Minimal Work Required for Arbitrary Computation

David Wolpert

SFI WORKING PAPER: 2015-08-032

SANTA FE INSTITUTE

Minimal work required for arbitrary computation

David H. Wolpert
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM, 87501
http://davidwolpert.weebly.com
(Dated: August 31, 2015)

Recent studies have analyzed the minimal thermodynamic work required for a given logical map to be implemented on any physical system. These studies have focused on maps whose output does not depend on the input, e.g., bit erasure in a digital computer. In addition, they have considered physical systems whose design varies depending on the distribution of inputs to the map. However very often we are interested in implementing a map whose output depends on its input. In addition, we often want our system to implement the same map even if the system's environment changes, so that the distribution over map inputs changes. Here I introduce a thermodynamic engine that satisfies both of these desiderata. I then calculate how much work it requires, deriving an additive correction to the "generalized Landauer bound" of previous studies. I also calculate the Bayes-optimal engine for any given distribution over environments. I end with a short discussion on how these results relate the free energy flux incident on an organism / robot / biosphere to the maximal amount of (noisy) computation that the organism / robot / biosphere can do per unit time.

Much has been learned in the past century about the connection between measurement, thermodynamics and computation [2, 6, 11, 17, 18, 27–29, 36, 37, 44, 45, 47, 49–51]. A breakthrough was made with the seminal analyses of Landauer and his colleagues who concluded that at least $kT \ln[2]$ of work needs to be used by any system that implements a 2-to-1 map [1–4, 12, 18, 24–26, 31, 33, 40, 43]. This implies that whenever a bit is erased in a computer heat must be dumped into the computer's environment, a conclusion that is starting to be confirmed experimentally [5, 13, 22, 23, 38]. A related conclusion was that a 1-to-2 map can act as a *refrigerator* rather than a heater, *removing* heat from the environment [2–4, 24]. For example, this occurs during adiabatic demagnetization of an Ising spin system [24].

This seminal early work leaves many issues unresolved however. To give a simple example, say we have a map over the set $\{0, 1, 2, 3\}$ that sends both 0 and 1 to 0, as in bit erasure, but also sends the value 2 to either 0 or 3, with probabilities .8 and .2, respectively. So it is a heater for inputs 0 or 1, and a refrigerator for input 2. How does the total required work depend on what bin 3 gets mapped to?

More recently, there has been dramatic progress in our understanding of non-equilibrium statistical physics and its relation to information-processing [7, 9, 10, 12, 14–16, 19, 21, 32, 34, 35, 40–42, 46, 48]. Much of this recent literature has analyzed the minimal work required to drive a physical system's (fine-grained, microstate) dynamics during the interval from $t = 0$ to $t = 1$ in such a way that the conditional distribution of the coarse-grained state of the system at $t = 1$, given the coarse-grained state at $t = 0$, is some desired π. In particular, there has been detailed analysis of the minimal work needed when there are only two coarse-grained bins, $v = 0$ and $v = 1$, and we require that both get mapped to the bin $v = 0$ during the interval $[0, 1)$ [15, 32, 39]. By identifying the bins v as Information Bearing Degrees of Freedom (IBDF [4]) of an information-processing device like a digital computer, these analyses can be seen as elaborations of the analyses of Landauer et al. on the thermodynamics of bit erasure.

Many of the work-minimizing systems considered in this recent literature proceed in two stages. First, they transform an initial, non-equilibrium distribution over microstates w to the equilibrium distribution, $\rho^{eq}(w)$. All information concerning the initial microstate is lost from the distribution over w by the end of this first stage. So in particular all information is lost about what the initial bin v_0 was. Then in the second stage $\rho^{eq}(w)$ is transformed to an ending (non-equilibrium) distribution over w, with an associated distribution over the ending coarse-grained bin, v_1. However since all information about v_0 has been lost by the beginning of the second stage, v_0 cannot have any effect on the distribution over v_1 produced in the second stage. So although such a system can be used to implement a many-to-one map over the IBDF (i.e., the bins) in a digital computer, it cannot be used to implement any computational map whose output varies with its input.

Another important aspect of these work-minimizing systems is that their design depends not only on the map π that they are implementing, but also on $\mathcal{P}_0(v)$, the precise distribution of values that are fed into π at $t = 0$. In particular, suppose we are considering a system like a digital computer implementing a map π over its IBDF (i.e., "running a step of a program" with the program's input given by v_0). The distribution over the initial IBDF of the computer is set by the human user of the computer. The analyses in the recent literature implicitly assume that if a new user of the system comes along, initializing the computer's $t = 0$ state according to a new $\mathcal{P}_0(v)$, then the system's dynamics is somehow modified to account for that new user's distribution.

In practice though, we are often interested in systems whose dynamics does *not* change if there is a change in the initial distribution over coarse-grained states. For example, we often use the same digital computer, even if the user of that computer changes. Similarly, biological organisms can be viewed as systems that perform a "computation" starting with an "input" that is set by their environment (e.g., via a sensor the organism uses, or more intrusively, via a direct perturbation of the state of the organism). Often on short-enough timescales

the dynamics of those systems does not change even if the distribution over "inputs" to the organism changes.

In this paper I show how to construct a system that is guaranteed to implement any given conditional distribution π, for all input distributions, and even if π maps different v_0 to different final states v_1. Like the systems in the literature, the system considered here is optimized for some "guessed" distribution over the inputs, $\mathscr{G}_0(v)$. However I allow the system to be used with an input distribution $\mathcal{P}_0$ that differs from $\mathscr{G}_0$, and calculate the expected work such a system requires to implement π, as a function of $\pi, \mathscr{G}_0$ and $\mathcal{P}_0$. I then concentrate on the case where there is a prior distribution over $\mathcal{P}_0$ (e.g., as occurs if the system will be used with multiple users) and $\mathscr{G}_0$ is the Bayes-optimal distribution for that prior. I end by using these results to relate the free energy flux incident on an organism (robot, biosphere) to the maximal "rate of computation" that the organism (resp., robot, biosphere) can achieve.

Problem setup.— I write $|X|$ for the number of elements x in any finite space X, and write the Shannon entropy of a distribution p over X as $S_p(X) = S(p) = -\sum_x p(x) \ln[p(x)]$, or even just $S(X)$ when p is implicit. I also write the Kullbach-Leibler (KL) divergence between two distributions p and q both defined over X as $D(p(X) \parallel q(X)) \equiv \sum_x p(x) \ln\left[p(x)/q(x)\right]$ or sometimes just $D(p \parallel q)$ for short [8, 30].

Let W be the (finite) space of all possible microstates of a system. Let $\mathcal{V}$ be a partition of W, i.e., a coarse-graining. For example, $\mathcal{V}$ could be the set of all possible states of the memory of a digital computer (i.e., the computer's IBDF). I will write $\mathcal{V}(w) \in V$ to refer to the partition element that contains w, and assume that one of the elements of V is labelled "0".

I assume that the evolution of the elements of $\mathcal{V}$ during $t \in [0, 1)$ results in a conditional distribution $\pi(v_1 \mid v_0)$. In addition, as in the analyses of computers in [2–4, 24], there is a "user" of the system who intervenes in its dynamics at or before $t = 0$, which results in v_0 being set by sampling a **user distribution** $\mathcal{P}_0(v_0)$. As examples, $\mathcal{P}_0$ could model randomness in how a single user of a computer initializes the computer at $t = 0$, or randomness in how an environment of an organism perturbs the organism at $t = 0$.

I write the Hamiltonian over W at $t = 0$ as H^0_{sys}, with associated equilibrium (Boltzmann) distribution ρ^{eq}. I assume that for any $v \in V$, at both $t = 0$ and $t = 1$, the conditional distribution of elements in W, given that w is in partition element v, is some fixed distribution $q^v_{in}(w)$. Since they are set by the designer of the system, I take π and the conditional distributions q^v_{in} to be fixed and known to that designer. However I allow the designer to be uncertain what $\mathcal{P}_0$ is. I write the (potentially non-equilibrium) distribution over W at $t = 0$ as

$$\mathcal{P}_0(w) \equiv \sum_{v_0} \mathcal{P}_0(v_0) q^{v_0}_{in}(w) \tag{1}$$

As shorthand, I write

$$\mathcal{P}_1(v_1) \equiv \sum_{v_0} \mathcal{P}_0(v_0) \pi(v_1 \mid v_0) \tag{2}$$

To reduce notation, I assume that for all v, $\sum_w q^v_{in}(w) H^0_{sys}(w)$ is the same constant, h^0_{sys}, and that this is also the expected value of $H^0_{sys}(w)$ under ρ^{eq}. (In particular, this would be true if $H^0_{sys}(w)$ were a constant.)

Overview of the system.— Following [15, 20, 32, 39, 40, 44], I consider a system that implements π in two stages. Broadly speaking, at the start of the first stage, for each possible value v_0 the system sends $\mathcal{P}_0(w \mid v_0) = q^{v_0}_{in}(w)$ to ρ^{eq}. It does this using the quench-then-quasi-statically relax procedure (QTR) described in [15, 32]. In the current context, that means it runs two successive processes. First we replace H^0_{sys} with a **quenching Hamiltonian**

$$H^{v_0}_{in}(w) \equiv -kT \ln[q^{v_0}_{in}(w)] \tag{3}$$

chosen such that $q^{v_0}_{in}$ is an equilibrium distribution for that Hamiltonian. (While there is no change to w in this quench, in general work is required, if $H^{v_0}_{in} \neq H^0_{sys}$.) Next we isothermally and quasi-statically relax $H^{v_0}_{in}$ back to H^0_{sys}, so that the distribution over W becomes ρ^{eq}.

Again speaking in broad terms, in the second stage the system sends $\rho^{eq}_{H^0_{sys}}$ to the non-equilibrium distribution

$$\mathcal{P}_1(w \mid v_0) \equiv \sum_{v_1} \pi(v_1 \mid v_0) q^{v_1}_{in}(w) \tag{4}$$

thereby sampling $\pi(v_1 \mid v_0)$. The system does this by running another QTR, just "in reverse", using a quenching Hamiltonian

$$H^{v_0}_{out}(w) \equiv -kT \ln[q^{v_0}_{out}(w)] \tag{5}$$

where

$$q^{v_0}_{out}(w) \equiv \sum_{v_1} q^{v_1}_{in}(w) \pi(v_1 \mid v_0) \tag{6}$$

Details of the system.— Neither of the two quenching Hamiltonians depends on $\mathcal{P}$. This is necessary to ensure that the system faithfully executes π no matter what $\mathcal{P}$ actually is. However to compensate for this $\mathcal{P}$-independence, both quenching Hamiltonians depend on v_0. Therefore we need to extend the system, to include a way to "measure" v_0.

This is accomplished with an error-free **measurement apparatus**, patterned after the one introduced in [32, 39, 40]. As in those studies, the "measurement" is a process that correlates the value of v with the coarse-grained state of an external, symmetric memory, without changing v (or even w). I write the (coarse-grained) states of that memory as $m \in V$, with associated microstates $u \in U$. For simplicity I suppose that the conditional distribution of u given any m at $t = 0$, $Q^m(u)$, is also the conditional distribution of u given that m at $t = 1$.

For simplicity, I assume that the memory and system are both always in contact with a heat bath at temperature T. In addition, I make the inductive hypothesis that the starting value of the memory is $m_0 = 0$ with probability 1. The system performs the following four steps (see Fig. 1):

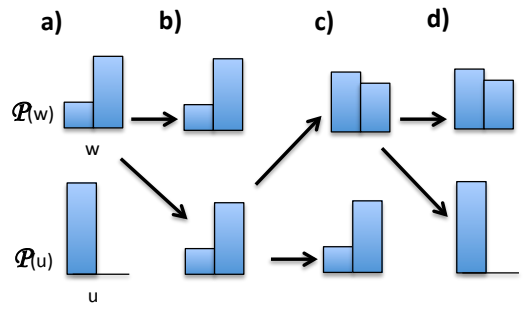

FIG. 1. Example of dynamics of the marginal distributions of a system with a binary coarse-graining, where the bins have the same size and both $q^v_{in}(w)$ and $Q^m(u)$ are uniform for all v, m. Fig. (a) shows the $t = 0$ state, with the right bin of the system more probable than the left bin, and the memory in its initialized bin. Fig. (d) shows the $t = 1$ state, where the relative probabilities of the system bins have changed according to π, and the memory has been returned to its initialized bin. The transition (a)→(b) illustrates step 1 of the dynamics, requiring no work. (b)→(c) illustrates the combination of steps 2 and 3, requiring expected work $-ktS_{\mathcal{P}}(\mathcal{V}_1 \mid \mathcal{V}_0)$. (c)→(d) illustrates step 4, requiring expected work $-kT \sum_{v_0,v_1} \mathcal{P}(v_0, v_1) \ln \Big[\mathcal{G}(v_0 \mid v_1)\Big]$. Arrows indicate dependencies in the microstate dynamics.

1 — In the first step the current value of v_0 is measured and its value is stored in the memory by setting $m_0 = v_0$. This step is done without any change to w_0, so in particular $\mathcal{P}_0(w)$ is unchanged. As described in [32], since the measurement is error-free and the memory is symmetric, this step can be done with no thermodynamic work.

2 — Next the QTR described in [15, 32] is run with the quenching Hamiltonian $H^{m_0}_{in}$, where the beginning (and ending) Hamiltonian is H^0_{sys}. This is done without changing m.

3 — Next we use the fact that $m_0 = v_0$ to run a QTR in reverse, using the quenching Hamiltonian $H^{v_0}_{out}$. This procedure begins by isothermally and quasi-statically sending H^0_{sys} to $H^{v_0}_{out}$. After that $H^{v_0}_{out}$ is replaced by H^0_{sys}, with no change to w. As in step (2), there is no change to m in step (3).

4 — Finally, we need to reset m to 0, so that the system can be run again, thereby satisfying the inductive hypothesis.[1]

Note that since the system samples $\pi(v_1 \mid v_0)$ in step 3, with v_0 chosen by randomly sampling $\mathcal{P}_0(v_0)$, the set of four steps implement π even if that map's output depends on its input, and no matter what the user distribution actually is. So it meets our two desiderata. Moreover, as shown in the Supplementary Material (SM), the expected work expended in the first three steps is

$$-kT\Big(S_{\mathcal{P}}(\mathcal{V}_1 \mid \mathcal{V}_0) + \sum_v S(q^v_{in}(W))\Big[\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big]\Big) \quad (7)$$

Resetting the memory.—To calculate the work done in the fourth step, suppose it runs a QTR in one stage followed by a reverse QTR in a second stage, only now applied to the distribution over U, not W. In addition, the value v_1 will be exploited to set the quenching Hamiltonian for the first stage. (Intuitively, v_1 is a "noisy measurement" of m_0.) It is here that the (possibly erroneous) assumption for the user distribution will arise.

To describe these two stages in detail, suppose that we guess that the distribution over the initial states of the coarse-grained variable is $\mathcal{G}_0(v_0)$. Since by the time of step (4) we know that $m_0 = v_0$, this distribution serves as a prior probability over the values of m_0. The associated likelihood of v_1 given m_0 is $\pi(v_1 \mid m_0)$. So the posterior probability of m_0 given v_1, $\mathcal{G}(m_0 \mid v_1)$, is proportional to $\mathcal{G}_0(m_0)\pi(v_1 \mid m_0)$. This then fixes the (guessed) posterior probability over memory microstates,

$$\mathcal{G}(u_0 \mid v_1) = \sum_{m_0} \mathcal{G}(m_0 \mid v_1) Q^{m_0}(u_0) \quad (8)$$

In contrast, the actual posterior distribution $\mathcal{P}(m_0 \mid v_1)$ is given by the actual prior $\mathcal{P}_0$, and is proportional to $\mathcal{P}_0(m_0)\pi(v_1 \mid m_0)$. This fixes a posterior distribution

$$\mathcal{P}(u_0 \mid v_1) = \sum_{m_0} \mathcal{P}(m_0 \mid v_1) Q^{m_0}(u_0) \quad (9)$$

The central supposition of this paper is that when resetting the memory, the system uses the quenching Hamiltonian that would result in minimal work if in fact $\mathcal{G}_0$ equalled $\mathcal{P}_0$, so that $\mathcal{G}(u_0 \mid v_1)$ were the actual posterior:

$$H^{v_1}_{mem}(u_0) \equiv -kT \ln \mathcal{G}(u_0 \mid v_1) \quad (10)$$

(In the special case that $\mathcal{G}_0 = \mathcal{P}_0$, running the QTR with this quenching Hamiltonian would result in minimal total expended work.) This QTR stage run using $H^{v_1}_{mem}(u_0)$ does not change w_1, just as measurement of v_0 did not change w_0.[2]

To complete the resetting of the memory we now run the second, reverse QTR stage, that takes u from the uniform distribution over all U to the distribution that equals 0 with probability 1. As shown in the SM, averaging the amount of work generated in these two stages, and adding it to the expression in Eq. (7), gives the following expression for the minimal expected work expended when running π:

$$\Omega_{\mathcal{G}_0,\mathcal{P}_0} \equiv -kT\Big(\sum_{v_0,v_1} \mathcal{P}(v_0, v_1) \ln\big[\mathcal{G}(v_0 \mid v_1)\big] + S_{\mathcal{P}}(\mathcal{V}_1 \mid \mathcal{V}_0) + \sum_v S(q^v_{in})\Big[\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big]\Big) \quad (11)$$

[1] This resetting of the memory is analogous to the resetting of the mind of the demon in modern analyses of Maxwell's demon [4].

[2] Note that if $\mathcal{G}_0 \neq \mathcal{P}_0$, then the actual posterior $\mathcal{P}(u_0 \mid v_1)$ is not the equilibrium distribution for $H^{v_1}_{mem}$. In such a situation immediately after the quenching process, as the Hamiltonian over U begins to quasi-statically relax, the distribution over U will first settle, in a thermodynamically irreversible process, to the equilibrium distribution for $H^{v_1}_{mem}$.

We can rewrite this as

$$kT\Big(\sum_{v_1}\mathcal{P}_1(v_1)D\Big[\mathcal{P}(\mathcal{V}_0 \mid v_1) \parallel \mathcal{G}(\mathcal{V}_0 \mid v_1)\Big] + S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1) + \sum_v S(q^v_{in})\Big[\mathcal{P}_0(v) - \mathcal{P}_1(v)\Big]\Big) \quad (12)$$

If $\mathcal{G}_0 = \mathcal{P}_0$, then this reduces to

$$kT\Big(S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1) + \sum_v S(q^v_{in})\Big[\mathcal{P}_0(v) - \mathcal{P}_1(v)\Big]\Big) \quad (13)$$

which is sometimes called "generalized Landauer cost" in the literature. So due to the $S(q^v_{in})$ terms, even when $\mathcal{G}_0 = \mathcal{P}_0$ the work required to implement π varies from one system to another. On the other hand, Eq. (22) in the SM shows that Eq. (13) equals $kT[S_{\mathcal{P}_0}(W) - S_{\mathcal{P}_1}(W)]$, the change in entropy over W, regardless of the choice of coarse-grained variable. So when $\mathcal{G}_0 = \mathcal{P}_0$, if $\mathcal{P}(w_0, w_1)$ is not changed but a change is made to $\mathcal{V}$ (which induces changes in $\mathcal{P}_0$, $\mathcal{G}_0$, π, and $\{q^v_{in} : v \in V\}$), then the expected work does not change. Note as well that when $\mathcal{G}_0 = \mathcal{P}_0$ the mutual information between v_1 and v_0 has no effect on $\Omega_{\mathcal{G}_0,\mathcal{P}_0}$. Even if the output of π *does* depends on its input, that dependence ends up being irrelevent; only the marginals over v_1 and over v_0 matter.

How we use π. — $\Omega_{\mathcal{G}_0,\mathcal{P}_0}$ is the expected work if we start with v_0 already set to a sample of $\mathcal{P}_0$ and then run π once. Suppose instead that we run π a total of N times starting from that v_0, e.g., as when we run a computer program for a total of N steps. If $S(q^v_{in})$ is independent of v (e.g., as when $|V| = |W|$), then the total expected work is the sum of N chained instances of Eq. (12). This equals $S(\mathcal{V}_0) - S(\mathcal{V}_N)$ plus a non-negative term, given by adding N expected KL divergences. So if the starting and ending distributions over $\mathcal{V}$ are the same, total work cannot be negative, even though the work in any one of the N iterations may be. In this sense, the second law is embodied in Eq. (12).

Refer to a scenario in which we sample $\mathcal{P}_0$ once and then iterate π one or more times as a **hermit** computer. Sometimes we will instead want to use a **calculator** computer, in which we sample $\mathcal{P}_0$ at the end of each iteration, before running π again. In calculators, after step (4) the value v_1 is copied to an external system via a measurement apparatus (e.g., in order to determine some physical action). Then a different external system (e.g., a sensor) samples $\mathcal{P}_0(v'_0)$, and v_1 gets replaced by v'_0. After these two new steps we have completed a single iteration. At this point we can run another iteration, to apply π again — but to the value v'_0 rather than v_1.

Since no work is required in the new step where we measure v_1, the total work in an iteration is given by adding Eq. (12) to the additional average work required to map $v = v_1$ to $v = v'_0$. Since both the values v_1 and v'_0 exist outside of W, they can be used to specify the two quenching Hamiltonians that implement this map. So the additional average work is

$$kT\sum_{v,v'}\Big[S(q^v_{in})\mathcal{P}_1(v) - S(q^{v'}_{in})\mathcal{P}_0(v')\Big] \quad (14)$$

and the total expected work of the iteration is

$$kT\Big(\sum_{v_1}\mathcal{P}_1(v_1)D\Big[\mathcal{P}(\mathcal{V}_0 \mid v_1) \parallel \mathcal{G}(\mathcal{V}_0 \mid v_1)\Big] + S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1)\Big) \quad (15)$$

So the expected work of a calculator has no dependence on the values $S(q^v_{in})$; in contrast to hermits, in calculators the work to implement π is independent of the system we use. On the other hand, if instead of holding π fixed we hold $P(w_0, w_1)$ fixed, then the work *does* depend on $\mathcal{V}$, even if $\mathcal{G}_0 = \mathcal{P}_0$ (again, in contrast to the case with hermit computers).

Optimal $\mathcal{G}_0$ — There are several ways one can set $\mathcal{G}_0$ if we do not know $\mathcal{P}_0$ (as would be the case for example if we were not certain who will use a digital computer we are designing). As an example, suppose we are going to sample $\mathcal{P}_0$ once and then run π once. Then one way to set $\mathcal{G}_0$ would be a minimax approach:

$$\mathcal{G}_0 = \text{argmin}_{\mathcal{G}'_0}\text{max}_{\mathcal{P}_0}[\Omega_{\mathcal{G}'_0,\mathcal{P}_0}] \quad (16)$$

As an alternative, suppose we are given a prior distribution over users, $Pr(\mathcal{P}_0)$, and a loss function $L(\mathcal{G}_0, \mathcal{P}_0) = \Omega_{\mathcal{G}_0,\mathcal{P}_0}$. Define

$$\langle\mathcal{P}_0\rangle \equiv \int d\mathcal{P}_0 Pr(\mathcal{P}_0)\mathcal{P}_0 \quad (17)$$

By Eq. (11), the average under $Pr(\mathcal{P}_0)$ of $\Omega_{\mathcal{G}_0,\mathcal{P}_0}$ equals $\Omega_{\mathcal{G}_0,\langle\mathcal{P}_0\rangle}$. Applying this equality to Eq. (12), and using the fact that KL divergence is minimized (at zero) when its arguments are equal, we see that the Bayes-optimal $\mathcal{G}_0$ for our loss function is $\langle\mathcal{P}_0\rangle$, and the associated expected work is

$$\Omega^* \equiv S_{\langle\mathcal{P}\rangle}(\mathcal{V}_0) - S_{\langle\mathcal{P}\rangle}(\mathcal{V}_1) \quad (18)$$

Note that in general Ω^* exceeds

$$\langle S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1)\rangle \equiv \int d\mathcal{P}_0 Pr(\mathcal{P}_0)\,[S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1)] \quad (19)$$

The expected work would be $\langle S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1)\rangle$ instead of Ω^* only if we could somehow re-optimize $\mathcal{G}_0$ for each $\mathcal{P}_0$.

Finally, note that no matter what $\mathcal{G}_0$ we choose, in general there is nonzero variance over the users of how much expected work is required to implement π for each of them. So choosing $\mathcal{G}_0 = \langle\mathcal{P}_0\rangle$ results in expected work that is greater than Ω^* for some users but less than Ω^* for others. Sometimes we care more about the former than the latter, e.g., if we have a maximal amount of expected work we can use to drive the computation. In such a case we should use a loss function $L(\mathcal{G}_0, \mathcal{P}_0)$ that increases more than linearly in $\Omega_{\mathcal{G}_0,\mathcal{P}_0}$, which results in a Bayes-optimal $\mathcal{G}_0$ different from $\langle\mathcal{P}_0\rangle$.

Implications for biology — By conservation of energy, any work expended on the system must first be acquired as free energy taken from the system's environment. However in many situations there is a limit on the flux of free energy through

a system's immediate environment. Combined with the analysis above, such limits provide upper bounds on the "rate of (potentially noisy) computation" that can be achieved by a biological organism in that environment.

As an example, these results upper bound the rate of computation a human brain can achieve. Given the fitness cost of such computation (the human brain uses $\sim 20\%$ of the calories used by the human body), this bound could contribute significantly to the natural selective pressures on humans, in the limit that operational inefficiencies of the brain have already been minimized. In other words, these bounds suggest that natural selection imposes a tradeoff between the fitness quality of a brain's decisions, and how much computation is required to make those decisions.

As a second example, the rate of free energy incident upon the earth as sunlight provides an upper bound on the rate of computation that can be achieved by the biosphere. In particular it provides an upper bound on the rate of computation that can be achieved by human civilization, if we remain on the surface of the earth, and ultimately only uses sunlight to power our computation.

Acknowledgements – I would like to thank Josh Grochow, Cris Moore, Daniel Polani, Eric Libby, and especially Sankaran Ramakrishnan for many stimulating discussions, and the Santa Fe Institute for helping to support this research. This paper was made possible through the support of Grant No. TWCF0079/AB47 from the Templeton World Charity Foundation and Grant No. FQXi-RHl3-1349 from the FQXi foundation. The opinions expressed in this paper are those of the author and do not necessarily reflect the view of Templeton World Charity Foundation.

[1] Charles H Bennett, *Logical reversibility of computation*, IBM journal of Research and Development **17** (1973), no. 6, 525–532.

[2] ———, *The thermodynamics of computation—a review*, International Journal of Theoretical Physics **21** (1982), no. 12, 905–940.

[3] ———, *Time/space trade-offs for reversible computation*, SIAM Journal on Computing **18** (1989), no. 4, 766–776.

[4] ———, *Notes on landauer's principle, reversible computation, and maxwell's demon*, Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics **34** (2003), no. 3, 501–510.

[5] Antoine Bérut, Artak Arakelyan, Artyom Petrosyan, Sergio Ciliberto, Raoul Dillenschneider, and Eric Lutz, *Experimental verification of landauer/'s principle linking information and thermodynamics*, Nature **483** (2012), no. 7388, 187–189.

[6] L. Brillouin, *Science and information theory*, Academic Press, 1962.

[7] Farid Chejne Janna, Fadl Moukalled, and Carlos Andrés Gómez, *A simple derivation of crooks relation*, International Journal of Thermodynamics **16** (2013), no. 3, 97–101.

[8] T. Cover and J. Thomas, *Elements of information theory*, Wiley-Interscience, New York, 1991.

[9] Gavin E Crooks, *Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems*, Journal of Statistical Physics **90** (1998), no. 5-6, 1481–1487.

[10] ———, *Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences*, Physical Review E **60** (1999), no. 3, 2721.

[11] Lidia Del Rio, Johan Åberg, Renato Renner, Oscar Dahlsten, and Vlatko Vedral, *The thermodynamic meaning of negative entropy*, Nature **474** (2011), no. 7349, 61–63.

[12] Raoul Dillenschneider and Eric Lutz, *Comment on"minimal energy cost for thermodynamic information processing: Measurement and information erasure"*, Physical review letters **104** (2010), no. 19, 198903.

[13] Jörn Dunkel, *Thermodynamics: Engines and demons*, Nature Physics **10** (2014), no. 6, 409–410.

[14] Massimiliano Esposito and Christian Van den Broeck, *Three faces of the second law. i. master equation formulation*, Physical Review E **82** (2010), no. 1, 011143.

[15] ———, *Second law and landauer principle far from equilibrium*, EPL (Europhysics Letters) **95** (2011), no. 4, 40004.

[16] Philippe Faist, Frédéric Dupuis, Jonathan Oppenheim, and Renato Renner, *A quantitative landauer's principle*, arXiv preprint arXiv:1211.1037 (2012).

[17] Edward Fredkin, *An informational process based on reversible universal cellular automata*, Physica D: Nonlinear Phenomena **45** (1990), no. 1, 254–270.

[18] Edward Fredkin and Tommaso Toffoli, *Conservative logic*, Springer, 2002.

[19] H-H Hasegawa, J Ishikawa, K Takara, and DJ Driebe, *Generalization of the second law for a nonequilibrium initial state*, Physics Letters A **374** (2010), no. 8, 1001–1004.

[20] Jordan M Horowitz and Juan MR Parrondo, *Designing optimal discrete-feedback thermodynamic engines*, New Journal of Physics **13** (2011), no. 12, 123019.

[21] Christopher Jarzynski, *Nonequilibrium equality for free energy differences*, Physical Review Letters **78** (1997), no. 14, 2690.

[22] Yonggun Jun, Momčilo Gavrilov, and John Bechhoefer, *High-precision test of landauer's principle in a feedback trap*, Physical review letters **113** (2014), no. 19, 190601.

[23] JV Koski, VF Maisi, JP Pekola, and DV Averin, *Experimental realization of a szilard engine with a single electron*, arXiv preprint arXiv:1402.5907 (2014).

[24] Rolf Landauer, *Irreversibility and heat generation in the computing process*, IBM journal of research and development **5** (1961), no. 3, 183–191.

[25] ———, *Minimal energy requirements in communication*, Science **272** (1996), no. 5270, 1914–1918.

[26] ———, *The physical nature of information*, Physics letters A **217** (1996), no. 4, 188–193.

[27] Harvey S Leff and Andrew F Rex, *Maxwell's demon: entropy, information, computing*, Princeton University Press, 2014.

[28] Seth Lloyd, *Use of mutual information to decrease entropy: Implications for the second law of thermodynamics*, Physical Review A **39** (1989), no. 10, 5378.

[29] ———, *Ultimate physical limits to computation*, Nature **406** (2000), no. 6799, 1047–1054.

[30] D.J.C. Mackay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[31] O.J.E. Maroney, *Generalizing landauer's principle*, Physical Review E **79** (2009), no. 3, 031105.

[32] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa, *Thermodynamics of information*, Nature Physics **11** (2015), no. 2, 131–139.

[33] Martin B Plenio and Vincenzo Vitelli, *The physics of forgetting:*

Landauer's erasure principle and information theory, Contemporary Physics **42** (2001), no. 1, 25–60.

[34] Blake S Pollard, *A second law for open markov processes*, arXiv preprint arXiv:1410.6531 (2014).

[35] Mikhail Prokopenko and Itai Einav, *Information thermodynamics of near-equilibrium computation*, Physical Review E **91** (2015), no. 6, 062143.

[36] Mikhail Prokopenko and Joseph T Lizier, *Transfer entropy and transient limits of computation*, Nature Scientific reports **4** (2014).

[37] Mikhail Prokopenko, Joseph T Lizier, and Don C Price, *On thermodynamic interpretation of transfer entropy*, Entropy **15** (2013), no. 2, 524–543.

[38] É Roldán, Ignacio A Martinez, Juan MR Parrondo, and Dmitri Petrov, *Universal features in the energetics of symmetry breaking*, Nature Physics (2014).

[39] Takahiro Sagawa, *Thermodynamic and logical reversibilities revisited*, Journal of Statistical Mechanics: Theory and Experiment **2014** (2014), no. 3, P03025.

[40] Takahiro Sagawa and Masahito Ueda, *Minimal energy cost for thermodynamic information processing: measurement and information erasure*, Physical review letters **102** (2009), no. 25, 250602.

[41] ———, *Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics*, Physical review letters **109** (2012), no. 18, 180602.

[42] Udo Seifert, *Stochastic thermodynamics, fluctuation theorems and molecular machines*, Reports on Progress in Physics **75** (2012), no. 12, 126001.

[43] Kousuke Shizume, *Heat generation required by information erasure*, Physical Review E **52** (1995), no. 4, 3495.

[44] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks, *Thermodynamics of prediction*, Physical review letters **109** (2012), no. 12, 120604.

[45] Leo Szilard, *On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings*, Behavioral Science **9** (1964), no. 4, 301–310.

[46] K Takara, H-H Hasegawa, and DJ Driebe, *Generalization of the second law for a transition between nonequilibrium states*, Physics Letters A **375** (2010), no. 2, 88–92.

[47] Tommaso Toffoli and Norman H Margolus, *Invertible cellular automata: A review*, Physica D: Nonlinear Phenomena **45** (1990), no. 1, 229–253.

[48] Hugo Touchette and Seth Lloyd, *Information-theoretic approach to the study of control systems*, Physica A: Statistical Mechanics and its Applications **331** (2004), no. 1, 140–172.

[49] Karoline Wiesner, Mile Gu, Elisabeth Rieper, and Vlatko Vedral, *Information-theoretic lower bound on energy cost of stochastic computation*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science **468** (2012), no. 2148, 4058–4066.

[50] W. H. Zurek, *Algorithmic randomness and physical entropy*, Phys. Rev. A **40** (1989), 4731–4751.

[51] Wojciech H Zurek, *Thermodynamic cost of computation, algorithmic complexity and the information metric*, Nature **341** (1989), no. 6238, 119–124.

SUPPLEMENTARY MATERIAL

Derivation of Eq. (7)

In this section I evaluate the expected work required to implement the first three steps of the system for initial distribution $\mathcal{P}_0$. To do this, it will be convenient to calculate the expected work to perform those steps conditioned on a particular v_0, and then average over all v_0 according to $\mathcal{P}_0(v_0)$.

As in [32], I assume that after step 1 the interaction Hamiltonian between W and U is negligible. Also as in that work, I assume that the quench step at the beginning of step 2 is an instantaneous change to the energy of every w, $\Delta E(w)$. This process does not actually w (such changes are associated with transfer of heat). Since the quenching Hamiltonian depends on the value of m_0 (which due to step 1 depends on v_0), the value of $\Delta E(w)$ for each w also depends on m_0. That change in the energy of w is identified as the work done on the system in the quench step when it starts (and stays) in that state w.

Now due to the fact that step 1 did not change w, at the beginning of the quench step the posterior probability of w given a current value m_0 is $q_{in}^{m_0}(w)$. Therefore the expected work done in this quench step conditioned on a particular value m_0 is $\sum_w q_{in}^{m_0}(w)[H_{in}^{m_0}(w) - H_{sys}^0(w)]$. As shorthand, define S_{sys}^0 as the Shannon entropy over W for the Boltzmann distribution with temperature T and Hamiltonian H_{sys}^0. Then conditioned on a value v_0 at the beginning of step 1, the work to perform the entire QTR in step 2 is

$$\sum_w q_{in}^{m_0}(w)[H_{in}^{m_0}(w) - H_{sys}^0(w)] + \mathcal{F}(H_{sys}^0) - \mathcal{F}(H_{in}^{m_0}) \quad (20)$$

where $m_0 = v_0$ and

$$\mathcal{F}(H_{sys}^0) = h_{sys}^0 - kTS_{sys}^0 \quad (21)$$

is the equilibrium free energy of H_{sys}^0 at temperature T.

By definition of $H_{in}^{m_0}$, $\mathcal{F}(H_{in}^{m_0}) = 0$. So the expression in Eq. (20) just equals $kT\left[S(q_{in}^{m_0}) - S_{sys}^0\right]$. Note that this amount of work is negative, since work is extracted by sending $q_{in}^{m_0}$ to the equilibrium distribution for H_{sys}^0.

Similarly, to implement step 3 requires work of at least $kT\left[S_{sys}^0 - S(q_{out}^{m_0})\right]$.[3] Now for any distribution $Pr(w)$, with some abuse of notation, we can write $S_{Pr}(v \mid w) = 0$, since w sets v uniquely. Therefore

$$\begin{aligned} S_{Pr}(w) &= S_{Pr}(v \mid w) + S_{Pr}(w) \\ &= S_{Pr}(v, w) \\ &= S_{Pr}(v) + S_{Pr}(w \mid v) \end{aligned} \quad (22)$$

[3] In steps 2 and 3 the usual convention was followed by quasi-statically sending $H_{in}^{m_0}$ to H_{sys}^0 and then sending H_{sys}^0 to $H_{out}^{v_0}$. The same total work would arise if we instead quasi-statically send $H_{in}^{m_0}$ to $H_{out}^{v_0}$ directly.

So if we write the Shannon entropy of the distribution over values v_1 conditioned on a particular value of v_0 as

$$S_\pi(\mathcal{V}_1 \mid v_0) \equiv -\sum_{v_1} \pi(v_1 \mid v_0) \ln[\pi(v_1 \mid v_0)] \tag{23}$$

then we can write

$$\begin{aligned} S(q_{out}^{v_0}) &= S_\pi(\mathcal{V}_1 \mid v_0) - \sum_{w_1, v_1} \pi(v_1 \mid v_0) q_{in}^{v_1}(w_1) \ln(q_{in}^{v_1}(w_1)) \\ &= S_\pi(\mathcal{V}_1 \mid v_0) + \sum_{v_1} \pi(v_1 \mid v_0) S(q_{in}^{v_1}) \end{aligned} \tag{24}$$

Accordingly, the total amount of work in the first three steps, conditioned on a value v_0, is

$$\begin{aligned} & kT\Big[S(q_{in}^{v_0}) - S(q_{out}^{v_0})\Big] \\ &\quad = kT\Big[S(q_{in}^{v_0}) - S_\pi(\mathcal{V}_1 \mid v_0) - \sum_{v_1} \pi(v_1 \mid v_0) S(q_{in}^{v_1})\Big] \end{aligned} \tag{25}$$

Combining and averaging under $\mathcal{P}_0(v_0)$, the expected work required to complete the first three steps is

$$-kT\Big[S_\pi(\mathcal{V}_1 \mid \mathcal{V}_0) + \sum_{v} S(q_{in}^{v}) \Big(\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big)\Big] \tag{26}$$

(The analogous expression in much of the literature has $S_\pi(\mathcal{V}_1)$ instead of $S_\pi(\mathcal{V}_1 \mid \mathcal{V}_0)$; the difference is due to the requirement that π govern the coarse-grained dynamics even if its output depends on its input, a requirement that means that we must measure the value v_0.)

The expected work to reset the measurement apparatus memory

The QTR in resetting the memory is run at $t = 1$, using $H_{mem}^{v_1}(u_0)$. It does not change w_1, just as measurement of v_0 did not change w_0.[4] Accordingly, the minimal amount of work in this QTR is

$$\begin{aligned} & \sum_{u_0} \mathcal{P}(u_0 \mid v_1)[H_{mem}^{v_1}(u_0) - H_{mem}^{0}(u_0)] + \mathcal{F}(H_{mem}^{0}) \\ &\quad = kT\Big(-\sum_{u_0} \mathcal{P}(u_0 \mid v_1) \ln\Big[\mathcal{G}(u_0 \mid v_1)\Big] - \ln|V|\Big) \end{aligned} \tag{27}$$

To complete the resetting of the memory we now run a reverse QTR that takes u from the uniform distribution over all U to the distribution that equals 0 with probability 1. That ending distribution has entropy of $S(Q^0)$ over U. Using the fact that $m_0 = v_0$, this means that for the given value of v_1, the total work required to reset m to 0 is

$$-kT\Big(\sum_{u_0} \mathcal{P}(u_0 \mid v_1) \ln\Big[\mathcal{G}_0(u_0 \mid v_1)\Big] \; - \; S(Q^0)\Big) \tag{28}$$

Multiply and divide the argument of the logarithm in the summand by $\mathcal{P}(u_0 \mid v_1)$. Next use the same kind of decomposition as in Eq. (22), and then use the chain-rule for KL divergence. This transforms our expression into

$$-kT\Big(\sum_{v_0} \mathcal{P}(v_0 \mid v_1) \ln\Big[\mathcal{G}_0(v_0 \mid v_1)\Big] - \sum_{v_0} \mathcal{P}(v_0 \mid v_1) S(Q^{v_0}) + S(Q^0)\Big) \tag{29}$$

Averaging this according to $\mathcal{P}_1(v_1)$ gives

$$-kT\Big(\sum_{v_0, v_1} \mathcal{P}(v_0, v_1) \ln\Big[\mathcal{G}_0(v_0 \mid v_1)\Big] - \sum_{v_0} \mathcal{P}(v_0) S(Q^{v_0}) + S(Q^0)\Big) \tag{30}$$

Note though that we assumed that the states of the memory are symmetric. (This is why there is no expected work in step 1.) So $S(Q^v)$ is independent of v, and Eq. (30) reduces to

$$-kT \sum_{v_0, v_1} \mathcal{P}(v_0, v_1) \ln\Big[\mathcal{G}_0(v_0 \mid v_1)\Big] \tag{31}$$

Adding Eq. (31) to Eq. (7) then gives Eq. (11), as claimed.

[4] Note though that if $\mathcal{G}_0 \neq \mathcal{P}_0$, then the actual posterior $\mathcal{P}(u_0 \mid v_1)$ is not the equilibrium distribution for $H_{mem}^{v_1}$. This means that immediately after the quenching process, as the Hamiltonian over U begins to quasi-statically relax, the distribution over U will first settle, in a thermodynamically irreversible process, to the equilibrium distribution for $H_{mem}^{v_1}$.