S_N, then determines the stationary distribution of Y'_N. However, if Y_N and X_N are known, Y'_N is independent of S_N.

3.7.2 Ratchet Energetics: Specific Expressions

The symbol-labeled transition matrices for the noisy period-2 input process are given by:

$$T^{(0)} = \begin{pmatrix} 0 & 0 & 0 \\ .5 & c & 1-c \\ 0 & 0 & 0 \end{pmatrix} \begin{matrix} D \\ E \\ F \end{matrix}$$

$$T^{(1)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ .5 & 1-c & c \end{pmatrix} .$$

The transducer form of the ratchet M shown in Fig. 4 is given by the four conditional symbol-labeled transition matrices:

$$M^{(0|0)} = \begin{pmatrix} 0 & \frac{1-\delta}{e} & \gamma \\ 1-\delta & 0 & 0 \\ \delta & 0 & 1-\gamma \end{pmatrix} \begin{matrix} A \\ B \\ C \end{matrix}$$

$$M^{(1|0)} = \begin{pmatrix} 0 & 1-\frac{1-\delta}{e} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$M^{(0|1)} = \begin{pmatrix} 0 & 0 & 0 \\ 1-\frac{1-\delta}{e} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$M^{(1|1)} = \begin{pmatrix} 0 & 1-\delta & 0 \\ \frac{1-\delta}{e} & 0 & \gamma \\ 0 & \delta & 1-\gamma \end{pmatrix} ,$$

where we switched to the transducer representation of the joint Markov process $M_{x\otimes y\to x'\otimes y'} = M^{(y'|y)}_{x\to x'}$ [1, 13].

To find the stationary distribution over the causal states of the input bit and the internal states of the ratchet (step 1), we calculate the output process $T'^{(y')}_{x\otimes s\to x'\otimes s'} = \sum_y M^{(y'|y)}_{x\to x'} T^{(y)}_{s\to s'}$ and sum over output symbols to get the Markov dynamic over the hidden

states:

$$T' = T'^{(0)} + T'^{(1)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & c & \bar{c} & 0.5\gamma & c\gamma & \gamma\bar{c} \\ 0 & 0 & 0 & 0.5\bar{\delta} & \bar{c}\bar{\delta} & c\bar{\delta} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5\bar{\delta} & c\bar{\delta} & \bar{c}\bar{\delta} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & \bar{c} & c & 0 & 0 & 0 & 0.5\gamma & \gamma\bar{c} & c\gamma \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5\delta & c\delta & \delta\bar{c} & 0 & 0 & 0 & 0.5\bar{\gamma} & c\bar{\gamma} & \bar{c}\bar{\gamma} \\ 0 & 0 & 0 & 0.5\delta & \delta\bar{c} & c\delta & 0.5\bar{\gamma} & \bar{c}\bar{\gamma} & c\bar{\gamma} \end{pmatrix} \begin{matrix} A \otimes D \\ A \otimes E \\ A \otimes F \\ B \otimes D \\ B \otimes E \\ B \otimes F \\ C \otimes D \\ C \otimes E \\ C \otimes F \end{matrix},$$

where $\bar{c} = 1 - c$, $\bar{\delta} = 1 - \delta$, and $\bar{\gamma} = 1 - \gamma$.

Then, we find the stationary state π' over the joint hidden states (step 2), which solves $T'\pi' = \pi'$:

$$\pi' = \begin{pmatrix} \pi'_{A\otimes D} \\ \pi'_{A\otimes E} \\ \pi'_{A\otimes F} \\ \pi'_{B\otimes D} \\ \pi'_{B\otimes E} \\ \pi'_{B\otimes F} \\ \pi'_{C\otimes D} \\ \pi'_{C\otimes E} \\ \pi'_{C\otimes F} \end{pmatrix} = \begin{pmatrix} 0 \\ \gamma(\delta + c - \delta c)/\nu \\ \gamma(c - \delta c)/\nu \\ 0 \\ \gamma(c - \delta c)/\nu \\ \gamma(\delta + c - \delta c)/\nu \\ 0 \\ \delta c/\nu \\ \delta c/\nu \end{pmatrix},$$

where $\nu = 2(c\delta + \gamma(\delta + 2c - 2\delta c))$.

And, we find the stationary distribution over the ratchet state input bit by plugging

in to the equation $\pi_{x\otimes y} = \sum_{s,s'} T^{(y)}_{s\to s'} \pi'_{x\otimes s}$. The result is:

$$\pi = \begin{pmatrix} \pi_{A\otimes 0} \\ \pi_{A\otimes 1} \\ \pi_{B\otimes 0} \\ \pi_{B\otimes 1} \\ \pi_{C\otimes 0} \\ \pi_{C\otimes 1} \end{pmatrix} = \begin{pmatrix} \gamma c/\nu \\ \gamma(\delta + c - 2\delta c)/\nu \\ \gamma(\delta + c - 2\delta c)/\nu \\ \gamma c/\nu \\ \delta c/\nu \\ \delta c/\nu \end{pmatrix} .$$

Substituting this stationary distribution into the work expression (step 3) in Eq. (3.22), we find an explicit expression for the ratchet's work production rate:

$$\langle W \rangle = k_B T \frac{(1-\delta)(\delta + c - 2\delta c - ec)}{ec\delta/\gamma + e(\delta + 2c - 2\delta c)} . \tag{3.23}$$

3.7.2.1 Period-2 Input

To restrict to period-2 input sequences with no phase slips we set $c = 0$. Then, T' has the stationary distribution:

$$\pi'_{A\otimes E} = \pi'_{B\otimes F} = 0.5 ,$$

and all other elements vanish. The ratchet is fully synchronized to the internal states of the input process. Substituting $c = 0$ into Eq. (3.23) gives the work production rate when synchronized:

$$\langle W \rangle = k_B T \frac{1-\delta}{e} .$$

3.7.2.2 Noisy Period-2 Input

What happens when the environment fluctuates, generating input sequence phase slips with probability c? Consider the optimal parameter settings at which the ratchet generates work. When the ratchet behaves as an engine, the optimal setting is $\gamma = 1$, which follows from the partial derivative of the work production:

$$\frac{\partial \langle W \rangle}{\partial \gamma} = \langle W \rangle \frac{ec\delta}{\gamma^2(ec\delta/\gamma + e(\delta + 2c(1-\delta))} ,$$

which is always positive when the engine produces work. This means that it is always possible to enhance our engine's power by increasing γ to its maximum value at $\gamma = 1$. And so, to build an optimal engine that leverages the noisy period-2 input process, we set $\gamma = 1$, yielding:

$$\langle W \rangle(\delta, c, \gamma = 1) = k_B T \frac{(1-\delta)[\delta + c - c(2\delta + e)]}{2ec + \delta e(1-c)} \ . \tag{3.24}$$

3.7.2.3 Period-2 Input Entropy Rates

To check that the period-2 input process obeys Eq. (3.6), we calculate the entropy rate:

$$\Delta h_\mu = h'_\mu - h_\mu \ .$$

The entropy rate h_μ of a period-2 process is:

$$\begin{aligned} h_\mu &= \lim_{N \to \infty} \frac{\mathrm{H}[Y_{0:N}]}{N} \\ &= \lim_{N \to \infty} \frac{1}{N} \\ &= 0 \ . \end{aligned}$$

The entropy rate h'_μ of the output process generated by T' can be calculated using the uncertainty in the next symbol given the hidden state since T' is unifilar [12]:

$$\begin{aligned} h'_\mu &= \lim_{N \to \infty} \mathrm{H}[Y'_N | S'_N] \\ &= \lim_{N \to \infty} \sum_{s'} \mathrm{H}[Y'_N | S'_N = s'] \Pr(S'_N = s') \ . \end{aligned}$$

(No such general expressions hold for nonunifilar transducers.)

For the period-2 process, $c = 0$, and we see that the stationary state consists of two states with nonzero probability: $\pi'_{A \otimes E} = \pi'_{B \otimes F} = 0.5$. These states transition back and forth between each other periodically, so the current hidden state and output uniquely determine the next hidden state, meaning this representation is unifilar. Thus, we can use our calculated output HMM for the entropy rate h'_μ.

$A \otimes E$ has probability $\frac{1-\delta}{e}$ of generating a 1 and $B \otimes F$ has probability $\frac{1-\delta}{e}$ of generating

a 0. Thus, the uncertainty in emitting the next bit from either causal state is:

$$\begin{aligned} \mathrm{H}[Y'_N|S'_N = A \otimes E] &= \mathrm{H}[Y'_N|S'_N = B \otimes F] \\ &= \mathrm{H}\left(\frac{1-\delta}{e}\right) \ . \end{aligned}$$

Thus, their entropy rates are the same and we find:

$$\Delta h_\mu = \mathrm{H}\left(\frac{1-\delta}{e}\right) \ . \tag{3.25}$$

Chapter 4

Leveraging Environmental Correlations: The Thermodynamics of Requisite Variety

The mid-twentieth century witnessed an efflorescence in information and control and, in particular, the roles they play in biological adaptation [116]. Norbert Wiener's linear prediction theory [117, 118] and Claude Shannon's mathematical theory of communication [75, 119, 120, 121] stood out as the technical underpinnings. It was Wiener, though, who advocated most directly for a broad development of a new calculus of control and adaptation, coining the term "cybernetics" [122, 123]. The overall vision and new methods of information theory and linear stochastic processes stimulated a tremendous enthusiasm and creativity during this period.

It must be said that, despite substantial efforts throughout the 1950s and 1960s to develop "general systems" theories and the like [124, 125], at best, only modest successes transpired which addressed Wiener's challenges for cybernetics [126]. Historians of science claimed, in fact, that progress was inhibited by the political tensions between the West and East during the Cold War [127]. More practically, one cause was the immodest complicatedness of the systems targeted—weather control, the brain, and social design. In short, there simply were not the powerful computational and mathematical tools required

to understand such large-scale, complex systems. This all said, we must not forget that the intellectual fallouts from this period—the development of communication, coding, computation, and control theories—substantially changed the landscape of the engineering disciplines and irrevocably modified modern society.

Now, at the beginning of the 21st century, it seems time to revisit the broad and ambitious goals these early pioneers laid out. For, indeed, the challenges they introduced are still with us and are evidently starting to reveal dire consequences of our failure to understand the dynamics and emergent properties of large-scale complex systems, both natural and man-made. Optimistically, very recent developments in nonlinear dynamics [77] and nonequilibrium thermodynamics [128] give hope to finally achieving several of their goals, including reframing them in ways that will facilitate physical implementation. Here, we elucidate cybernetics' Law of Requisite Variety in light of these recent advances.

W. Ross Ashby was one of cybernetics's best expositors [18], having an impact that rivaled Wiener's advocacy. Principle to Ashby's approach was his concept of *requisite variety*. The requisite variety that confronts an adaptive system is the set of accessible, detectable, and controllable states in its environment. In its most elementary form, Ashby re-interpreted Shannon's notion of information-as-surprise, retooling it for broader application to biological and cognitive systems [125]. In this, though, he was anticipated by 30 years by Leo Szilard's successful purging of Maxwell Demon [32, 3]: "... a simple inanimate device can achieve the same essential result as would be achieved by the intervention of intelligent beings. We have examined the 'biological phenomena' of a nonliving device and have seen that it generates exactly that quantity of entropy which is required by thermodynamics". In laying out the thermodynamic costs of measurement, and so showing any demon is consistent with the Second Law of Thermodynamics, Szilard not only anticipates by two decades Shannon's quantitative measure of information but also Wiener's conception of cybernetics in which stored information plays a *functional role*.

The conceptual innovation in Szilard's analysis, still largely underappreciated, is his identifying two distinct kinds of information. On the one hand, there is surprisal; Shannon's notion that later on lead to an algorithmic foundation for randomness and prob-

ability [129, 130, 131, 132]. Its parallel in physics is a system's thermodynamic entropy [81]. The Demon monitors statistical fluctuations in its heat-bath environment. On the other hand, there is information stored as historical contingency and memory. It is this latter kind that explains the thermodynamic functionality of Maxwell's Demon, as it uses stored information about the thermal fluctuations to convert them to useful work [42]. This recognition handily resolves Maxwell's Second Law paradox. This information dichotomy was recently laid bare by mapping Szilard's single-molecule engine to chaotic dynamical system; a mapping so simple that all questions can be analytically addressed [78]. The role of both informative measurement and its use, when stored, for control illustrates the complementary role and functional consequences of both kinds of information in an adaptive system.

In this way, the now-familiar physical setting of Maxwell's paradox highlights how the distinction between *information-as-surprise* and *stored actionable-information* motivated Ashby's emphasizing requisite variety in adaptation. Detecting environmental fluctuations and acting on their structure (such as temporal correlations) are critical to the Demon's functioning. Appealing to new results in nonlinear dynamics and nonequilibrium thermodynamics, the distinction similarly motivates our re-addressing this central concern in cybernetics, so basic to the operation of adaptive systems, but in a fully thermodynamic setting: What requisite variety (range of historical contexts) must an adaptive agent recognize in its environment to realize thermodynamic benefits?

In the following, we first give an overview of our contributions (Sec. 4.1). We mention how Ashby's law of requisite variety is faithfully reflected in the behavior of *information engines*—autonomous versions of Maxwell's Demon. This close connection follows from the bounds set by the Second Law of Thermodynamics for information processing [7, 13, 6]. Important for engineered and biological implementations, we note that these bounds, and so those specified by Landauer's Principle [4, 34], are not generally achievable. The subsequent sections form the technical components of the development, key to which is representing an information reservoir in terms of the outcomes of a hidden Markov process.

Section 4.2 considers (i) the meaning of memory for the input processes of information

engines and for the engines themselves, (ii) their energetics, and (iii) the role of memory in information thermodynamics more generally [5, 36]. It is the thermodynamics of memory that establishes the correspondence between Ashby's law and the behavior of information engines. Section 4.3 addresses the limits on information engines achieving the informational Second Law bounds. We see that the bounds are not saturated even by optimal, finite-state engines. We also mention the curious case of infinite-memory information engines that can achieve and then go beyond these bounds, essentially by leveraging their internal infinite "negentropy" to generate work [133]. These results bear directly on the description of Maxwell's original Demon and, more contemporarily, stochastic universal Turing machines built out of information engines. Finally, we conclude with a summary of our results and their implications for biological physics and engineering.

4.1 Synopsis of main results

Szilard's Engine and related Maxwellian Demons are instances of thermal agents processing environmental information in order to convert thermal energy into work. Turning disordered thermal energy into work (ordered energy) was long thought to violate the Second Law of Thermodynamics [134]. However, the past century resolved the apparent violation by recognizing that information processing has unavoidable energy costs. Rolf Landauer was one of the first to set bounds on information processing—specifically, erasing a bit—such that the work production over a thermodynamic cycle cannot be positive, satisfying the Second Law of thermodynamics [4, 34].

However, if the Demon accesses an information reservoir in its environment, it can use the reservoir's statistics as a resource to convert thermal energy into work. This view of a Demon taking advantage of a structured environment connects back to cybernetics. Just as Ashby asked how a controller's variety should match that of its inputs, we ask how the Demon's internal structure should match the structure of an input process, which characterizes the information reservoir, in order to generate work. In contrast to cybernetics, though, we consider the variety inherent in "information ratchets" viewed as thermodynamic systems and, by implication, the variety they can detect and then leverage in their

environments.

An information ratchet is an explicit construction of an autonomous Maxwellian Demon that uses an input symbol sequence to turn thermal energy into work energy [7, 63]. The ratchet steadily transduces the input symbols into an output sequence, processing the input information into an output while effecting thermodynamic transformations—implementing a physically embedded, real-time computation. This is accomplished by driving the ratchet along the input symbol sequence unidirectionally, so that the ratchet (with states in set $\mathcal{X}$) interacts once with each symbol (with values in alphabet $\mathcal{Y}$). This is described by a thermally activated Markov transition matrix

$$\begin{aligned} & M_{x \otimes y \to x' \otimes y'} \\ & \quad = \Pr(X_{N+1} = x', Y'_N = y' | X_N = x, Y_N = y) \ , \end{aligned}$$

from $x \otimes y \in \mathcal{X} \otimes \mathcal{Y}$ to $x' \otimes y'$, which has detailed balance, as described in the previous chapters. The transition matrix M determines the energetics as well as the ratchet's information processing capacity.

Recent work introduces a general computational mechanics [77, 1] framework for analyzing thermodynamic devices that transduce an input process into an output process [13, 1]. Figure 4.1 depicts the relative roles of the input process specified by a finite-state hidden Markov model (HMM), the ratchet as transducer operating on the input process, and the resulting output process, also given by an HMM.

The tools of computational mechanics were developed to quantitatively analyze how a ratchet's structure should match that of its input for maximum efficacy, since they use a consistent notion of structure for general processes and transformations. In particular, using them we recently established a general information processing Second Law (IPSL) for thermodynamically embedded information processing by finite ratchets that bounds the asymptotic work per cycle $\langle W \rangle$ in terms of the difference in entropy rates of the input and output processes, h_μ and h'_μ, respectively [13] (see App. 4.4.2 for an alternate proof to that presented in Chapter 2):

$$\langle W \rangle \leq k_B T \ln 2 \left(h'_\mu - h_\mu \right) \ . \tag{4.1}$$

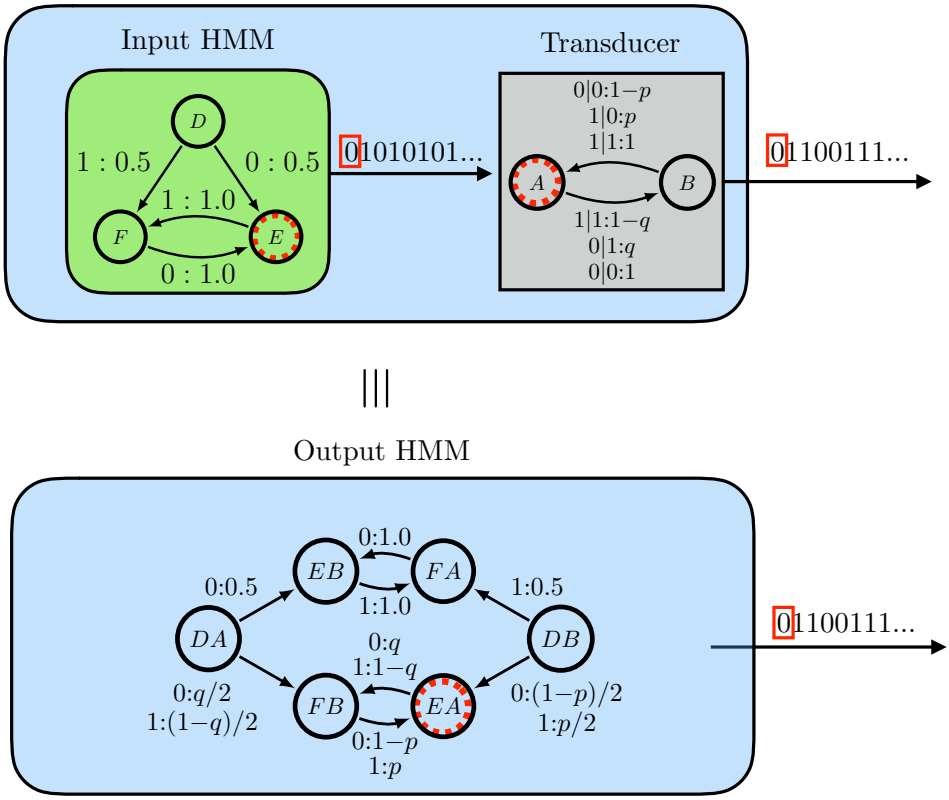


Figure 4.1. Computational mechanics view of an information ratchet: The input signal (environment) is described by a hidden Markov model (HMM) that generates the input symbol sequence. The ratchet itself acts as a transducer, using its internal states or memory to map input symbols to output symbols. The resulting output sequence is described by an HMM that results from composing the transducer with the input HMM. The current internal state of the input HMM, transducer, and output HMM are each highlighted by a dashed red circle. These are the states achieved after the last output symbol (highlighted by a red box) of each machine. We see that the internal state of the output HMM is the direct product of the internal state of the transducer and the input HMM.

(Definitions are given shortly in Sec. 4.2.) Employing entropy rates—the Shannon entropy rate of the symbol sequence or, equivalently here, the Kolmogorov-Sinai entropy of its generative dynamical system—the bound accounts for all temporal correlations in the input and output processes as well as the single-symbol biases. While this bound appears similar to that $\langle W \rangle \leq \langle I \rangle - \Delta F$ [39] on work production in a system with feedback control, $\langle I \rangle$ quantifies correlations between the controller and environment rather than temporal correlations induced in the environment.

Two uses of Eq. (4.1)'s IPSL suggest themselves. First, it sets an informational upper bound on the maximum average work production $\langle W \rangle$ per thermodynamic cycle. Here, W is the flow of work from the ratchet to an external driver. Second, and complementarily, it places an energetic lower bound on the minimal work $\langle W_d \rangle$ required to drive a given amount (Δh_μ) of computation forward. Here, $W_d = -W$ is the flow of work from the driver into the ratchet. In this second use, the IPSL is a substantial extension of Landauer's Principle. The latter says that erasing a bit of information requires a minimum energy expenditure of $k_{\mathrm{B}} T \ln 2$ while the IPSL applies to any kind of computational processing that transforms an input process to an output process, not simply erasure. The first use appears, in this light, as a rough converse to Landauer's limit: There is a potential thermodynamic benefit of "destroying variety" in the form of work [4, 34].

Practically, computational mechanics gives a means to partition the ratchet and input process into different cases: memoryful and memoryless. Whether or not the input process or ratchet have memory substantially changes the bound on work production. And so, we can examine how environmental and demon varieties interact. For example, in the case in which temporal correlations (varieties) vanish, the difference between the input's single-symbol entropy H_1 and the output's H'_1 gives an analogous bound [10]:

$$\langle W \rangle \leq k_B T \ln 2 \left(\mathrm{H}'_1 - \mathrm{H}_1 \right) \ , \tag{4.2}$$

Using the single-symbol approximation H_1 of the true entropy rate h_μ can be quite convenient since H_1 is much easier to calculate than h_μ, as the latter requires asymptotic (long-range) sequence statistics. (Again, definitions are given shortly in Sec. 4.2.) Likely, this is why the H_1-bound has appeared frequently to describe ratchet information pro-

cessing [7, 60, 61, 58, 36]. Also, Eq. (4.2) is a rather direct generalization of the Landauer limit, since the input entropy $\mathrm{H}_1 = 1$ bit and the output $\mathrm{H}'_1 = 0$ bits saturate the bound on the work required to drive erasing a binary symbol. However, a key difference is that Eq. (4.1)'s entropy rates are dynamical invariants; unchanged by smooth transformations [135, 136]. The single-symbol Shannon entropies are not dynamical invariants. In addition, the single-symbol bound does not properly account for the temporal correlations in the input process or those created by the ratchet in the output process and so leads to several kinds of error in thermodynamic analysis. Let us explore these.

First, the average total temporal correlation in a process can be quantified by the difference between the single-symbol entropy and the entropy rate, known as a process' length-1 *redundancy* [12]:

$$\mathrm{H}_1 - h_\mu \geq 0 \ . \tag{4.3}$$

This is the extent to which single-symbol entropy-rate estimates (H_1) exceed the actual per-symbol uncertainty (h_μ); and it is always nonnegative. This measure describes a type of structure distinct from statistical autocorrelations. Unless stated otherwise, going forward, the informational temporal correlations quantified in Eq. (4.3) are what we mean by *correlations*.

How inputs or ratchets create or destroy these correlations determines the relative strength and validity of the Eq. (4.1) and Eq. (4.2) work bounds. These bounds, in turn, suggest that memoryless ratchets are best for leveraging memoryless inputs and memoryful ratchets are best for leveraging memoryful inputs and generating work. However, it is not clear if and when the bounds are achievable. So, more effort is required to establish this thermodynamic version of Ashby's Law of Requisite Variety.

To address achievability, we turn to a general energetic framework for calculating ratchet work production [6]. There it was shown that memoryful ratchets can leverage temporal correlations which memoryless ratchets cannot. In short, memoryful ratchets are indeed best for leveraging memoryful inputs. This gives an explicit violation of Eq. (4.2). However, for memoryless ratchets both Eqs. (4.2) and (4.1) are valid bounds [10]. We show, with proof given in App. 4.4.1, that memoryless ratchets are the best among

all finite ratchets at leveraging statistical biases in memoryless inputs to produce work. Notably, these ratchets do not achieve the derived upper bounds on work production, demonstrating fundamental inefficiencies in the information-to-work conversion in this class of an autonomous Maxwellian Demon.

To approach the bounds described by Eqs. (4.1) and (4.2) it is necessary to go beyond the information processing paradigm of a single finite-memory ratchet that interacts with a single symbol at a time. For instance, consider a "swarm" of finely tuned ratchets that work in a sequence, the output of one acting as the input of the next, and each ratchet being optimized with respect to its own input. This stepwise, sequential processing of the information reservoir is more efficient than the single-ratchet paradigm and is able to approach the upper bounds on information processing as the number of ratchets in the army grows. (This is reminiscent of the higher efficiency of quasistatic thermodynamic processes compared to finite-time, irreversible processes.) We reserve the detailed analysis of this phenomenon for a later work since the framework for collective thermodynamics is less developed than the single-ratchet setting we focus on here.

While the IPSL and related bounds on work are suggestive of how the structure of the input matches the output, the fact that they are unachievable for single information ratchets means we must reach further to solidify the relationship between input statistics and ratchet thermodynamics. Exact calculations here for the work production verify the intuition that the memory of an optimal ratchet must match the memory of the input. This leads to a variation on Ashby's Law of Requisite Variety: "memory leverages memory".

In this way, the transducer framework for information ratchets gives insight into how adaptive agents leverage structure. Its importance extends far beyond, however, to general computation. On the one hand, transducers describe mappings from input sequences to distributions over output sequences [1, 137] and do so in real time. Turing machines, on the other, map individual input sequences to individual output sequences with no particular reference to physical time. In this sense, Turing machines are a subclass of transducers, emphasizing that transducers are a general model for physical computation

and information processing. However, to do universal computation, as properly configured Turing machines can, requires infinitely many states [137]. And, this suggests examining the thermodynamics of infinite-memory ratchets.

It turns out that infinite ratchets with states having finite energy differences are pathological in that they violate both the IPSL and its single-symbol sister bounds on work production—Eqs. (4.1) and (4.2), respectively. The proof of Eq. (4.1) assumes a stationary distribution over the ratchet state and input symbol. This need not exist for infinite ratchets [13]. In this case structure in the ratchet's memory, rather than structure in the information reservoir, can be used as an additional thermodynamic resource to produce work. And, this means that a framework for general computation requires more detailed analysis to set bounds on work production that account for the ratchet's memory. While we leave this for upcoming work, it does call into question any discussion of the thermodynamics of universal computation.

Following are the major contributions:

1. To address the role of memory in ratchets, we introduce thermodynamically predictive definitions of memory for *both* the input string and the ratchet which performs a computation on the input.

2. The validity of IPSL bounds changes depending on whether or not the input or ratchet are separately or together memoryful or memoryless.

3. The memory dependence of IPSLs implies that, if the IPSL derived in our previous work [13] is achievable, then we arrive at a thermodynamic Law of Requisite Variety [125].

4. However, our exact analysis of memoryless ratchets driven by memoryless inputs tells us that to achieve IPSL bounds we must go beyond the class of individual autonomous ratchets that operate on a single bit at a time. The Law of Requisite Variety may not hold for such generalized ratchets.

5. Fortunately, we also complete a suite of results about information ratchets that show memoryful ratchets are best for leveraging memoryful inputs and memoryless ratch-

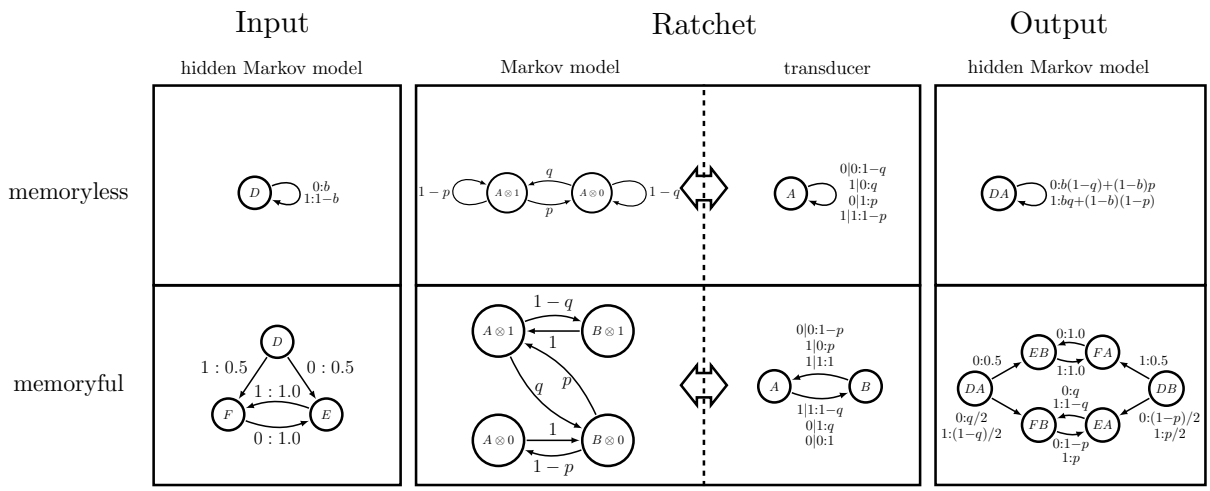


Figure 4.2. Ratchets and input and output signals can be either memoryful or memoryless. For the input or output signal to be memoryless, the generating (minimal) HMM must have more than one internal state. The action of a ratchet can be represented in two different ways: either by a detailed Markov model involving the joint state space of the ratchet and an input symbol or by a symbol-labeled Markov dynamic on the ratchet's state space. We call the latter the *transducer representation* [1]. Similar to the input and output signals, if the (minimal) transducer has more than one internal state, then the ratchet is memoryful.

ets are best for leveraging memoryless inputs. This confirms the Law of Requisite Variety from a dynamical perspective, independent of IPSLs.

6. While the results hold for finite ratchets, for the potentially unphysical case of infinite ratchets, both IPSLs and the Law of Requisite Variety fail: an infinite ratchet can violate IPSL bounds, extracting more energy than a memoryless input allows.

With this overview laid out, with the goals and strategy stated, we now are ready to delve into memory's role in information-engine thermodynamics and the achievability of the IPSL and its related bounds.

4.2 Memory

To explore how a ratchet's structure "matches" (or not) that of an environmental signal requires quantifying what is meant by structure. In terms of their structure, both ratchets and environmental inputs can be either memoryless or memoryful and this distinction

delineates a ratchet's thermodynamic functioning via the IPSL. This section introduces what we mean by the distinction, describes how it affects identifying temporal correlations, and shows how it determines bounds on work production and functionality. The results, though, can be concisely summarized. Figure 4.2 presents a tableau of memoryless and memoryful ratchets and inputs in terms of example HMM state-transition diagrams. Figure 4.3 then summarizes IPSL bounds for the possible cases.

4.2.1 Process memory

The amount of memory in the input or output processes is determined by the number of states in the minimal representative dynamics that generates the associated sequence probability distributions. We make this definition clear in the following section and show that it is relevant to predicting thermodynamic bounds.

While there are many ways to generate a process, HMMs are a particularly useful representation of generating mechanisms. For example, they describe a broader class of processes than finite-order Markov models, since they can generate infinite Markov-order processes using only a finite number of hidden states [12].

Here, we use the Mealy representation of HMMs [138, 139, 140, 141], which consists of a set $\mathcal{S}$ of internal states and an alphabet $\mathcal{A}$ of symbols that are emitted. As with a Markov chain, transitions between hidden states in $\mathcal{S}$ are made according to conditional probabilities. However, the generated symbols in $\mathcal{Y}$ are emitted during transitions between hidden states, rather than when entering states [142]. The Mealy HMM dynamic is specified by a set of symbol-labeled transition matrices:

$$T^{(y_N)}_{s_N \to s_{N+1}} = \Pr(Y_N = y_N, S_{N+1} = s_{N+1} | S_N = s_N) \ ,$$

which give the joint probability of emitting y_N and transitioning to hidden state s_{N+1} given that the current hidden state is s_N. For the special class of *unifilar* HMMs the current hidden state s and emitted symbol y uniquely determine the next hidden state $s'(s, y)$. Helpfully, for unifilar HMMs the generated process' entropy rate h_μ is exactly given by the state-averaged uncertainty in the emitted symbols given the current state

[12]:

$$\begin{aligned} h_\mu &= \lim_{N\to\infty} \mathrm{H}[Y_N|Y_{0:N}] \\ &= \lim_{N\to\infty} \mathrm{H}[Y_N|S_N] \\ &= \sum_{s\in\mathcal{S}} \pi_s \lim_{N\to\infty} \mathrm{H}[Y_N|S_N = s] \\ &= -\sum_{s\in\mathcal{S}} \pi_s \sum_{y\in\mathcal{Y}} T^{(y)}_{s\to s'(s,y)} \log_2 T^{(y)}_{s\to s'(s,y)} \ , \end{aligned}$$

where π_s is the steady-state distribution over the hidden states. A process' *ϵ-machine* is its minimal unifilar HMM generator, where minimality is determined by having the smallest internal-state Shannon entropy [77]:

$$\begin{aligned} \lim_{N\to\infty} \mathrm{H}[S_N] &= -\lim_{N\to\infty} \sum_{s\in\mathcal{S}} \Pr(S_N = s) \log_2 \Pr(S_N = s) \\ &= -\sum_{s\in\mathcal{S}} \pi_s \log_2 \pi_s \\ &\equiv C_\mu \ . \end{aligned}$$

where C_μ in the last line is the process' *statistical complexity* C_μ. Since h_μ gives an exact expression for process entropy rate and C_μ a unique definition of process memory [143], throughout we represent processes by their ϵ-machines. An ϵ-machine's internal states are called *causal states*.

Broadly, the memory of an ϵ-machine refers to its hidden states. As shown in Fig. 4.2, memoryless input processes have ϵ-machines with a single state: $|\mathcal{S}| = 1$. The sequence distributions for such processes are given by a product of single-symbol marginal distributions. For a stationary process, the single-symbol marginal entropy H_1 is the same for every symbol:

$$\mathrm{H}_1 \equiv \mathrm{H}[Y_N] \text{ for all } N \in \mathbb{N}. \tag{4.4}$$

For memoryless processes, the entropy rate is the same as the single-symbol entropy:

$$\begin{aligned} h_\mu &= \lim_{N\to\infty} \mathrm{H}[Y_N|Y_{0:N}] \\ &= \lim_{N\to\infty} \mathrm{H}[Y_N] \\ &= \mathrm{H}_1 \ . \end{aligned}$$

This means that their difference vanishes:

$$\mathrm{H}_1 - h_\mu = 0 \ . \tag{4.5}$$

Thus, there are no temporal correlations in the symbol string, since $\mathrm{H}_1 - h_\mu$ quantifies the informational correlation of individual input symbols with past inputs:

$$\begin{aligned} \mathrm{H}_1 - h_\mu &= \lim_{N\to\infty} (\mathrm{H}[Y_N] - \ H[Y_N|Y_{0:N}]) \\ &= \lim_{N\to\infty} \mathrm{I}[Y_N : Y_{0:N}] \ , \end{aligned} \tag{4.6}$$

where $\mathrm{I}[W : Z]$ is the mutual information of random variables W and Z [11].

For memoryful input processes, as shown in Fig. 4.2, there are multiple causal states for the ϵ-machine: $|\mathcal{S}| > 1$. In other words, sequence probabilities cannot be broken into a product of marginals. And so, in general, we have:

$$\mathrm{H}_1 > h_\mu \ .$$

Thus, there are temporal correlations in the input process:

$$\mathrm{H}_1 - h_\mu > 0 \ . \tag{4.7}$$

This means that individual symbols of the input sequence share information with past inputs. In the maximally correlated case, every symbol is exactly predictable from its past. As a result the entropy rate vanishes and the temporal correlation measure in Eq. (4.6) is equal to the single-symbol entropy.

To summarize, memoryless input signals have a single causal state and, thus, do not exhibit temporal correlations, since they have no way to store information from the past. Meanwhile, memoryful inputs have multiple hidden states that are used to transmit information from the past to the present and so express temporal correlations.

4.2.2 Ratchet memory

From the perspective of information processing, the ratchet is a transducer that interacts with each symbol in the input sequence in turn, converting it into a output symbol stored in the output sequence [1, 13]. The ratchet is a form of communication channel [11]. One that is determined by a detailed-balanced Markov dynamic:

$$\begin{aligned} & M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} \\ & = \Pr(Y'_N = y'_N, X_{N+1} = x_{N+1} | X_N = x_N, Y_N = y_N) \end{aligned}$$

over the ratchet's state space $\mathcal{X}$ and a symbol alphabet $\mathcal{Y}$. This is the probability that the ratchet ends in state x_{N+1} and writes a symbol y'_N to the output sequence, given that the input symbol was y_N and the ratchet's state was x_N before the symbol-state interaction interval.

The Markovian dynamic describes the behavior of the joint event (ratchet-state $\otimes$ symbol-value) during the interaction transition and leads to the transducer representation of the ratchet's functionality, illustrated in Fig. 4.2. As we use the terms, the *ratchet* refers to the physical device implementing the Markovian dynamic, whereas *transducer* refers to the computational mechanics state-transition machine (ϵ-transducer) that captures its information-theoretic functionalities in a compact way [1]. The form of the transducer is:

$$M_{x_N \to x_{N+1}}^{(y'_N | y_N)} = M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} \ . \tag{4.8}$$

The distinction between the Markov dynamic and the transducer representation is best illustrated graphically, as in the second column of Fig. 4.2.

The definition of a ratchet's memory involves its ϵ-transducer representation. In other words, memory is related to the size of the ratchet's causal state space $|\mathcal{X}|$ in its ϵ-transducer representation. (The very definition of ϵ-machines and ϵ-transducers entails that they have the minimal set of states for a given input, output, or input-output process.) As seen in the top middle of Fig. 4.2, memoryless ratchets have only a single internal (hidden) state: $|\mathcal{X}| = 1$. Thus, the ratchet behaves as a memoryless channel from input

to output [11]. And, in this, it reduces temporal correlations in the input signal:

$$\mathrm{H}'_1 - h'_\mu \leq \mathrm{H}_1 - h_\mu \ , \tag{4.9}$$

according to the Data Processing Inequality [11]. As shown by Eq. (4.6) the difference between the single symbol entropy and the entropy rate is the mutual information between the infinite past symbols and the current symbol. If the channel is memoryless, then $Y_N \to Y'_N$ and $Y_{0:N} \to Y'_{0:N}$, so by applying the Data Processing Inequality twice, we find $I[Y'_N; Y'_{0:N}] \leq I[Y_N; Y_{0:N}]$. Thus, the change in single-symbol entropy is a lower bound for the change in entropy rates [10]. In contrast, a memoryful ratchet has more than one state, $|\mathcal{X}| > 1$, and behaves as a memoryful channel [1]; bottom right of Fig. 4.2.

How the ratchet transduces the current input to the current output depends on in which state it is. As a result, the ratchet can create correlations in the output such that, regardless the input process:

$$\mathrm{H}'_1 - h'_\mu \geq 0 \ . \tag{4.10}$$

Several explicit constructions of the output process based on given input and ratchet are shown in the last column of Fig. 4.2.

4.2.3 Thermodynamics of memory

This section considers the role of memory in the thermodynamic efficacy of information engines. In particular, we consider the average work production per cycle $\langle W \rangle$. The role can be explored in two complementary ways: either following the IPSL and related bounds, Eqs. (4.1) and (4.2), or from the exact expression of $\langle W \rangle$.

4.2.3.1 Information Processing Second Law bounds

The thermodynamics of memory is summarized in Fig. 4.3's table, where each row considers a different combination of input process and ratchet. This section addresses each cell in the table individually.

Consider the case of memoryless input and a memoryless ratchet. In Eq. (4.9), we saw that the temporal correlations in the input signal cannot be increased by such ratchets. Since the input signal is memoryless, the output signal must also be memoryless. For

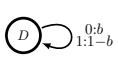
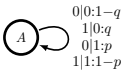
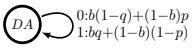
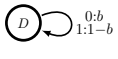
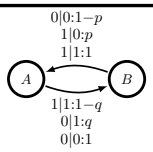
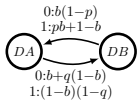
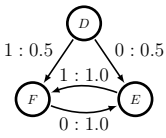
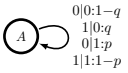
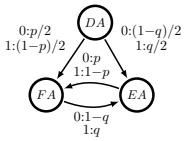
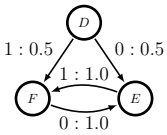
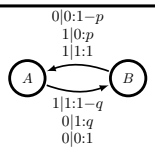
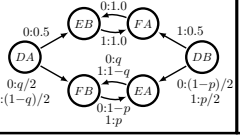
	Input Process	Ratchet Transducer	Output Process	Thermal Relations
memoryless input memoryless ratchet				$0 = H_1 - h_\mu = H'_1 - h'_\mu$ $\langle W \rangle \leq \Delta h_\mu = \Delta H_1$
memoryless input memoryful ratchet				$0 = H_1 - h_\mu \leq H'_1 - h'_\mu$ $\langle W \rangle \leq \Delta h_\mu \leq \Delta H_1$
memoryful input memoryless ratchet				$0 \leq H'_1 - h'_\mu \leq H_1 - h_\mu$ $\langle W \rangle \leq \Delta H_1 \leq \Delta h_\mu$
memoryful input memoryful ratchet				$H'_1 - h'_\mu \stackrel{?}{=} H_1 - h_\mu$ $\langle W \rangle \leq \Delta h_\mu$ ~~$\langle W \rangle \leq \Delta H_1$~~

Figure 4.3. The informational (IPSL) bounds on work that use Δh_μ or $\Delta\,\mathrm{H}(1)$ depend critically on input signal and ratchet memory. In all finite memory cases, Δh_μ is a valid bound on $\langle W \rangle / k_B T \ln 2$, but the same is not true of $\Delta\,\mathrm{H}_1$, as indicated in the far right column on thermal relations. The bounds shown in column have $k_B T \ln 2$ set to unity, so that the relations can be shown in compact form. If the ratchet is memoryless, then $\Delta\,\mathrm{H}_1$ is a valid and stronger bound than Δh_μ, because these channels decrease the temporal temporal correlations in transducing input to output. For a memoryless input with memoryful ratchet, $\Delta\,\mathrm{H}_1$ is still a valid bound, but it is a weaker bound than Δh_μ, because a memoryful ratchet typically creates temporal correlations in its output. However, in the case where both input and output are memoryful, the $\Delta\,\mathrm{H}_1$ bound is invalid. It is violated by systems that turn temporal correlations into work by using ratchet memory to synchronize to the input memory.

memoryless signals, however, we saw via Eq. (4.5) that the entropy rate h_μ is the same as the single-symbol entropy H_1. We conclude that the temporal correlations vanish in both the input and output and, thus, that the single-symbol entropy input-to-output difference difference is the same as the entropy-rate difference:

$$\mathrm{H}'_1 - h'_\mu = \mathrm{H}_1 - h_\mu \ .$$

As a result both Eqs. (4.1) and (4.2) give the same bound on the the average rate of work

production:

$$\langle W \rangle \leq k_B T \ln 2 \ \Delta \mathrm{H}_1 \tag{4.11}$$

$$= k_B T \ln 2 \ \Delta h_\mu \ , \tag{4.12}$$

This is noted at the right column in the table's first row.

Consider now the case of memoryful input with, again, a memoryless ratchet. A memoryful input contains temporal correlations that are decreased by the memoryless ratchet, from Eq. (4.9). The same equation implies that the single-symbol entropy difference is an upper bound on the entropy-rate difference. As a result, Eq. (4.2) provides a quantitatively tighter bound on the work production compared to the IPSL of Eq. (4.1) [10]:

$$\begin{aligned} \langle W \rangle &\leq k_B T \ln 2 \ \Delta \mathrm{H}_1 \\ &\leq k_B T \ln 2 \ \Delta h_\mu \ , \end{aligned}$$

These observations suggest that memoryless ratchets cannot leverage temporal correlations, since the stricter bound (single symbol) on work production stays fixed as we hold the single-symbol entropy fixed but vary the temporal correlations in the input. It appears that to leverage temporal correlations, one must use a memoryful ratchet.

We now address the case of memoryful ratchets. First, consider the case of memoryless inputs (no temporal correlations: $h_\mu = \mathrm{H}_1$). From Eq. (4.10), we know that memoryful ratchets can create correlations in the output. In other words, the output signal is generally memoryful, implying $\mathrm{H}'_1 - h'_\mu \geq 0$. As a result, $\mathrm{H}'_1 - h'_\mu \geq \mathrm{H}_1 - h'_\mu$, which implies $\Delta h_\mu \leq \Delta \mathrm{H}_1$. And so, the change in entropy rate is a stricter bound than the change in single-symbol entropy:

$$\begin{aligned} \langle W \rangle &\leq k_B T \ln 2 \ \Delta h_\mu \\ &\leq k_B T \ln 2 \ \Delta \mathrm{H}_1 \ , \end{aligned}$$

as seen in Table 4.3's second row. We explored this in some detail in Chapter 2 [13]. By calculating Δh_μ, we found a novel type of functionality in which the ratchet used stored work energy to increase temporal correlations in the input while simultaneously *increasing*

the single-symbol uncertainty. The above relations also imply that memoryless ratchets may be best suited for leveraging memoryless input processes, since the bounds on work production for memoryless ratchets are higher than the bounds for memoryful ratchets.

Consider now a memoryful input driving a memoryful ratchet. In this case, memory in the ratchet is useful for work production. Chapter 3 [6] considers a maximally correlated, period-2 input process, that has no single-symbol negentropy to leverage ($\mathrm{H}_1 = 1$ bit of information), but that has maximal temporal correlations ($\mathrm{H}_1 - h_\mu = 1$ bit). Notably, the single-symbol bound indicates that no work can be produced, since $\Delta \mathrm{H}_1 \leq 0$ regardless of the output. Critically, though, the IPSL bound indicates that work production is possible, since $h'_\mu - h_\mu > 0$ as long as the output has some uncertainty in each sequential symbol. Indeed, Ref. [6] constructs a ratchet that produces positive work: $\langle W \rangle = k_\mathrm{B} T \frac{1-\delta}{e}$, where $\delta \in (0, 1)$. Thus, the single-symbol bound is violated, but the IPSL bound is satisfied, as shown in Fig. 4.3's last row.

The final case to consider, in fact, is left out of Fig. 4.3: infinite-memory ratchets. This is because infinite memory ratchets do not necessarily have a steady state, so the IPSL bound in Ref. [13] does not hold. There are, as yet, no predictions for infinite-memory ratchets based on the information measures of the input or output processes. However, this is an intriguing case. And so, we turn to infinite ratchets in Sec. 4.3.3.

Stepping back, Fig. 4.3's table details a constructive thermodynamic parallel to Ashby's Law of Requisite Variety: *Memory can leverage memory.* However, the bounds do not constitute existence proofs, since it is not yet known if the specified bounds are achievable. Though, we constructed an example of a temporally correlated process that is best leveraged by memoryful ratchets, it is possible that there is an alternative temporally correlated input process that is best leveraged by a memoryless ratchet. Similarly, we see that the bounds on memoryless inputs are stricter for memoryful ratchets than for memoryless ratchets. If these bounds are not achievable, however, then this does not translate into a statement about the ratchet's actual efficiency in producing work.

Before addressing the puzzle of achievability, we need to determine the work production.

4.2.3.2 Exact work production

An exact expression for the average work production rate was introduced in Ref. [6], as discussed in Chapter 3:

$$\langle W \rangle = k_B T \sum_{\substack{x,x' \in \mathcal{X} \\ y,y' \in \mathcal{Y}}} \pi_{x \otimes y} M_{x \otimes y \to x' \otimes y'} \ln \frac{M_{x' \otimes y' \to x \otimes y}}{M_{x \otimes y \to x' \otimes y'}}, \tag{4.13}$$

where $\{\pi_{x \otimes y}\}$ is the steady-state joint probability distribution of the ratchet and the input symbol before interaction. Heuristically, the formula can be understood in the following way. At the beginning of the interaction interval, the ratchet and the incoming bit have probability $\pi_{x \otimes y}$ to be in state $x \otimes y$. Thus, the joint system has the probability $\pi_{x \otimes y} M_{x \otimes y \to x' \otimes y'}$ to make the transition $x \otimes y \to x' \otimes y'$. Since M specifies a detailed-balanced thermal dynamic, the amount of energy extracted from the reservoir in each transition is given by the log-ratio $\ln(M_{x' \otimes y' \to x \otimes y} / M_{x \otimes y \to x' \otimes y'})$. The right-hand side of Eq. (4.13) therefore gives the average energy extracted from the heat reservoir every thermodynamic cycle. From the First Law of Thermodynamics, this must be the ratchet's average work production, since its energy is fixed in the steady state. Not only does the expression confirm our physical law of requisite memory, it also expands our understanding of the validity of IPSL-like bounds, as we see below.

Irrespective of the nature of the input, consider the case of memoryless ratchets for which we have:

$$\begin{aligned} \pi_{x \otimes y} &= \lim_{N \to \infty} \Pr(X_N = x, Y_N = y) \\ &= \lim_{N \to \infty} \Pr(Y_N = y) \\ &= \Pr(Y_N = y) \ , \end{aligned}$$

simply the single-symbol probabilities of the input process. This follows since there is only a single ratchet state x. Thus, from Eq. (4.13), the only dependence the work has on the input process is on the latter's single-symbol distribution. In short, memoryless ratchets are insensitive to correlations in the inputs. To leverage correlations beyond single symbols in the input process it is necessary to add memory to the ratchet, as discussed in the previous section and in our companion work [6].

Conversely, as App. 4.4.1 establishes, if the input process is memoryless, there is no energetic advantage of using finite memoryful ratchets for binary input processes. For any finite memoryful ratchet that extracts work using the input process, there exists a memoryless ratchet that extracts at least as much work.

These two results confirm the intuition that to be thermodynamically optimal a ratchet's memory must match that of the input: Memoryful ratchets best leverage memoryful inputs and memoryless ratchets best leverage memoryless inputs.

4.3 Achievability of Bounds

The IPSL bound on average work production rate was derived based on the Second Law of Thermodynamics applied to the joint evolution of the ratchet, the input-output symbol sequence, and the heat reservoir. Since the Second Law is merely an inequality, it does not guarantee that the bounds are actually achievable for the nonequilibrium class of information engines considered here. In point of fact, we saw that the bound cannot be saturated by memoryless ratchets. A somewhat opposite picture is presented by infinite-memory ratchets. And, understanding these is a necessity if we wish to build a thermodynamics of general computation; that is, of physically embedded universal Turing machines. As we will show shortly, infinite-memory ratchets can violate the IPSL bound since they can leverage the steady increase in their own entropy to reduce the entropy of the heat reservoir, in addition to the contributions from an input signal. The following analyzes these cases individually.

4.3.1 Memoryless ratchets

This section applies the work expression of Eq. (4.13) to find optimal memoryless nonequilibrium ratchets, which operate through a sequence of driven and thermally activated transitions as described in Chapter 3, and then compares their optimal work production to the preceding information thermodynamics bounds to determine their achievability. Understanding the relationships between the memory of the ratchet and that of the input process, as discussed above, deepens the interpretation of the analysis. Since memoryless ratchets are insensitive to correlations, our calculated work productions are not only the

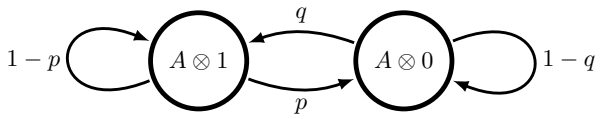


Figure 4.4. All possible memoryless ratchets that operate on a binary input, which are parametrized by transition probabilities p and q.

work productions for memoryless inputs, but the work productions for all inputs with the same single-symbol statistical biases.

A memoryless ratchet's memory consists of a single state. As a result, the Markovian dynamic M acts only on individual input symbols. Thus, the work for any input process is a function only of its single-symbol distribution $\pi_y = \Pr(Y_N = y)$ (given M):

$$\langle W \rangle = k_B T \sum_{y,y' \in \mathcal{Y}} \pi_y M_{y \to y'} \ln \frac{M_{y' \to y}}{M_{y \to y'}} \ .$$

Here, we discuss in detail the particular case of a memoryless ratchet driven by binary inputs. The relevant class of transducers comprises all two-state HMMs over the state space $\{A\} \otimes \{0, 1\}$, where A is the ratchet's sole state. Since the transducers' state space is two-dimensional, the Markovian dynamic M is guaranteed to be detailed balanced. Moreover, we can parametrize this class by two transition probabilities p and q, as shown in Fig. 4.4. This, then, allows us to optimize over p and q to maximize work production.

For the ratchet shown in Fig. 4.4 driven by a process with single-symbol probabilities $\Pr(Y_N = 0) = b$ and $\Pr(Y_N = 1) = 1 - b$, the average work done is a function of b, p, and q:

$$\langle W \rangle(b, p, q) = k_B T (b - b') \ln \frac{p}{q} \ , \tag{4.14}$$

where $b' = b'(b, p, q) = (1 - q)b + (1 - b)p$ is the probability $\Pr(Y'_N = 0)$ of symbol 0 in the output. The expression for b' follows from the dynamic depicted in Fig. 4.4, whereas Eq. (4.14) follows from the fact that work $\ln(p/q)$ is gained for each transformation $0 \to 1$. For a given input bias b, optimization of the ratchet's transducer dynamic to produce maximal work yields ratchet parameters $p_{\max}(b)$ and $q_{\max}(b)$:

$$[q_{\max}(b), p_{\max}(b)] = \begin{cases} [\frac{1-b}{b\,\Omega(e(1-b)/b)}, 1], & 1/2 \leq b \leq 1 \\ [1, \frac{b}{(1-b)\,\Omega(eb/(1-b))}], & 0 \leq b < 1/2 \end{cases} ,$$

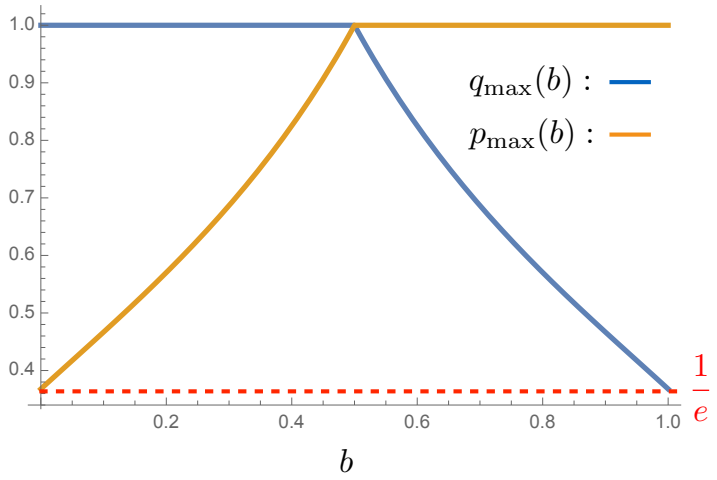


Figure 4.5. Optimal ratchet parameters $p_{\max}(b)$ (solid orange line) and $q_{\max}(b)$ (dashed blue line) are mirror images about $b = 1/2$. For $b < 1/2$, we set $p_{\max}(b) < 1$ and $q_{\max} = 1$ so that the interaction transition $1 \to 0$ has a positive energy change $\Delta E_{1\to 0} = k_B T \ln(q/p)$ and, thus, absorbs heat from the thermal reservoir. The same reasoning applies to $b > 1/2$, where $p_{\max}(b) = 1$ and $q_{\max} < 1$. In the unique case where the input is all 1s, the most effective ratchet for generating work has $p_{\max} = 1/e$. Both functions realize a minimum value of $1/e$, as shown.

where the function $\Omega(\cdot)$ is defined implicitly as $\Omega(ze^z) = z$. To confirm that these are indeed the maximal parameters for a given input $b \in [0, 1]$, note that $\langle W \rangle(b, p, q)$ is concave down over the physical parameter range $p, q \in [0, 1]$. For $b < 1/2$, plugging $q = 1$ and $p = b/(1 - b)\Omega(eb/(1 - b))$ into partial derivatives of the work yields $\partial\langle W \rangle/\partial q \geq 0$ and $\partial\langle W \rangle/\partial p = 0$. However, since q is already at its maximum, this a local maximum for the allowed parameter range. And, since the work is concave down, we know that this local maximum is the global maximum. The same can be shown for $b \geq 1/2$ by using the symmetry with respect the simultaneous exchanges $\{p \leftrightarrow q, b \leftrightarrow 1 - b\}$. Figure 4.5 shows how the optimal parameters depend on input bias $\Pr(Y_N = 0) = b$.

Substituting $q_{\max}$ and $p_{\max}$ into the work production expression, we find the maximum:

$$\langle W \rangle_{\max}(b) = \langle W \rangle(b, p_{\max}(b), q_{\max}(b)) \ ,$$

yielding the solid (blue) curve in Fig. 4.6. The curve is the maximum work production $\langle W \rangle_{\max}(b)$ of a memoryless ratchet for an input with bias b. This may seem like a limited result at first, since it was calculated by driving a memoryless ratchet with memoryless inputs. However, memoryless ratchets are insensitive to temporal correlations, and finite

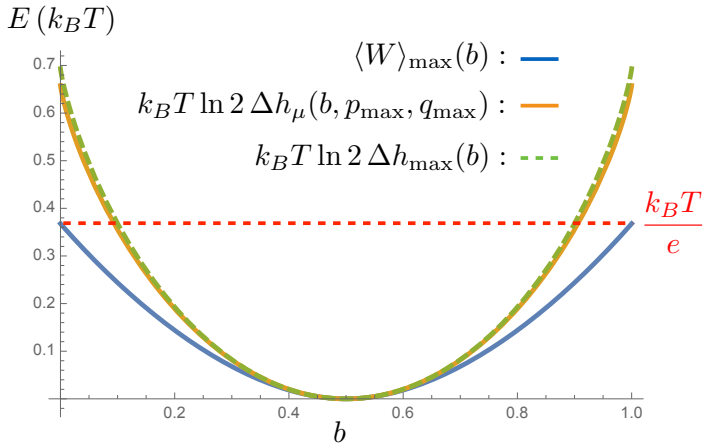


Figure 4.6. Maximum work production $\langle W\rangle_{\max}$ for any input bias b is $k_{\mathrm{B}}T/e$ (horizontal dashed line) and so ratchets do not achieve the IPSL upper bound $\langle W\rangle \leq k_BT\Delta h_\mu(b, p_{\max}, q_{\max})$ that derives from pure informational properties of the input and output processes. Also, $\Delta h_\mu(b, p_{\max}, q_{\max})$ itself is slightly less than the absolute maximum possible change in entropy $\Delta h_{\max}(b)$ given an input bias b. This means that a memoryless ratchet does not leverage all of the single-symbol statistical order in the input. This is also true of memoryful ratchets when driven by memoryless processes of the same input bias. Appendix 4.4.1 tells us that the blue curve is also the maximal work production for these ratchets with memoryless inputs.

memory ratchets are no better than memoryless ratchets when driven by memoryless inputs, as discussed in the next section. Thus, the blue curve also represents the maximum work production of memoryless ratchets with memoryful inputs as well as of memoryful ratchets with memoryless inputs, as long as the probability of an input 0 is b.

To compare work production directly with the IPSL and related bounds, Eqs. (4.1) and (4.2), we need to calculate the changes in single-symbol entropy difference $\Delta\,\mathrm{H}_1$ and entropy-rate difference Δh_μ. Reminding ourselves that the ratchet is memoryless, these differences are the same if we assume the input to be memoryless. We find:

$$\begin{aligned}\Delta\,\mathrm{H}_1 &= \Delta h_\mu(b, p, q) \\ &= \mathrm{H}_{\mathrm{B}}(b') - \mathrm{H}_{\mathrm{B}}(b)\ ,\end{aligned}$$

with $\mathrm{H}_{\mathrm{B}}(z) = \mathrm{H}(\{z, 1-z\})$ for $z \in [0, 1]$, the binary entropy function [11]. We obtain the bounds for an optimal ratchet, for a given input bias b, by substituting $p_{\max}$ and $q_{\max}$ for p and q, respectively. We plot this optimal bound as the dashed line (orange) in Fig. 4.6. Even though we maximized over the memoryless ratchet's parameters (p and

q), the output work $\langle W \rangle_{\max}(b)$ falls far short of the bounds set on it, as the solid (blue) curve lies below the dashed (orange) curve except exactly at $b = 1/2$, where there is zero work production. This demonstrates that there are inherent inefficiencies in memoryless information ratchets.

There is a second source of inefficiency for memoryless ratchets. The maximum possible bound for the generated work comes from the case where there are no statistical biases and no correlations left in the output sequence, so that the output has maximum Shannon entropy. In this case we have $b' = 1/2$, the maximal entropy change being:

$$\Delta h_{\max}(b) = 1 - \mathrm{H}_B(b) \ .$$

Figure 4.6 plots the corresponding bound as a dashed line (green), showing that it lies above the actual change in entropy for an optimal ratchet. Thus, not all of the order in the input sequence is being leveraged to generate work. In fact, the output bias $b'(b, p_{\max}, q_{\max})$ for a nonequilibrium optimal ratchet is generally not equal to $1/2$, meaning that another ratchet could use its output as fuel.

4.3.2 Optimal memoryless ratchets versus memoryful ratchets

At this point we may ask: Is it possible to surpass the optimal memoryless ratchet in terms of work production with a memoryful ratchet? The answer seems to be negative for memoryless inputs. More to the point, Appendix 4.4.1 proves the following statement:

> For memoryless, binary inputs work production by the optimal memoryless ratchet cannot be surpassed by any memoryful ratchet.

Thus, by optimizing over memoryless ratchets, we can actually determine the optimum work production over all finite memoryful ratchets. Appendix 4.4.1 proves that for sequences of binary symbols, memoryless ratchets are optimal for producing work.

This has a number of implications. First of all, it means that the dashed (blue) curve in Fig. 4.6 is not only a bound on the work production of a memoryless ratchet for any input with bias b, but it is also a bound on the work production of *any* finite memory ratchet with a memoryless input with the same bias. Second, in particular, the work production is at most $k_B T/e$, as shown by the dashed (red) horizontal line.

Third, importantly, this line is less than the conventional Landauer bound of $k_B T \ln 2$. It may seem counterintuitive that no single ratchet can autonomously achieve the Landauer bound, but this is a natural result of dynamics of this class of autonomous Maxwellian Demon. Between each interaction interval, as the ratchet switches between inputs, the Hamiltonian changes instantaneously and discontinuously. As a result, the ratchet and bit exist in a nonequilibrium distribution, which dissipates unrecoverable heat as it relaxes towards the Boltzmann distribution.

Finally, to achieve entropic bounds, the joint state of the ratchet and bit should follow the Boltzmann distribution during the interaction. However, to do this while also performing meaningful computation, it is necessary to implement some form of adiabatic isothermal protocol [144]. The latter can be implemented either by dynamically controlling the ratchet's energies over the interaction interval, as we discuss in Chapters 5 and 6, or by stringing together a series of ratchets, each one gradually updating the distribution of it's input infinitesimally.

Appendix 4.4.1's observation also suggests that multiple ratchets in series—the output sequence of one is input to the next—cannot be represented as a single finite-memory ratchet that interacts with one bit at a time and only once. This is because we can surpass the work production of an optimal memoryless ratchet with multiple ratchets interacting with multiple symbols at a time, as we noted already. Ratchets composed in series form a fundamentally different construction than a single memoryful ratchet; a topic of some biological importance to which we will return elsewhere.

4.3.3 Infinite-memory ratchets

We emphasized that the very general IPSL bound on information processing based on input-output entropy rate change holds for finite-state ratchets. What happens if infinite memory is available to a ratchet? This section constructs infinite-memory ratchets that can violate both Eqs. (4.1) and (4.2) and, by implication, Landauer's bound. The intuition behind this is that, due to the infinite memory, the ratchet can continue indefinitely to store information that need not be written to the output. In effect, an apparent violation of the IPSL bound arises since the hidden degrees of freedom of the ratchet's

memory are not accounted for.

Nonetheless, infinite-memory ratchets offer intriguing possibilities for thermodynamic and computational functionality. While finite-memory ratchets can do meaningful computations and can even be appropriate models for, say, biological organisms that have finite information-processing capacities and response times, they cannot be computationally universal in the current architecture [145, 68]. More precisely, a one-way universal Turing machine (UTM), like our ratchet, that reads its input once and never again, requires an "internal" infinite work tape to read from and write on. So, an infinite-state ratchet of our type is needed to emulate the infinite bidirectional read-write tape of the usual UTM [137].

Appendix 4.4.1 shows that memoryless ratchets are able to extract the most work from memoryless binary input processes, under the assumption that the ratchet's memory is finite. Without finiteness the proof breaks down, since an asymptotic state distribution may not exist over infinite states [146]. In addition, the proof of Eq. (4.1) fails for the same reason. Thus, we turn to other tools for understanding the behavior in this case. The expression for work production still holds, so despite not having general informational bounds on work production, we can still calculate the exact work production for a prototype infinite ratchet.

Here, we present an infinite-state ratchet with finite energy-differences between all states. Our main result is that it produces more work than any finite memory ratchet for a given input. More to the point, it violates both the bounds in Eqs. (4.1) and (4.2). This demonstrates the need for the finite-memory assumption in developing Landauer and IPSL bounds. Consider, for example, an input process of all 1s. According to Sec. 4.3.1, the maximum amount of work that can be extracted from this input by a memoryless ratchet is given by:

$$\langle W \rangle_{\max} = \frac{k_{\mathrm{B}} T}{e} \ .$$

The discussion in App. 4.4.1 indicates that this should be the maximum amount of work that can be extracted by any finite-memory ratchet (for the same input). However, the

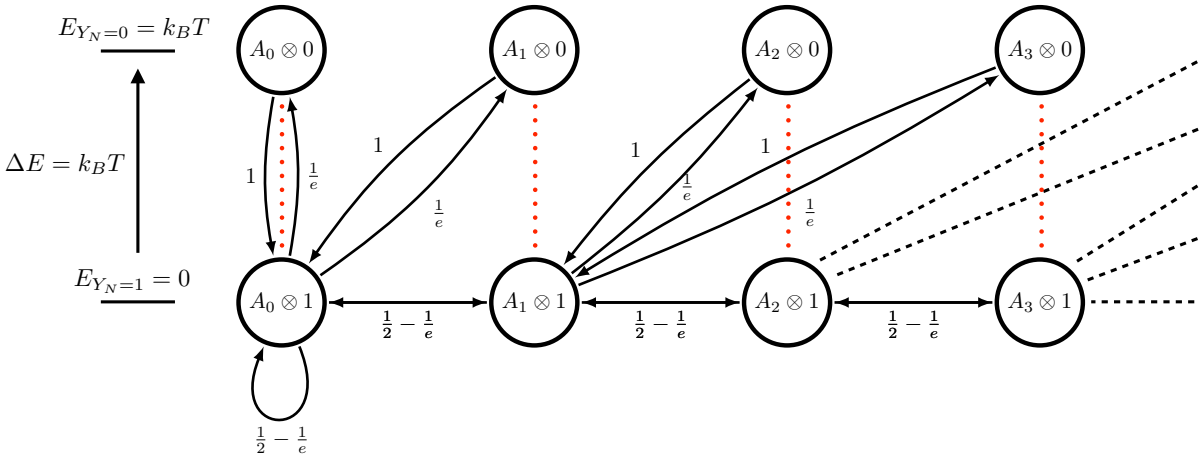


Figure 4.7. Infinite-state ratchet that violates the IPSL and single-symbol bounds, Eqs. (4.1) and (4.2), respectively. The ratchet state-space is given by $\mathcal{X} = \{A_0, A_1, A_2, \ldots\}$: all states effectively have the same energy. The symbol values $\mathcal{Y} = \{0, 1\}$ differ by energy $\Delta E = k_B T$, with 0 having higher energy. The black arrows indicate the possible *interaction transitions* among the shown joint states of the ratchet and symbol during the interaction interval. For example, transitions $A_0 \otimes 1 \leftrightarrow A_1 \otimes 1$ are allowed whereas transitions $A_0 \otimes 0 \leftrightarrow A_1 \otimes 0$ are not. The dashed black lines show interaction transitions between the shown joint states and joint states that could not be shown. Briefly, for $i \geq 1$, there can be only the following interaction transitions: $A_i \otimes 1 \to \{A_{i\pm1} \otimes 1, A_{2i} \otimes 0, A_{2i+1} \otimes 0\}$ and $A_i \otimes 0 \to A_{j(i)} \otimes 1$ with $j(i) = i/2$ for even i and $(i-1)/2$ for odd i. For the $i = 0$ transitions, see the diagram. Every interaction transition is followed by a switching transition and vice versa. The red dotted lines are the possible paths for driven *switching transitions* between the joint states, which correspond to the production or dissipation of work. During the switching interval, the only allowed transitions are the vertical transitions between energy levels $A_i \otimes 0 \leftrightarrow A_i \otimes 1$. The probability of these transitions depends on the input bias.

infinite-state ratchet shown in Fig. 4.7 produces twice as much work, as we now show:

$$\langle W \rangle^{\infty} = \frac{2 k_{\mathrm{B}} T}{e} \ .$$

The infinite-state ratchet also violates both of the IPSL and single-symbol bounds, Eqs. (4.1) and (4.2), since $k_B T \ln 2$ is an upper bound for the work generation in all binary input processes according to these bounds, whereas $2/e > \ln 2$.

Let's describe the structure and dynamics of the infinite-state ratchet in Fig. 4.7 in detail. This ratchet has a countably infinite number of states A_i, with $i \in \{0, 1, 2, \ldots\}$. In other words, the ratchet state space is $\mathcal{X} = \{A_0, A_1, A_2, \ldots\}$. The joint dynamic of the ratchet and the interacting symbol is shown in Fig. 4.7, where the arrows indicate allowed transitions and the number along the arrow, the associated transition probabilities. Apart

from the case $i = 0$, only the following transitions are allowed: $A_i \otimes 1 \to \{A_{i\pm1} \otimes 1, A_{2i} \otimes 0, A_{2i+1} \otimes 0\}$ and $A_i \otimes 0 \to A_j \otimes 1$ with $j = i/2$ for even i and $(i-1)/2$ for odd i. If the incoming symbol is 0, the only transition allowed involves a simultaneous change in the ratchet state and symbol, switching over to state $A_{j(i)}$ if it started in state A_i and the symbol switching to 1. If the incoming symbol is 1, there are generally four possible transitions: $A_i \otimes 1 \to A_{i\pm1} \otimes 1$, $A_i \otimes 1 \to A_{2i} \otimes 0$, and $A_i \otimes 1 \to A_{2i+1} \otimes 0$. The first two transitions occur with equal probabilities $1/2 - 1/e$, while the third and fourth transition occurs with probability $1/e$. For $i = 0$, there are four transitions possible: $A_0 \otimes 1 \to \{A_0 \otimes 1\,(\text{self-loop}), A_1 \otimes 1, A_0 \otimes 0, A_1 \otimes 1\}$. The transition probabilities are shown in the figure.

We can assign relative energy levels for the joint states $A_i \otimes \{0, 1\}$ based on the transition probabilities. Since the (horizontal) transitions $A_i \otimes 1 \leftrightarrow A_{i+1} \otimes 1$ have equal forward and reverse transition probabilities, all the joint states $A_i \otimes 1$ have the same energy. Any state $A_i \otimes 0$ is higher than the state $A_{j(i)\otimes 1}$ by an energy:

$$\begin{aligned}\Delta E_{A_i\otimes 1 \to A_j \otimes 0} &= k_B T \ln \frac{1}{1/e} \\ &= k_B T \ .\end{aligned}$$

As a result, all states $A_i \otimes 0$ have the same energy, higher than that of the states $A_i \otimes 1$ by $k_B T$. This energy difference is responsible for producing the work. When the ratchet is driven by the all-1s process, if it is in an $A_i \otimes 0$ state after the previous interaction transition, then the switching transition changes the state to $A_i \otimes 1$ gaining $\Delta E_{A_i\otimes 0 \to A_i \otimes 1} = k_B T$ in work. The probability of being in a $Y_N = 0$ state after an interaction interval is $2/e$, so the work production is $\langle W \rangle = 2k_B T/e$, as stated above.

The reason this infinite-state ratchet violates the information-theoretic bounds is that those bounds ignore the asymptotic entropy production in the ratchet's internal state space. There is no steady state over the infinite set of states and this leads to continual entropy production within the ratchet's state space $\mathcal{X}$. For the specific case of the all-1s input process note that, before the interaction interval, the joint state-space distribution of the ratchet and the incoming symbol must be positioned over only $A_i \otimes 1$ states.

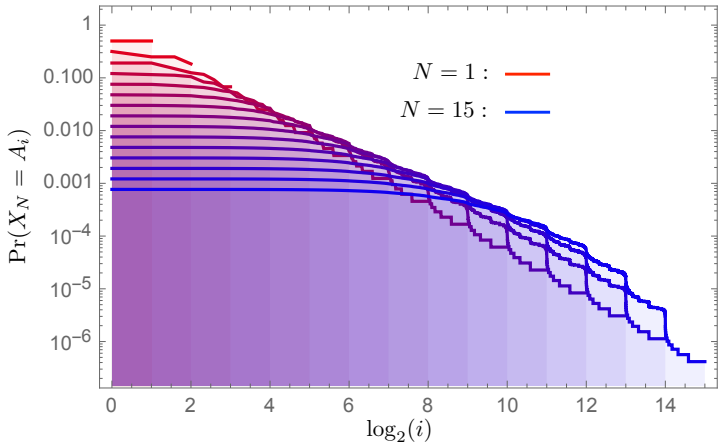


Figure 4.8. Evolution of infinite-ratchet state distribution starting from an initial distribution peaked over the set $\mathcal{X}$, whose states are indexed A_i. The state distribution curves are plotted over 15 time steps, starting in red at time step $N = 1$ and slowly turning to blue at time step $N = 15$. With each sequential step, the support of the ratchet state distribution doubles in size, leading to increasing uncertainty in the ratchet state space and so increasing state entropy.

This is due to the fact that the switching transition always changes the symbol value to 1. From a distribution $\{\Pr(X_N = A_i, Y_N = 1)\}_{i\in\{0,1,\ldots\}}$ over the $A_i \otimes 1$ states at time N, the interaction interval spreads the joint distribution to both $A_i \otimes 0$ and $A_i \otimes 1$ states. However, they are reset to a new distribution over the $A_i \otimes 1$ states $\{\Pr(X_{N+1} = A_i, Y_{N+1} = 1)\}_{i\in\{0,1,\ldots\}}$ after the following switching transition. This leads to a spreading of the probability distribution—and, therefore, to an increase in entropy—in the ratchet space $\mathcal{X}$ after each time step.

Figure 4.8 demonstrates the spreading by setting the initial joint ratchet-symbol state $X_0 \otimes Y_0$ to $A_0 \otimes 0$ and letting the distribution evolve for $N = 15$ time steps over the ratchet states. The ratchet states are indexed by i and the time steps are indexed by N, going from 1 to 15. The curves show the probabilities $\Pr(X_N = A_i)$ of the ratchet at time step N being in the ith ratchet state. By filling the area under each distribution curve and plotting the ratchet-state index in logarithm base 2, we see that the distribution's support doubles in size after every time step. This indicates an increase in the ratchet's internal entropy at each time step. This increase in internal entropy is responsible for the violation of the IPSL bounds in Eqs. (4.1) and (4.2).

We have yet to discover a functional form for a steady state that is invariant—that

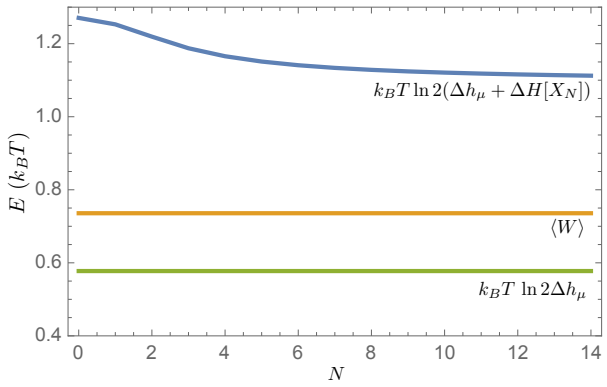


Figure 4.9. The dashed (orange) line indicates average work production $\langle W \rangle$ per time step. It lies above the dotted (green) curve that indicates the IPSL entropy-rate bound on $\langle W \rangle$ (Eq. (4.1)), indicating a violation of the latter. The interpretation of the violation comes from the solid (blue) curve that indicates the joint entropy production of the input process and the ratchet together. We see a violation of the entropy-rate bound since there is continuous entropy production in the ratchet's (infinite) state space.

maps to itself under one time-step. We made numerical estimates of the ratchet's entropy production, though. From the distributions shown in Fig. 4.8, we calculated the ratchet's state entropies at each time step N. The entropy production $\Delta \mathrm{H}[X_N] = \mathrm{H}[X_{N+1}] - \mathrm{H}[X_N]$ at the Nth step is shown in Fig. 4.9. We see that the sum $\Delta \mathrm{H}[X_N] + \Delta h_\mu$ of the changes in ratchet entropy and symbol entropy upper bounds the work production. Note that only the Δh_μ curve lies below the work production. Thus, while this infinite ratchet violates the IPSL bounds of Eqs. (4.1) and (4.2), it still satisfies a more general version of the Second Law of Thermodynamics for information ratchets—Eq. (A7) of Ref. [13]:

$$\langle W_N \rangle \leq k_B T \ln 2 \left(\mathrm{H}_{N+1} - \mathrm{H}_N \right) \ , \tag{4.15}$$

where W_N is the work gain at the Nth time step and $\mathrm{H}_N = \mathrm{H}[X_N, Y_{N:\infty}, Y'_{0:N}]$ is the joint Shannon entropy of the ratchet and the input and output symbol sequences $Y_{N:\infty}$ and $Y'_{0:N}$, respectively, at time $t = N$. As we can see, this bound is based on not only the input and output process statistics, but also the ratchet memory. We will explore how the memory of the ratchet plays a role in thermodynamic bounds in the next two chapters.

Conclusion

How an agent interacts with and leverages it's environment is a topic of broad interest, from engineering and cybernetics to biology and now physics [18, 106]. General principles for how the structure of an agent must match that of its environment will become essential tools for understanding how to take thermodynamic advantage of correlations in structured environments, whether the correlations are temporal or spatial. Ashby's Law of Requisite Variety—a controller must have at least the same variety as its input so that the whole system can adapt to and compensate that variety and achieve homeostasis [18]—was an early attempt at such a general principle of regulation and control. In essence, a controller's variety should match that of its environment. Above, paralleling this, we showed that a near-optimal thermal agent (information engine) interacting with a structured input (information reservoir) obeys a similar variety-matching principle.

For an efficient finite-state information ratchet, the ratchet memory should reflect the memory of the input process. More precisely, memoryless ratchets are optimal for leveraging memoryless inputs, while memoryful ratchets are optimal for leveraging memoryful inputs. This can be appreciated in a two different ways.

On the one hand, the first comes from information processing properties of the ratchet and input and the associated IPSL bounds on work. The operation of memoryless ratchets can only destroy temporal correlations. These ratchets' work production is still bounded by single-symbol entropy changes, as in Eq. (4.2). And, since memoryless input processes only produce single-symbol correlations (statistical biases), the memoryless ratchet bound of Eq. (4.2) allows for maximal work production. Thus, according their bounds, memoryless ratchets and inputs produce the most work *when paired*.

On the other hand, in the second view memoryful input processes exhibit multiple-symbol temporal correlations. And, the entropy rate bound of Eq. (4.1) suggests that the memoryful input processes can be used to produce work in a memoryful ratchet, but not a memoryless one. More precisely, we can conceive of memoryful input processes whose single-symbol statistics are unbiased (equal proportions of 0s and 1s, in case of binary alphabet) but the entropy rate is smaller than the single-symbol entropy: $h_\mu < \mathrm{H}_1$. In

this case, since the single-symbol entropy is already at its maximum possible value, memoryless ratchets are unable to extract any work. Since the memoryful ratchets satisfy the IPSL bound of Eq. (4.1), however, they can extract work from such memoryful processes. One such example is studied in detail by Ref. [6]. (For a quantum-mechanical ratchet, compare Ref. [73].) Thus, memoryful ratchets are best paired with memoryful inputs. This and its complement result—memoryless inputs are optimally used by memoryless ratchets—is biologically suggestive. If one observes memory (temporal correlations) in the transduction implemented by a biomolecular assembly, for example, then it has adapted to some structured environment.

We summarized the role of memory in thermodynamic processes in Fig. 4.3 which considers each of the four possible combinations of memoryful or memoryless ratchets with memoryful or memoryless input.

While the Second Law of Thermodynamics determines the IPSL and related bounds discussed here, it does not follow that the bounds are achievable for the class of nonequilibrium information ratchets considered. Based on an exact method for calculating the average work production [6], we saw that there are indeed situations where the bounds are not achievable (Fig. 4.6). In Sec. 4.3.1, we saw that memoryless ratchets cannot generally saturate their bound (Eq. (4.2)). Furthermore, based on the results of App. 4.4.1 we could prove that finite memoryful ratchets fare no better than memoryless ratchets at leveraging memoryless inputs. Thus, not even memoryful ratchets can extract the maximum amount of work possible from a memoryless input. There are some hints, though, as to what the architecture of information engines should be to extract the maximum possible work allowed by the Second Law. We alluded to one such situation in Sec. 4.3.1 involving isothermal or adiabatic control of the energy landscape, which we discuss more in the following chapters.

Another path to addressing the unattainability of the IPSL bound observed above pertains to the architecture of the nonequilibrium information engines where there is only a single ratchet that interacts with one environmental signal value at a time. This leads one to speculate that multiple ratchets interacting with different signals—say, chained

together so that the output of one is the input of another—will lead to a closer approach to the bound. Simply having multiple copies of the optimal memoryless ratchets one after another, however, will not necessarily address unattainability. Interestingly, depending on input bias b, there may be oscillations in the amount of work that is gained per cycle. And, even with infinitely many ratchets chained together sequentially, we may still be far from the IPSL bound. Based on our intuition about thermodynamically reversible processes, we postulate that to approach the bound more closely we need increasingly many memoryless ratchets, each optimized with respect to *its own input*. We leave the verification of this intuition for a future investigation. This does suggest, though, architectural trade-offs that should manifest themselves in evolved biological thermodynamic processes.

To complete our exploration of the role of memory in thermodynamic processes, we considered infinite-state ratchets, which are necessary if we wish to physically implement universal Turing machines with the unidirectional information ratchets. Infinite ratchets, however, pose a fundamental challenge since the IPSL entropy-rate bound on work production does not apply to them. The proof of the bound (Eq. (4.1) [6]) is based on the assumption that the ratchet reaches a steady state after interacting with sufficiently many input symbols. This need not be the case for infinite-state ratchets. In fact, the numerical investigations of Sec. 4.3.3 indicate that the probability distribution in the state space of an infinite ratchet can continue to spread indefinitely, without any sign of relaxing to a steady state; recall Fig. 4.8. By calculating both the average work production per time step and the amount of change in the entropy rate, Fig. 4.9 showed that there is a violation of the IPSL and related bounds. This necessitates a modification of the IPSL for infinite-state ratchets. The appropriate bound, though, has already been presented in a previous work [6], which we quoted in Eq. (4.15). This relation shows that the work production is still bounded by the system's entropy production; only, we must include the contribution from the ratchet's internal state space on top of the entropy-rate difference of the input and the output HMMs.

We close by highlighting the close correspondence between information ratchets and biological enzymes. Most directly, it is possible to model the biomimetic enzymes following

the design of information ratchets [85]. The correspondence goes further, though. In Sec. 4.1, we discussed how a swarm of ratchets acting cooperatively may be more efficient than individual information ratchets, even if they are quite sophisticated. A similar phenomenon holds for enzymes where the enzymes along a metabolic pathway assemble to form a multi-enzyme complex—a "swarm"—to affect faster, efficient reaction turnover, known as *substrate channeling* [147].

4.4 Appendices

4.4.1 Optimally Leveraging Memoryless Inputs

It is intuitively appealing to think that memoryless inputs are best utilized by memoryless ratchets. In other words, the optimal ratchet for a memoryless input is a memoryless ratchet. We prove the validity of this intuition in the following. We start with the expression of work production per time step:

$$\begin{aligned}
\beta\langle W\rangle &= \sum_{x,x',y,y'} \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ln \frac{M_{x'\otimes y'\to x\otimes y}}{M_{x\otimes y\to x'\otimes y'}} \\
&= \sum_{x,x',y,y'} \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ln \frac{\pi_{x'\otimes y'} M_{x'\otimes y'\to x\otimes y}}{\pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'}} \\
&\quad - \sum_{x,x',y,y'} \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ln \frac{\pi_{x'\otimes y'}}{\pi_{x\otimes y}},
\end{aligned}$$

with $\beta = 1/k_B T$. The benefit of the decomposition in the second line will be clear in the following. Let us introduce several quantities that will also be useful in the following:

$$\begin{aligned}
p(x,y,x',y') &= \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \ , \\
p_R(x,y,x',y') &= \pi_{x'\otimes y'} M_{x'\otimes y'\to x\otimes y} \ , \\
\pi_x^X &= \sum_y \pi_{x\otimes y} \ , \\
\pi_y^Y &= \sum_x \pi_{x\otimes y} \ , \\
p^X(x,x') &= \sum_{y,y'} p(x,y,x',y') \ , \text{ and} \\
p^Y(y,y') &= \sum_{x,x'} p(x,y,x',y').
\end{aligned}$$

For a memoryless input process, sequential inputs are statistically independent. This implies Y_N and X_N are independent, so the stationary distribution $\pi_{x\otimes y}$ can be written as a product of marginals:

$$\pi_{x\otimes y} = \pi_x^X \pi_y^Y . \tag{4.16}$$

In terms of the above quantities, we can rewrite work for a memoryless input process as:

$$\begin{aligned}\beta\langle W\rangle = &-D_{KL}(p\|p_R) \\ &- \sum_{y,y'} p^Y(y,y') \ln \frac{\pi_{y'}^Y}{\pi_y^Y} - \sum_{x,x'} p^X(x,x') \ln \frac{\pi_{x'}^X}{\pi_x^X} \ ,\end{aligned}$$

where $D_{KL}(p\|p_R)$ is the relative entropy of the distribution p with respect to p_R [11]. Note that the last term in the expression vanishes, since the ratchet state distribution is the same before and after an interaction interval:

$$\sum_x p^X(x,x') = \sum_x p^X(x',x) = \pi_{x'}^X, \tag{4.17}$$

and so:

$$\begin{aligned}&\sum_{x,x'} p^X(x,x') \ln \frac{\pi_{x'}^X}{\pi_x^X} \\ &\quad = \sum_{x,x'} p^X(x,x') \ln \pi_{x'}^X - \sum_{x,x'} p^X(x,x') \ln \pi_x^X \\ &\quad = \sum_{x'} \pi_{x'}^X \ln \pi_{x'}^X - \sum_x \pi_x^X \ln \pi_x^X \\ &\quad = 0 \ .\end{aligned}$$

Thus, we find find the average work production to be:

$$\beta\langle W\rangle = -D_{KL}(p\|p_R) - \sum_{y,y'} p^Y(y,y') \ln \frac{\pi_{y'}^Y}{\pi_y^Y} \ . \tag{4.18}$$

Let us now use the fact that the coarse graining of any two distributions, say p and q, yields a smaller relative entropy between the two [11, 148]. In the work formula, p^Y is a coarse graining of p and p_R^Y is a coarse graining of p^R, implying:

$$D_{KL}(p^Y\|p_R^Y) \leq D_{KL}(p\|p_R) \ . \tag{4.19}$$

Combining the above relations, we find the inequality:

$$\beta\langle W\rangle \leq -D_{KL}(p^Y\|p_R^Y) - \sum_{y,y'} p^Y(y,y')\ln\frac{\pi_{y'}^Y}{\pi_y^Y} \ .$$

Now, the marginal transition probability $p^Y(y,y')$ can be broken into the product of the stationary distribution over the input variable π_y^Y and a Markov transition matrix $M_{y\to y'}^Y$ over the input alphabet:

$$p^Y(y,y') = \pi_y^Y M_{y\to y'}^Y \ ,$$

which for any ratchet M is:

$$\begin{aligned} M_{y\to y'}^Y &= \frac{1}{\pi_y^Y} p^Y(y,y') \\ &= \frac{1}{\pi_y^Y}\sum_{x,x'} \pi_{x\otimes y} M_{x\otimes y\to x'\otimes y'} \\ &= \frac{1}{\pi_y^Y}\sum_{x,x'} \pi_x^X \pi_y^Y M_{x\otimes y\to x'\otimes y'} \\ &= \sum_{x,x'} \pi_x^X M_{x\otimes y\to x'\otimes y'} \ . \end{aligned}$$

We can treat the Markov matrix M^Y as corresponding to a ratchet in the same way as M. Note that M^Y is effectively a memoryless ratchet since we do not need to refer to the internal states of the corresponding ratchet. See Fig. 4.2. The resulting work production for this ratchet $\langle W^Y\rangle$ can be expressed as:

$$\begin{aligned} \beta\langle W^Y\rangle &= \sum_{y,y'} \pi_y^Y M_{y\to y'}^Y \ln\frac{M_{y'\to y}^Y}{M_{y\to y'}^Y} \\ &= -D_{KL}(p^Y\|p_R^Y) - \sum_{y,y'} p^Y(y,y')\ln\frac{\pi_{y'}^Y}{\pi_y^Y} \\ &\geq \beta\langle W\rangle \ . \end{aligned}$$

Thus, for any memoryful ratchet driven by a memoryless input we can design a memoryless ratchet that extracts at least as much work as the memoryful ratchet.

There is, however, a small caveat. Strictly speaking, we must assume the case of binary input. This is due to the requirement that the matrix M be detailed balanced

(see Sec. 4.1) so that the expression of work used here is appropriate. More technically, the problem is that we do not yet have a proof that if M is detailed balanced then so is M^Y, a critical requirement above. In fact, there are examples where M^Y does not exhibit detailed balance. We do, however, know that M^Y is guaranteed to be detailed balanced if $\mathcal{Y}$ is binary, since that means M^Y only has two states and all flows must be balanced. Thus, for memoryless binary input processes, we established that there is little point in using finite memoryful ratchets to extract work: memoryless ratchets extract work optimally from memoryless binary inputs.

4.4.2 An IPSL for Information Engines

Reference [13] proposed a generalization of the Second Law of Thermodynamics to information processing systems (IPSL, Eq. (4.1)) under the premise that the Second Law can be applied even when the thermodynamic entropy of the information bearing degrees of freedom is taken to be their Shannon information entropy. This led to a consistent prediction of the thermodynamics of information engines. It was also validated through numerical calculations. This appendix proves this assertion for the class of information engines considered here. The key idea is to use the irreversibility of the Markov chain dynamics followed by the engine and by the information bearing degrees of freedom to derive the IPSL inequality.

For the sake of presentation, we introduce new notation here. We refer to the engine as the demon D, following the original motivation for information engines. We refer to the information-bearing two-state systems as the bits B. According to our set up, D interacts with an infinite sequence of bits, $\mathrm{B}_0\mathrm{B}_1\mathrm{B}_2\ldots$ as shown in Fig. 4.10. The figure also explains the connection of the current terminology to that in the main text. In particular, we show two snapshots of our setup, at times $t = N$ and $t = N + 1$. During that interval D interacts with bit $\mathrm{B_N}$ and changes it from (input) symbol Y_N to (output) symbol Y'_N. The corresponding dynamics is governed by the Markov transition matrix $M_{\mathrm{D}\otimes\mathrm{B}_N}$ which acts *only* on the joint subspace of D and B_N.

Under Markov dynamics the relative entropy of the current distribution with respect to the asymptotic steady-state distribution is a monotonically decreasing function of time.

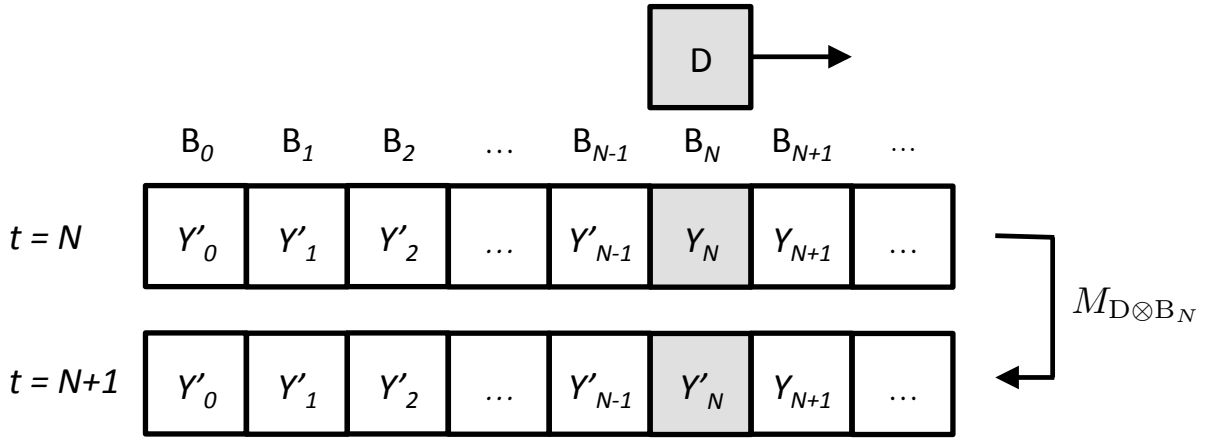


Figure 4.10. The demon D interacts with one bit at a time for a fixed time interval; for example, with bit B_N for the time interval $t = N$ to $t = N + 1$. During this, the demon changes the state of the bit from (input) Y_N to (output) Y'_N. There is an accompanying change in D's state as well, not shown. The joint dynamics of D and B_N is governed by the Markov chain $M_{\mathrm{D}\otimes\mathrm{B}_N}$.

We now use this property for the transition matrix $M_{\mathrm{D}\otimes\mathrm{B}_N}$ to derive the IPSL. Denote the distribution of D's states and the bits B at time t by $P_{\mathrm{DB}_{0:\infty}}(t)$. Here, $\mathrm{B}_{0:\infty}$ stands for all the information-bearing degrees of freedom [1]. The steady-state distribution corresponding to the operation of $M_{\mathrm{D}\otimes\mathrm{B}_N}$ is determined via:

$$\lim_{n\to\infty} M^n_{\mathrm{D}\otimes\mathrm{B}_N} P_{\mathrm{DB}_{0:\infty}}(N) = \pi^{\mathrm{eq}}_{\mathrm{DB}_N} P_{\mathrm{B}_{0:\infty/N}}(N) \tag{4.20}$$

$$\equiv \pi^{\mathrm{s}}(N) \ , \tag{4.21}$$

where $\pi^{\mathrm{eq}}_{\mathrm{DB}_N}$ denotes the steady-state distribution:

$$M_{\mathrm{D}\otimes\mathrm{B}_N} \pi^{\mathrm{eq}}_{\mathrm{DB}_N} = 0$$

and $P_{\mathrm{B}_{0:\infty/N}}(N)$ the marginal distribution of all the bits other than the N-th bit at time $t = N$. We introduce $\pi^{\mathrm{s}}(N)$ in Eq. (4.21) for brevity.

[1] Probability distributions over infinitely many degrees of freedom ultimately require a measure-theoretic treatment. This is too heavy a burden in the current context. The difficulties can be bypassed by assuming that the number of bits in the infinite information reservoir is a large, but positive finite integer L. And so, instead of infinities in $\mathrm{B}_{0:\infty}$ and $\mathrm{B}_{0:\infty/N}$ we use $\mathrm{B}_{0:L}$ and $\mathrm{B}_{0:L/N}$, respectively, and take the appropriate limit when needed.

The rationale behind the righthand side of Eq. (4.20) is that the matrix $M_{\mathrm{D}\otimes\mathrm{B}_N}$ acts only on D and B_N, sending to their joint distribution to the stationary distribution $\pi^{\mathrm{eq}}_{\mathrm{DB}_N}$ (on repeated operation), while leaving intact the marginal distribution of the rest of B. The superscript eq emphasizes the fact the distribution $\pi^{\mathrm{eq}}_{\mathrm{DB}_N}$ is an equilibrium distribution, as opposed to a nonequilibrium steady-state distribution, due to the assumed detailed-balance condition on $M_{\mathrm{D}\otimes\mathrm{B}_N}$. In other words, $\pi^{\mathrm{eq}}_{\mathrm{DB}_N}$ follows the Boltzmann distribution:

$$\pi^{\mathrm{eq}}_{\mathrm{DB}_N}(\mathrm{D}=x,\mathrm{B}_N=y)=e^{\beta[F_{\mathrm{DB}_N}-E_{\mathrm{DB}_N}(x,y)]} \tag{4.22}$$

for inverse temperature β, free energy F_{DB_N}, and energy $E_{\mathrm{DB}_N}(x,y)$. In the current notation we express the monotonicity of relative entropy as:

$$D(P_{\mathrm{DB}_{0:\infty}}(N)\|\pi^{\mathrm{s}}(N))\geq D(P_{\mathrm{DB}_{0:\infty}}(N+1)\|\pi^{\mathrm{s}}(N)), \tag{4.23}$$

where $D(p\|q)$ denotes the relative entropy of the distribution p with respect to q:

$$D(p\|q)=\sum_i p(i)\ln\left[\frac{p(i)}{q(i)}\right]$$

over D's states i. The IPSL is obtained as a consequence of inequality Eq. (4.23), as we now show [2].

First, we rewrite the lefthand side of Eq. (4.23) as:

$$\begin{aligned}
&D(P_{\mathrm{DB}_{0:\infty}}(N)\|\pi^{\mathrm{s}}(N))\\
&=-\mathrm{H}_{\mathrm{DB}_{0\infty}}(N)\ln 2-\sum_{\mathrm{DB}_{0:\infty}}P_{\mathrm{DB}_{0:\infty}}(N)\ln\pi^{\mathrm{s}}(N)\\
&=-\mathrm{H}_{\mathrm{DB}_{0\infty}}(N)\ln 2-\sum_{\mathrm{DB}_N}P_{\mathrm{DB}_N}(N)\ln\pi^{\mathrm{eq}}_{\mathrm{DB}_N}\\
&\quad-\sum_{\mathrm{DB}_{0:\infty/N}}P_{\mathrm{DB}_{0:\infty/N}}(N)\ln P_{\mathrm{B}_{0:\infty/N}}(N)\\
&=-\mathrm{H}_{\mathrm{DB}_{0\infty}}(N)\ln 2-\beta F_{\mathrm{DB}_N}+\beta\langle E_{\mathrm{DB}_N}\rangle(N)\\
&\quad+\mathrm{H}_{\mathrm{B}_{0:\infty/N}}(N)\ln 2\ .
\end{aligned} \tag{4.24}$$

[2] For a somewhat similar approach see N. Merhav. J. Stat. Mech.: Th. Expt. (2017) 023207.

The first line applies the definition of relative entropy. Here, $\mathrm{H_X}$ denotes the Shannon entropy of random variable X in *information units* of bits (base 2). The second line employs the expression of $\pi^{\mathrm{s}}(N)$ given in Eq. (4.21). The final line uses the Boltzmann form of $\pi^{\mathrm{eq}}_{\mathrm{DB}_N}$ given in Eq. (4.22). Here, $\langle E_{\mathrm{DB}_N}\rangle(N)$ denotes the average energy of D and the interacting bit B_N at time $t = N$.

Second, in a similar way, we have the following expression for the righthand side of Eq. (4.23):

$$\begin{aligned} D(P_{\mathrm{DB}_{0:\infty}}(N+1)\|\pi^{\mathrm{s}}_N) & \\ &= -\mathrm{H}_{\mathrm{DB}_{0:\infty}}(N+1)\ln 2 + \mathrm{H}_{\mathrm{B}_{0:\infty/N}}(N)\ln 2 \\ &\quad + \beta\langle E_{\mathrm{DB}_N}\rangle(N+1) - \beta F_{\mathrm{DB}_N}\ . \end{aligned} \tag{4.25}$$

Note that the marginal distribution of the noninteracting bits $\mathrm{B}_{0:\infty/N}$ does not change over the time interval $t = N$ to $t = N+1$ since the matrix $M_{\mathrm{D}\otimes\mathrm{B}_N}$ acts only on D and B_N, and the Shannon entropy of the noninteracting bits remains unchanged over the interval.

Third, combining Eqs. (4.23), (4.24), and (4.25), we get the inequality:

$$\ln 2\Delta\,\mathrm{H}_{\mathrm{DB}_{0;\infty}} - \beta\Delta\langle E_{\mathrm{DB}_N}\rangle \geq 0\ , \tag{4.26}$$

where $\Delta\,\mathrm{H}_{\mathrm{DB}_{0;\infty}}$ is the change in the Shannon entropy of D and B and $\Delta\langle E_{\mathrm{DB}_N}\rangle$ is the change in the average energy of D and B over the interaction interval.

Fourth, according to the ratchet's design, D and B are decoupled from the work reservoir during the interaction intervals. (The work reservoir is connected only at the end points of intervals, when one bit is replaced by another.) From the First Law of Thermodynamics, the increase in energy $\Delta\langle E_{\mathrm{DB}_N}\rangle$ comes from the heat reservoir. In other words, we have the relation:

$$\Delta\langle E_{\mathrm{DB}_N}\rangle = \langle\Delta Q\rangle\ , \tag{4.27}$$

where ΔQ is the heat given to the system. (In fact, Eq. (4.27) is valid for each realization of the dynamics, not just on the average, since the conservation of energy holds in each realization.)

Finally, combining Eqs. (4.26) and (4.27), we get:

$$\ln 2\Delta \, \mathrm{H}_{\mathrm{DB}_{0:\infty}} - \beta \langle \Delta Q \rangle \geq 0 \ , \tag{4.28}$$

which is the basis of the IPSL as demonstrated in Ref. [13]; see, in particular, Eq. (A7) there.

Chapter 5

Transient Structural Costs of Physical Information Processing

5.1 Introduction.

Classical thermodynamics and statistical mechanics appeal to various reservoirs–reservoirs of heat, work, particles, and chemical species–each characterized by unique, idealized thermodynamic properties. A heat reservoir, for example, corresponds to a physical system with a large specific heat and short equilibration time. A work reservoir accepts or gives up energy without a change in entropy. Arising naturally in recent analyses of Maxwellian demons and information engines [30, 39, 42, 7, 60, 56, 61, 62, 63, 78, 73, 10, 64, 109, 9, 149, 6], *information reservoirs* have come to play a central role as idealized physical systems that exchange information but not energy [5, 150, 36]. Their inclusion led rather directly to an extended Second Law of Thermodynamics for complex systems: The total physical (Clausius) entropy of the universe and the Shannon entropy of its information reservoirs cannot decrease in time [151, 7, 5, 34, 4]. We refer to this generalization as the Information Processing Second Law (IPSL) [17].

A specific realization of an information reservoir is a tape of symbols where information is encoded in the symbols' values [1]. To understand the role that information processing plays in the efficiencies and bounds on thermodynamic transformations, we have thoroughly explored information ratchets in the previous chapters. By increasing

[1] See Appendices: Information reservoir implementations.

the tape's Shannon entropy, the ratchet can steadily transfer energy from the heat to the work reservoirs [7]. This violates the conventional formulation of the Second Law of Thermodynamics but is permitted by the IPSL.

These transducers are memoryful communication channels from input to output symbol sequences [1]. Information transducers are also similar to Turing machines in design [69], except that a Turing machine need not move unidirectionally. More importantly, an information transducer is a physical thermodynamic system and so is typically stochastic [2]. Despite this difference, like a Turing machine a transducer can perform any computation, if allowed any number of internal states.

Previous analyses on the thermodynamics of information processing largely focused on the minimal asymptotic entropy production *rate* for a given information transduction; see Eq. (5.2) below. The minimal rate is completely specified by the information transduction; there is no mention of any cost due to the transducer itself. In contrast, the Letter [19] first derives an exact expression for the minimal *transient* entropy production required for information transduction; see Eq. (5.3). This transient dissipation is the cost incurred by a system as it adapts to its environment. It is related to the excess heat in transitions between nonequilibrium steady states [152, 153, 154]. Moreover, hidden in this minimal transient dissipation, we identify the minimal cost associated with the transducer's construction; Eq. (5.4) below. Among all possible constructions that support a given computational task, there is a minimal, finite cost due to the physical *implementation*.

We go on to consider the specific case of structured pattern generation from a structureless information reservoir—a tape of independent and identically distributed (IID) symbols. While the transducer formalism for information ratchets naturally includes inputs with temporal structure, most theory so far has considered structureless inputs [20, 13, 7, 60, 61, 58]. This task requires designing a transducer that reads a tape of IID symbols as its input and outputs a target pattern. Employing the algebra of Shannon measures [79] and the structure-analysis tools of computational mechanics [155, 77],

[2]See Appendices: Turing machines versus transducers.

we show that the minimum implementation-dependent cost is determined by the mutual information between the transducer and the output's "past"—that portion of the output tape already generated. The result is that a maximally efficient implementation is achieved with a "retrodictive" model of the structured pattern transducer. Since the retrodictor's states depend only on the output future, it only contains as much information about the output's past as is required to generate the future. As a result it has a minimal cost proportional to the tape's *excess entropy* [77]. Such thermodynamic costs affect information processing in physical and biological systems that undergo finite-time transient processes when adapting to a complex environment.

5.2 Information Processing Second Law.

Consider a discrete-time Markov process involving the transducer's current state X_N and the current state of the information reservoir $\mathbf{Y}_N$ it processes. The latter is a semi-infinite chain of variables over the set $\mathcal{Y}$ that the transducer processes sequentially. As discussed in previous chapters Y_N is the Nth tape element, if the transducer has not yet processed that symbol; it is denoted Y'_N, if the transducer has. We call Y_N an input and Y'_N an output. The current tape $\mathbf{Y}_N = Y'_{0:N}Y_{N:\infty}$ concatenates the input tape $Y_{N:\infty} = Y_N Y_{N+1} Y_{N+2} \ldots$ and output tape $Y'_{0:N} = Y'_0 Y'_1 \ldots Y'_{N-2} Y'_{N-1}$. The information ratchet performs a computation by steadily transducing the input tape process $\Pr(Y_{0:\infty})$ into the output tape process $\Pr(Y'_{0:\infty})$.

The IPSL sets a bound on the average heat dissipation $Q_{0\to N}$ into the thermal reservoir over the time interval $t \in [0, N]$ in terms of the change in state uncertainty of the information ratchet and information reservoir [13, App. A]:

$$\frac{\langle Q_{0\to N}\rangle}{k_{\mathrm{B}}T \ln 2} \geq \mathrm{H}[X_0, \mathbf{Y}_0] - \mathrm{H}[X_N, \mathbf{Y}_N]. \tag{5.1}$$

Until now, our studies of such information engines developed the IPSL's asymptotic-rate form:

$$\lim_{N\to\infty} \frac{1}{N} \frac{\langle Q_{0\to N}\rangle}{k_{\mathrm{B}}T \ln 2} \geq -(h'_\mu - h_\mu) \ , \tag{5.2}$$

where h'_μ (h_μ) is the Shannon entropy rate of the output (input) tape [3] and, in addition,

[3] See Supplementary Materials: Information generation in physical systems.

we assume the transducer has a finite number of states [13].

The asymptotic IPSL in Eq. (5.2) says that thermal fluctuations from the environment can be rectified to either perform work or refrigerate (on average) at the cost of randomizing the information reservoir ($\langle Q \rangle < 0$ when $h'_\mu > h_\mu$). Conversely, an information reservoir can be refueled or 'charged' back to a clean slate by erasing its Shannon-entropic information content at the cost of emitting heat.

Equation (5.2), however, does not account for correlations between input and output tapes nor those that arise between the transducer and the input and output. As we now show, doing so leads directly to predictions about the relative effectiveness of transducers that perform the same information processing on a given input, but employ different physical implementations. Specifically, the retrodictive generator is the thermodynamically simplest implementation, *not* computational mechanics' optimal predictor—the ϵ-machine [155], which lies in contrast with Ref. [20]'s optimality claim for the ϵ-machine.

Subtracting the IPSL's asymptotic-rate version (Eq. (5.2)) from the IPSL's original (Eq. (5.1)) leads to a lower bound on the *transient* thermodynamic cost $\langle Q^{\text{tran}} \rangle$ of information transduction, this chapter's central result:

$$\begin{aligned} \frac{\langle Q^{\text{tran}} \rangle_{\text{min}}}{k_{\text{B}} T \ln 2} &\equiv \lim_{N \to \infty} \left[\frac{\langle Q_{0 \to N} \rangle_{\text{min}}}{k_{\text{B}} T \ln 2} + N(h'_\mu - h_\mu) \right] \\ &= -\mathbf{E}' + \mathrm{I}[\overleftarrow{Y}'; \overrightarrow{Y}] + \mathrm{I}[X_0; \overleftarrow{Y}', \overrightarrow{Y}] \ . \end{aligned} \tag{5.3}$$

Here, $\mathbf{E}'$ is the output sequence's *excess entropy* [12], which is the mutual information $I[\overleftarrow{Y}'; \overrightarrow{Y}']$ between the output past and output future. In these expressions we use arrows to describe past and future random variables. $\overleftarrow{Y}'$ is the random variable for output past—the sequence of output symbols that have been produced by the transducer—and $\overrightarrow{Y}'$ is the output future—the sequence of output symbols that have not yet been produced. Similarly, the input past random variable $\overleftarrow{Y}$ is the sequence of input variables that have already interacted with the transducer and have been transformed into an output. While the input future random variable $\overrightarrow{Y}$ is the sequence of input variables that have yet to interact. Finally, X_0 is the random variable for the transducer's state after sufficiently long time, such that $\overleftarrow{Y}'$ and $\overleftarrow{Y}$ are both effectively semi-infinite. The expression itself

comes from shifting to the ratchet's reference frame, so that at time N state X_N becomes X_0 and the currently interacting tape symbol is relabeled Y_0, rather than Y_N. (Equation (5.3) is proved in the Appendices.)

From it we conclude that the minimum transient cost has three components. However, they are subtly interdependent and so we cannot minimize them piece-by-piece to maximize thermodynamic efficiency. For instance, the first term in the transient cost is a benefit of having correlation between the output past and output future, quantified by $\mathbf{E}'$. Without further thought, one infers that outputs that are more predictable from their past, given a fixed entropy production rate, are easier to produce thermodynamically. However, as we see below when analyzing process generation, the other terms cancel this benefit, regardless of the output process. Perhaps counterintuitively, the most important factor is the output's intrinsic structure.

The remaining two terms in the transient cost are the cost due to correlations between the input and the output, quantified by $\mathrm{I}[\overleftarrow{Y}'; \overrightarrow{Y}]$, and the cost due to correlations between the transducer and the entire input-output sequence, quantified by $\mathrm{I}[X_0; \overleftarrow{Y}', \overrightarrow{Y}]$. The last term, which through X_0 depends explicitly on the transducer's structure, shows how different implementations of the same computation change energetic requirements. Said differently, we can alter transducer states as well as their interactions with tape symbols, all the while preserving the computation—the joint-input output distribution—and this *only* affects the last term in Eq. (5.3). For this reason, we call it the *minimal implementation energy cost* Q^{impl} given a transducer:

$$(k_{\mathrm{B}}T \ln 2)^{-1} \langle Q^{\mathrm{impl}} \rangle_{\mathrm{min}} = \mathrm{I}[X_0; \overleftarrow{Y}', \overrightarrow{Y}] \ . \tag{5.4}$$

This bound on the cost applies beyond predictively generating an output process [33] to any type of input-output transformation. The minimum implementation energy cost is not guaranteed to be achievable, but, like Landauer's bound on the cost of erasure [4, 34], it provides a guidepost for an essential physical cost of information processing. By choosing an implementation with the least information shared between the transducer's state and the joint state of the output past and input future, we minimize the unavoidable cost of computation. Moreover, we discuss how to achieve the minimum dissipation with

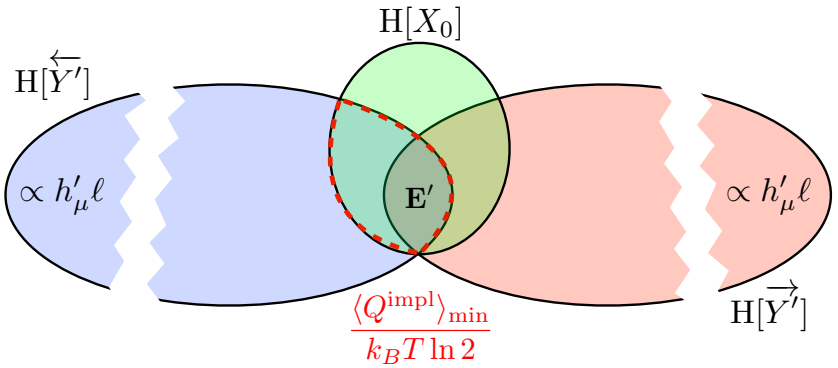


Figure 5.1. Shannon measures for physical information transduction—general case of nonunifilar transducers: Transducer output past $\overleftarrow{Y'}$ and output future $\overrightarrow{Y'}$ left (blue) and right (red) ellipses, respectively; shown broken since the future and past entropies $\mathrm{H}[\overleftarrow{Y'}]$ and $\mathrm{H}[\overrightarrow{Y'}]$ diverge as $h_\mu \ell$, with ℓ being the length of past or future, respectively. $\mathrm{H}[X_0]$ illustrates the most general relationship the generating transducer state X_0 must have with the process future and past. Implementation cost $\mathrm{I}[X_0; \overleftarrow{Y'}] = \langle Q^{\text{impl}} \rangle_{\text{min}} / k_B T \ln 2$ is highlighted by a dashed (red) outline.

the physical implementations of pattern generators developed in Ref. [20].

5.3 Generating Structured Patterns.

Paralleling Ref. [13], we now consider the thermodynamic cost of generating a sequential pattern of output symbols from a sequence of IID input symbols. Since the latter are uncorrelated and we restrict ourselves to nonanticipatory transducers (i.e., transducers with no direct access to future input [1]), the input future is statistically independent of both the current transducer state and the output past: $\mathrm{I}[X_0; \overleftarrow{Y'}, \overrightarrow{Y}] = \mathrm{I}[X_0; \overleftarrow{Y'}]$ and $\mathrm{I}[\overleftarrow{Y'}; \overrightarrow{Y}] = 0$. As a result, we have the following simplifications for the minimal transient dissipation and implementation costs (from Eq. (5.3)):

$$(k_{\mathrm{B}} T \ln 2)^{-1} \langle Q^{\text{tran}} \rangle_{\text{min}} = \mathrm{I}[X_0; \overleftarrow{Y'}] - \mathbf{E}' \tag{5.5}$$

$$(k_{\mathrm{B}} T \ln 2)^{-1} \langle Q^{\text{impl}} \rangle_{\text{min}} = \mathrm{I}[X_0; \overleftarrow{Y'}] \ . \tag{5.6}$$

The fact that the input is IID tells us that the transducer's states are also the internal states of the hidden Markov model (HMM) generator of the output process [1, 13]. This means that the transducer variable X_0 must contain all information shared between the output's past $\overleftarrow{Y'}$ and future $\overrightarrow{Y'}$ [156, 12], as shown in the *information diagram* in Fig. 5.1. (Graphically, the $\mathbf{E}'$ atom is entirely contained within $\mathrm{H}[X_0]$.) There, an ellipse depicts a

variable's Shannon entropy, an intersection of two ellipses denotes the mutual information between variables, and the exclusive portion of an ellipse denotes a variable's conditional entropy. For example, $\mathbf{E}' = \mathrm{I}[\overleftarrow{Y}'; \overrightarrow{Y}']$ is the intersection of $\mathrm{H}[\overleftarrow{Y}']$ and $\mathrm{H}[\overrightarrow{Y}']$. And, the leftmost crescent in Fig. 5.1 is the conditional Shannon entropy $\mathrm{H}[\overleftarrow{Y}'|X_0]$ of the output past $\overleftarrow{Y}'$ conditioned on transducer state X_0. The diagram also notes that this information atom, which is in principle infinite, scales as $h_\mu \ell$, where ℓ is the sequence length.

As stated above, Fig. 5.1 also shows that the ratchet state statistically shields past from future, since the ratchet-state entropy $\mathrm{H}[X_0]$ (green ellipse) contains the information $\mathbf{E}'$ shared between the output past and future (overlap between (left) blue and right (red) ellipses). Thus, the implementation cost $\mathrm{I}[X_0; \overleftarrow{Y}']$, highlighted by dashed (red) outline, necessarily contains the mutual information between the past and future. As discussed in the Appendices, both the transient and asymptotic bounds (Eqs. 5.2 and 5.5, respectively) are achievable through an alternating adiabatic and quasistatic protocol. This is distinct from the local isothermal protocol we discuss in the next chapter, though similar in principle. We are now ready to find the most efficient thermodynamic implementations for a given computation in the transient regime.

Consider first the class of predictive, *unifilar* information transducers; denote their states $\mathcal{R}_0^+$. Unifilarity here says that the current state $\mathcal{R}_0^+$ is restricted to be a function of the semi-infinite output past: the ratchet's next state $\mathcal{R}_0^+$ is unambiguously determined by $\overleftarrow{Y}'$. It should be noted that the alternating adiabatic and quasistatic protocol described in the Appendices is not local, and thus allows unifilar generators to not dissipate asymptotically, as discussed in the next chapter. Instead, there are information theoretic properties of the global distribution that leads to additional energetic costs.

A unifilar information transducer corresponds to the case where the transducer state entropy $\mathrm{H}[X_0 = \mathcal{R}_0^+]$ has no area outside that of the output past's entropy $\mathrm{H}[\overleftarrow{Y}']$. (See Fig. 5.2.) As evident there, the implementation cost $\mathrm{I}[X_0; \overleftarrow{Y}']$ is the same as the transducer's *state uncertainty*—the Shannon entropy $\mathrm{H}[X_0 = \mathcal{R}_0^+]$. Thus, according to Eq. (5.6) the thermodynamically most efficient unifilar transducer is that with minimal state-uncertainty $\mathrm{H}[X_0 = \mathcal{S}_0^+]$—the entropy of the *$\epsilon$-machine causal states* $\mathcal{S}_0^+$ of com-

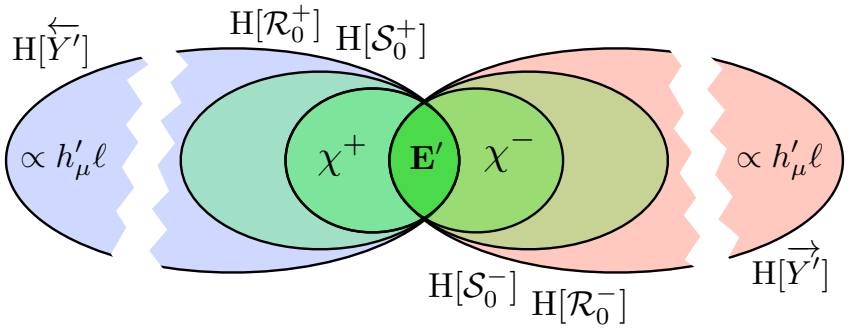


Figure 5.2. Optimal physical information transducers–predictive and retrodictive process generators: Process generator variables, predictive states $\mathcal{R}$ and causal states $\mathcal{S}$, denoted with green ellipses. Being the minimal set of predictive states, causal states $\mathcal{S}$ are contained within the set of general predictive states $\mathcal{R}$. A given process has alternative unifilar ($\mathcal{R}_0^+$ or $\mathcal{S}_0^+$) and co-unifilar generators ($\mathcal{R}_0^-$ or $\mathcal{S}_0^-$). Component areas are the sigma-algebra *atoms*: conditional entropies—entropy rate h_μ and cryptticities χ^+ and χ^-—and a mutual information—the excess entropy $\mathbf{E}'$. Since the state random variables $\mathcal{R}_0^+$ and $\mathcal{S}_0^+$ are functions of the output past $\overleftarrow{Y'}$, their entropies are wholly contained within the past entropy $\mathrm{H}[\overleftarrow{Y'}]$. Similarly, co-unifilar generators, denoted by the random variables $\mathcal{R}_0^-$ and $\mathcal{S}_0^-$, are functions of output future $\overrightarrow{Y'}$. Thus, their entropies are contained within the output future entropy $\mathrm{H}[\overrightarrow{Y'}]$. The ϵ-machine generator with causal states $\mathcal{S}_0^+$ is the unifilar generator with minimal Shannon entropy (area). The random variable $\mathcal{R}_0^-$ realizes the current state of the minimal co-unifilar generator, which is the time reversal of the ϵ-machine for the time-reversed process [2]. Transducers taking the form of any of these generators produce the same process, but structurally distinct generators exhibit different dissipations and thermodynamic implementation costs.

putational mechanics [77], which comprise the minimal set of predictive states [4]. This confirms the result that, if one is restricted to *predictive* generators, simpler is better [20].

Critically, there are further connections with computational mechanics that, by removing the restriction, lead to substantial generalizations. For ϵ-machine information transducers with causal states $\mathcal{S}_0^+$, the mutual information between the transducer and the output past is the output process' statistical complexity: $\mathrm{I}[\mathcal{S}_0^+; \overleftarrow{Y'}] = C'_\mu$. In other words, the minimal transient implementation cost of a pattern generated by a unifilar information transducer is the pattern's statistical complexity. The transient dissipation that occurs when generating a structured pattern, given in Eq. (5.5), is then the output's

[4] One might conclude that simpler (smaller forward causal-state entropy) is thermodynamically better (more efficient) [20]. Our development shows when this holds and when not, leading to a broader and more incisive view of optimal adaptive systems.

crypticity $\chi^+ = C'_\mu - \mathbf{E}'$ [156], as Ref. [2] concluded previously and Ref. [20] more recently.

Now, consider the more general case in which we allow the transducer implementation to be nonunifilar; see Fig. 5.1 again. From the Data Processing Inequality [11], it follows that the mutual information between X_0 and $\overleftarrow{Y}'$ cannot be less than the output's excess entropy:

$$\mathrm{I}[X_0; \overleftarrow{Y}'] \geq \mathbf{E}' \ . \tag{5.7}$$

Thus, the minimum structural cost over alternate pattern-generator implementations is the output pattern's excess entropy.

Figure 5.1 suggests how to find this minimum. The implementation cost highlighted by the dashed (red) line can be minimized by choosing a transducer whose states are strictly functions of the future. In this case, the transducer's mutual information with the output past is simply $\mathbf{E}'$, achieving the bound on implementation cost given by Eq. (5.7). (Refer now to Fig. 5.2.) Constructed using states that are functions of the future, such a ratchet is a generator with retrodictive (as opposed to predictive) states, denoted $\mathcal{R}_0^-$ or $\mathcal{S}_0^-$ [157]. This means that the generator is *counifilar*, as opposed to unifilar [158, 80]. These generators have the same states as the unifilar generators of the time-reversed process, but generally are nonunifilar. These *retrodictive generators* produce the same output process by running along the information reservoir in the same way as the predictive generators, but rather than store all of the information in the past outputs required to *predict* the future, they only store just enough to *generate* it. This affords them a fundamental energetic advantage.

Critically, too, any such retrodictive implementation is maximally efficient, dissipating zero transient heat $\langle Q^{\mathrm{tran}} \rangle_{\mathrm{min}} = 0$, even though the state uncertainty varies across implementations: $\mathrm{H}[\mathcal{R}_0^-] > \mathrm{H}[\mathcal{S}_0^-]$. Unlike unifilar transducers, for a given output process there are infinitely many counifilar information transducers of varying state-complexity that are all maximally thermodynamically efficient. In other words, simpler is not necessarily thermodynamically better for optimized transducers, as we demonstrate using the example of the $(3, 2)$ Golden Mean Process in the Supplementary Materials. Figure 5.3 there demonstrates the distinct thermodynamic advantages of retrodictive represen-

tations. This shows, as a practical matter, that both the design and evolution of efficient biological computations have a wide latitude when it comes to physical instantiations.

To summarize, we identified the transient and structural thermodynamic costs of physical information transduction, generalizing the recent Information Process Second Law. These bound the energetic costs incurred by any physically embedded adaptive system as it comes to synchronize with the states of a structured environment. Physical organization in the environment is a thermal resource for adaptive biological agents. To take advantage of that resource, however, the agent's internal state must reflect the hidden structure of its input [6, 17, 20]. Similarly, when producing an organized output, the agent must transition among the recurrent hidden states that are capable of generating the desired structure. Information transducing agents such as these involve a transient phase during which additional costs are incurred due to the agent adapting to its asymptotic behavior. When asking about which physical implementations have the least transient dissipation we showed that they can be bounded by the information shared by the agent, output past, and input future. This led us to see that the most efficient generators of organization are retrodictive, not necessarily ϵ-machines which are generative but predictive.

These results apply strictly to the transient behavior of thermodynamically reversible computations. In the next section, we quantify structural costs of computation that affect asymptotic behavior as well, and correspond to irreversible dissipation. The difference is that the transient costs come from globally integrated computations, which can be reversed, while the next chapter considers localized modular computations that are potentially irreversible.

5.4 Appendices

Derivations, further discussion and interpretation, and an explicit comparison of the thermodynamics of generators and predictors for an example system.

5.4.1 Shannon versus Kolmogorov-Sinai Entropy Rates

On the one hand, it is now commonplace shorthand in physics to describe a symbol-based process in terms of Shannon's information theory and so measure its intrinsic randomness

via the Shannon entropy rate. On the other, properly capturing the information processing behavior of physically embedded systems is more subtle. There are key conceptual problems. For one, the symbol-based Shannon entropy rate may not be well defined for a given continuous physical system. In the present setting we consider the entire engine as a physical system. Then, h_μ and h'_μ are the Kolmogorov-Sinai entropies of the associated reservoir dynamical systems [135, 136]. They are well defined suprema over all coarse-grainings of the system's state space.

5.4.2 Information Reservoirs Beyond Tapes

Rather than implement the information reservoir as a tape of symbols, one can simply employ a one-dimensional lattice of Ising spins. Moreover, the reservoir need *not* be 1D, but this is easiest to analyze since the total entropy of a 1D sequence is related to (but not equal to) the Shannon entropy rate and an engine simply accesses information by moving sequentially along the tape. Higher-dimension reservoirs, even with nonregular topologies connecting the information-bearing degrees-of-freedom, can be a thermodynamic resource when there is large total correlation among its degrees of freedom, at the cost of decorrelating the information-bearing degrees of freedom [159].

5.4.3 Achievability and Implementability

A protocol that changes the joint Hamiltonian of the ratchet and information reservoir quasistatically in time is thermodynamically reversible. The joint distribution evolves according to the time-dependent Boltzmann distribution [153]. It is possible to achieve the bounds of Eqs. (3) and (4) with the help of such protocols. While optimal protocols for general information transduction has not been formulated yet, a recent work describes a protocol for process generation that achieves the information-theoretic bounds, both the asymptotic and the transient [33]. One has to follow an alternating sequence of adiabatic (fast) and then quasistatic control of energy levels. It is noteworthy that the efficient quasistatic protocols described in [33] requires the ratchet to interact the entire sequence of bits simultaneously, to take advantage of all correlations between the bits and ratchet. If we limit ourselves to ratchets that interact with one bit at a time, as is the

standard implementation, the bounds may not be achievable even for process generation, as discussed in the next chapter.

5.4.4 Thermodynamics of General Computation

We focused on spatially-unidirectional information transduction due to the stronger thermodynamic results. However, the thermodynamic results are valid much more broadly, applicable to Turing-equivalent machines as well as non-1D information transduction, as just noted.

First, Turing machines that move unidirectionally, reading input tape cells once and writing results only once to an output tape, are equivalent to the transducers used here. However, unidirectional Turing machines employ internal tapes as scratch storage [160] and this now-internal memory must be taken into account when assessing thermodynamic resources.

Second, the choice of *implementation* of a particular computation implies a transient thermodynamic cost above the asymptotic implementation-independent work rate. The general result is:

$$\langle Q_{0\to N}^{\text{tran}} \rangle / k_{\text{B}} T \ln 2 \geq \text{H}[X_0, \mathbf{Y}_0] - \text{H}[X_N, \mathbf{Y}_N] + N\Delta h \ ,$$

where $\mathbf{Y}_N$ is the random variable for the information-bearing degrees of freedom at time N and Δh is the difference in the extensive component of the entropy density of the output and input tape processes. In short, *the transient cost due to an implementation stems from the correlation built up between the device's state and the pattern on which it acts, discounted by the intensive part of the output pattern's entropy.*

5.4.5 Origin of Transient Information Processing Costs

We demonstrate how the transient IPSL of Eq. (5.3) arises. The steps give additional insight.

Assuming that we are able to achieve asymptotic IPSL bounds—say, as in Ref. [20]—the cumulative transient cost of information processing over the interval $t \in [0, N]$ is given

by:

$$\langle Q^{\text{tran}}_{0\to N}\rangle \equiv \langle Q_{0\to N}\rangle - N(h_\mu - h'_\mu)k_{\text{B}}T\ln 2\ . \tag{5.8}$$

Combining with Eq. (5.1), yields:

$$\begin{aligned}
\frac{\langle Q^{\text{tran}}_{0\to N}\rangle}{k_{\text{B}}T\ln 2} &\geq \text{H}[X_0, \boldsymbol{Y}_0] - \text{H}[X_N, \boldsymbol{Y}_N] + N(h'_\mu - h_\mu) \\
&= \text{H}[X_0, Y_{0:\infty}] - \text{H}[X_N, Y'_{0:N}, Y_{N,\infty}] + N(h'_\mu - h_\mu) \\
&= (\text{H}[X_0] + \text{H}[Y_{0:\infty}] - \text{I}[X_0; Y_{0:\infty}]) \\
&\quad - (\text{H}[X_N] + \text{H}[Y'_{0:N}, Y_{N:\infty}] - \text{I}[X_N; Y'_{0:N}, Y_{N:\infty}]) + N(h'_\mu - h_\mu)\ .
\end{aligned}$$

The last line used the standard identity $\text{H}[A, B] = \text{H}[A] + \text{H}[B] - \text{I}[A; B]$ for random variables A and B. Since we are interested in the purely transient cost and not spurious costs arising from arbitrary initial conditions, we start the engine in its stationary state, resulting in stationary behavior, so that $\text{H}[X_0]$ is the same as $\text{H}[X_N]$. Furthermore, we assume that the engine's initial state is uncorrelated with the incoming symbols and so disregard $\text{I}[X_0; Y_{0:\infty}]$. We then decompose the terms $\text{H}[Y'_{0:N}, Y_{N:\infty}]$ and $\text{H}[Y_{0:\infty}] \equiv \text{H}[Y_{0:N}, Y_{N:\infty}]$ according to the above. These assumptions and decompositions lead to:

$$\begin{aligned}
\frac{\langle Q^{\text{tran}}_{0\to N}\rangle}{k_{\text{B}}T\ln 2} &\geq \text{H}[Y_{0:N}] + \text{H}[Y_{N:\infty}] - \text{I}[Y_{0:N}; Y_{N:\infty}] - \text{H}[Y'_{0:N}] - \text{H}[Y_{N:\infty}] + \text{I}[Y'_{0:N}; Y_{N:\infty}] \\
&\quad + \text{I}[X_N; Y'_{0:N}, Y_{N:\infty}] + N(h'_\mu - h_\mu)\ .
\end{aligned} \tag{5.9}$$

In the limit of large N, in which the transducer has interacted with a sufficiently large number of input symbols, we can invoke the following definitions of excess entropy:

$$\begin{aligned}
\mathbf{E} &= \lim_{N\to\infty} (\text{H}[Y_{0:N}] - Nh_\mu) \\
&= \lim_{N\to\infty} I[Y_{0:N}; Y_{N:\infty}] \\
\mathbf{E}' &= \lim_{N\to\infty} (H[Y'_{0:N}] - Nh'_\mu)\ .
\end{aligned}$$

Upon shifting to the ratchet's reference frame and switching back to the more intuitive notation: $X_N \to X_0$, $Y_{0:N} \to \overleftarrow{Y}$, $Y'_{0:N} \to \overleftarrow{Y}'$, and $Y_{N:\infty} \to \overrightarrow{Y}$, in which a left arrow means

the past and a right arrow the future, and invoking the above definitions, new notation, and these definitions, we rewrite the inequality Eq. (5.9), after some cancellation, as:

$$\frac{\langle Q^{\text{tran}} \rangle}{k_{\text{B}} T \ln 2} \geq -\mathbf{E}' + \mathrm{I}[\overleftarrow{Y}'; \overrightarrow{Y}] + \mathrm{I}[X_0; \overleftarrow{Y}', \overrightarrow{Y}] \ ,$$

where $\langle Q^{\text{tran}} \rangle$ is the total transient cost over infinite time: $\langle Q^{\text{tran}} \rangle = \lim_{N \to \infty} \langle Q^{\text{tran}}_{0 \to N} \rangle$. This is our main result, Eq. (5.3), and the starting point for the other results.

5.4.6 Efficiency of Different Generators of the Same Process

We now explore the consequences of the bound on implementation energy cost of generating a process with an example. The $(3, 2)$ Golden Mean Process consists of strings where 0s only appear in sequences of three bookended by 1s, and 1s only appear in sequences of two or more, bookended by 0s. There are infinitely many HMMs that generate this process, four of which are shown in the first row in Fig. S1. Here, the circles denote the hidden states of the HMMs and the arrows the transitions between the hidden states. Each transition is associated with an output. We use this Mealy representation of HMMs, because it is useful in computational mechanics for describing a wide variety of processes and calculating the their information theoretic properties [36]. The transitions in Mealy HMMs are described by a symbol-labeled transition matrix, with elements $T^{(y')}_{x \to x'}$ that describe the probability of transitioning to the hidden state x' and outputting y' given the current hidden state x:

$$T^{(y')}_{x \to x'} = \Pr(Y'_N = y', X_{N+1} = x' | X_N = x). \tag{5.10}$$

Pictorially, this is described by labeling the arrow from x to x' with $T^{(y')}_{x \to x'}$. For instance, in Fig. S1 we see that the ϵ-machine generator (first row, second column) has a transition from hidden state A to hidden state B with label $T^{(0)}_{A \to B} = 0.5$. This means, if the $\epsilon-$machine is in the hidden state A, it has the probability 0.5 to make a transition to hidden state B and produce an output 0.

It is easy to see how the HMMs in the first row lead to the (3,2) Golden Mean Process. Consider, for instance, the $\epsilon-$machine HMM. If we happen to be in state A, with probability 0.5 we make a transition to state B and output a 0. Afterwards, we are

forced to output two more 0's as we make the only possible transitions B $\rightarrow$ C $\rightarrow$ D. Then, we produce 1's until we arrive at A again. Clearly, 0's are produced in threes and 1's in pairs or more.

Each of the different HMMs corresponds to a different information ratchet operating on an IID input process. As described in [26], a ratchet is characterized by a Markov transition $M_{x\otimes y\rightarrow x'\otimes y'}$ over the joint state space of the ratchet and interaction bit

$$M_{x\otimes y\rightarrow x'\otimes y'} = \Pr(Y'_N = y', X_{N+1} = x' | X_N = x, Y_N = y). \tag{5.11}$$

To implement a particular HMM generator of a process from an IID input, we choose a ratchet which ignores the input y

$$M_{x\otimes y\rightarrow x'\otimes y'} = T^{(y')}_{x\rightarrow x'}. \tag{5.12}$$

The result is an output process generated by $T^{(y')}_{x\rightarrow x'}$ where the states of the ratchet $\mathcal{X}$ take the place of hidden states. For a device which can quasistatically implement the HMM process generator, as described in Ref. [20], taking advantage of all correlations to achieve the information-theoretic bounds, $\langle Q^{\mathrm{impl}}\rangle_{\mathrm{min}}$ is the implementation cost of generating the process. In the following, we calculate $\langle Q^{\mathrm{impl}}\rangle_{\mathrm{min}} = \mathrm{I}[X_0; \overleftarrow{Y}']$ for each of the four HMMs.

Figure 5.3 compares the implementation costs for both unifilar (predictive) and counifilar (retrodictive) generators, as discussed in the main text. The table shows that for the unifilar generators (shown in the first two columns), which have states that are a function of the output past, the implementation energy cost is proportional to the state uncertainty of the ratchet-generator states $H[X_0]$. That is, a simpler ratchet with fewer states is more thermodynamically efficient. Thus, the most efficient predictive generator is the ϵ-machine (second column), which dissipates $k_B T \ln 2$ less when synchronizing than the larger unifilar model presented in the first column. However, both the simple and more complex co-unifilar generators (shown in the last two columns) outperform the ϵ-machine in terms of efficiency. They achieve the theoretical bound on implementation cost, which is $k_B T \ln 2$ times the excess entropy of the output process $\mathbf{E}' = 1.5850$. Thus, even extremely complex generators can be maximally efficient as long as they are counifilar and their internal states are a function of future outputs.

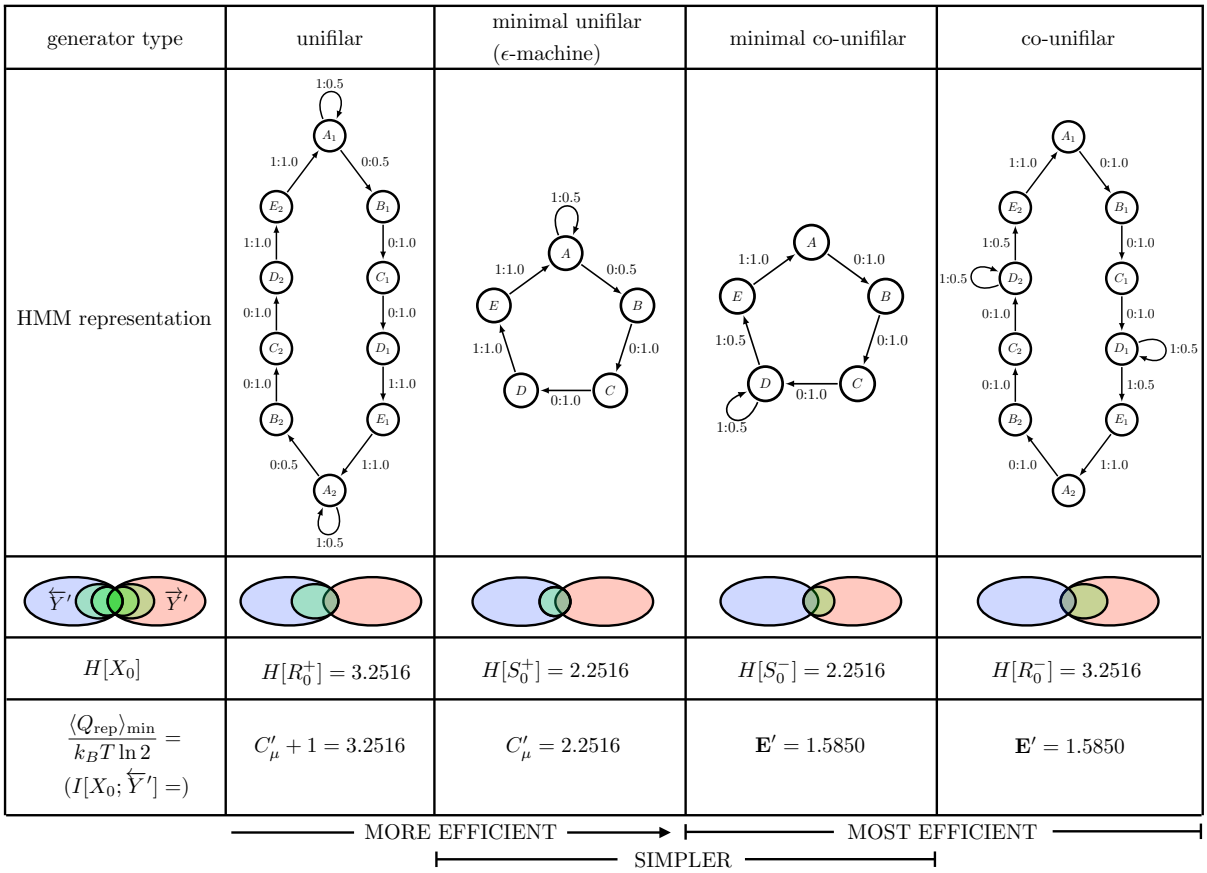


Figure 5.3. Four hidden Markov models that generate the same $(3, 2)$ Golden Mean Process. On the one hand, in the first two columns the representations are unifilar and, thus, have states that are functions of the past. For these two HMMs, the ratchet state uncertainty is proportional to the representation dissipation. Thus, the HMM in the first column, with an extra bit of state uncertainty above the ϵ-machine in the second column, also has additional representation dissipation. On the other hand, the two co-unifilar models in the last two columns minimize dissipation, but have different state uncertainties. They are more thermodynamically efficient than the first two HMMs.

Chapter 6

Thermodynamics of Modularity: Structural Costs Beyond the Landauer Bound

6.1 Introduction

Physically embedded information processing operates via thermodynamic transformations of the supporting material substrate. The thermodynamics is best exemplified by Landauer's principle: erasing one bit of stored information at temperature T must be accompanied by the dissipation of at least $k_{\mathrm{B}}T \ln 2$ amount of heat [4] into the substrate. While the Landauer cost is only time-asymptotic and not yet the most significant energy demand in everyday computations—in our cell phones, tablets, laptops, and cloud computing—there is a clear trend and desire to increase thermodynamic efficiency. Digital technology is expected, for example, to reach the vicinity of the Landauer cost in the near future; a trend accelerating with now-promising quantum computers. This seeming inevitability forces us to ask if the Landauer bound can be achieved for more complex information processing tasks than writing or erasing a single bit of information.

In today's massive computational tasks, in which vast arrays of bits are processed in sequence and in parallel, a task is often broken into modular components to add flexibility and comprehensibility to hardware and software design. This holds far beyond the arenas of today's digital computing. Rather than tailoring processors to do only the task spec-

ified, there is great benefit in modularly deploying elementary, but universal functional components—e.g., NAND, NOR, and perhaps Fredkin [161] logic gates, biological neurons [162], or similar units appropriate to other domains [163]—that can be linked together into circuits which perform any functional operation. This leads naturally to hierarchical design, perhaps across many organizational levels. In these ways, the principle of modularity reduces the challenges of designing, monitoring, and diagnosing efficient processing considerably [164, 165]. Designing each modular component of a complex computation to be efficient is vastly simpler than designing and optimizing the whole. Even biological evolution seems to have commandeered prior innovations, remapping and reconnecting modular functional units to form new organizations and new organisms of increasing survivability [166].

There is, however, a potential thermodynamic cost to modular information processing. For concreteness, recall the stochastic computing paradigm in which an input (a sequence of symbols) is sampled from a given probability distribution and the symbols are correlated to each other. In this setting, a modularly designed computation processes only the *local* component of the input, ignoring the latter's *global* structure. This inherent locality necessarily leads to irretrievable loss of the global correlations during computing. Since such correlations are a thermal resource [6, 167], their loss implies an energy cost—a thermodynamic *modularity dissipation*. Employing stochastic thermodynamics and information theory, we show how modularity dissipation arises by deriving an exact expression for dissipation in a generic localized information processing operation. We emphasize that this dissipation is above and beyond the Landauer bound for losses in the operation of single logical gates. It arises solely from the modular architecture of complex computations. One immediate consequence is that the additional dissipation requires investing additional work to drive computation forward.

In general, to minimize work invested in performing a computation, we must leverage the global correlations in a system's environment. Globally integrated computations can achieve the minimum dissipation by simultaneous control of the whole system, manipulating the joint system-environment Hamiltonian to follow the desired joint distribution. Not

only is this level of control difficult to implement physically, but designing the required protocol poses a considerable computational challenge in itself, with so many degrees of freedom and a potentially complex state space. Genetic algorithm methods have been proposed, though, for approximating the optimum [168]. Tellingly, they can find unusual solutions that break conventional symmetries and take advantage of the correlations between the many different components of the entire system [169, 170]. However, as we will show, it is possible to rationally design local information processors that, by accounting for these correlations, minimized modularity dissipation.

The following shows how to design optimal modular computational schemes such that useful global correlations are not lost, but stored in the structure of the computing mechanism. Since the global correlations are not lost in these optimal schemes, the net processing can be thermodynamically reversible (dissipationless). Utilizing the tools of information theory and computational mechanics—Shannon information measures and optimal hidden Markov generators—we identify the informational system structures that can mitigate and even nullify the potential thermodynamic cost of modular computation.

A brief tour of our main results will help orient the reader. It can even serve as a complete, but approximate description for the approach and technical details, should this be sufficient for the reader's interests.

Section 6.2 considers the thermodynamics of a composite information reservoir, in which only a subsystem is amenable to external control. In effect, this is our model of a localized thermodynamic operation. We assume that the information reservoir is coupled to an ideal heat bath, as a source of randomness and energy. Thus, external control of the information reservoir yields random Markovian dynamics over the informational states, heat flows into the heat bath, and work investment from the controller. Statistical correlations may exist between the controlled and uncontrolled subsystems, either due to initial or boundary conditions or due to an operation's history.

To highlight the information-theoretic origin of the dissipation and to minimize the energetic aspects, we assume that the informational states have equal internal (free) energies. Appealing to stochastic thermodynamics and information theory, we then show that

the minimum irretrievable *modularity dissipation* over the duration of an operation due to the locality of control is proportional to the reduction in mutual information between the controlled and uncontrolled subsystems; see Eq. (6.7). We deliberately refer to "operation" here instead of "computation" since the result holds whether the desired task is interpreted as computation or not. The result holds so long as free-energy uniformity is satisfied at all times, a condition natural in computation and other information processing settings.

Section 6.4 applies this analysis to *information engines*, an active subfield within the thermodynamics of computation in which information effectively acts as the fuel for driving physically embedded information processing [7, 13, 10, 17, 6]. The particular implementations of interest—*information ratchets*—process an input symbol string by interacting with each symbol in order, sequentially transforming it into an output symbol string, as shown in Fig. 2.1. This kind of *information transduction* [1, 13] is information processing in a very general sense: with properly designed dynamics over an infinite reservoir of internal states, the devices can implement a universal Turing machine [137]. Since information engines rely on localized information processing, reading in and manipulating one symbol at a time in their original design [7], the measure of irretrievable dissipation applies directly. The exact expression for the modularity dissipation is given in Eq. (6.15).

Sections 6.5 and 6.6 specialize information transducers further to the cases of pattern extractors and pattern generators. Section 6.5's *pattern extractors* use structure in their environment to produce work and *pattern generators* use stored work to create structure from an unstructured environment. The irreversible relaxation of correlations in information transduction can then be curbed by intelligently designing these computational processes. While there are not yet general principles for designing implementations for arbitrary computations, the measure of modularity dissipation that we develop in the following shows how to construct energy-efficient extractors and generators. For example, efficient extractors consume complex patterns and turn them into sequences of independent and identically distributed (IID) symbols.

We show that extractor transducers whose states are predictive of their inputs are

optimal, with zero minimal modularity dissipation. This makes immediate intuitive sense since, by design, such transducers can anticipate the next input and adapt accordingly. This observation also emphasizes the principle that thermodynamic agents should requisitely match the structural complexity of their environment to leverage those informational correlations as a thermodynamic fuel [17]. We illustrate this result in the case of the Golden Mean pattern in Fig. 6.3.

Conversely, Section 6.6 shows that when generating patterns from unstructured IID inputs, transducers whose states are retrodictive of their output are most efficient—i.e., have minimal modularity dissipation. This builds on the results of the last chapter, which showed that retrodictive generators have the least transient dissipation. This is also intuitively appealing in that pattern generation may be viewed as the time reversal of pattern extraction. Since predictive transducers are efficient for pattern extraction, retrodictive transducers are expected to be efficient pattern generators; see Fig. 6.5. This also allows one to appreciate that pattern generators previously thought to be asymptotically efficient are actually quite dissipative [20]. Taken altogether, these results provide guideposts for designing efficient, modular, and complex information processors—guideposts that go substantially beyond Landauer's principle for localized processing.

6.2 Global versus Localized Processing

If a physical system, denote it $\mathcal{Z}$, stores information as it behaves, it acts as an information reservoir. Then, a wide range of physically-embedded computational processes can be achieved by connecting $\mathcal{Z}$ to an ideal heat bath at temperature T and externally controlling the system's physical parameters, its Hamiltonian. Coupling with the heat bath allows for physical phase-space compression and expansion, which are necessary for useful computations and which account for the work investment and heat dissipation dictated by Landauer's bound. However, the bound is only achievable when the external control is precisely designed to harness the changes in phase-space. This may not be possible for modular computations. The modularity here implies that control is localized and potentially ignorant of global correlations in $\mathcal{Z}$. This leads to uncontrolled changes

in phase-space.

Most computational processes unfold via a sequence of local operations that update only a portion of the system's informational state. A single step in such a process can be conveniently described by breaking the whole informational system $\mathcal{Z}$ into two constituents: the informational states $\mathcal{Z}^{\mathrm{i}}$ that are controlled and evolving and the informational states $\mathcal{Z}^{\mathrm{s}}$ that are not part of the local operation on $\mathcal{Z}^{\mathrm{i}}$. We call $\mathcal{Z}^{\mathrm{i}}$ the *interacting* subsystem and $\mathcal{Z}^{\mathrm{s}}$ the *stationary* subsystem. As shown in Fig. 6.1, the dynamic over the joint state space $\mathcal{Z} = \mathcal{Z}^{\mathrm{i}} \otimes \mathcal{Z}^{\mathrm{s}}$ is the product of the identity over the stationary subsystem and a local Markov channel over the interacting subsystem. The informational states of the noninteracting stationary subsystem $\mathcal{Z}^{\mathrm{s}}$ are fixed over the immediate computational task, since this information should be preserved for use in later computational steps.

Such classical computations are described by a global Markov channel over the joint state space:

$$\begin{aligned} M^{\mathrm{global}}_{z^{\mathrm{i}}_t, z^{\mathrm{s}}_t \to z^{\mathrm{i}}_{t+\tau}, z^{\mathrm{s}}_{t+\tau}} \\ = \Pr(Z^{\mathrm{i}}_{t+\tau} = z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_{t+\tau} = z^{\mathrm{s}}_{t+\tau} | Z^{\mathrm{i}}_t = z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t = z^{\mathrm{s}}_t), \end{aligned} \tag{6.1}$$

where $Z_t = Z^{\mathrm{i}}_t \otimes Z^{\mathrm{s}}_t$ and $Z_{t+\tau} = Z^{\mathrm{i}}_{t+\tau} \otimes Z^{\mathrm{s}}_{t+\tau}$ are the random variables for the informational state of the joint system before and after the computation, with Z^{i} describing the $\mathcal{Z}^{\mathrm{i}}$ subspace and Z^{s} the $\mathcal{Z}^{\mathrm{s}}$ subspace, respectively. (Lowercase variables denote values their associated random variables realize.) The righthand side of Eq. (6.1) gives the conditional transition probability over the time interval $(t, t+\tau)$ from joint state $(z^{\mathrm{i}}_t, z^{\mathrm{s}}_t)$ to state $(z^{\mathrm{i}}_{t+\tau}, z^{\mathrm{s}}_{t+\tau})$. The fact that $\mathcal{Z}^{\mathrm{s}}$ is fixed means that the global dynamic can be expressed as the product of a local Markov computation on $\mathcal{Z}^{\mathrm{i}}$ with the identity over $\mathcal{Z}^{\mathrm{s}}$:

$$M^{\mathrm{global}}_{(z^{\mathrm{i}}_t, z^{\mathrm{s}}_t) \to (z^{\mathrm{i}}_{t+\tau}, z^{\mathrm{s}}_{t+\tau})} = M^{\mathrm{local}}_{z^{\mathrm{i}}_t \to z^{\mathrm{i}}_{t+\tau}} \delta_{z^{\mathrm{s}}_t, z^{\mathrm{s}}_{t+\tau}} \ , \tag{6.2}$$

where the local Markov computation is the conditional marginal distribution:

$$M^{\mathrm{local}}_{z^{\mathrm{i}}_t \to z^{\mathrm{i}}_{t+\tau}} = \Pr(Z^{\mathrm{i}}_{t+\tau} = z^{\mathrm{i}}_{t+\tau} | Z^{\mathrm{i}}_t = z^{\mathrm{i}}_t) \ . \tag{6.3}$$

When the processor is in contact with a heat bath at temperature T, the average entropy production $\langle \Sigma_{t \to t+\tau} \rangle$ of the universe over the time interval $(t, t+\tau)$ can be expressed

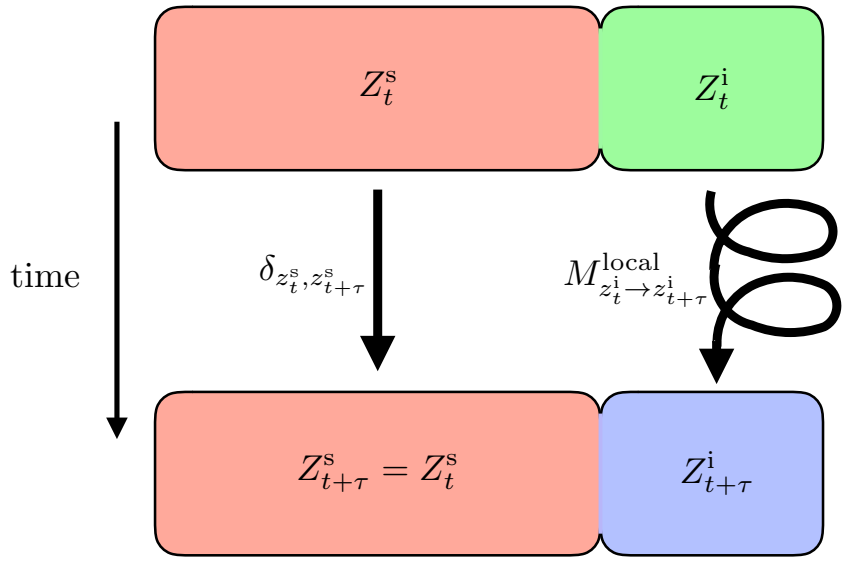


Figure 6.1. Local computations operate on only a subset $\mathcal{Z}^{\mathrm{i}}$ of the entire information reservoir $\mathcal{Z} = \mathcal{Z}^{\mathrm{i}} \otimes \mathcal{Z}^{\mathrm{s}}$. The Markov channel that describes the global dynamic is the product of a local operation with the identity operation: $M^{\mathrm{global}}_{(z_t^{\mathrm{i}},z_t^{\mathrm{s}})\to(z_{t+\tau}^{\mathrm{i}},z_{t+\tau}^{\mathrm{s}})} = M^{\mathrm{local}}_{z_t^{\mathrm{i}}\to z_{t+\tau}^{\mathrm{i}}}\delta_{z_t^{\mathrm{s}},z_{t+\tau}^{\mathrm{s}}}$, such that the stationary noninteracting portion $\mathcal{Z}^{\mathrm{s}}$ of the information reservoir remains invariant, but the interacting portion $\mathcal{Z}^{\mathrm{i}}$ changes.

in terms of the work done minus the change in nonequilibrium free energy F^{neq}:

$$\langle \Sigma_{t\to t+\tau} \rangle = \frac{\langle W_{t\to t+\tau} \rangle - (F^{\mathrm{neq}}_{t+\tau} - F^{\mathrm{neq}}_t)}{T} \ .$$

In this chapter, the sign of the work $\langle W_{t\to t+\tau} \rangle$ has flipped in the equation for entropy production, because it is the work *invested* by the controller, as opposed to the work *produced* by the controlled system as we describe in the previous chapters. The nonequilibrium free energy F^{neq}_t at any time t can be expressed as the weighted average of the internal (free) energy U_z of the joint informational states minus the uncertainty in those states:

$$F^{\mathrm{neq}}_t = \sum_z \Pr(Z_t = z) U_z - k_{\mathrm{B}} T \ln 2 \, \mathrm{H}[Z_t] \ . \tag{6.4}$$

Here, $\mathrm{H}[Z]$ is the Shannon information of the random variable Z that realizes the state of the joint system $\mathcal{Z}$ [9]. When the information bearing degrees of freedom support an information reservoir, where all states z and z' have the same internal energy $U_z = U_{z'}$. This is the situation we have considered in the following. In this situation, the first term on the right of Eq. (6.4) does not change even when there is a change in the probability

distribution $\Pr(Z_t = z)$. The entropy production $\langle \Sigma_{t \to t+\tau} \rangle$ then reduces to the work minus a change in Shannon information of the information-bearing degrees of freedom [8, 9]:

$$\langle \Sigma_{t \to t+\tau} \rangle = \frac{\langle W_{t \to t+\tau} \rangle}{T} + k_{\mathrm{B}} \ln 2 (\mathrm{H}[Z_{t+\tau}] - \mathrm{H}[Z_t]) \ . \tag{6.5}$$

Essentially, this is an expression of a generalized Landauer Principle: entropy increase guarantees that work production is bounded by the change in Shannon entropy of the informational variables [4].

App. 6.8.1 describes an isothermal method for implementing a Markov channel, in this case either M^{global} or M^{local}. By controlling the energy landscape, we exactly specify the form of the computation from input to output. Thus, if one is concerned with implementing deterministic logical operations, we can exponentially reduce any thermal randomness in the computation by making linear changes in energies. Our work, however, is closer in spirit to modern random computation [171, 172, 173], where the outcome of a computation is not a deterministic variable but a random one. In the natural (e.g., biological or molecular) setting, information processing in the presence of noise and stochasticity is the rule, not the exception. Rarely are noise-free discrete computation theory concepts applicable there. We developed thermodynamic principles that apply in this setting, too. In fact, a more general perspective upon the current work is to look at it as a study on the computation of thermodynamic systems, much in the spirit of computational mechanics (interpreted as the mechanics of computation [143]). Our work considers the generation, storage, dissipation, and transmission of information as a thermodynamic system performs its dynamics. This outlook provides a broader perspective on the thermodynamics of computation.

For the particular case of a globally integrated isothermal operation, the energy landscape over the whole system space $\mathcal{Z}$ is controlled simultaneously. In this case we can achieve zero entropy production. And, the globally integrated work done on the system achieves the theoretical minimum:

$$\langle W^{\mathrm{global}}_{t \to t+\tau} \rangle_{\mathrm{min}} = -k_{\mathrm{B}} T \ln 2 (\mathrm{H}[Z_{t+\tau}] - \mathrm{H}[Z_t]) \ .$$

The process is reversible since the change in system Shannon entropy balances the change

in the reservoir's physical entropy due to heat dissipation. Note that we do not assume the initial and final probability distributions of a thermodynamic operation to be equilibrium distributions. In fact, to have any meaningful computation we need to make transitions between *nonequilibrium* distributions. This is because equilibrium distributions are uniform distributions, since the internal energies of the information bearing degrees of freedom are uniform [5]. The transitions we consider are between nonequilibrium, metastable states with a decay time much longer than the experimental timescale. This time scale separation is necessary if information has to be stored reliably over long periods of time. We can achieve reversibility for transitions between nonequilibrium metastable states if the control time scale is much larger than the time scale of the internal dynamics of metastable states but much smaller than the time scale of global equilibration dynamics. This is the regime we have considered in our discussions. Since the internal energy is uniform, the system cannot store the work and must dissipate it as heat to the surrounding environment. This may not be the case for a generic modular operation.

There are two consequences of the locality of control. First, since Z^{s} is kept fixed—that is, $Z^{\mathrm{s}}_t = Z^{\mathrm{s}}_{t+\tau}$—the change in uncertainty $\mathrm{H}[Z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_{t+\tau}] - \mathrm{H}[Z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t]$ of the joint informational variables during the operation, which is the second term in lefthand side of Eq. (6.5), simplifies to:

$$\mathrm{H}[Z_{t+\tau}] - \mathrm{H}[Z_t] = \mathrm{H}[Z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_t] - \mathrm{H}[Z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t] \ .$$

Second, since we operate locally on Z^{i}, with hamiltonian control limited to that subsystem, and because the joint system Z is an information reservoir [5], which is energetically mute at the beginning and end of the computation, there is no energetic coupling to the non-interacting subsystem of Z^s during control. The lack of coupling to the stationary subsystem Z^{s} implies that the interacting subsystem on its own fits the framework for an open driven system described by Esposito and Van den Broeck, and thus the entropy production estimated from just the interacting system alone $\langle \Sigma^i \rangle = \langle W \rangle - \Delta F^i$ must be non-negative, where

$$F^i_t = \sum_{z \in \mathcal{Z}^i} \Pr(Z^i_t = z) U_z - k_B T \ln 2 \, H[Z^i_t]$$

is the marginalized estimate of the nonequilibrium free energy in *just* the interacting system [8]. As a result the work investment is bounded by the change this marginalized estimate of the nonequilibrium free energy, which implies the generalized Landauer Principle corresponding to the marginal distribution over Z^{i}; see Eq. (6.2). In other words, in absence of any control over the noninteracting subsystem Z^s, which remains stationary over the local computation on Z^{i}, the minimum work performed on Z^{i} is given by:

$$\begin{aligned}\langle W_{t\to t+\tau}\rangle &\geq \langle W^{\mathrm{local}}_{t\to t+\tau}\rangle_{\mathrm{min}} \\ &= k_{\mathrm{B}}T\ln 2(\mathrm{H}[Z^{\mathrm{i}}_t] - H[Z^{\mathrm{i}}_{t+\tau}) \ . \end{aligned} \tag{6.6}$$

This information theoretic bound on the work is achievable, as described in App. 6.8.1, by an isothermal process which is composed of slow manipulations of the energy landscape of the interacting subsystem, which takes the whole system between nonequilibrium metastable distributions.

Combining the last two relations with the expression for entropy production in Eq. (6.5) gives the *modularity dissipation* Σ^{mod}, which is the minimum irretrievable dissipation of a modular computation that comes from local interactions:

$$\begin{aligned}\frac{\langle \Sigma^{\mathrm{mod}}_{t\to t+\tau}\rangle_{\mathrm{min}}}{k_{\mathrm{B}}\ln 2} &= \frac{\langle W^{\mathrm{local}}_{t\to t+\tau}\rangle_{\mathrm{min}}}{k_BT\ln 2} + \mathrm{H}[Z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_t] - \mathrm{H}[Z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t] \\ &= \mathrm{I}[Z^{\mathrm{i}}_t; Z^{\mathrm{s}}_t] - I[Z^{\mathrm{i}}_{t+\tau}; Z^{\mathrm{s}}_t] \ , \end{aligned} \tag{6.7}$$

where $\mathrm{I}[X;Y]$ is the mutual information between the random variables X and Y.

This is the central result of this chapter: there is a thermodynamic cost above and beyond the Landauer bound for modular operations. It is a thermodynamic cost arising from a computation's *implementation architecture*. Specifically, the minimum entropy production is proportional to the minimum additional work that must be done to execute a computation modularly:

$$\langle W^{\mathrm{local}}_{t\to t+\tau}\rangle_{\mathrm{min}} - \langle W^{\mathrm{global}}_{t\to t+\tau}\rangle_{\mathrm{min}} = T\langle \Sigma^{\mathrm{mod}}_{t\to t+\tau}\rangle_{\mathrm{min}} \ .$$

App. A describes how to achieve this minimum dissipation through isothermal protocols, and any other protocol, perhaps done in finite time [15], or with unobserved coarse-grained

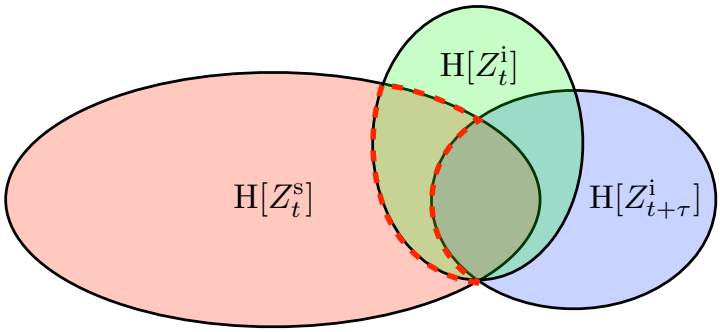


Figure 6.2. Information diagram for a local computation: Information atoms of the noninteracting subsystem $\mathrm{H}[Z_t^{\mathrm{s}}]$ (red ellipse), the interacting subsystem before the computation $\mathrm{H}[Z_t^{\mathrm{i}}]$ (green circle), and the interacting subsystem after the computation $\mathrm{H}[Z_{t+\tau}^{\mathrm{i}}]$ (blue circle). The initial state of the interacting subsystem shields the final state from the noninteracting subsystem; graphically the blue and red ellipses only overlap within the green ellipse. The modularity dissipation is proportional to the difference between information atoms $\mathrm{I}[Z_t^{\mathrm{i}}; Z_t^{\mathrm{s}}]$ and $\mathrm{I}[Z_{t+\tau}^{\mathrm{i}}; Z_t^{\mathrm{s}}]$. Due to statistical shielding, it simplifies to the information atom $\mathrm{I}[Z_t^{\mathrm{i}}; Z_t^{\mathrm{s}} | Z_{t+\tau}^{\mathrm{i}}]$, highlighted by a red dashed outline.

variables [174], would necessarily require more work to implement. The following draws out the implications.

Using the fact that the local operation M^{local} ignores Z^{s}, we see that the joint distribution over all three variables $Z_t^{\mathrm{i}}, Z_t^{\mathrm{s}}$, and $Z_{t+\tau}^{\mathrm{i}}$ can be simplified to:

$$\begin{aligned}\Pr(Z_{t+\tau}^{\mathrm{i}} = z_{t+\tau}^{\mathrm{i}}, Z_t^{\mathrm{i}} = z_t^{\mathrm{i}}, Z_t^{\mathrm{s}} = z_t^{\mathrm{s}}) \\ = \Pr(Z_{t+\tau}^{\mathrm{i}} = z_{t+\tau}^{\mathrm{i}} | Z_t^{\mathrm{i}} = z_t^{\mathrm{i}}) \Pr(Z_t^{\mathrm{i}} = z_t^{\mathrm{i}}, Z_t^{\mathrm{s}} = z_t^{\mathrm{s}}) \ .\end{aligned}$$

Thus, Z_t^{i} *shields* $Z_{t+\tau}^{\mathrm{i}}$ from Z_t^{s}. A consequence is that the mutual information between $Z_{t+\tau}^{\mathrm{i}}$ and Z_t^{s} conditioned on Z_t^{i} vanishes. This is shown in Fig. 6.2 via an information diagram. Figure 6.2 also shows that the modularity dissipation, highlighted by a dashed red outline, can be re-expressed as the mutual information between the noninteracting stationary system Z^{s} and the interacting system Z^{i} before the computation that is not shared with Z^{i} after the computation:

$$\langle \Sigma_{t \to t+\tau}^{\mathrm{mod}} \rangle_{\mathrm{min}} = k_{\mathrm{B}} \ln 2 \, \mathrm{I}[Z_t^{\mathrm{i}}; Z_t^{\mathrm{s}} | Z_{t+\tau}^{\mathrm{i}}] \ . \tag{6.8}$$

The conditional mutual information on the right bounds how much entropy is produced when performing a local computation. It quantifies the irreversibility of modular information processing.

6.3 Thermodynamics of Correlations Revisited with Modularity Costs

The new concept of thermodynamics of modularity lets us revisit some existing results in new light. The cost in Eq. (6.8) has been recognized in the context in the context of copying/measurement [175], and has relevance to biological push-pull systems [176]. It was shown that these biological processes can reversibly implement many computations, but are unable to decorrelate readouts with their receptors thermodynamically efficiently. Treating the readout as the interacting subsystem and the receptor as the stationary subsystem, this inefficiency can be predicted from the modularity dissipation. If the non-interacting stationary subsystem is uniformly distributed, such that $H[Z_t^s] = \log_2 |\mathcal{Z}^s|$, and the interacting subsystem is a copy of that system, then all structure in the information reservoir comes in the form of correlations between the subsystems, such that $I[Z_t^i; Z_t^s] = H[Z_t^s]$. If we perform a decorrelation operation, mapping the interacting system to a uniform distribution, and decorrelating the two subsystems such that $I[Z_{t+\tau}^i; Z_t^s] = 0$, we could potentially recover $k_B T \ln 2 H[Z_t^s]$ of work from the system with globally integrated control and energetic coupling between subsystems. But, if the control is local, all these correlations are dissipated in the decorrelation operation, as reflected by the modularity dissipation

$$\begin{aligned} \langle \Sigma^{\text{mod}}_{t \to t+\tau} \rangle_{\text{min}} &= k_B T \ln 2 (I[Z_t^i; Z_t^s] - I[Z_{t+\tau}^i; Z_t^s]) \\ &= k_B T \ln 2 H[Z_t^s], \end{aligned}$$

because energetic coupling is impossible in modular computations.

The analytic form of the modularity dissipation shown in Eq. (6.7), a difference of mutual informations, has been seen before in different contexts [177, 178]. These works show that the un-utilized change in free energy corresponds to dissipated work. In the context of data representations, Eq. (6.7)'s bound is analogous to the expression for the minimum work required for data representation, with Z_t^{i} being the work medium, $Z_{t+\tau}^{\text{i}}$ the work extraction device, and Z_t^{s} the data representation device [177]. And Still et al.'s investigation of the thermodynamics of prediction in a system driven by an input signal

shows that the irretrievable work dissipation is the same as the modularity dissipation if the driving signal is treated as the interacting subsystem, and the driven system is treated as the stationary subsystem [178]. While similar mathematically, the setup is subtly different from the cost of local modular control of information processing, and the frameworks lead to different results. This becomes especially clear in the context of signal transduction.

The next section unpacks the implications of Eqs. (6.7) and (6.8) for information transducers–information processing architectures in which the processor sequentially takes one input symbol at a time and performs localized computation on it, much as a Turing machine operates. These devices respond to an input signal much as the driven systems discussed by Still et al. do. The irretrievable dissipation that they derive for their driven systems can be minimized by ensuring that the driven system not store any unwarranted information about the input, beyond what's require to predict [178]. However, this means that thermodynamic efficiency can be achieved by having the driven system store zero information about the input. For a structured input to an information transducer, on the other hand, we see distinctly different behavior. Such memoryless systems are thermodynamically inefficient, as we discuss in the section 6.5, which discusses the necessity of prediction for efficient extractors. Moreover, the transducer framework allows us to extend beyond prediction. We see results reminiscent of Still et al. with pattern generators in section 6.6, in that they are thermodynamically efficient when they store as little unwarranted information as possible. However, these devices are retrodicting rather than predicting. The language and mathematics used in the thermodynamics of modularity is very similar to driven systems, but the modularity dissipation more directly speaks to the design of efficient controllers.

Various other models of information-driven processes have been considered in the literature. For example, Toyabe *et al.* [40] have considered a situation where a system is subject to an external force depending on the instantaneous position of the system. In this case, there is a modified expression of the second law of thermodynamics that is obeyed by the system [39]. Horowitz, Sagawa, and Parrondo have considered a situation where

a chemical force replaces the role of information in driving the system out of equilibrium and extracting work [57]. McGrath *et al.* have considered a situation where the mutual information between two spatial degrees of freedom in a biochemical context acts as a thermodynamic resource [179]. It will be an interesting future exercise to explore the implications of the current results for these previous works as well.

While we have identified the inherent dissipation due to modular computations, suggesting that globally integrated control would lead to more thermodynamically efficient computations, we see in the context of transducers that there are other paths to thermodynamic efficiency. The modularity dissipation can be minimized by designing the computation such that the modular components store the relevant global correlations, preventing dissipation. Locality and modularity are natural parts of complex computations, so rather than rely on the ability to simultaneously control the global energy landscape, we use the modularity dissipation as a structural guide to design modular computational architectures to be thermodynamically efficient.

Modular approach is not unique to in silico computation. Modularity is displayed in the structure and functionality of biological organisms as well [180]. Our work can be viewed as providing the information theoretic and thermodynamic backdrop to understand modular operations of biological functionality such as memory [17], self-correction [6], and pattern formation [19], among others.

6.4 Information Transducers: Localized processors

Information ratchets [63, 13] are thermodynamic implementations of information transducers [1] that sequentially transform an input symbol string, described by the chain of random variables $Y_{0:\infty} = Y_0 Y_1 Y_2, \ldots$, into an output symbol string, described by the chain of random variables $Y'_{0:\infty} = Y'_0 Y'_1 Y'_2, \ldots$. The ratchet traverses the input symbol string unidirectionally, processing each symbol in turn to yield the output sequence. As shown in Fig. 2.1, at time $t = N\tau$ the information reservoir is described by the joint distribution over the ratchet state X_N and the symbol string $\mathbf{Y}_N = Y'_{0:N} Y_{N:\infty}$, the concatenation of the first N symbols of the output string and the remaining symbols of the input string. (This

differs slightly from treatments in the previous chapters [6] in which only the symbol string is the information reservoir. The information processing and energetics are the same, however.) Including the ratchet state in present definition of the information reservoir allows us to directly determine the modularity dissipation of information transduction.

Going from time $t = N\tau$ to $t + \tau = (N + 1)\tau$ preserves the state of the current output history $Y'_{0:N}$ and the input future, excluding the Nth symbol $Y_{N+1:\infty}$, while changing the Nth input symbol Y_N to the Nth output symbol Y'_N and the ratchet from its current state X_N to its next X_{N+1}. In terms of the previous section, this means the noninteracting stationary subsystem $\mathcal{Z}^{\mathrm{s}}$ is the entire semi-infinite symbol string *without* the Nth symbol:

$$Z_t^{\mathrm{s}} = (Y_{N+1:\infty}, Y'_{0:N}) \ . \tag{6.9}$$

The ratchet and the Nth symbol constitute the interacting subsystem $\mathcal{Z}^{\mathrm{i}}$ so that, over the time interval $(t, t + \tau)$, only two variables change:

$$Z_t^{\mathrm{i}} = (X_N, Y_N) \tag{6.10}$$

and

$$Z_{t+\tau}^{\mathrm{i}} = (X_{N+1}, Y'_N) \ . \tag{6.11}$$

Despite the fact that only a small portion of the system changes on each time step, the physical device is able to perform a wide variety of physical and logical operations. Ignoring the probabilistic processing aspects, Turing showed that a properly designed (very) finite-state transducer can compute any input-output mapping [181] [1]. Such machines, even those with as few as two internal states and a sufficiently large symbol alphabet [182] or with as few as a dozen states but operating on a binary-symbol strings, are *universal* in that sense [145].

Information ratchets—physically embedded, probabilistic Turing machines—are able to facilitate energy transfer between a thermal reservoir at temperature T and a work

[1] Space limitations here do not allow a full digression on possible implementations. Suffice it to say that for unidirectional tape reading, the ratchet state requires a storage register or an auxiliary internal working tape as portrayed in Fig. 3 of Ref. [143].

reservoir by processing information in symbol strings. In particular, they can function as an eraser by using work to create structure in the output string [7, 13] or act as an engine by using the structure in the input to turn thermal energy into useful work energy [13]. They are also capable of much more, including detecting, adapting to, and synchronizing to environment correlations [19, 17] and correcting errors [6].

Information transducers are a novel form of information processor from a different perspective, that of communication theory's channels [1]. They are memoryful channels that map input stochastic processes to output processes using internal states which allow them to store information about the past of both the input and the output. With sufficient hidden states, as just noted from the view of computation theory, information transducers are Turing complete and so able to perform any computation on the information reservoir [69]. Similarly, the physical steps that implement a transducer as an information ratchet involve a series of modular local computations.

The ratchet operates by interacting with one symbol at a time in sequence, as shown in Fig. 2.1. The Nth symbol, highlighted in yellow to indicate that it is the interacting symbol, is changed from the input Y_N to output Y'_N over time interval $(N\tau, (N+1)\tau)$. The ratchet and interaction symbol change together according to the local Markov channel over the ratchet-symbol state space:

$$M^{\text{local}}_{(x,y)\to(x',y')} = \Pr(X_{N+1}=x', Y'_N=y' | X_N=x, Y_N=y).$$

This determines how the ratchet transduces inputs to outputs [13].

Each of these localized operations keeps the remaining noninteracting symbols in the information reservoir fixed. If the ratchet only has energetic control of the degrees of freedom it manipulates, then, as discussed in the previous section and App. 6.8.1, the ratchet's work production in the Nth time step is bounded by the change in uncertainty of the ratchet state and interaction symbol:

$$\langle W^{\text{local}}_N \rangle_{\text{min}} = k_{\text{B}} T \ln 2 (\text{H}[X_N, Y_N] - \text{H}[X_{N+1}, Y'_N]). \tag{6.12}$$

This bound has been recognized in previous investigations of information ratchets [7, 183]. Here, we make a key, but important and compatible observation: If we relax the condition

of local control of energies to allow for global control of all symbols simultaneously, then it is possible to extract more work.

That is, foregoing localized operations—abandoning modularity—allows for (and acknowledges the possibility of) globally integrated interactions. Then, we can account for the change in Shannon information of the information reservoir—the ratchet and the entire symbol string. This yields a looser upper bound on work production that holds for both modular and globally integrated information processing. Assuming that all information reservoir configurations have the same free energies, the change in the nonequilibrium free energy during one step of a ratchet's computation is proportional to the global change in Shannon entropy:

$$\Delta F^{\text{neq}}_{N\tau\to(N+t)\tau} = k_{\text{B}}T \ln 2(\text{H}[X_N, \mathbf{Y}_N] - \text{H}[X_{N+1}, \mathbf{Y}_{N+1}]).$$

Recalling the definition of entropy production $\langle \Sigma \rangle = \langle W \rangle - \Delta F^{\text{neq}}$ reminds us that for entropy to increase, the minimum work investment must match the change in free energy:

$$\begin{aligned} &\langle W_N^{\text{global}} \rangle_{\text{min}} \\ &\quad = k_{\text{B}}T \ln 2(\text{H}[X_N, \mathbf{Y}_N] - \text{H}[X_{N+1}, \mathbf{Y}_{N+1}]) \ . \end{aligned} \tag{6.13}$$

This is the minimum work investment that can be achieved through globally integrated isothermal information processing. And, in turn, it can be used to bound the asymptotic work production in terms of the entropy rates of the input and output processes [13]:

$$\lim_{N\to\infty} \langle W_N \rangle \geq k_{\text{B}}T \ln 2(h_\mu - h'_\mu) \ . \tag{6.14}$$

This is known as the *Information Processing Second Law* (IPSL).

Reference [6] already showed that this bound is not necessarily achievable by information ratchets. This is due to ratchets operating locally. The local bound on work production of modular implementations in Eq. (6.12) is less than or equal to the global bound on integrated implementations in Eq. (6.13), since the local bound ignores correlations between the interacting system $\mathcal{Z}^{\text{i}}$ and noninteracting elements of the symbol string in $\mathcal{Z}^{\text{s}}$. Critically, though, if we design the ratchet such that its states store the relevant correlations in the symbol string, then we can achieve the global bounds. This was hinted

at in the fact that the gap between the work done by a ratchet and the global bound can be closed by designing a ratchet that matches the input process' structure [6]. However, comparing the two bounds now allows us to be more precise.

The difference between the two bounds represents the amount of additional work that could have been performed by a ratchet, if it was not modular and limited to local interactions. If the computational device is globally integrated, with full access to all correlations between the information bearing degrees of freedom, then all of the nonequilibrium free energy can be converted to work, zeroing out the entropy production. Thus, the minimum entropy production for a modular transducer (or information ratchet) at the Nth time step can be expressed in terms of the difference between Eq. (6.12) and the entropic bounds in Eq. (6.13):

$$\begin{aligned} \frac{\langle \Sigma_N^{\text{mod}} \rangle_{\text{min}}}{k_{\text{B}} \ln 2} &= \frac{\langle W_N^{\text{local}} \rangle_{\text{min}} - \Delta F_{N\tau \to (N+1)\tau}^{\text{neq}}}{k_B T \ln 2} \\ &= \text{I}[Y_{N+1:\infty}, Y'_{0:N}; X_N, Y_N] \\ &\quad - \text{I}[Y_{N+1:\infty}, Y'_{0:N}; X_{N+1}, Y'_N] \\ &= \text{I}[Y_{N+1:\infty}, Y'_{0:N}; X_N, Y_N | X_{N+1}, Y'_N] \ . \end{aligned} \tag{6.15}$$

This can also be derived directly by substituting our interacting variables $(X_N, Y_N) = Z_t^{\text{i}}$ and $(X_{N+1}, Y'_N) = Z_{t+\tau}^{\text{i}}$ and stationary variables $(Y_{N+1:\infty}, Y'_{0:N}) = Z^{\text{s}}$ into the expression for the modularity dissipation in Eqs. (6.7) and (6.8) in Sec. 6.2. Even if the energy levels are controlled so slowly that entropic bounds are reached, Eq. (6.16) quantifies the amount of lost correlations that cannot be recovered. And, this leads to the entropy production and irreversibility of the transducing ratchet. This has immediate consequences that limit the most thermodynamically efficient information processors.

While previous bounds, such as the IPSL, demonstrated that information in the symbol string can be used as a thermal fuel [7, 13]—leveraging structure in the inputs symbols to turn thermal energy into useful work—they largely ignore the structure of information ratchet states X_N. The transducer's hidden states, which can naturally store information about the past, are critical to taking advantage of structured inputs. Until now, we only used informational bounds to predict transient costs to information processing [20, 19].

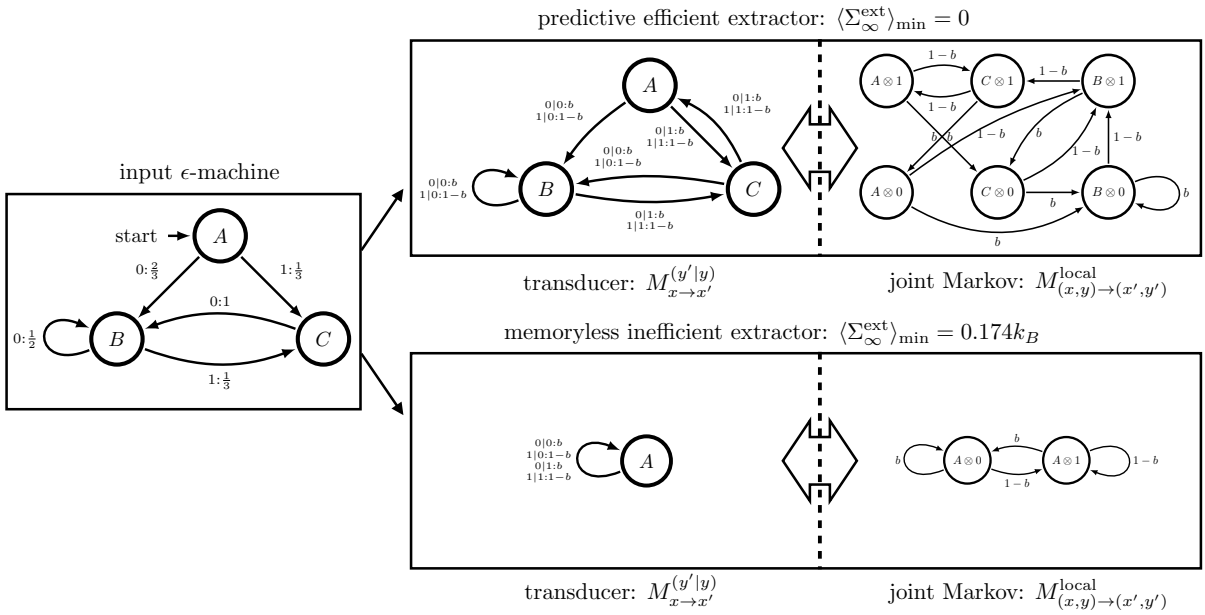


Figure 6.3. Multiple physical methods for transforming the Golden Mean Process input, whose ϵ-machine generator is shown in the far left box, into a sequence of uncorrelated symbols. The ϵ-machine is a Mealy hidden Markov model that produces outputs along the edges, with $y : p$ denoting that the edge emits symbol y and is taken with probability p. (Top row) Ratchet whose internal states match the ϵ-machine states and so it is able to minimize dissipation—$\langle \Sigma_\infty^{\text{ext}} \rangle_{\text{min}} = 0$—by making transitions such that the ratchet's states are synchronized to the ϵ-machine's states. The transducer representation to the left shows how the states remain synchronized: its edges are labeled $y'|y : p$, which means that if the input was y, then with probability p the edge is taken and it outputs y'. The joint Markov representation on the right depicts the corresponding physical dynamic over the joint state space of the ratchet and the interaction symbol. The label p along an edge from the state $x \otimes y$ to $x' \otimes y'$ specifies the probability of transitioning between those states according to the local Markov channel $M_{(x,y)\to(x',y')}^{\text{local}} = p$. (Bottom row) In contrast to the efficient predictive ratchet, the memoryless ratchet shown is inefficient, since it's memory cannot store the predictive information within the input ϵ-machine, much less synchronize to it.

With the expression for the modularity dissipation of information ratchets in Eq. (6.16), however, we now have bounds that apply to the ratchet's asymptotic functioning. In short, this provides the key tool for designing thermodynamically efficient transducers. We will now show that it has immediate implications for pattern generation and pattern extraction.

6.5 Predictive Extractors

A *pattern extractor* is a transducer that takes in a structured process $\Pr(Y_{0:\infty})$, with correlations among the symbols, and maps it to a series of independent identically distributed (IID), uncorrelated output symbols. An output symbol can be distributed however we wish individually, but each must be distributed with an identical distribution and independently from all others. The result is that the joint distribution of the output process symbols is the product of the individual marginals:

$$\Pr(Y'_{0:\infty}) = \prod_{i=0}^{\infty} \Pr(Y'_i) \ . \tag{6.16}$$

If implemented efficiently, this device can use temporal correlations in the input as a thermal resource to produce work. The modularity dissipation of an extractor $\langle \Sigma_N^{\text{ext}} \rangle_{\text{min}}$ can be simplified by noting that the output symbols are uncorrelated with any other variable and, thus, the Y' terms fall out of the mutual information expression for dissipation in Eq. 6.15, yielding:

$$\frac{\langle \Sigma_N^{\text{ext}} \rangle_{\text{min}}}{k_B \ln 2} = \mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] - \mathrm{I}[Y_{N+1:\infty}; X_{N+1}] \ .$$

Minimizing this irreversibility, as shown in App. 6.8.2, leads directly to a fascinating conclusion that relates thermodynamics to prediction: the states of maximally thermodynamically efficient extractors are perfect predictors of the input process. Other work has used predictive methods for extracting additional work from temporal correlations [73, 6, 184], but the modularity dissipation provides the first proof of the need for predictive states. Moreover, it can be applied to any extractor, to determine the dissipation of an imperfect predictor.

To take full advantage of the temporal structure of an input process, the ratchet's states X_N must be able to predict the future of the input $Y_{N:\infty}$ from the input past $Y_{0:N}$. Thus, the ratchet shields the input past from the output future such that there is no information shared between the past and future which is not captured by the ratchet's states:

$$\mathrm{I}[Y_{N:\infty}; Y_{0:N}|X_N] = 0 \ . \tag{6.17}$$

Additionally, transducers cannot anticipate the future of the inputs beyond their correlations with past inputs [1]. This means that there is no information shared between the ratchet and the input future when conditioned on the input past:

$$\mathrm{I}[Y_{N:\infty}; X_N|Y_{0:N}] = 0 \ . \tag{6.18}$$

As shown in Appendix 6.8.2, Eqs. (6.17) and (6.18), which together are equivalent to the state X_N being predictive, can be used to prove that the modularity dissipation vanishes $\langle \Sigma_N^{\mathrm{ext}} \rangle_{\mathrm{min}} = 0$. Moreover, setting the modularity dissipation to zero guarantees that the state shields the past and future of the input from each other, as shown in Eq. (6.17). Thus, because Eq. (6.18) is a given for transducers, this proves that the ratchet being predictive is equivalent to zero modularity dissipation, and thus perfect thermodynamic efficiency. The efficiency of predictive ratchets suggests that predictive generators, such as the ϵ-machine [12], are useful in designing efficient information engines that can leverage temporal structure in an environment.

For example, consider an input string that is structured according to the Golden Mean Process, which consists of binary strings in which 1's always occur in isolation, bookended by 0's. Figure 6.3 gives two examples of ratchets, described by different local Markov channels $M^{\mathrm{local}}_{(x,y)\to(x',y')}$, that each map the Golden Mean Process to a biased coin. The input process' ϵ-machine, shown in left box, provides a template for how to design a thermodynamically efficient local Markov channel, since its states are predictive of the process. The Markov channel is a transducer [13]:

$$M^{(y'|y)}_{x\to x'} \equiv M^{\mathrm{local}}_{(x,y)\to(x',y')} \ . \tag{6.19}$$

By designing transducer states to stay synchronized to the states of the process' ϵ-machine, we can minimize the modularity dissipation to zero. For example, the efficient transducer shown in Fig. 6.3 has almost the same topology as the Golden Mean ϵ-machine, with an added transition between states C and A corresponding to a disallowed word in the input. This transducer is able to harness all structure in the input because it synchronizes to the input process and so is able to optimally predict the next input.

The efficient ratchet shown in Fig. 6.3 (top row) comes from a general method for constructing an optimal extractor given the input's ϵ-machine. The ϵ-machine is represented by a Mealy hidden Markov model (HMM) with the symbol-labeled state-transition matrices:

$$T^{(y)}_{s \to s'} = \Pr(Y_N = y, S_{N+1} = s' | S_N = s) \ , \tag{6.20}$$

where S_N is the random variable for the hidden state reading the Nth input Y_N. If we design the ratchet to have the same state space as the input process' hidden state space—$\mathcal{X} = \mathcal{S}$—and if we want the IID output to have bias $\Pr(Y_N = 0) = b$, then we set the local Markov channel over the ratchet and interaction symbol to be:

$$M^{\text{local}}_{(x,y) \to (x',y')} = \begin{cases} b, & \text{if } T^{(y)}_{x \to x'} \neq 0 \text{ and } y' = 0 \\ 1 - b, & \text{if } T^{(y)}_{x \to x'} \neq 0 \text{ and } y' = 1 \ . \end{cases}$$

These constraints on the channel, combined with normalized transition probabilities, does not uniquely specify M^{local}, since there can be forbidden words in the input that, in turn, lead to ϵ-machine causal states which always emit a single symbol. This means that there are joint ratchet-symbol states (x, y) such that $M_{(x,y) \to (x',y')}$ is unconstrained. For these states, we may make any choice of transition probabilities from (x, y), since this state will never be reached by the combined dynamics of the input and ratchet. The end result is that, with this design strategy, we construct a ratchet whose memory stores all information in the input past that is relevant to the future, since the ratchet remains synchronized to the input's causal states. In this way, it leverages all temporal order in the input. This is characteristic of any efficient extractor, and confirms the thermodynamic principle of requisite variety [17]. The fact that the ratchet states must synchronize to

the ϵ-machine's causal states, implies that the uncertainty in the ratchet's memory must at least match the uncertainty in the causal states of the input, which is its *statistical complexity*

$$H[X_N] \geq H[S_N] \tag{6.21}$$

$$= C_\mu. \tag{6.22}$$

Thus, we have not only proved the thermodynamic principle of requisite variety in general, but also refined it to a principle of requisite *complexity*.

By way of contrast, consider a memoryless transducer, such as that shown in Fig. 6.3 (bottom row). It has only a single state and so cannot store any information about the input past. As discussed in previous explorations, ratchets without memory are insensitive to correlations [17, 6]. This result for stationary input processes is subsumed by the measure of modularity dissipation. Since there is no uncertainty in X_N, the asymptotic dissipation of memoryless ratchets simplifies to:

$$\begin{aligned} \langle \Sigma_\infty^{\text{ext}} \rangle_{\text{min}} &= \lim_{N \to \infty} k_{\text{B}} \ln 2 \ \text{I}[Y_{N+1:\infty}; Y_N] \\ &= k_{\text{B}} \ln 2 \ (\text{H}_1 - h_\mu) \ , \end{aligned}$$

where in the second step we used input stationarity—every symbol has the same marginal distribution—and so the same single-symbol uncertainty $\text{H}_1 = \text{H}[Y_N] = \text{H}[Y_M]$. Thus, the modularity dissipation of a memoryless ratchet is proportional to the length-1 redundancy $\text{H}_1 - h_\mu$ [12]. This is the amount of additional uncertainty that comes from ignoring temporal correlations. As Fig. 6.3 shows, this means that a memoryless extractor driven by the Golden Mean Process dissipates $\langle \Sigma_\infty^{\text{ext}} \rangle_{\text{min}} \approx 0.174 k_B$ with every bit, which lies in contrast Still et al.'s claim that "unwarranted retention of past information is fundamentally equivalent to energetic inefficiency," because such a memoryless ratchet minimizes the instantaneous nonpredictive information, which is the measure of dissipation in a driven system [178]. Moreover, it should be noted that the predictive model shown in Fig. 6.3 could include duplicate states and still be predictive and thus maximally efficient, further conflicting with Still et al.'s thermodynamics of prediction. Despite the fact that

both of these ratchets perform the same computational process—converting the Golden Mean Process into a sequence of IID symbols—the simpler model requires more energy investment to function, due to its irreversibility.

6.6 Retrodictive Generators

Pattern generators are rather like time-reversed pattern extractors, in that they take in an uncorrelated input process:

$$\Pr(Y_{0:\infty}) = \prod_{i=0}^{\infty} \Pr(Y_i) \ , \tag{6.23}$$

and turn it into a structured output process $\Pr(Y_{0:\infty})$ that has correlations among the symbols. The modularity dissipation of a generator $\langle \Sigma_N^{\text{gen}} \rangle_{\text{min}}$ can also be simplified by removing the uncorrelated input symbols:

$$\frac{\langle \Sigma_N^{\text{gen}} \rangle_{\text{min}}}{k_{\text{B}} \ln 2} = \text{I}[Y'_{0:N}; X_N] - \text{I}[Y'_{0:N}; X_{N+1} Y'_N] \ .$$

Paralleling extractors, App. 6.8.2 shows that retrodictive ratchets minimize the modularity dissipation to zero.

Retrodictive generator states carry as little information about the output past as possible. Since this ratchet generates the output, it must carry all the information shared between the output past and future. Thus, it shields output past from output future just as a predictive extractor does for the input process:

$$\text{I}[Y'_{N:\infty}; Y'_{0:N} | X_N] = 0 \ .$$

However, unlike the predictive states, the output future shields the retrodictive ratchet state from the output past:

$$\text{I}[X_N; Y'_{0:N} | Y'_{N:\infty}] = 0 \ . \tag{6.24}$$

As show in App. 6.8.2, these two conditions mean that X_N is retrodictive and imply that the modularity dissipation vanishes. While we have not established the equivalence of retrodictiveness and efficiency for pattern generators, as we have for predictive pattern extractors, there are easy-to-construct examples demonstrating that diverging from efficient retrodictive implementations leads to modularity dissipation at every step.

Consider once again the Golden Mean Process. Figure 6.4 shows that there are alternate ways to generate such a process from a hidden Markov model. The ϵ-machine, shown on the left, is the minimal predictive model, as discussed earlier. It is unifilar, which means that the current hidden state S_N^+ and current output Y'_N uniquely determine the next hidden state S_{N+1}^+ and that once synchronized to the hidden states one stays synchronized to them by observing only output symbols. Thus, its states are a function of past outputs. This is corroborated by the fact that the information atom $\mathrm{H}[S_N^+]$ is contained by the information atom for the output past $\mathrm{H}[Y'_{0:N}]$. The other hidden Markov model generator shown in Fig. 6.4 (right) is the time reversal of the ϵ-machine that generates the reverse process. This is much like the ϵ-machine, except that it is retrodictive instead of predictive. The recurrent states B and C are co-unifilar as opposed to unifilar. This means that the next hidden state S_{N+1}^- and the current output Y'_N uniquely determine the current state S_N^-. The hidden states of this minimal retrodictive model are a function of the semi-infinite future. And, this can be seen from the fact that the information atom for $\mathrm{H}[S_N^-]$ is contained by the information atom for the future $\mathrm{H}[Y'_{N:\infty}]$.

These two different hidden Markov generators both produce the Golden Mean Process, and they provide a template for constructing ratchets to generate that process. For a hidden Markov model described by symbol-labeled transition matrix $\{T^{(y)}\}$, with hidden states in $\mathcal{S}$ as described in Eq. (6.20), the analogous generative ratchet has the same states $\mathcal{X} = \mathcal{S}$ and is described by the joint Markov local interaction:

$$M^{\text{local}}_{(x,y)\to(x',y')} = T^{(y')}_{x\to x'} \ .$$

Such a ratchet effectively ignores the IID input process and obeys the same informational relationships between the ratchet states and outputs as the hidden states of hidden Markov model with its outputs.

Figure 6.5 shows both the transducer and joint Markov representation of the minimal predictive generator and minimal retrodictive generator. The retrodictive generator is potentially perfectly efficient, since the process' minimal modularity dissipation vanishes: $\langle \Sigma_N^{\text{gen}} \rangle_{\text{min}} = 0$ for all N. However, despite being a standard tool for generating an output, the predictive ϵ-machine is necessarily irreversible and dissipative. The ϵ-machine-based

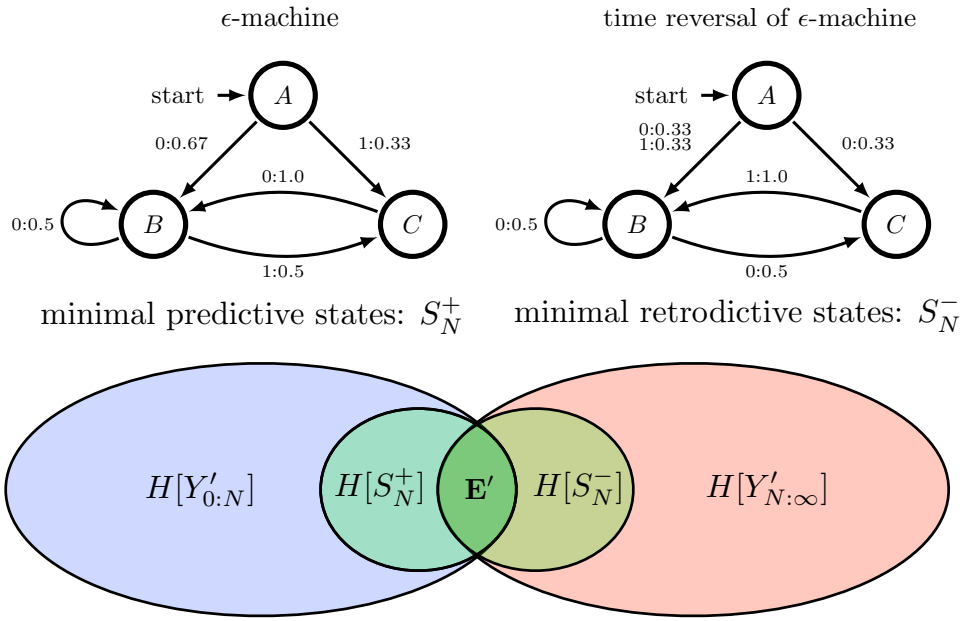


Figure 6.4. Alternate minimal generators of the Golden Mean Process: predictive and retrodictive. (Left) The ϵ-machine has the minimal set of causal states S^+ required to predictively generate the output process. As a result, the uncertainty $\mathrm{H}[S_N^+]$ is contained by the uncertainty $\mathrm{H}[Y'_{0:N}]$ in the output past. (Right) The time reversal of the reverse-time ϵ-machine has the minimal set of states required to retrodictively generate the output. Its states are a function of the output future. Thus, its uncertainty $\mathrm{H}[S_N^-]$ is contained by the output future's uncertainty $\mathrm{H}[Y'_{N:\infty}]$.

ratchet, as shown in Fig. 6.5(bottom row), approaches an asymptotic dynamic where the current state X_N stores more than it needs to about the past output past $Y'_{0:N}$ in order to generate the future $Y'_{N:\infty}$. As a result, it irretrievably dissipates:

$$\begin{aligned}
\langle \Sigma_\infty^{\mathrm{gen}} \rangle_{\mathrm{min}} &= k_{\mathrm{B}} \ln 2 \lim_{N\to\infty} (\mathrm{I}[Y'_{0:N}; X_N] - \mathrm{I}[Y'_{0:N}; X_{N+1}, Y'_N]) \\
&= \tfrac{2}{3} k_{\mathrm{B}} \ln 2 \\
&\approx 0.462 \ k_{\mathrm{B}} \ .
\end{aligned}$$

With every time step, this predictive ratchet stores information about its past, but it also erases information, dissipating 2/3 of a bit worth of correlations without leveraging them. Those correlations could have been used to reverse the process if they had been turned into work. They are used by the retrodictive ratchet, though, which stores just enough information about its past to generate the future.

The previous chapter showed that storing unnecessary information about the past

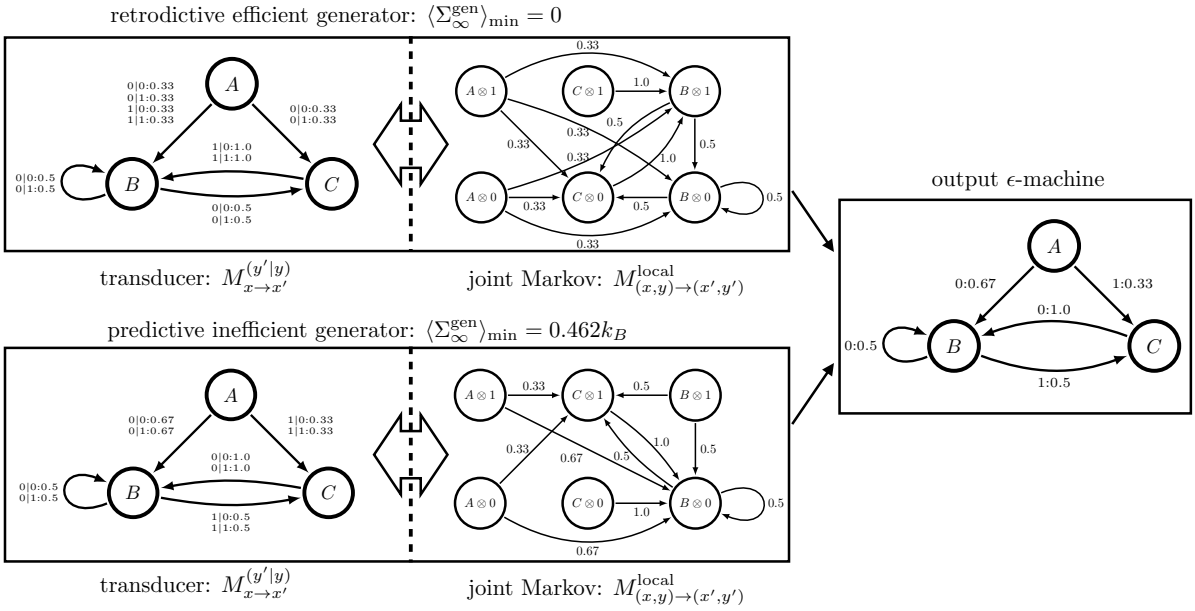


Figure 6.5. Alternative generators of the Golden Mean Process: (Right) The process' ϵ-machine. (Top row) Optimal generator designed using the topology of the minimal retrodictive generator. It is efficient, since it stores as little information about the past as possible, while still storing enough to generate the output. (Bottom row) The predictive generator stores far more information about the past than necessary, since it is based off the predictive ϵ-machine. As a result, it is far less efficient. It dissipates at least $\frac{2}{3} k_{\text{B}} T \ln 2$ extra heat per symbol and requires that much more work energy per symbol emitted.

leads to additional transient dissipation when generating a pattern [19, 20]. This cost also arises from implementation. However, our measure of modularity dissipation shows that there are implementation costs that persist through time. The two locally-operating generators of the Golden Mean Process perform the same computation, but have different bounds on their dissipation per time step. Thus, the additional work investment required to generate the process grows linearly with time for the ϵ-machine implementation, but is zero for the retrodictive implementation.

Moreover, we can consider generators that fall in-between these extremes using the parametrized HMM shown in Fig. 6.6 (top). This HMM, parametrized by z, produces the Golden Mean Process at all $z \in [.5, 1]$, but the hidden states share less and less information with the output past as z increases, as shown by Ref. [2]. One extreme $z = 0.5$ corresponds to the minimal predictive generator, the ϵ-machine. The other at $z = 1$ corresponds to the minimal retrodictive generator, the time reversal of the reverse-time ϵ-machine. The graph

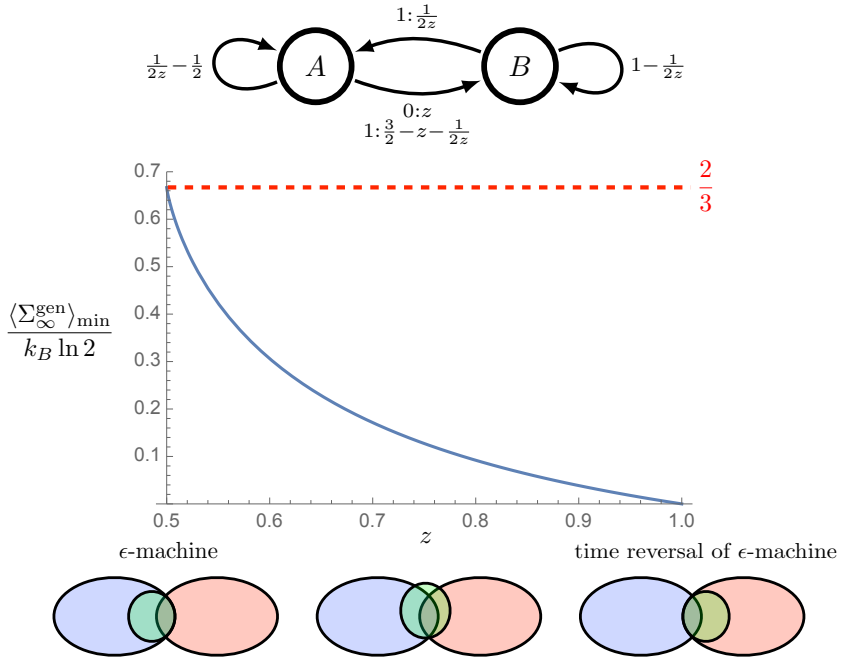


Figure 6.6. (Top) A parametrized family of HMMs that generate the Golden Mean Process for $z \in [.5, 1]$. (Middle) As parameter z increases, the information stored in the hidden states about the output past decreases. At $z = 0.5$ the HMM is the ϵ-machine, whose states are a function of the past. At $z = 1.0$, the HMM is the time reversal of the reverse-time ϵ-machine, whose states are a function of the future. The modularity dissipation decreases monotonically as z increases and the hidden states' memory of the past decreases. (Bottom) Information diagrams corresponding to the end cases and a middle case. Labeling as in Fig. 6.4.

there plots the modularity dissipation as a function of z. It decreases with z, suggesting that the unnecessary memory of the past leads to additional dissipation, and echoing the claim that "unwarranted retention of past information is fundamentally equivalent to energetic inefficiency" in the particular context of pattern generation [178]. So, while we have only proved that retrodictive generators are maximally efficient, this demonstrates that extending beyond that class can lead to unnecessary dissipation and that there may be a direct relationship between unnecessary memory and dissipation.

Taken altogether, we see that the thermodynamic consequences of localized information processing lead to direct principles for efficient information transduction. Analyzing the most general case of transducing arbitrary structured processes into other arbitrary structured processes remains a challenge. That said, pattern generators and pattern extractors have elegantly symmetric conditions for efficiency that give insight into the range

of possibilities. Pattern generators are effectively the time-reversal of pattern extractors, which turn structured inputs into structureless outputs. As such they are most efficient when retrodictive, which is the time-reversal of being predictive. Figure 6.4 illustrated graphically how the predictive ϵ-machine captures past correlations and stores the necessary information about the past, while the retrodictive ratchet's states are analogous, but store information about the future instead. This may seem unphysical—as if the ratchet is anticipating the future. However, since the ratchet generates the output future, this anticipation is entirely physical, because the ratchet controls the future, as opposed to mysteriously predicting it, as an oracle would.

6.7 Conclusion

Modularity is a key design theme in physical information processing, since it gives the flexibility to stitch together many elementary logical operations to implement a much larger computation. Any classical computation can be composed from local operations on a subset of information reservoir observables. Modularity is also key to biological organization, its functioning, and our understanding of these [163].

However, there is an irretrievable thermodynamic cost, the modularity dissipation, to this localized computing, which we quantified in terms of the global entropy production. This modularity-induced entropy production is proportional to the reduction of global correlations between the local and interacting portion of the information reservoir and the fixed, noninteracting portion. This measure forms the basis for designing thermodynamically efficient information processing. It is proportional to the additional work investment required by the modular form of the computation, beyond the work required by a globally integrated and reversible computation.

Turing machine-like information ratchets provide a natural application for this new measure of efficient information processing, since they process information in a symbol string through a sequence of local operations. The modularity dissipation allows us to determine which implementations are able to achieve the asymptotic bound set by the IPSL which, substantially generalizing Landauer's bound, says that any type of structure

in the input can be used as a thermal resource and any structure in the output has a thermodynamic cost. There are many different ratchet implementations that perform a given computation, in that they map inputs to outputs in the same way. However, if we want an implementation to be thermodynamically efficient, the modularity dissipation, monitored by the global entropy production, must be minimized. Conversely, we now appreciate why there are many implementations that dissipate and are thus irreversible. This establishes modularity dissipation as a new thermodynamic cost, due purely to an implementation's architecture, that complements Landauer's bound on isolated logical operations.

We noted that there are not yet general principles for designing devices that minimize modularity dissipation and thus work investment for arbitrary information transduction. However, for the particular cases of pattern generation and pattern extraction we find that there are prescribed classes of ratchets that are guaranteed to be dissipationless, if operated isothermally. The ratchet states of these devices are able to store and leverage the global correlations among the symbol strings, which means that it is possible to achieve the reversibility of globally integrated information processing but with modular computational design. Thus, while modular computation often results in dissipating global correlations, this inefficiency can be avoided when designing processors by employing the tools of computations mechanics outlined here.

6.8 Appendices

6.8.1 Isothermal Markov Channels

To satisfy information-theoretic bounds on work dissipation, we describe an isothermal channel where we change system energies in slow steps to manipulate the distribution over $\mathcal{Z}$'s states. Precisely, our challenge is to evolve an input distribution $\Pr(Z_t = z_t)$ over the time interval $(t, t+\tau)$ according to the Markov channel M, so that the system's conditional probability at time $t+\tau$ is:

$$\Pr(Z_{t+\tau} = z_{t+\tau} | Z_t = z_t) = M_{z_t \to z_{t+\tau}}.$$

The Markov channel M specifies the form of the computation, as it probabilistically maps inputs to outputs. While we don't need to know the input distribution $\Pr(Z_t = z_t)$ to implement a computation, it is necessary to know this distribution in order to design a thermodynamically efficient computation. Making this as efficient as possible in a thermal environment at temperature T means ensuring that the work invested in the evolution achieves the lower bound:

$$\langle W \rangle \geq k_B T \ln 2 (\mathrm{H}[Z_t] - \mathrm{H}[Z_{t+\tau}]) \ .$$

This expresses the Second Law of Thermodynamics for the system in contact with a heat bath.

To ensure the appropriate conditional distribution, we introduce an ancillary system $\mathcal{Z}'$, which is a copy of $\mathcal{Z}$, as is proposed by Garner et al. in their implementation of thermodynamic pattern manipulation [20]. This is necessary, because continuous time Markov chains, which are the probabilistic rate equations that underly thermodynamic stochastic dynamics, have limits in the logical functions they can implement. Some logical functions, such as flipping a bit (0 to 1 and 1 to 0) must be implemented with ancillary/hidden states [185]. By including an ancillary system which is a copy of $\mathcal{Z}$, we allow ourselves to implement any probabilistic channel $M_{z_t \to z_{t+\tau}}$, and thus any logical operation on $\mathcal{Z}$. So that it is efficient, we take τ to be large with respect to the system's relaxation time scale and break the overall process into three steps that occur over the time intervals $(t, t + \tau_0)$, $(t + \tau_0, t + \tau_1)$, and $(t + \tau_1, t + \tau)$, where $0 < \tau_0 < \tau_1 < \tau$.

Our method of manipulating $\mathcal{Z}$ and $\mathcal{Z}'$ is to control the energy $E(t, z, z')$ of the joint state $z \otimes z' \in \mathcal{Z} \otimes \mathcal{Z}'$ at time t. We also control whether or not probability is allowed to flow in $\mathcal{Z}$ or $\mathcal{Z}'$. This corresponds to raising or lowering energy barriers between system states.

At the beginning of the control protocol we choose $\mathcal{Z}'$ to be in a uniform distribution uncorrelated with $\mathcal{Z}$. This means the joint distribution can be expressed:

$$\Pr(Z_t = z_t, Z'_t = z'_t) = \frac{\Pr(Z_t = z_t)}{|\mathcal{Z}'|} \ . \tag{6.25}$$

Since we are manipulating an energetically mute information reservoir, we also start with

the system in a uniformly zero-energy state over the joint states of $\mathcal{Z}$ and $\mathcal{Z}'$:

$$E(t, z, z') = 0 \ . \tag{6.26}$$

While this energy and the distribution change when executing the protocol, we return $\mathcal{Z}'$ to the independent uniform distribution and the energy to zero at the end of the protocol. This means that the starting and ending distributions are usually out of equilibrium, but because we limit the flow between informational states, they are metastable, and don't relax to the uniform equilibrium distribution. In this way, the information reservoir reliably stores and processes many different nonequilibrium states.

The three steps that evolve this system to isothermally implement the Markov channel M are as follows:

1. Over the time interval $(t, t + \tau_0)$, continuously change the energy such that the energy at the end of the interval $E(t + \tau_0, z, z')$ obeys the relation:

   $$e^{-(E(t+\tau_0,z,z')-F(t+\tau_0))/k_BT} = \Pr(Z_t = z)M_{z\to z'} \ ,$$

   while allowing state space and probability to flow in $\mathcal{Z}'$, but not in $\mathcal{Z}$. Since the protocol is done slowly, $\mathcal{Z}'$ follows the Boltzmann distribution and at time $t + \tau_0$ the distribution over $\mathcal{Z} \otimes \mathcal{Z}'$ is:

   $$\Pr(Z_{t+\tau_0} = z, Z'_{t+\tau_0} = z') = \Pr(Z_t = z)M_{z\to z'} \ .$$

   This yields the conditional distribution of the current ancillary variable $Z'_{t+\tau}$ on the initial system variable Z_t:

   $$\Pr(Z'_{t+\tau_0} = z'|Z_t = z) = M_{z\to z'} \ ,$$

   since the system variable Z_t remains fixed over the interval. This protocol effectively applies the Markov channel M to evolve from $\mathcal{Z}$ to $\mathcal{Z}'$. However, we want the Markov channel to apply strictly to $\mathcal{Z}$.

   Because the protocol is slow and isothermal, there is no entropy production and the

work flow is simply the change in nonequilibrium free energy:

$$\begin{aligned}\langle W_1 \rangle &= \Delta F^{\text{neq}} \\ &= \Delta \langle E \rangle - T \Delta S[Z, Z'] \ .\end{aligned}$$

Since the average initial energy is uniformly zero, the change in average energy is the average energy at time $t + \tau_0$. And so, we can express the work done:

$$\begin{aligned}\langle W_1 \rangle &= \langle E(t + \tau_0) \rangle - T \Delta S[Z, Z'] \\ &= \langle E(t + \tau_0) \rangle \\ &\quad + k_{\text{B}} T \ln 2 (\text{H}[Z_t, Z'_t] - \text{H}[Z_{t+\tau_0}, Z'_{t+\tau_0}]) \ .\end{aligned}$$

2. Now, swap the states of $\mathcal{Z}$ and $\mathcal{Z}'$ over the time interval $(t + \tau_0, t + \tau_1)$. This is logically reversible. Thus, it can be done without any work investment over the second time interval:

$$\langle W_2 \rangle = 0 \ . \tag{6.27}$$

The result is that the energies and probability distributions are flipped with regard to exchange of the system $\mathcal{Z}$ and ancillary system $\mathcal{Z}'$:

$$\begin{aligned}E(t + \tau_1, z, z') &= E(t + \tau_0, z', z) \\ \Pr(Z_{t+\tau_1} = z, Z'_{t+\tau_1} = z') &= \Pr(Z_{t+\tau_0} = z', Z'_{t+\tau_0} = z) \ .\end{aligned}$$

Most importantly, however, this means that the conditional probability of the current system variable is given by M:

$$\begin{aligned}Pr(Z_{t+\tau_1} = z' | Z_t = z) &= \Pr(Z'_{t+\tau_0} = z' | Z_t = z) \\ &= M_{z \to z'} \ .\end{aligned}$$

The ancillary system must still be reset to a uniform and uncorrelated state and the energies must be reset.

3. Finally, we again hold $\mathcal{Z}$'s state fixed while allowing $\mathcal{Z}'$ to change over the time interval $(t + \tau_1, t + \tau)$ as we change the energy, ending at $E(t + \tau, z, z') = 0$. This

isothermally brings the joint distribution to one where the ancillary system is uniform and independent of $\mathcal{Z}$:

$$\Pr(Z_{t+\tau} = z, Z'_{t+\tau} = z') = \frac{\Pr(Z_{t+\tau} = z)}{|\mathcal{Z}'|} \ . \tag{6.28}$$

Again, the invested work is the change in average energy plus the change in thermodynamic entropy of the joint system:

$$\begin{aligned}
\langle W_3 \rangle &= \langle \Delta E \rangle \\
&\quad + k_{\mathrm{B}} T \ln 2 (\mathrm{H}[Z_{t+\tau_1}, Z'_{t+\tau_1}] - \mathrm{H}[Z_{t+\tau}, Z'_{t+\tau}]) \\
&= -\langle E(t+\tau_1) \rangle \\
&\quad + k_{\mathrm{B}} T \ln 2 (\mathrm{H}[Z_{t+\tau_1}, Z'_{t+\tau_1}] - \mathrm{H}[Z_{t+\tau}, Z'_{t+\tau}]) \ .
\end{aligned}$$

This ends this three-step protocol.

Summing up the heat terms, gives the total dissipation:

$$\begin{aligned}
\langle W_{\mathrm{total}} \rangle &= \langle W_t \rangle + \langle W_2 \rangle + \langle W_3 \rangle \\
&= k_{\mathrm{B}} T \ln 2 (\mathrm{H}[Z_t, Z'_t] - \mathrm{H}[Z_{t+\tau_0}, Z'_{t+\tau_0}]) \\
&\quad + k_{\mathrm{B}} T (\mathrm{H}[Z_{t+\tau_1}, Z'_{t+\tau_1}] - \mathrm{H}[Z_{t+\tau}, Z'_{t+\tau}]) \\
&\quad + \langle E(t+\tau_0) \rangle - \langle E(t+\tau_1) \rangle \ .
\end{aligned}$$

Recall that the distributions $\Pr(Z_{t+\tau_1}, Z'_{t+\tau_1})$ and $\Pr(Z_{t+\tau_0}, Z'_{t+\tau_0})$, as well as $E(t+\tau_0, z, z')$ and $E(t+\tau_1, z, z')$, are identical under exchange of $\mathcal{Z}$ and $\mathcal{Z}'$, so $H[Z_{t+\tau_1}, Z'_{t+\tau_1}] = H[Z_{t+\tau_0}, Z'_{t+\tau_0}]$ and $\langle E(t+\tau_0) \rangle = \langle E(t+\tau_1) \rangle$. Additionally, we know that both the starting and ending distributions have a uniform and uncorrelated ancillary system, so their entropies can be expressed:

$$\mathrm{H}[Z_t, Z'_t] = \mathrm{H}[Z_t] + \log_2 |\mathcal{Z}'| \tag{6.29}$$

$$\mathrm{H}[Z_{t+\tau}, Z'_{t+\tau}] = \mathrm{H}[Z_{t+\tau}] + \log_2 |\mathcal{Z}'| \ . \tag{6.30}$$

Substituting this in to the above expression for total invested work, we find that we achieve the lower bound with this protocol:

$$\langle W_{\mathrm{total}} \rangle = k_{\mathrm{B}} T \ln 2 (\mathrm{H}[Z_t] - \mathrm{H}[Z_{t+\tau}]) \ . \tag{6.31}$$

Thus, the protocol implements a Markov channel that achieves the information-theoretic bounds. It is similar to that described in Ref. [20].

The basic principle underlying the protocol is that when manipulating system energies to change state space, choose the energies so that there is no instantaneous probability flow. That is, if one stops changing the energies, the distribution will not change. However, there are cases in which it is impossible prevent instantaneous flow. Then, there are necessarily inefficiencies that arise from the dissipation of the distribution flowing out of equilibrium.

6.8.2 Transducer Dissipation

6.8.2.1 Predictive Extractors

For a pattern extractor, being reversible means that the transducer is predictive of the input process. More precisely, an extracting transducer that produces zero entropy is equivalent to it being a predictor of its input.

As discussed earlier, a reversible extractor satisfies:

$$\mathrm{I}[Y_{N+1:\infty}; X_{N+1}] = \mathrm{I}[Y_{N+1:\infty}; X_N Y_N] \ ,$$

for all N, since it must be reversible at every step to be fully reversible. The physical ratchet being predictive of the input means two things. It means that X_N shields the past $Y_{0:N}$ from the future $Y_{N:\infty}$. This is equivalent to the mutual information between the past and future vanishing when conditioned on the ratchet state:

$$\mathrm{I}[Y_{0:N}; Y_{N:\infty} | X_N] = 0 \ .$$

Note that this also implies that any subset of the past or future is independent of any other subset conditioned on the ratchet state:

$$\mathrm{I}[Y_{a:b}; Y_{c:d} | X_N] = 0 \text{ where } b \leq N \text{ and } c \geq N \ .$$

The other feature of a predictive transducer is that the past shields the ratchet state from the future:

$$\mathrm{I}[X_N; Y_{N:\infty} | Y_{0:N}] = 0 \ .$$

This is guaranteed by the fact that transducers are nonanticipatory: they cannot predict future inputs outside of their correlations with past inputs.

We start by showing that if the ratchet is predictive, then the entropy production vanishes. It is useful to note that being predictive is equivalent to the extractor storing as much information about the input future as the input past does:

$$\mathrm{I}[X_N; Y_{N:\infty}] = \mathrm{I}[Y_{0:N}; Y_{N:\infty}] \ ,$$

which can be seen by subtracting $\mathrm{I}[Y_{0:N}; Y_{N:\infty}; X_N]$ from each side of the immediately preceding expression and noting that the ratchet is nonanticipatory. Thus, it is sufficient to show that the mutual information between the partial input future $Y_{N+1:\infty}$ and the joint distribution of the predictive variable X_N and next output Y_N is the same as mutual information with the joint variable $(Y_{0:N}, Y_N) = Y_{0:N+1}$ of the past inputs and the next input:

$$\mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] = \mathrm{I}[Y_{N+1:\infty}; Y_{0:N}, Y_N] \ .$$

To show this for a predictive variable, we use Fig. 6.7, which displays the information diagram for all four variables with the information atoms of interest labeled.

Assuming that X_N is predictive zeros out a number of information atoms, as shown below:

$$\begin{aligned}
\mathrm{I}[X_N; Y_N, Y_{N+1:\infty} | Y_{0:N}] &= b + c + h = 0 \\
\mathrm{I}[X_N; Y_N | Y_{0:N}] &= b + h = 0 \\
\mathrm{I}[Y_{0:N}; Y_N, Y_{N+1:\infty} | X_N] &= i + f + g = 0 \\
\mathrm{I}[Y_{0:N}; Y_N | X_N] &= i + f = 0 \ .
\end{aligned}$$

These four equations make it clear that $g = c = 0$. Thus, substituting $\mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] = a + b + c + d + e + f$ and $\mathrm{I}[Y_{N+1:\infty}; Y_{0:N}, Y_N] = a + b + d + e + f + g$, we find that their difference vanishes:

$$\begin{aligned}
\mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; Y_{0:N}, Y_N] &= c - g \\
&= 0 \ .
\end{aligned}$$

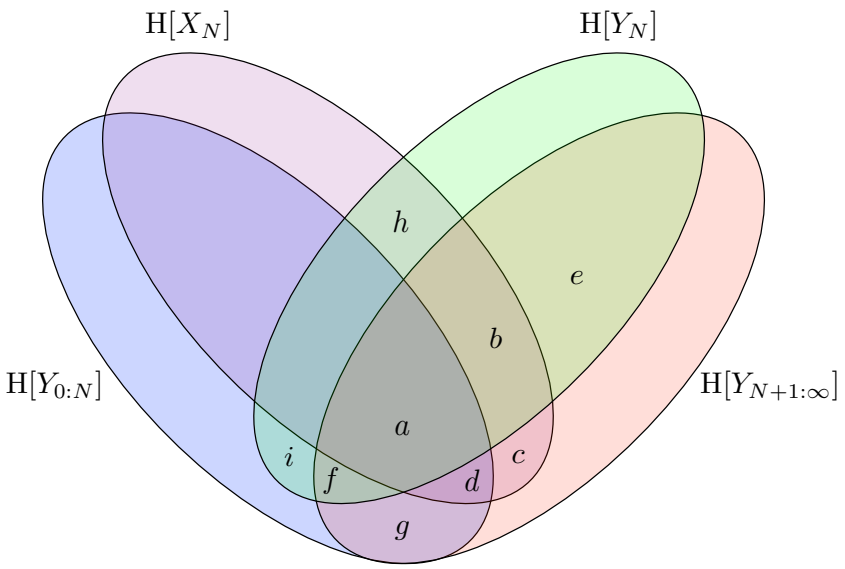


Figure 6.7. Information diagram for dependencies between the input past $Y_{0:N}$, next input Y_N, current ratchet state X_N, and input future $Y_{N+1:\infty}$, excluding the next input. We label certain information atoms to help illustrate the algebraic steps in the associated proof.

There is zero dissipation, since X_{N+1} is also predictive, meaning $\mathrm{I}[Y_{N+1:\infty}; Y_{0:N}, Y_N] = \mathrm{I}[Y_{N+1:\infty}; X_{N+1}]$, so:

$$\begin{aligned}\frac{\langle \Sigma_N^{\mathrm{ext}} \rangle_{\mathrm{min}}}{k_{\mathrm{B}} T \ln 2} &= \mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; X_{N+1}] \\ &= \mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; Y_{0:N+1}] \\ &= 0 \ .\end{aligned}$$

Going the other direction, using zero entropy production to prove that X_N is predictive for all N is now simple.

We already showed that $\mathrm{I}[Y_{N+1:\infty}; X_N, Y_N] = \mathrm{I}[Y_{N+1:\infty}; Y_{0:N}, Y_N]$ if X_N is predictive. Combining with zero entropy production ($\mathrm{I}[Y_{N+1:\infty}; X_{N+1}] = \mathrm{I}[Y_{N+1:\infty}; X_N, Y_N]$) immediately implies that X_{N+1} is predictive, since $\mathrm{I}[Y_{N+1:\infty}; X_{N+1}] = \mathrm{I}[Y_{N+1:\infty}; Y_{0:N}, Y_N]$ plus the fact that X_{N+1} is nonanticipatory is equivalent to X_{N+1} being predictive.

With this recursive relation, all that is left to establish is the base case, that X_0 is predictive. Applying zero entropy production again we find the relation necessary for

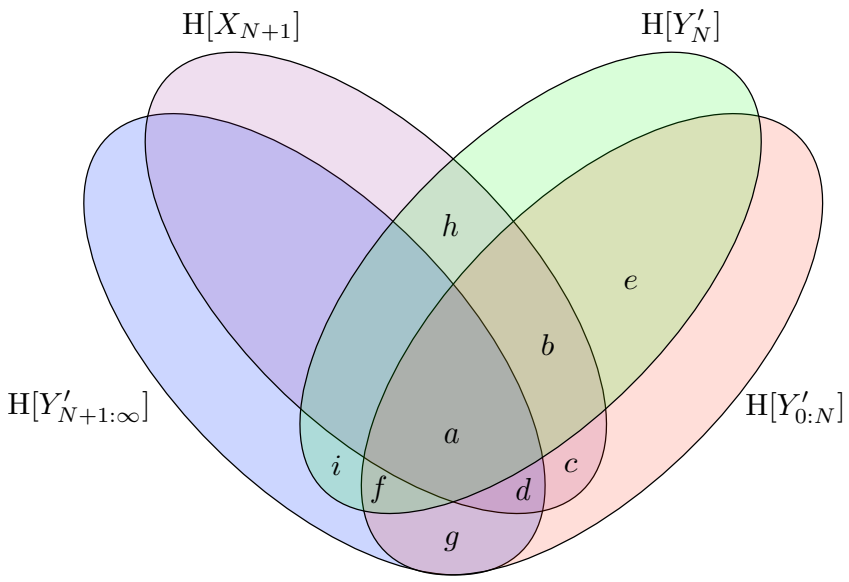


Figure 6.8. Information shared between the output past $Y'_{0:N}$, next output Y'_N, next ratchet state X_{N+1}, and output future $Y'_{N+1:\infty}$, excluding the next input. Key information atoms are labeled.

prediction:

$$\begin{aligned} \mathrm{I}[Y_{1:\infty}; X_1] &= \mathrm{I}[Y_{1:\infty}; X_0, Y_0] \\ &= \mathrm{I}[Y_{1:\infty}; Y_0] \ , \end{aligned}$$

From this, we find the equivalence $\mathrm{I}[Y_{1:\infty}; Y_0] = \mathrm{I}[Y_{1:\infty}; X_0, Y_0]$, since X_0 is independent of all inputs, due to it being nonanticipatory. Thus, zero entropy production is equivalent to predictive ratchets for pattern extractors.

6.8.2.2 Retrodictive Generators

An analogous argument can be made to show the relationship between retrodiction and zero entropy production for pattern generators, which are essentially time reversed extractors.

Efficient pattern generators must satisfy:

$$\mathrm{I}[Y'_{0:N}; X_N] = \mathrm{I}[Y'_{0:N}; X_{N+1} Y'_N] \ .$$

The ratchet being retrodictive means that the ratchet state X_N shields the past $Y'_{0:N}$ from

the future $Y'_{N:\infty}$ and that the future shields the ratchet from the past:

$$\begin{aligned} \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}|X_N] &= 0 \\ \mathrm{I}[Y'_{0:N}; X_N|Y'_{N:\infty}] &= 0 \ . \end{aligned}$$

Note that generators necessarily shield past from future $\mathrm{I}[Y'_{0:N}; Y'_{N:\infty}|X_N] = 0$, since all temporal correlations must be stored in the generator's states. Thus, for a generator, being retrodictive is equivalent to:

$$\mathrm{I}[Y'_{0:N}; X_N] = \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}] \ .$$

This can be seen by subtracting $\mathrm{I}[Y'_{0:N}; X_N; Y'_{N:\infty}]$ from both sides, much as done with the extractor.

First, to show that being retrodictive implies zero minimal entropy production, it is sufficient to show that:

$$\mathrm{I}[Y'_{0:N}; X_{N+1}, Y'_N] = \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}] \ ,$$

since we know that $\mathrm{I}[Y'_{0:N}; X_N] = \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}]$. To do this, consider the information diagram in Fig. 6.8. There we see that the difference between the two mutual informations of interest reduce to the difference between the two information atoms:

$$\mathrm{I}[Y'_{0:N}; X_{N+1}Y'_N] - \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}] = c - g \ .$$

The fact that the ratchet state X_{N+1} shields the past $Y'_{0:N+1}$ from the future $Y'_{N+1:\infty}$ and the future shields the ratchet from the past gives us the following four relations:

$$\begin{aligned} \mathrm{I}[Y'_{0:N}Y'_N; Y'_{N+1:\infty}|X_{N+1}] &= i + f + g = 0 \\ \mathrm{I}[Y'_N; Y'_{N+1:\infty}|X_{N+1}] &= i + f = 0 \\ \mathrm{I}[Y'_{0:N}Y'_N; X_{N+1}|Y'_{N+1:\infty}] &= h + b + c = 0 \\ \mathrm{I}[Y'_N; X_{N+1}|Y'_{N+1:\infty}] &= h + b = 0 \ . \end{aligned}$$

These equations show that that $c = g = 0$ and thus:

$$\frac{\langle \Sigma_N^{\mathrm{gen}} \rangle_{\mathrm{min}}}{k_{\mathrm{B}} T \ln 2} = 0 \ .$$

Going the other direction—zero entropy production implies retrodiction—requires that we use $\mathrm{I}[Y'_{0:N}; X_N] = \mathrm{I}[Y'_{0:N}; X_{N+1}, Y'_N]$ to show $\mathrm{I}[Y'_{0:N}; X_N] = \mathrm{I}[Y'_{0:N}; Y'_{N:\infty}]$. If X_{N+1} is retrodictive, then we can show that X_N must be as well. However, this makes the base case of the recursion difficult, since there is not yet a reason to conclude that X_∞ is retrodictive. While we conjecture the equivalence of optimally retrodictive generators and efficient generators, at this point we can only conclusively say that retrodictive generators are also efficient.

References

[1] 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.

[2] C. J. Ellison, J. R. Mahoney, R. G. James, J. P. Crutchfield, and J. Reichardt. Information symmetries in irreversible processes. *CHAOS*, 21(3):037107, 2011.

[3] H. Leff and A. Rex. *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*. Taylor and Francis, New York, 2002.

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

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

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

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

[8] M. Esposito and C. van den Broeck. Second law and Landauer principle far from eqilibrium. *Europhys. Lett*, 95:40004, 2011.

[9] Takahiro Sagawa Juan M. R. Parrondo, Jordan M. Horowitz. Thermodynamics of information. *Nature Physics*, 11(2):131–139, February 2015.

[10] N. Merhav. Sequence complexity and work extraction. *J. Stat. Mech.*, page P06037, 2015.

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

[12] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

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

[14] D. Mandal, A. B. Boyd, and J. P. Crutchfield. Memoryless thermodynamics? A reply. 2015. arxiv.org:1508.03311 [cond-mat.stat- mech].

[15] P. R. Zulkowski and M. R. DeWeese. Optimal finite-time erasure of a classical bit. *Phys. Rev. E*, 89:052140, May 2014.

[16] S. Lahiri, J. Sohl-Dickstein, and S. Ganguli. A universal tradeoff between power, precision and speed in physical communication. *arXiv:1603.07758*, 2016.

[17] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *Journal of Statistical Physics*, 167(6):1555–1585, 2017.

[18] W. R. Ashby. *An Introduction to Cybernetics*. John Wiley and Sons, New York, second edition, 1960.

[19] Paul M. Riechers James P. Crutchfield Alexander B. Boyd, Dibyendu Mandal. Transient dissipation and structural costs of physical information transduction. *Phys Rev Lett*, 118(22), 2017.

[20] Andrew J. P. Garner, Jayne Thompson, Vlatko Vedral, and Mile Gu. When is simpler thermodynamically better? *arXiv*, 1510.00010, 2015.

[21] Wolfgang Lohr. Predictive models and generative complexity. *Journal of Systems Science and Complexity*, 25(1):30–45, 2012.

[22] James P. Crutchfield Alexander B. Boyd, Dibyendu Mandal. Above and beyond the landauer bound: Thermodynamics of modularity. *arXiv:1708.03030 [cond-mat.stat-mech]*, 2017.

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

[24] D. J. Evans and D. J. Searles. Equilibrium microstates which generate second law violating steady states. *Phys. Rev. E*, 50:1645, 1994.

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

[26] J. Kurchan. Fluctuation theorem for stochastic dynamics. *J. Phys. A: Math. Gen.*, 31:3719, 1998.

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

[28] J. L. Lebowitz and H. Spohn. A Gallavotti-Cohen-type symmetry in the large deviation functional for stochastic dynamics. *J. Stat. Phys.*, 95:333, 1999.

[29] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco Jr., and C. Bustamante. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature*, 437:231, 2005.

[30] K. Maruyama, F. Nori, and V. Vedral. Colloquium: The physics of Maxwell's demon and information. *2009*, 81:1, Rev. Mod. Phys.

[31] M. v. Smoluchowski. Drei vorträge über diffusion, etc. *Physik. Zeit.*, XVII:557–571, 1916.

[32] L. Szilard. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.*, 53:840–856, 1929.

[33] O. Penrose. *Foundations of statistical mechanics; a deductive treatment*. Pergamon Press, Oxford, 1970.

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

[35] T. Sagawa. Thermodynamics of information processing in small systems. *Prog. Theo. Phys.*, 127:XXX, 2012.

[36] A. C. Barato and U. Seifert. Stochastic thermodynamics with information reservoirs. *Phys. Rev. E*, 90:042150, 2014.

[37] H. Touchette and S. Lloyd. Information-theoretic limits of control. *Phys. Rev. Lett.*, 84:1156, 2000.

[38] F. J. Cao, L. Dinis, and J. M. R. Parrondo. Feedback control in a collective flashing ratchet. *Phys. Rev Lett.*, 93:040603, 2004.

[39] T. Sagawa and M. Ueda. Generalized Jarzynski equality under nonequilibrium feedback control. *Phys. Rev. Lett.*, 104:090602, 2010.

[40] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Physics*, 6:988–992, 2010.

[41] M. Ponmurugan. Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys. Rev. E*, 82:031129, 2010.

[42] J. M. Horowitz and S. Vaikuntanathan. Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys. Rev. E*, 82:061120, 2010.

[43] J. M. Horowitz and J. M. R. Parrondo. Thermodynamic reversibility in feedback processes. *Europhys. Lett.*, 95:10005, 2011.

[44] L. Granger and H. Krantz. Thermodynamic cost of measurements. *Phys. Rev. E*, 84:061110, 2011.

[45] D. Abreu and U. Seifert. Extracting work from a single heat bath through feedback. *Europhys. Lett.*, 94:10001, 2011.

[46] S. Vaikuntanathan and C. Jarzynski. Modeling Maxwell's demon with a microcanonical szilard engine. *Phys. Rev. E*, 83:061120, 2011.

[47] A. Abreu and U. Seifert. Thermodynamics of genuine nonequilibrium states under feedback control. *Phys. Rev. Lett.*, 108:030601, 2012.

[48] A. Kundu. Nonequilibrium fluctuation theorem for systems under discrete and continuous feedback control. *Phys. Rev. E*, 86:021107, 2012.

[49] T. Sagawa and M. Ueda. Fluctuation theorem with information exchange: role of correlations in stochastic thermodynamics. *Phys. Rev. Lett.*, 109:180602, 2012.

[50] L. B. Kish and C. G. Granqvist. Energy requirement of control: Comments on Szilard's engine and Maxwell's demon. *Europhys. Lett.*, 98:68001, 2012.

[51] S. Ito and T. Sagawa. Information thermodynamics on causal networks. *Phys. Rev. Lett.*, 111:180603, 2013.

[52] D. Hartich, A. C. A. C. Barato, and U. Seifert. Stochastic thermodynamics of bipartite systems: transfer entropy inequalities and a maxwell's demon interpretation. *J. Stat. Mech.: Theor. Exp.*, 2013:P02016, 2014.

[53] J. M. Horowitz and M. Esposito. Thermodynamics with continuous information flow. *Phys. Rev. X*, 4:031015, 2014.

[54] J. M. Horowitz. Multipartite information flow for multiple Maxwell demons. *J. Stat. Mech.: Theor. Exp.*, 2015:P03006, 2015.

[55] M. Esposito and G. Schaller. Stochastic thermodynamics for "Maxwell demon" feedbacks. *Europhys. Lett.*, 99:30003, 2012.

[56] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito. Thermodynamics of a physical model implementing a Maxwell demon. *Phys. Rev. Lett.*, 110:040601, 2013.

[57] J. M. Horowitz, T. Sagawa, and J. M. R. Parrondo. Imitating chemical motors with mptimal information motors. *Phys. Rev. Lett.*, 111:010602, 2013.

[58] A. C. Barato and U. Seifert. Unifying three perspectives on information processing in stochastic thermodynamics. *Phys. Rev. Lett.*, 112:090601, 2014.

[59] J. M. Horowitz and H. Sandberg. Second-law-like inequalities with information and their interpretations. *New J. Phys.*, 16:125007, 2014.

[60] D. Mandal, H. T. Quan, and C. Jarzynski. Maxwell's refrigerator: an exactly solvable model. *Phys. Rev. Lett.*, 111:030602, 2013.

[61] A. C. Barato and U. Seifert. An autonomous and reversible Maxwell's demon. *Europhys. Lett.*, 101:60001, 2013.

[62] J. Hoppenau and A. Engel. On the energetics of information exchange. *Europhys. Lett.*, 105:50002, 2014.

[63] Z. Lu, D. Mandal, and C. Jarzynski. Engineering Maxwell's demon. *Physics Today*, 67(8):60–61, January 2014.

[64] J. Um, H. Hinrichsen, C. Kwon, and H. Park. Total cost of operating an information engine. *arXiv:1501.03733 [cond-mat.stat-mech]*, 2015.

[65] J. R. Dorfman. *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*. Cambridge University Press, Cambridge, United Kingdom, 1999.

[66] B. Alberts, A. Johnson, J. Lewis, D. Morgan, and M. Raff. *Molecular Biology of the Cell*. Garland Science, New York, sixth edition, 2014.

[67] L. Brillouin. Life, thermodynamics, and cybernetics. *Am. Scientist*, 37:554–568, 1949.

[68] H. R. Lewis and C. H. Papadimitriou. *Elements of the Theory of Computation*. Prentice-Hall, Englewood Cliffs, N.J., second edition, 1998.

[69] P. Strasberg, J. Cerrillo, G. Schaller, and T. Brandes. Thermodynamics of stochastic Turing machines. arXiv:1506.00894, 2015.

[70] C. Moore. Unpredictability and undecidability in dynamical systems. *Phys. Rev. Lett.*, 64:2354, 1990.

[71] *The Entropy of Functions of Finite-state Markov Chains*, volume 28, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957. Held at Liblice near Prague from November 28 to 30, 1956.

[72] B. Marcus, K. Petersen, and T. Weissman, editors. *Entropy of Hidden Markov Process and Connections to Dynamical Systems*, volume 385 of *Lecture Notes Series*. London Mathematical Society, 2011.

[73] A. Chapman and A. Miyake. How can an autonomous quantum maxwell demon harness correlated information? *Phys. Rev. E*, 92:062125, 2015.

[74] P. M. Riechers and J. P. Crutchfield. Broken reversibility: Fluctuation theorems, path entropies, and exact results for complex nonequilibrium systems. *in preparation*, 2015.

[75] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

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

[77] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012.

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

[79] R. W. Yeung. *Information Theory and Network Coding*. Springer, New York, 2008.

[80] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James. Synchronization and control in intrinsic and designed computation: An information-theoretic analysis of competing models of stochastic computation. *CHAOS*, 20(3):037105, 2010. Santa Fe Institute Working Paper 10-08-015; arxiv.org:1007.5354 [cond-mat.stat-mech].

[81] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[82] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck. Dissipation: The phase-space perspective. *Phys. Rev. Lett.*, 98:080602, 2007.

[83] Irreversibility and the arrow of time in a quenched quantum system. *Phys. Rev. Lett.*, 115:190601, 2015.

[84] J. C. Maxwell. *Theory of Heat*. Longmans, Green and Co., London, United Kingdom, ninth edition, 1888.

[85] Y. Cao, Z. Gong, and H. T. Quan. Thermodynamics of information processing based on enzyme kinetics: An exactly solvable model of an information pump. *Phys. Rev. E*, 91:062117, 2015.

[86] N. Shiraishi, S. Ito, K. Kawaguchi, and T. Sagawa. Role of measurement-feedback separation in autonomous maxwell's demon. *New J. Phys.*, 17:045012, 2015.

[87] G. Diana, G. B. Bagci, and M. Esposito. Finite-time erasing of information stored in fermionic bits. *Phys. Rev. E*, 87:012111, 2013.

[88] S. K. Park and K. W. Miller. Random number generators: Good ones are hard to find. *Comm. ACM*, 31:1192–1201, 1988.

[89] F. James. A review of pseudorandom number generators. *Comp. Phys. Comm.*, 60:329–344, 1990.

[90] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong. Monte carlo simulations: Hidden errors from "good" random number generators. *Phys. Rev. Lett.*, 69:3382–3384, 1992.

[91] J. Oppenheim, M. Horodecki, P. Horodecki, and R. Horodecki. Thermodynamical approach to quantifying quantum correlations. *Phys. Rev. Lett.*, 89(18):180402, oct 2002.

[92] W. H. Zurek. Quantum discord and Maxwell's demons. *Phys. Rev. A*, 67(1):012320, jan 2003.

[93] K. Maruyama, F. Morikoshi, and V. Vedral. Thermodynamical detection of entanglement by Maxwell's demons. *Phys. Rev. A*, 71(1):012108, 2005.

[94] R. Dillenschneider and E. Lutz. Energetics of quantum correlations. *EPL (Europhysics Lett.*, 88(5):50003, dec 2009.

[95] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral. Inadequancy of von neumann entropy for characterizing extractable work. *New. J. Phys.*, 13:053015, 2011.

[96] S. Jevtic, D. Jennings, and T. Rudolph. Maximally and minimally correlated states attainable within a closed evolving system. *Phys. Rev. Lett.*, 108(11):110403, mar 2012.

[97] K. Funo, Y. Watanabe, and M. Ueda. Thermodynamic work gain from entanglement. *Phys. Rev. A*, 88(5):052319, 2013.

[98] H. C. Braga, C. C. Rulli, T. R. De Oliveira, and M. S. Sarandy. Maxwell's demons in multipartite quantum correlated systems. *Phys. Rev. A*, 90(4):042338, 2014.

[99] Extractable work from correlations. *Phys. Rev. X*, 5:041011, 2015.

[100] T. McGrath, N. S. Jones, P. R. ten Wolde, and T. E. Ouldridge. Biochemical machines for the interconversion of mutual information and work. *Phys. Rev Lett.*, 118(028101), 2017.

[101] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Many roads to synchrony: Natural time scales and their algorithms. *Phys. Rev. E*, 89:042135, 2014.

[102] Y. Izumida, H. Kori, and U. Seifert. Energetics of synchronization in coupled oscillators. *Arxiv*, 126001(2012):126001, 2016.

[103] D. Andrieux and P. Gaspard. Nonequilibrium generation of information in copolymerization processes. *Proc. Natl. Acad. Sci. USA*, 105:9516–9521, 2008.

[104] J. Hopfield. Kinetic proofreading - new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA*, 71:4135–4139, 1974.

[105] J. Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57:587–595, 1975.

[106] M. Ehrenberg and C. Blomberg. Thermodynamic contstraints on kinetic proofreading in biosynthetic pathways. *Biophys. J.*, 31:333–358, 1980.

[107] C. H. Bennett. Dissipation error tradeoff in proofrading. *BioSystems*, 11:85–91, 1979.

[108] A. Murugan, D. A. Huse, and S. Leibler. Speed, dissipation, and error in kinetic proofreading. *Proc Natl. Acad. Sci. USA*, 109:12034, 2012.

[109] J. V. Koski, A. Kutvonen, I. M. Khaymovich, T. Ala-Nissila, and J. P. Pekola. On-chip maxwell's demon as an information-powered refrigerator. *Phys. Rev. Lett.*, 115(26):260602, dec 2015.

[110] R. B. Karabalin, M. H. Matheny, X. L. Feng, E. Defa, G. Le Rhun, C. Marcoux, S. Hentz, P. Andreucci, and M. L. Roukes. Piezoelectric nanoelectromechanical resonators based on aluminum nitride thin films. *Appl. Phys. Lett.*, 95(10):103111, 2009.

[111] 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(1):014101, jan 2014.

[112] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell. Observation of Bose-Einstein condensation in a dilute atomic vapor. *Science*, 269(5221):198–201, 1995.

[113] K. B. Davis, M. O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle. Bose-Einstein condensation in a gas of sodium atoms. *Phys. Rev. Let.*, 75:3969, 1995.

[114] C. C. Bradley, C. A. Sackett, J. J. Tollett, and R. G. Hulet. Evidence of Bose-Einstein condensation in an atomic gas with attractive interactions. *Phys. Rev. Let.*, 75:1687, 1995.

[115] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, second edition, 1992.

[116] C. E. Shannon and J. McCarthy, editors. *Automata Studies*. Number 34 in Annals of Mathematical Studies. Princeton University Press, Princeton, New Jersey, 1956.

[117] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York, 1949.

[118] N. Wiener. Nonlinear prediction and dynamics. In *Norbert Wiener, Collected Works III*. MIT Press, Cambridge, 1981.

[119] C. E. Shannon. Communication theory of secrecy systems. *Bell Sys. Tech. J.*, 28:656–715, 1949.

[120] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, 7:142–163, 623–656, 1959.

[121] C. E. Shannon. Two-way communication channels. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 611–644. Univ. of Calif. Press, 1961.

[122] N. Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT, Cambridge, 1948.

[123] N. Wiener. *The Human Use Of Human Beings: Cybernetics And Society*. Da Capo Press, Cambridge, 1988.

[124] L. von Bertalanffy. *General System Theory: Foundations, Development, Applications*. Penguin University Books, New York, revised edition, 1969.

[125] W. R. Ashby. *Design for a Brain: The Origin of Adaptive Behavior*. Chapman and Hall, New York, second edition, 1960.

[126] H. Quastler. The status of information theory in biology—A roundtable discussion. In H. P. Yockey, editor, *Symposium on Information Theory in Biology*. Pergamon Press, 1958.

[127] F. Conway and J. Siegelman. *Dark Hero of the Information Age: In Search of Norbert Wiener The Father of Cybernetics*. Basic Books, New York, New York, 2006.

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

[129] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[130] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29–40, 1983.

[131] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[132] P. M. B. Vitanyi. *Introduction to Kolmogorov complexity and its applications*. ACM Press, Reading, 1990.

[133] L. Brillouin. Maxwell's demon cannot operate: Information and entropy I. *J. Appl. Phys.*, 22:334–337, 1951.

[134] C. H. Bennett. Demons, engines and the Second Law. *Sci. Am.*, 257(5):108–116, 1987.

[135] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[136] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[137] J. G. Brookshear. *Theory of computation: Formal languages, automata, and complexity*. Benjamin/Cummings, Redwood City, California, 1989.

[138] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January, 1986.

[139] L. R. Rabiner. A tutorial on hidden Markov models and selected applications. *IEEE Proc.*, 77:257, 1989.

[140] R. J. Elliot, L. Aggoun, and J. B. Moore. *Hidden Markov Models: Estimation and Control*, volume 29 of *Applications of Mathematics*. Springer, New York, 1995.

[141] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002.

[142] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[143] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[144] A. del Campo, J. Goold, and M. Parenostro. More bang for your buck: Superadiabatic quantum engines. *Scientific Reports*, 4:6208, 2014.

[145] M. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[146] J. G. Kemeny and J. L. Snell. *Denumerable Markov Chains*. Springer, New York, second edition, 1976.

[147] R. Phillips, J. Kondev, J. Theriot, and N. Orme. *Physical biology of the cell*. Garland Science, New York, USA, 2008.

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

[149] S. Rana and A. M. Jayannavar. A multipurpose information engine that can go beyond the carnot limit. *J. Stat. Mech.*, page 103207, 2016.

[150] Takahiro Sagawa and Masahito Ueda. Information thermodynamics: Maxwell's demon in nonequilibrium dynamics. *Nonequilibrium Stat. Phys. Small Syst. Fluct. Relations Beyond*, pages 181–211, nov 2013.

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

[152] T. Hatano and S. Sasa. Steady-state thermodynamics of langevin systems. *Phys. Rev. Lett.*, 86, 2001.

[153] D. Mandal and C. Jarzynski. Analysis of slow transitions between nonequilibrium steady states. *J. Stat. Mech.*, 6(063204), 2016.

[154] P. M. Riechers and J. P. Crutchfield. Fluctuations when driving between nonequilibrium steady states. *arXiv:1610.09444 [cond-mat.stat-mech]*, 2016.

[155] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[156] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.*, 103(9):094101, 2009. SFI Working Paper 09-02-002; arxiv.org:0902.1209 [cond-mat.stat-mech]; DOI 10.1103/PhysRevLett.103.094101.

[157] 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.

[158] J. R. Mahoney, C. J. Ellison, R. G. James, and J. P. Crutchfield. How hidden are hidden processes? a primer on crypticity and entropy convergence. *CHAOS*, 21(3):037112, 2011.

[159] P. M. Riechers. *Exact Results Regarding the Physics of Complex Systems via Linear Algebra, Hidden Markov Models, and Information Theory*. PhD thesis, University of California, Davis, 2016.

[160] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Prentice-Hall, New York, third edition, 2006.

[161] E. Fredkin and T. Toffoli. Conservative logic. *Intl. J. Theo. Phys.*, 21(3/4):219–253, 1982.

[162] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. Bradford Book, New York, 1999.

[163] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, 2007.

[164] Karl Ulrich. *Management of Design*, chapter Fundamentals of Modularity. Number 219-231. Spring Netherlands, 1994.

[165] Nathan Crilly Chih-Chun Chen. From modularity to emergence: a primer on design and science of complex systems. Department of Engineering, University of Cambridge, September 2016.

[166] J. Maynard-Smith and E. Szathmary. *The Major Transitions in Evolution*. Oxford University Press, Oxford, reprint edition, 1998.

[167] S. Lloyd. Use of mutual information to decrease entropy: Implications for the second law of thermodynamics. *Phys. Rev. A*, 39:5378–5386, 1989.

[168] Kenneth Price Rainer Storn. Differential evolution - a simple and efficient heuristic for global optimization over continuous space. *Journal of Global Optimization*, 11(341-359), 1997.

[169] Derek S. Linden Jason D. Lohn, Gregory S. Hornby. *Genetic Programming Theory and Practice II*, chapter An Evolved Antenna for Deployment on Nasa's Space Technology 5 Mission. Number 301-315. Springer US, 2005.

[170] J. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Bradford Book, New York, 1992.

[171] R. Motwani and P. Raghavan. *Randomized algortihms*. Cambridge University Press, New York, 1995.

[172] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, 2005.

[173] C. Aghamohammadi and J. P. Crutchfield. Thermodynamics of random number generation. *Physical Review E*, 95:062139, 2017.

[174] Massimiliano Esposito. Stochastic thermodynamics under coarse graining. *Phys. Rev. E*, 85(041125), 2012.

[175] F. N. Fahn. Maxwell's demon and the entropy cost of information. *Found. Physics*, 26:71–93, 1996.

[176] Pieter Rein ten Wolde Thomas E Ouldridge, Christopher C Govern. Thermodynamics of computational copying in biochemical systems. *Phys. Rev. X*, 7(2):021004, 2017.

[177] Susanne Still. Thermodynamic cost and benefit of data representations. *arXiv:1705.00612v1*, 2017.

[178] Anthony J. Bell Gavin E. Crooks Susanne Still, David A. Sivak. Thermodynamics of prediction. *Phys Rev Lett*, 109(12):120604, 2012.

[179] Thomas McGrath, Nick S. Jones, Pieter Rein ten Wolde, and Thomas E. Ouldridge. Biochemical machines for the interconversion of mutual information and work. *Physical Review Letters*, 118:028101, 2017.

[180] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:c47, 1999.

[181] A. Turing. On computable numbers, with an application to the Entschiedungsproblem. *Proc. Lond. Math. Soc.*, 42, 43:230–265, 544–546, 1937.

[182] C. E. Shannon. A universal Turing machine with two internal states. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*, number 34 in Annals of Mathematical Studies, pages 157–165. Princeton University Press, Princeton, New Jersey, 1956.

[183] Neri Merhav. Relations between work and entropy production for general information-driven, finite-state engines. *Journal of Statistical Mechanics: Theory and Experiment*, 2017.

[184] Philipp Strasberg. *Thermodynamics and information processing at the nanoscale*. PhD thesis, Technische Universität Berlin, 2016.

[185] Jeremy A. Owen David H. Wolpert, Artemy Kochinsky. *The minimal hidden computer needed to implement a visible computation*. PhD thesis, arXiv:1708.08494 [cond-mat.stat-mech], 2017.