Title

Identifying functional thermodynamics in autonomous Maxwellian ratchets

Journal

New Journal of Physics, 18(2)

ISSN

1367-2630

Authors

Boyd, Alexander B
Mandal, Dibyendu
Crutchfield, James P

Publication Date

2016-02-01

DOI

10.1088/1367-2630/18/2/023049

Peer reviewed

Identifying functional thermodynamics in autonomous Maxwellian ratchets

2016 New J. Phys. 18 023049

(http://iopscience.iop.org/1367-2630/18/2/023049)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
26 October 2015

REVISED
21 January 2016

ACCEPTED FOR PUBLICATION
28 January 2016

PUBLISHED
22 February 2016

PAPER

Identifying functional thermodynamics in autonomous Maxwellian ratchets

Alexander B Boyd[1], **Dibyendu Mandal**[2] **and James P Crutchfield**[1,3]

[1] Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA
[2] Department of Physics, University of California, Berkeley, CA 94720, USA
[3] Author to whom any correspondence should be addressed.

E-mail: abboyd@ucdavis.edu, dibyendu.mandal@berkeley.edu and chaos@ucdavis.edu

Keywords: Maxwell's Demon, Maxwell's refrigerator, detailed balance, entropy rate, Second Law of thermodynamics

Abstract

We introduce a family of Maxwellian Demons for which correlations among information bearing degrees of freedom can be calculated exactly and in compact analytical form. This allows one to precisely determine Demon functional thermodynamic operating regimes, when previous methods either misclassify or simply fail due to approximations they invoke. This reveals that these Demons are more functional than previous candidates. They too behave either as engines, lifting a mass against gravity by extracting energy from a single heat reservoir, or as Landauer erasers, consuming external work to remove information from a sequence of binary symbols by decreasing their individual uncertainty. Going beyond these, our Demon exhibits a new functionality that erases bits not by simply decreasing individual-symbol uncertainty, but by increasing inter-bit correlations (that is, by adding temporal order) while increasing single-symbol uncertainty. In all cases, but especially in the new erasure regime, exactly accounting for informational correlations leads to tight bounds on Demon performance, expressed as a refined Second Law of thermodynamics that relies on the Kolmogorov–Sinai entropy for dynamical processes and not on changes purely in system configurational entropy, as previously employed. We rigorously derive the refined Second Law under minimal assumptions and so it applies quite broadly—for Demons with and without memory and input sequences that are correlated or not. We note that general Maxwellian Demons readily violate previously proposed, alternative such bounds, while the current bound still holds. As such, it broadly describes the minimal energetic cost of any computation by a thermodynamic system.

1. Introduction

The Second Law of thermodynamics is only statistically true: while the entropy production in any process is non-negative on the average, $\langle \Delta S \rangle \geqslant 0$, 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 [1–7]:

$$\frac{\Pr(\Delta S)}{\Pr(-\Delta S)} = \mathrm{e}^{\Delta S}, \tag{1}$$

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 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 [8, 9]. 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 [10].

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 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 [11]. Later, through the works of Landauer, Penrose, and 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 [12–14]. 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 [15]: the Demon's erasure of one bit of information at temperature TK requires at least $k_B T \ln 2$ 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 [16].)

Building on this, a modified Second Law was recently proposed that explicitly addresses information processing in a thermodynamic system [17, 18]:

$$\langle \Delta S \rangle + k_B \ln 2 \, \Delta H \geqslant 0, \tag{2}$$

where ΔH 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 [19]. 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 [17] 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 of information reservoirs 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 [20–33]; (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 [34–37]; and (iii) steady-state models that invoke time-scale separation to identify a portion of the overall entropy production as an information current [38, 39]. A unified approach to these perspectives was attempted in [40–42].

Recently, Maxwellian Demons have been proposed to explore plausible automated mechanisms that appeal to equation (2)'s modified Second Law to do useful work, by decreasing the physical entropy, at the expense of positive change in reservoir Shannon information [39, 43–48]. Paralleling the modified Second Law's development and the analysis 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 [19], 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 Second Law that refines equation (2) by properly accounting for intrinsic information processing reflected in temporal correlations via the overall dynamic's Kolmogorov–Sinai entropy [49]. Conversely, it bounds the thermodynamic cost of any computation. Helpfully, the hidden Markov model (HMM) representations we introduce afford a compact representation of a large class of information reservoirs, much broader than has been considered so far. Moreover, information reservoirs expressed as ϵ-machines lead to exact analytical treatment.

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

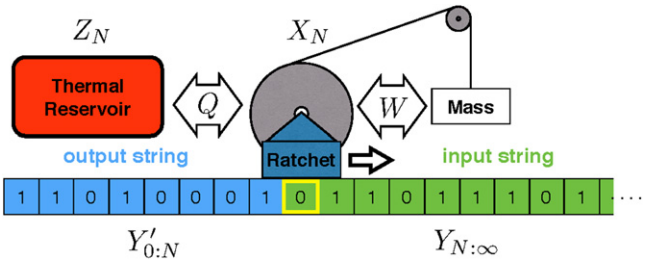

Figure 1. Information ratchet sequentially processing a bit string: at time step N, X_N is the random variable for the ratchet state and Z_N that for the thermal reservoir. $Y_{N:\infty}$ is the block random variable for the input bit string and $Y'_{0:N}$ that for the output bit string. The last bit Y_N of the input string, highlighted in yellow, interacts with the ratchet. 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.

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 environments.

2. Information ratchets

Our model consists of four components, see figure 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 [50] or that acts as an enzyme sequentially catalyzing otherwise unfavorable reactions [51]. 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 [52]. When the control logic is stochastic, this sometimes is referred to as 'Brownian computing' [53, and references therein]. Finally, we are reminded of the deterministic finite-state tape processor of [54] 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.

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 $\mathcal{X}_N \otimes \mathcal{Y}_N$, 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 figure 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 $\Pr(Y'_{0:\infty}) \equiv \{\Pr(Y'_{0:\infty} = v)\}_{v\in\{0,1\}^\infty}$. We assume that the word distributions we work with are stationary, meaning that $\Pr(Y_{a:a+b}) = \Pr(Y_{0:b})$ for all non-negative 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 $Y_{0:\ell}$ sequence is given by the *Shannon block entropy* [55]:

$$H[Y_{0:\ell}] \equiv -\sum_{w\in\{0,1\}^\ell} \Pr(Y_{0:\ell} = w)\log_2\Pr(Y_{0:\ell} = w). \tag{3}$$

Due to correlations in typical process sequences, the irreducible randomness per symbol is not the *single-symbol entropy* $H[Y_0]$. Rather, it is given by the *Shannon entropy rate* [55]:

$$h_\mu \equiv \lim_{\ell\to\infty} \frac{H[Y_{0:\ell}]}{\ell}. \tag{4}$$

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 behavior.

Note that these ways of monitoring information are quantitatively quite different. For large ℓ, $h_\mu\ell \ll H[Y_{0:\ell}]$ and, in particular, anticipating later use, $h_\mu \leqslant H[Y_0]$, 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 [56] and it remains a technical challenge [57]. The entropy rates of the output sequence and input sequence are $h'_\mu = \lim_{\ell\to\infty} H[Y'_{0:\ell}]/\ell$ and $h_\mu = \lim_{\ell\to\infty} H[Y_{0:\ell}]/\ell$, respectively.

The informational properties of the input and output word distributions set bounds on energy flows in the system. Appendix A 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[4]:

$$\begin{aligned}\langle W\rangle &\leqslant k_\mathrm{B}T\ \ln 2\ (h'_\mu - h_\mu)\\ &= k_\mathrm{B}T\ \ln 2\ \Delta h_\mu.\end{aligned} \tag{5}$$

A complementary interpretation of this new Second Law is that it places a lower bound on the thermodynamic cost (work) to perform information processing and intrinsic computation (change in Shannon entropy of the information reservoir).

In light of the preceding remarks on the basic difference between $H[Y_0]$ and h_μ, we can now consider more directly the differences between equations (2) and (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 is time invariant, even when the overall system's information production is positive. The entropies h'_μ and h_μ in equation (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 ΔH in the joint state space of all bits, the two equations are analogous in the current setup. However, often enough, a weaker version of equation (2) is considered in the discussions on Maxwell's Demon [41, 43–45, 58] and information reservoirs [18], wherein the statistical correlations between the bits are neglected, and one simply interprets ΔH to be the change in the marginal Shannon entropies $H[Y_0]$ of the individual bits. This implies the following relation in the current context:

$$\langle W\rangle \leqslant k_\mathrm{B}\ \ln 2\ \Delta H[Y_0], \tag{6}$$

where $\Delta H[Y_0] = H[Y'_0] - H[Y_0]$. While equation (6) is valid for the studies in [18, 41, 43–45, 58], it can be violated under certain scenarios [59]. In comparison, equation (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:

[4] Reference [43]'s appendix suggests equation (5) without any detailed proof. An integrated version appeared also in [58] for the special case of memoryless demons. Our appendix A gives a more general proof of equation (5) that, in addition, accounts for memory.

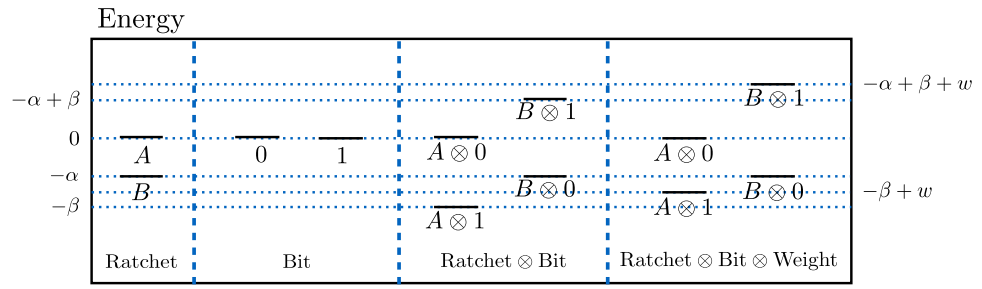

Figure 2. Energy levels of the Demon states, interacting bits, their joint system, and their joint system with a weight in units of $[k_B T]$.

$h_\mu = H[Y_0]$. However, the ratchet's memory can create correlations in the output bit string, so:

$$\begin{aligned}\Delta h_\mu &= h'_\mu - H[Y_0] \\ &\leqslant H[Y'_0] - H[Y_0] \\ &= \Delta H[Y_0]. \end{aligned} \tag{7}$$

In this case, equation (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 ratchet shown in figure 3 with parameters $\{p = 0.5, q = 0.1\}$ and input bias $\Pr(Y_i = 0) = 0.9$, 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 [55].

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 increased correlations; see section 5. As a result, we can now detect an eraser regime (previously unexpected) in the phase diagram of the system (figure 7). A similar regime may be present also in the model of [43] where the outgoing bits were observed to have small but finite correlations.

Note that a somewhat different situation was considered in [58], a memoryless channel (ratchet) driven by a correlated process. In this special case—ratchets unable to leverage or create temporal correlations—equation (6) is a tighter bound on work than equation (5). When a memoryless ratchet is driven by uncorrelated input, though, the bounds are equivalent. Critically, for memoryful ratchets driven by correlated input equation (6) can be violated. In all settings, equation (5) holds.

While we defer its development to a sequel, equation (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 [60].

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. 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 $E_B = -\alpha k_B T$, 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 $-\beta k_B T$ (or $+\beta k_B T$) if the bit is in state $Y = 1$ and the ratchet is in state A (or B). See figure 2 for a graphical depiction of the energy scheme under 'Ratchet ⊗ 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 $wk_B T$ of work. As a result, the energy of the composite system—Demon, interacting bit, and weight—increases by $wk_B T$ 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 figure 2 indicates this by raising the energy level of $A \otimes 1$ by $wk_B T$ compared to its previous value. Since the transitions between $A \otimes 1$

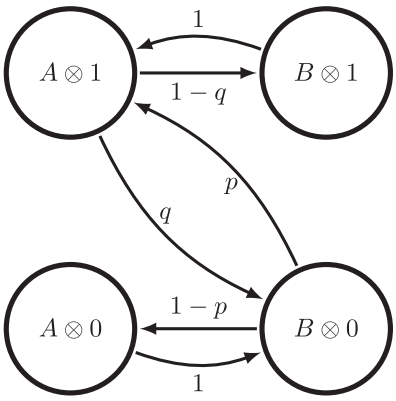

Figure 3. The Markovian, detailed-balance dynamic over the joint states of the ratchet and interacting bit.

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 wk_BT therefore implies the same increase in the energy of $B \otimes 1$. Again, see figure 2 for the energy scheme under 'Ratchet ⊗ Bit ⊗ 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 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 $\Pr(x_{N+1}, y'_N | x_N, y_N)$ 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:

$$\frac{M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N}}{M_{x_{N+1} \otimes y'_N \to x_N \otimes y_N}} = e^{(E_{x_{N+1} \otimes y'_N} - E_{x_N \otimes y_N})/k_B T}. \tag{8}$$

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 $0 \leqslant p \leqslant 1$ and $0 \leqslant q \leqslant 1$, as shown in figure 3.

The Markov transition matrix for this system is given by:

$$M = \begin{bmatrix} 0 & 1-p & 0 & 0 \\ 1 & 0 & q & 0 \\ 0 & p & 0 & 1 \\ 0 & 0 & 1-q & 0 \end{bmatrix}. \tag{9}$$

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:

$$\mathbf{p}((N-1)\tau + \tau^-) = M\mathbf{p}((N-1)\tau + 0^+), \tag{10}$$

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:

$$\alpha = -\ln(1-p), \tag{11}$$

$$\beta = -\frac{1}{2}\ln[(1-p)(1-q)], \text{ and} \tag{12}$$

$$w = \ln\left(\frac{q\sqrt{1-p}}{p\sqrt{1-q}}\right). \tag{13}$$

(Appendix B 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 [43, 47], 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:

$$\mathbf{p}^{\mathrm{s}}(0^+) = \frac{1}{2}\begin{bmatrix} b \\ b \\ 1-b \\ 1-b \end{bmatrix} \text{ and } \mathbf{p}^{\mathrm{s}}(\tau^-) = \frac{1}{2}\begin{bmatrix} b(1-p) \\ b+q-bq \\ bp+1-b \\ (1-b)(1-q) \end{bmatrix}. \tag{14}$$

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

To calculate net work extracted 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 figure 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:

$$\begin{aligned} \langle W_1 \rangle &= -\sum_{\substack{x \in \{A,B\} \\ y \in \{0,1\}}} p^{\mathrm{s}}_{x \otimes y}(0^+)(E_{x \otimes y} - E_x - E_y) \\ &= 0, \end{aligned} \tag{15}$$

where we used the probabilities in equation (14) and figure 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:

$$\langle W_3 \rangle = -\frac{k_{\mathrm{B}} T}{2}\beta[q + b(p-q)]. \tag{16}$$

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 figure 1. Since transitions $B \otimes 0 \to A \otimes 1$ raise the weight to give work $k_{\mathrm{B}} T w$ 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 $k_{\mathrm{B}} T w$ times the net probability transition along the former direction, which is $[T_{B \otimes 0 \to A \otimes 1} p^{\mathrm{s}}_{B \otimes 0}(0^+) - T_{A \otimes 1 \to A \otimes 1} p^{\mathrm{s}}_{A \otimes 1}(0^+)]$. This leads to the following expression:

$$\langle W_2 \rangle = \frac{k_{\mathrm{B}} T w}{2}[-q + b(p+q)], \tag{17}$$

where we used the probabilities in equation (14).

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

$$\begin{aligned} \langle W \rangle &= \langle W_1 \rangle + \langle W_2 \rangle + \langle W_3 \rangle \\ &= \frac{k_{\mathrm{B}} T}{2}[(pb - q + qb)\ln\left(\frac{q}{p}\right) + (1-b)q\ln(1-q) + pb\ln(1-p)]. \end{aligned} \tag{18}$$

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 section 5, the former is the thermodynamically more relevant quantity. A similar energetic scheme that incorporates the effects of interaction has also been discussed in [48].

In this way, we exactly calculated the work term in equation (5). We still need to calculate the entropy rate of the output and input strings to validate the proposed Second Law. For this, we introduce an information-theoretic formalism to monitor processing of the bit strings by the ratchet.

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

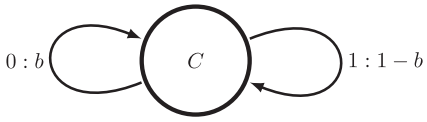

Figure 4. Biased coin input string as a unifilar hidden Markov model with bias $\Pr(Y = 0) = b$.

Markov model (HMM), since HMMs are strictly more powerful than Markov chains. 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 generated by any finite-state Markov chain. In short, HMMs give a compact representation for a wider range of memoryful processes.

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:

$$T^{(y_N)}_{s_N \to s_{N+1}} = \Pr(Y_N = y_N, S_{N+1} = s_{N+1} | S_N = s_N) \tag{19}$$

is the probability of outputting y_N for the Nth bit of the input string and transitioning to 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 [56, 57]. 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:

$$\Pr(Y_N = y_N | Y_{0:N} = w) = \Pr(Y_N = y_N | S_N = f(w)). \tag{20}$$

As a consequence, the entropy rate h_μ in its block-entropy form (equation (4)) can be re-expressed in terms of the transition matrices. First, recall the alternative, equivalent form for entropy rate: $h_\mu = \lim_{N\to\infty} H[Y_N|Y_{0:N}]$. Second, since S_N captures all the dependence of Y_N on the past, $h_\mu = \lim_{N\to\infty} H[Y_N|S_N]$. This finally leads to a closed-form expression for the entropy rate [55]:

$$\begin{aligned} h_\mu &= \lim_{N\to\infty} H[Y_N|S_N] \\ &= -\sum_{y_N, s_N, s_{N+1}} \pi_{s_N} T^{(y_N)}_{s_N \to s_{N+1}} \log_2 T^{(y_N)}_{s_N \to s_{N+1}}, \end{aligned} \tag{21}$$

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

Let us now put these observations to work. Here, we assume the ratchet's input string was generated by a memoryless biased coin. Figure 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 equation (21) and see that it is the single-symbol entropy of bias b:

$$\begin{aligned} h_\mu &= H(b) \\ &\equiv -b \log_2 b - (1 - b) \log_2 (1 - b), \end{aligned} \tag{22}$$

where $H(b)$ is the (base 2) binary entropy function [19].

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 is 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* [61]. In fact, this transducer is a communication channel in the sense of Shannon [62] that communicates the input bit sequence to the output bit sequence. Moreover, it is a channel with memory. Its internal states correspond to the ratchet's

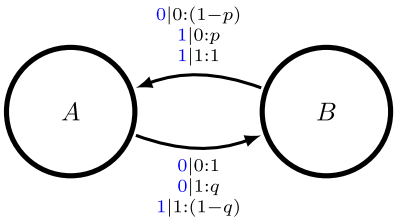

Figure 5. The Maxwellian ratchet's transducer.

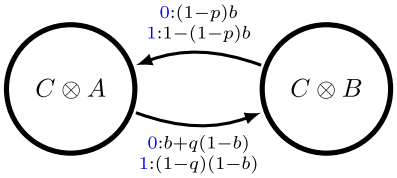

Figure 6. Unifilar HMM for the output string generated by the ratchet driven by a coin with bias b.

states. To work with M, we rewrite it componentwise as:

$$M^{(y'_N|y_N)}_{x_N \to x_{N+1}} = M_{x_N \otimes y_N \to x_{N+1} \otimes y'_N} \tag{23}$$

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:

$$M^{(y'_N|y_N)}_{x_N \to x_{N+1}} = \Pr(Y'_N = y'_N, X_{N+1} = x_{N+1} | Y_N = y_N, X_N = x_N). \tag{24}$$

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

$$T'^{(y'_N)}_{s_N \otimes x_N \to s_{N+1} \otimes x_{N+1}} = \sum_{y_N} M^{(y'_N|y_N)}_{x_N \to x_{N+1}} T^{(y_N)}_{s_N \to s_{N+1}}. \tag{25}$$

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 [43, 47]. 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 figure 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 5 shows our two-state ratchet's transducer. As noted above, its 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.

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 figure 4. The resulting unifilar HMM for the output string is shown in figure 6. The corresponding symbol-labeled transition matrices are:

$$T'^{(0)} = \begin{bmatrix} 0 & (1-p)b \\ b + q(1-b) & 0 \end{bmatrix}, \text{ and} \tag{26}$$

$$T'^{(1)} = \begin{bmatrix} 0 & 1-(1-p)b \\ (1-q)(1-b) & 0 \end{bmatrix}. \tag{27}$$

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

$$\begin{aligned} h'_\mu &= \lim_{N\to\infty} H[Y'_N|Y'_{0:N}] \\ &= \lim_{N\to\infty} H[Y'_N|S_N] \\ &= \frac{H(b(1-p))}{2} + \frac{H((1-b)(1-q))}{2}. \end{aligned} \tag{28}$$

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

$$\begin{aligned} h'_\mu &\leqslant H[Y'_0] \\ &= H((b(1-p) + (1-b)(1-q))/2). \end{aligned} \tag{29}$$

Any difference between h'_μ and single-symbol entropy $H[Y_0]$ 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.

5. Thermodynamic functionality

Let us 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:

$$\langle W\rangle \leqslant k_\mathrm{B}T \ln 2\, \Delta h_\mu, \tag{30}$$

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 [63] 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 and more incisive it is than earlier approximations.

We find diverse thermodynamic behaviors as shown in figure 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 $\langle W\rangle \leqslant k_\mathrm{B}T \ln 2\, \Delta h_\mu$. The engine regime is defined by (p, q) for which $k_\mathrm{B}T \ln 2\, \Delta h_\mu \geqslant \langle W\rangle > 0$ since work is positive. This is the only condition for which the ratchet extracts work. The eraser regime is defined by $0 > k_\mathrm{B}T \ln 2\, \Delta h_\mu \geqslant \langle W\rangle$, 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 $k_\mathrm{B}T \ln 2\, \Delta h_\mu \geqslant 0 \geqslant \langle W\rangle$, 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 [43], 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, figure 8 plots the single-symbol entropy difference $\Delta H[Y_0]$ 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 $\Delta H[Y_0]$ 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, $\Delta H[Y_0]$ 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 functional region, highlighted in orange in figures 7 and 8, 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 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 figure 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, too, 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,

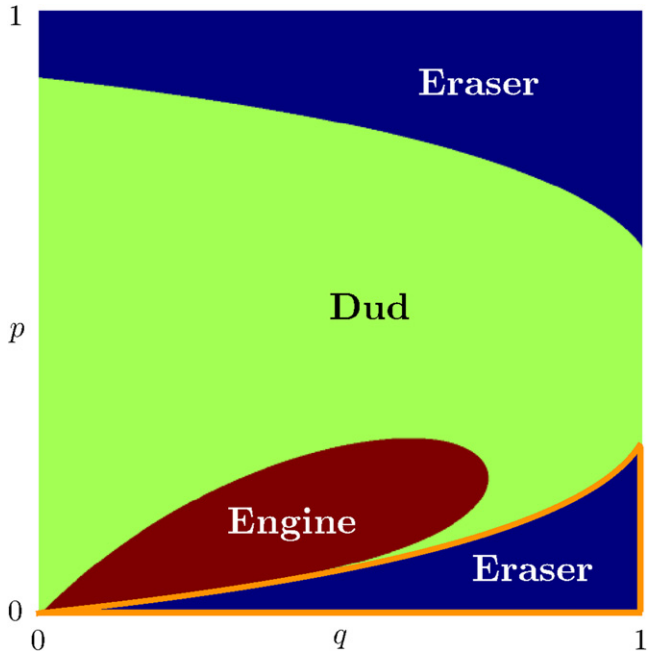

Figure 7. Information ratchet thermodynamic functionality at input bias $b = 0.9$: Engine: (p, q) such that $0 < \langle W \rangle \leqslant k_B T \ln 2 \, \Delta h_\mu$. Eraser: (p, q) such that $\langle W \rangle \leqslant k_B T \ln 2 \, \Delta h_\mu < 0$. Dud: (p, q) such that $\langle W \rangle \leqslant 0 \leqslant k_B T \ln 2 \, \Delta h_\mu$.

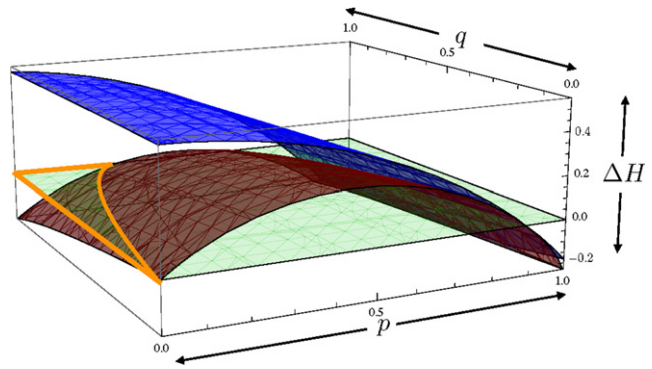

Figure 8. Exact entropy rate difference Δh_μ (red) is a much stricter bound on work than the difference in single-symbol entropy $\Delta H[Y_0]$ (blue). The zero surface (light green) highlights where both entropies are greater than zero and so is an aid to identifying functionalities.

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, elsewhere we will explore more intelligent ratchets that take advantage of input string correlations to do additional work.

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 contemporary advances in biomolecular chemistry and nanoscale physics and engineering. With the ability to manipulate thermal systems on energy scales closer and closer to the level of thermal fluctuations $k_B T$, 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. In particular, using HMMs greatly broadened the kind of information reservoir that can be considered; they are substantially more general than those used to date. Moreover,

expressing them as ϵ-machines leads to exact analytical treatment. Central to identifying thermodynamic functionality was our deriving equation (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 typically provides the tightest quantitative bound on work. This improvement comes directly from tracking Demon information production over system trajectories, not from time-local, configurational entropies.

Given the parallel relationship between the Kolmogorov–Chaitin complexity (size of the minimal universal Turing machine that generates a bit string) and the Kolmogorov–Sinai entropy [19], the entropy difference Δh_μ is seen to be a measure of the intrinsic computation performed by the ratchet in producing the output string from the input string [64]. Thus, the new Second Law has a complementary interpretation for thermodynamically embedded computations: it provides a lower bound on the energetic cost (work) required to drive the information processing performed by the ratchet as it generates the output string. As such, it is a substantial generalization of Landauer's approach [12] to the physical limits of computation.

Though its perspective and methods were not explicitly highlighted, *computational mechanics* [64] 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 explicated here and is left to sequels and sister works. Two complementary computational mechanics analyzes of information engines come to mind, in this light. The first is [16]'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 [60]. This suite of works constitutes the first successful synthesis of computational mechanics and thermodynamics.

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 [61, 64].

The theoretical innovations had a practical consequence. 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.

Acknowledgments

We thank M DeWeese and S Marzen for useful conversations. As an External Faculty member, JPC thanks the Santa Fe Institute for its hospitality during visits. This work was supported in part by the US Army Research Laboratory and the US Army Research Office under contracts W911NF-13-1-0390 and W911NF-12-1-0234.

Appendix A. Derivation of equation (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 non-negative 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:

$$\Delta S[X, Y, Z] \geqslant 0. \tag{A1}$$

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 $S[X, Y, Z] = k_B \ln 2\, H[X, Y, Z]$, we have the condition:

$$\Delta H[X, Y, Z] \geqslant 0. \tag{A2}$$

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:

$$\Delta H_N[X, Y, Z] = H[X_N, Y_{N:\infty}, Y'_{0:N}, Z_N] - H[X_0, Y_{0:\infty}, Z_0] \tag{A3}$$

$$= H[X_N, Y_{N:\infty}, Y'_{0:N}] + H[Z_N] - I[X_N, Y_{N:\infty}, Y'_{0:N}; Z_N] - H[X_0, Y_{0:\infty}] - H[Z_0] + I[X_0, Y_{0:\infty}; Z_0]. \tag{A4}$$

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 $I[X_N, Y_{N:\infty}, Y'_{0:N}; Z_N]$ and $I[X_0, Y_{0:\infty}; Z_0]$ 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:

$$\Delta H[Z] = H[Z_N] - H[Z_0] = Q_N / k_B T \ln 2. \tag{A5}$$

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:

$$\Delta H_N[X, Y, Z] = \frac{Q_N}{k_B T \ln 2} + H[X_N, Y_{N:\infty}, Y'_{0:N}] - H[X_0, Y_{0:\infty}]. \tag{A6}$$

Since $\Delta H \geqslant 0$, 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:

$$Q_N \geqslant k_B T \ln 2 (H[X_0, Y_{0:\infty}] - H[X_N, Y_{N:\infty}, Y'_{0:N}]). \tag{A7}$$

This bound is superficially similar to equation (6), but it is 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:

$$\langle Q \rangle \geqslant \lim_{N\to\infty} k_B T \ln 2 \frac{H[X_0, Y_{0:\infty}] - H[X_N, Y_{N:\infty}, Y'_{0:N}]}{N}. \tag{A8}$$

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 done is the opposite of the average dissipated heat: $\langle W \rangle = -\langle Q \rangle$. And so, it has the upper bound:

$$\langle W \rangle \leqslant k_B T \ln 2 \lim_{N\to\infty} \left(\frac{H[Y_{N:\infty}, Y'_{0:N}] - H[Y_{0:\infty}]}{N} + \frac{H[X_N] - H[X_0]}{N} + \frac{I[X_0; Y_{0:\infty}] - I[X_N; Y_{N:\infty}, Y'_{0:N}]}{N} \right), \tag{A9}$$

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 $C_\mu(N) = H[X_N]$ [64]. 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:

$$\langle W \rangle \leqslant k_B T \ln 2 \lim_{N\to\infty} \left(\frac{H[Y_{N:\infty}, Y'_{0:N}] - H[Y_{0:\infty}]}{N} \right). \tag{A10}$$

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. And, in a complementary way, we see that it is also a very general bound on the

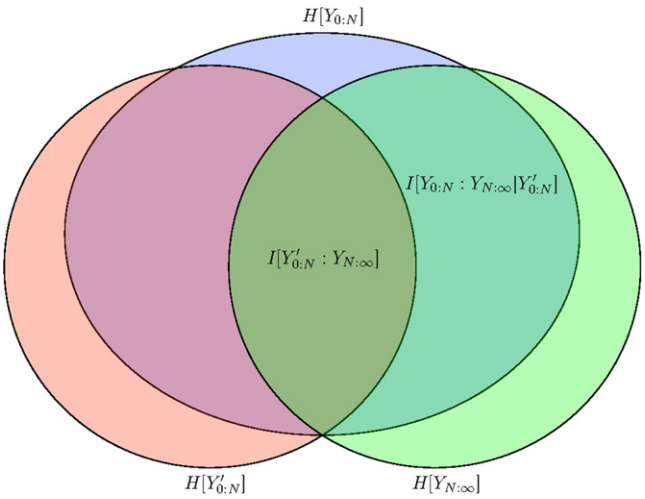

Figure 9. The N most recent variables of the input process shield the N variables of output from the rest of the input variables.

thermodynamic cost of intrinsic computing—the information processing performed by the ratchet in generating the output string from the input.

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

$$\langle W\rangle \geqslant k_{\mathrm{B}}T\ \ln 2 \lim_{N\to\infty}\left(\frac{H[Y_{N:\infty}] - H[Y_{0:\infty}]}{N} + \frac{H[Y'_{0:N}]}{N} - \frac{I[Y_{N:\infty};\ Y'_{0:N}]}{N}\right). \tag{A11}$$

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

$$\langle W\rangle \geqslant k_{\mathrm{B}}T\ \ln 2\left(h'_\mu - h_\mu - \lim_{N\to\infty}\frac{I[Y_{N:\infty};\ Y'_{0:N}]}{N}\right). \tag{A12}$$

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 $Y'_{0:N}$ from $Y_{N:\infty}$, as shown in information diagram [65] of figure 9. This means:

$$I[Y_{N:\infty};\ Y'_{0:N}] = I[Y_{N:\infty};\ Y_{0:N}] - I[Y_{N:\infty};\ Y_{0:N}|Y'_{0:N}]. \tag{A13}$$

This, in turn, gives: $I[Y_{N:\infty};\ Y_{0:N}] \geqslant I[Y_{N:\infty};\ Y'_{0:N}] \geqslant 0$. Thus, we find the input process's excess entropy **E** [55]:

$$\begin{aligned}\lim_{N\to\infty} I[Y_{N:\infty};\ Y'_{0:N}] &\leqslant \lim_{N\to\infty} I[Y_{N:\infty};\ Y_{0:N}]\\ &= \mathbf{E}.\end{aligned} \tag{A14}$$

However, dividing by N its contribution vanishes:

$$\begin{aligned}\lim_{N\to\infty}\frac{I[Y_{N:\infty};\ Y_{0:N}]}{N} &= \lim_{N\to\infty}\left(\frac{H[Y_{0:N}]}{N} - \frac{H[Y_{0:N}|Y_{N:\infty}]}{N}\right)\\ &= h_\mu - h_\mu\\ &= 0.\end{aligned} \tag{A15}$$

Thus, we are left with the inequality of equation (5):

$$\langle W\rangle \leqslant k_{\mathrm{B}}T\ \ln 2(h'_\mu - h_\mu); \tag{A16}$$

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 [66].

One key difference between equation (A16) (equivalently, equation (5)) and the more commonly used bound in equation (6), with the change in single-variable configurational entropy $H[Y'_0] - H[Y_0]$, 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. More generally, we do not look at single-step changes in configurational entropies—$H[X_{N-1},\ Y_{N-1},\ Z_{N-1}] \to H[X_N,\ Y_N,\ Z_N]$—but rather the rate of production of information

$H[W_N|\ldots W_{N-2}, W_{N-1}]$, where $W_N = (X_N, Y_N, Z_N)$. This global dynamical entropy rate has contributions from output rate h'_μ and input rate h_μ. This again indicates how equation (6) approximates equation (A16).

There are several special cases where the single-variable bound of equation (6) applies. In the case where the input is uncorrelated, it holds, but it is a weaker bound than equation (5) using entropy rates. Also, in the case when the ratchet has no internal states and so is memoryless, equation (6) is satisfied. Interestingly, either it or equation (A16) 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 [67]. 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 future works, but as a preview see [67].

Appendix B. Designing ratchet energetics

Figure 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 figure 3 first and, only then, introduced the associated energetic and thermodynamic quantities, as in figure 2. The introduction of energetic and thermodynamic quantities for an abstract transducer (as in figure 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 [43]. In our model, this implies $\beta = 0$, or equivalently, $p = q = 0$ (from equation (12)). In other words, we cannot describe our model, valid for all values $0 < p, q < 1$, by the energy scheme in figure 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 figure 3 if the following requirements of detailed balance are satisfied:

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

$$\frac{p}{q} = \frac{M_{B\otimes 0\to A\otimes 1}}{M_{A\otimes 1\to B\otimes 0}} = \mathrm{e}^{-\widetilde{W}_2}, \text{ and} \tag{B.2}$$

$$1-q = \frac{M_{A\otimes 1\to B\otimes 1}}{M_{B\otimes 1\to A\otimes 1}} = \mathrm{e}^{-\widetilde{W}_3}. \tag{B.3}$$

Existence of such an alternative scheme illustrates the fact that given the abstract model of figure 3, there is more than one possible consistent energy scheme. We suggest that this will allow for future engineering flexibility.

References

[1] Evans D J, Cohen E G D and Morriss G P 1993 Probability of second law violations in shearing steady flows *Phys. Rev. Lett.* **71** 2401–4
[2] Evans D J and Searles D J 1994 Equilibrium microstates which generate second law violating steady states *Phys. Rev.* E **50** 1645
[3] Gallavotti G and Cohen E G D 1995 Dynamical ensembles in nonequilibrium statistical mechanics *Phys. Rev. Lett.* **74** 2694–7
[4] Kurchan J 1998 Fluctuation theorem for stochastic dynamics *J. Phys. A: Math. Gen.* **31** 3719
[5] Crooks G E 1998 Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems *J. Stat. Phys.* **90** 1481–7
[6] Lebowitz J L and Spohn H 1999 A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics *J. Stat. Phys.* **95** 333
[7] Collin D, Ritort F, Jarzynski C, Smith S B, Tinoco I Jr and Bustamante C 2005 Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies *Nature* **437** 231
[8] Leff H and Rex A 2002 *Maxwell's Demon 2: Entropy, Classical and Quantum Information Computing* (New York: Taylor and Francis)
[9] Maruyama K, Nori F and Vedral V 2009 Colloquium: the physics of Maxwell's demon and information *Rev. Mod. Phys.* **81** 1
[10] Smoluchowski M v 1916 Drei vorträge über diffusion, etc *Phys. Z.* **17** 557–71
[11] Szilard L 1929 On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings *Z. Phys.* **53** 840–56
[12] Landauer R 1961 Irreversibility and heat generation in the computing process *IBM J. Res. Dev.* **5** 183–91
[13] Penrose O 1970 *Foundations of Statistical Mechanics; A Deductive Treatment* (Oxford: Pergamon)

[14] Bennett C H 1982 Thermodynamics of computation—a review *Int. J. Theor. Phys.* **21** 905
[15] Bérut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R and Lutz E 2012 Experimental verification of Landauer's principle linking information and thermodynamics *Nature* **483** 187
[16] Boyd A B and Crutchfield J P 2015 Demon dynamics: deterministic chaos, the Szilard map, and the intelligence of thermodynamic systems *SFI Working Paper* 15-06-019 (arxiv:1506.04327)
[17] Deffner S and Jarzynski C 2013 Information processing and the second law of thermodynamics: an inclusive, Hamiltonian approach *Phys. Rev.* X **3** 041003
[18] Barato A C and Seifert U 2014 Stochastic thermodynamics with information reservoirs *Phys. Rev.* E **90** 042150
[19] Cover T M and Thomas J A 2006 *Elements of Information Theory* 2nd edn (New York: Wiley)
[20] Touchette H and Lloyd S 2000 Information-theoretic limits of control *Phys. Rev. Lett.* **84** 1156
[21] Cao F J, Dinis L and Parrondo J M R 2004 Feedback control in a collective flashing ratchet *Phys. Rev Lett.* **93** 040603
[22] Sagawa T and Ueda M 2010 Generalized Jarzynski equality under nonequilibrium feedback control *Phys. Rev. Lett.* **104** 090602
[23] Toyabe S, Sagawa T, Ueda M, Muneyuki E and Sano M 2010 Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality *Nat. Phys.* **6** 988–92
[24] Ponmurugan M 2010 Generalized detailed fluctuation theorem under nonequilibrium feedback control *Phys. Rev.* E **82** 031129
[25] Horowitz J M and Vaikuntanathan S 2010 Nonequilibrium detailed fluctuation theorem for repeated discrete feedback *Phys. Rev.* E **82** 061120
[26] Horowitz J M and Parrondo J M R 2011 Thermodynamic reversibility in feedback processes *Europhys. Lett.* **95** 10005
[27] Granger L and Krantz H 2011 Thermodynamic cost of measurements *Phys. Rev.* E **84** 061110
[28] Abreu D and Seifert U 2011 Extracting work from a single heat bath through feedback *Europhys. Lett.* **94** 10001
[29] Vaikuntanathan S and Jarzynski C 2011 Modeling Maxwell's demon with a microcanonical Szilard engine *Phys. Rev.* E **83** 061120
[30] Abreu A and Seifert U 2012 Thermodynamics of genuine nonequilibrium states under feedback control *Phys. Rev. Lett.* **108** 030601
[31] Kundu A 2012 Nonequilibrium fluctuation theorem for systems under discrete and continuous feedback control *Phys. Rev.* E **86** 021107
[32] Sagawa T and Ueda M 2012 Fluctuation theorem with information exchange: role of correlations in stochastic thermodynamics *Phys. Rev. Lett.* **109** 180602
[33] Kish L B and Granqvist C G 2012 Energy requirement of control: comments on Szilard's engine and Maxwell's demon *Europhys. Lett.* **98** 68001
[34] Ito S and Sagawa T 2013 Information thermodynamics on causal networks *Phys. Rev. Lett.* **111** 180603
[35] Hartich D, Barato A C A C and Seifert U 2014 Stochastic thermodynamics of bipartite systems: transfer entropy inequalities and a Maxwell's demon interpretation *J. Stat. Mech.* P02016
[36] Horowitz J M and Esposito M 2014 Thermodynamics with continuous information flow *Phys. Rev.* X **4** 031015
[37] Horowitz J M 2015 Multipartite information flow for multiple Maxwell demons *J. Stat. Mech.* P03006
[38] Esposito M and Schaller G 2012 Stochastic thermodynamics for 'Maxwell demon' feedbacks *Europhys. Lett.* **99** 30003
[39] Strasberg P, Schaller G, Brandes T and Esposito M 2013 Thermodynamics of a physical model implementing a Maxwell demon *Phys. Rev. Lett.* **110** 040601
[40] Horowitz J M, Sagawa T and Parrondo J M R 2013 Imitating chemical motors with optimal information motors *Phys. Rev. Lett.* **111** 010602
[41] Barato A C and Seifert U 2014 Unifying three perspectives on information processing in stochastic thermodynamics *Phys. Rev. Lett.* **112** 090601
[42] Horowitz J M and Sandberg H 2014 Second-law-like inequalities with information and their interpretations *New J. Phys.* **16** 125007
[43] Mandal D and Jarzynski C 2012 Work and information processing in a solvable model of Maxwell's demon *Proc. Natl Acad. Sci. USA* **109** 11641–5
[44] Mandal D, Quan H T and Jarzynski C 2013 Maxwell's refrigerator: an exactly solvable model *Phys. Rev. Lett.* **111** 030602
[45] Barato A C and Seifert U 2013 An autonomous and reversible Maxwell's demon *Europhys. Lett.* **101** 60001
[46] Hoppenau J and Engel A 2014 On the energetics of information exchange *Europhys. Lett.* **105** 50002
[47] Lu Z, Mandal D and Jarzynski C 2014 Engineering Maxwell's demon *Phys. Today* **67** 60–1
[48] Um J, Hinrichsen H, Kwon C and Park H 2015 Total cost of operating an information engine *New J. Phys.* **17** 085001
[49] Dorfman J R 1999 *An Introduction to Chaos in Nonequilibrium Statistical Mechanics* (Cambridge: Cambridge University Press)
[50] Alberts B, Johnson A, Lewis J, Morgan D and Raff M 2014 *Molecular Biology of the Cell* 6th edn (New York: Garland Science)
[51] Brillouin L 1949 Life, thermodynamics, and cybernetics *Am. Sci.* **37** 554–68
[52] Lewis H R and Papadimitriou C H 1998 *Elements of the Theory of Computation* 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)
[53] Strasberg P, Cerrillo J, Schaller G and Brandes T 2015 Thermodynamics of stochastic Turing machines *Phys. Rev.* E **92** 042104
[54] Moore C 1990 Unpredictability and undecidability in dynamical systems *Phys. Rev. Lett.* **64** 2354
[55] Crutchfield J P and Feldman D P 2003 Regularities unseen, randomness observed: levels of entropy convergence *Chaos* **13** 25–54
[56] Blackwell D 1956 The entropy of functions of finite-state Markov chains *Transactions of the First Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes* (*Liblice near Prague, 28–30 November 1956*) vol 28 (Prague, Czechoslovakia: House of the Czechoslovak Academy of Sciences) pp 13–20
[57] *Entropy of Hidden Markov Process and Connections to Dynamical Systems* Marcus B, Petersen K and Weissman T (ed) 2011 (*London Mathematical Society Lecture Notes Series* vol 385) (Cambridge: Cambridge University Press)
[58] Merhav N 2015 Sequence complexity and work extraction *J. Stat. Mech.* P06037
[59] Chapman A and Miyake A 2015 How can an autonomous quantum Maxwell demon harness correlated information? *Phys. Rev.* E **92** 062125
[60] Riechers P M and Crutchfield J P 2015 Thermodynamics of adaptation: broken reversibility, fluctuations, and path entropies when controlling structured out-of-equilibrium systems (in preparation)
[61] Barnett N and Crutchfield J P 2015 Computational mechanics of input–output processes: structured transformations and the ϵ-transducer *J. Stat. Phys.* **161** 404–51
[62] Shannon C E 1948 A mathematical theory of communication *Bell Sys. Tech. J.* **27** 623–56
[63] Jarzynski C 1997 Nonequilibrium equality for free energy differences *Phys. Rev. Lett.* **78** 2690–3
[64] Crutchfield J P 2012 Between order and chaos *Nat. Phys.* **8** 17–24
[65] Yeung R W 2008 *Information Theory and Network Coding* (New York: Springer)
[66] Crutchfield J P, Ellison C J, Mahoney J R and James R G 2010 Synchronization and control in intrinsic and designed computation: an information-theoretic analysis of competing models of stochastic computation *Chaos* **20** 037105
[67] Mandal D, Boyd A B and Crutchfield J P 2015 Memoryless thermodynamics? A reply *SFI Working Paper* 15-08-031 (arxiv:1508.03311)