Dependence of dissipation on the initial distribution over states

Artemy Kolchinsky and David H. Wolpert
Santa Fe Institute, Santa Fe, NM

We consider the amount of dissipated work occurring in a given nonequilibrium process as a function of the initial distribution over microstates. We show that the extra dissipated work for any distribution, above the minimum achievable, has a simple information-theoretic form that depends only on the initial and final microstate distributions, and reflects the logical irreversibility of the process. We then consider the case where the process induces dynamics over a coarse-grained space that implement some specified computation. We show that our analysis of dissipated work still applies, though now stated in terms of distributions over the coarse-grained states. This extra dissipated work is a novel thermodynamic cost of computation, in addition to the well-known Landauer’s cost.

Background.—The past few decades have seen great advances in nonequilibrium statistical physics [1–16], resulting in many novel predictions and experiments [17–19]. This research has provided powerful new tools to analyze thermodynamic behavior in far from equilibrium systems.

Much of this research concerns the following general scenario. A physical system with microstates \( x \in \mathbb{X} \) is connected to a single external heat bath at temperature \( T \), and its Hamiltonian \( H_t \) evolves over time \( t \in [0,1) \). As the Hamiltonian evolves, the system’s distribution \( p_t(x) \) evolves according to a sequence of Hamiltonian-dependent transition-rate matrices due to interactions with the heat bath. These transition-rate matrices generate the conditional probability \( p(x_{0.1}|x_0) \) of observing trajectory \( x_{0.1} := \{x_0, \ldots, x_1\} \) given initial state \( x_0 \). We refer to the sequence of Hamiltonians \( H_t \) and the conditional probability distribution \( p(x_{0.1}|x_0) \) as a (thermodynamic) process operating on the system, indicated generically as \( \mathcal{P} \). Note that \( \mathcal{P} \) can run on different initial microstate distributions \( p_0(x) \), resulting in different distributions over trajectories \( x_{0.1} \). The non-equilibrium free energy of a system with Hamiltonian \( H_t \) and distribution \( p_t(x) \) is defined as

\[
\mathcal{F}(H_t, p_t) := \langle H_t \rangle_{p_t} - S(p_t)
\]

where \( S(\cdot) \) indicates Shannon entropy (in nats) and we choose units so that \( kT = 1 \) [11]. Note that the non-equilibrium free energy equals the conventional (equilibrium) Helmholtz free energy if \( p_t \) is a Boltzmann distribution for \( H_t \).

Suppose the system starts with some particular (possibly non-equilibrium) initial distribution \( p_0 \), and that \( \mathcal{P} \) transforms \( p_0 \) to \( p_1 \) while changing the Hamiltonian from \( H_0 \) to \( H_1 \). Write the work done on the system from \( t = 0 \) to \( t = 1 \) if it follows trajectory \( x_{0.1} \) as \( W(x_{0.1}) \). Then, expected work is lower bounded by

\[
\langle W(X_{0.1}) \rangle \geq \mathcal{F}(H_1, p_1) - \mathcal{F}(H_0, p_0) \tag{1}
\]

where the expectation is over \( p_0(x_0)p(x_{0.1}|x_0) \) [9, 11, 20, 21]. Eq. (1) reflects the modern understanding of the second law [11].

The portion of expected work that cannot be thermodynamically recovered from the heat bath after the process finishes is called dissipated work, or dissipation (not to be confused with the dissipated heat, which is the total energy transferred to the heat bath). Written as a function of the initial distribution \( p_0 \), the dissipation equals

\[
W_d(p_0) := \langle W(X_{0.1}) \rangle - \mathcal{F}(H_1, p_1) + \mathcal{F}(H_0, p_0) \tag{2}
\]

(See [11, 20].) Expanding,

\[
W_d(p_0) := S(p_1) - S(p_0) - \langle Q(X_0) \rangle_{p_0} \tag{3}
\]

where \( Q(x_0) \) is the expected total heat transferred from the bath to the system if it starts in \( x_0 \) [6]:

\[
Q(x_0) := \sum_{x_{0.1}} p(x_{0.1}|x_0) (H_1(x_1) - W(x_{0.1})) - \langle H_0(x_0) \rangle
\]

Thus, dissipated work is equal to the entropy change that does not correspond to heat exchanged with the heat bath. Unlike the minimal required work to run a process, dissipation is always non-negative, and is zero iff the process is thermodynamically reversible.

It is always possible to design a process \( \mathcal{P} \) that will transform any given pair \((p_0, H_0)\) to any pair \((p_1, H_1)\) without any dissipation, e.g., using an appropriately designed “quench-and-relax” process [11, 22–24]. However, if a quench-and-relax process designed to produce no dissipation for initial distribution \( p_0 \) is instead run with a different initial distribution, then in general it will dissipate work. This means that the work dissipated by such a process depends on the initial distribution [23, 24].

In this paper we analyze the generality of this phenomenon. Much of the research on the second law concerns how the dissipation that arises by running \( \mathcal{P} \) is affected by changes to \( \mathcal{P} \) without any changes to the initial microstate distribution [25, 26]. In contrast, in our results below we leave \( \mathcal{P} \) fixed, and analyze how the amount of dissipation varies with the initial distribution.

To start, we derive an expression for how the amount of dissipation incurred when running any fixed process \( \mathcal{P} \) varies with the initial distribution \( p_0 \). Specifically, let \( q_0 \) be the initial distribution which minimizes dissipation.
We show that if $\mathcal{P}$ is run with any initial distribution $r_0$, then the associated dissipation $W_d(r_0)$ equals $W_d(q_0)$ plus an extra cost. That extra cost is the drop in the value of the Kullback-Leibler (KL) divergence [27, 28] between $r(x)$ and $q(x)$ as the system evolves from $t = 0$ to $t = 1$, i.e., the KL divergence between $r_0$ and $q_0$ minus the KL divergence between $r_1$ and $q_1$.

Whereas much of the analysis of how changing $\mathcal{P}$ affects dissipation relies on the second law (e.g., Eq. (1)), this result describing how changes to the initial distribution affects dissipation does not. Furthermore, in contrast to most results that invoke the second law, all of our results are equalities, not inequalities. Finally, our results hold for any type of process (e.g., quasi-static, non-quasi-static, far-from-equilibrium processes, etc.), and they do not require that the dynamics of $\mathcal{P}$ obey local detailed balance [5, 6, 8, 29]. In fact, our results do not even require that the system be connected to a heat bath at the endpoints, $t = 0$ and $t = 1$. As a special case, our results apply to a maximally efficient Carnot engine; they specify the extra dissipated work that such an engine would produce if one changed its initial distribution. In all this, our results can be seen as a “new kind of second law”, reflecting entropic losses due to changes to the initial distribution, rather than due to changes to the subsequent thermodynamic process.

After deriving our first result, we use it to analyze the thermodynamics of computation, i.e., the thermodynamics of coarse-grained systems. Recent advances in nonequilibrium statistical physics [11, 30] have extended the pioneering analysis of of Landauer, Bennett and others [31–34]. In keeping with these analyses, we define a computer as a physical system with microstates $x \in X$ undergoing a thermodynamic process $\mathcal{P}$, together with a coarse-graining of $X$ into a set of Computational Macrostates (CMs) with labels $v \in V$ (sometimes called “information bearing degrees of freedom” [35]). $\mathcal{P}$ induces a stochastic dynamics over $X$, and the (possibly noisy) computation is identified with the associated dynamics over CMs. Typically it is assumed that CM dynamics are first-order Markovian. We write this dynamical process over CMs, implementing a single iteration of the computation that maps inputs $v_0$ to outputs $v_1$, as $\pi(v_1 | v_0)$. The canonical example is a single iteration of a laptop, modifying the bit pattern in its memory (i.e., its CM) [35].

Just as with quench-and-relax processes, it has been found that the amount of dissipation when a computer runs changes if the distribution over its initial CM is changed, in all models of computers where the issue has been analyzed [23, 24]. Here we analyze the fully general case, calculating how the dissipation of an arbitrary computer varies with changes in the initial distribution over its CMs. Specifically, we consider computers that are “thermodynamically cyclic”, in a special sense defined below. We show that for such computers, the exact same equations that determine how dissipation varies with the initial distribution over microstates also determine how dissipation in a computer varies with the initial distribution over computational macrostates.

As an example of what this means, suppose user Alice wishes to repeatedly run some single-iteration computation $\pi(v_1 | v_0)$, e.g., perform a calculation on a calculator. Suppose that Alice generates inputs by drawing IID samples from distribution $p_{0_{\text{Alice}}}$. Accordingly, she uses a specially-constructed computer that implements $\pi$ for all $v_0$, and which also achieves minimal dissipation when its inputs are truly sampled from $p_{0_{\text{Alice}}}$. (For example, she may have used the construction in [23, 24] to produce a computer that implements $\pi$ with zero dissipation when initialized with $p_{0_{\text{Alice}}}$. In general, however, that same computer would dissipate more work if it were instead run with its inputs generated by user Bob by sampling from some distribution $p_{0_{\text{Bob}}}$. (See [23] for a discussion of the implications of this dependence for how the evolutionary fitness of a biological cell changes when its environment changes.)

Interestingly, our results show that extra dissipation arising from changing the input distribution depends on the logical irreversibility of the process. As we show below, the extra dissipation is 0 if the computation is a single-valued invertible map (i.e., it is logically-reversible), and it is maximal if the process is input-independent (i.e., maximally logically-irreversible). Note though that this result has no direct implications for whether total dissipation and/or total work is small with a computation that is logically reversible [36].

Miscellaneous proofs used in the derivations of our results are found the Supplementary Material (SM) [37].

Notation.—We typically use $p$, $q$, and $r$ to refer to probability distributions. Upper case letters (e.g., $Z$) indicate either a random variable, or the set of possible outcomes of a random variable, depending on the context; lower case letters (e.g., $z$) indicate either a specific element of an associated set or a generic element (e.g., when we write $p(z)$ to indicate the distribution of the random variable $Z$).

As above, we indicate a (finite) space of microstates as $X$, and an associated space of CMs as $V$. The random variable $X_t$ indicates the microstate at time $t$. $p_{0_{\text{V}}}$ and $p_{1_{\text{V}}}$ are used to indicate microstate distribution $p(x_0)$ and $p(x_1)$ respectively, and similarly for $q_0$, $q_1$, $r_0$ and $r_1$. Subscript notation like $X_{0,1}$ indicates a random variable corresponding to a trajectory of states rather than a single-instant state: $X_{0,1} := \{X_0, \ldots, X_1\}$. We use $\Delta$ to indicate the $|X|$-dimensional unit simplex.

The random variable $V_t$ indicates the computational macrostate (CM) of the computer at time $t$. Distribution over CM are subscripted as $p_{V}$. We use $p_{0_{V}}$ and $p_{1_{V}}$. ...
to indicate $p_V(\nu_0)$ and $p_V(\nu_1)$ respectively, and similarly for $q^0_V, q^1_V, r^0_V$ and $r^1_V$. We use $\Delta^V$ to indicate the $|V|$-dimensional unit simplex.

The Kullback-Leibler divergence between any distributions $q$ and $p$ is $D(q//p) := -\sum_x q(x) \ln[p(x)/q(x)]$ and the cross-entropy between them is $C(q//p) := -\sum_x q(x) \ln[p(x)]$ [27, 28].

Throughout the analysis the units of time are arbitrary. Accordingly the equations we derive hold exactly no matter how long the process takes, and in particular even in the quasi-static limit.

**Dissipation due to incorrect priors.**—Fix a process $\mathcal{P}$. Define $q_0$ to be an initial state distribution that achieves minimum dissipation across $\Delta$, the $|X|$-dimensional unit simplex. We call this the **prior** distribution for $\mathcal{P}$. We make two assumptions [38]:

1. $W_d(p_0)$ is finite for all $p_0$ and that it has finite derivatives in the interior of the simplex;
2. $q_0$ has full support.

While $q_0$ is the initial distribution that results in minimal dissipation, in general initial states will be drawn from some environment distribution $r_0$, which need not equal $q_0$. To calculate the associated dissipation $W_d(r_0)$, first use Eq. (3) to write the $|X|$ components of $\nabla W_d(p_0)$ as

$$\frac{\partial W_d}{\partial p(x)}(p_0) = -Q(x) + \ln p(x) - \sum_{x_1} p(x_1|x_0) \ln p(x_1)$$

(These are understood to be one-sided derivatives if $p_0$ is on the edge of $\Delta$.) By Eq. (4), even though $W_d(p_0)$ is not a linear function of $p_0$, it is still true that for any $p_0$

$$W_d(p_0) = p_0 \cdot \nabla W_d(p_0) = \sum_{x_0} p(x_0) \frac{\partial W_d}{\partial p(x)}(p_0)$$

This gives our first result:

**Proposition 1.** Let $S$ be a convex set of distributions over $X$ and define $q_0 := \arg\min_{p_0 \in S} W_d(p_0)$. If $q_0$ is in the relative interior of $S$, then for all $r_0 \in S$

$$W_d(r_0) - W_d(q_0) = D(r_0//q_0) - D(r_1//q_1)$$

**Proof.** Because $q_0$ is in the relative interior of $S$ and minimizes $W_d$, for any $r_0 \in S$

$$\langle r_0 - q_0 \rangle \cdot \nabla W_d(q_0) = 0$$

So by Eq. (5),

$$r_0 \cdot \nabla W_d(q_0) = W_d(q_0)$$

Plugging in Eq. (4),

$$r_0 \cdot \nabla W_d(q_0) = -\langle Q \rangle r_0 - C(r_0//q_0) + C(r_1//q_1) = W_d(r_0) - [D(r_0//q_0) - D(r_1//q_1)]$$

Combining establishes the claim. □

In particular, if $S = \Delta$, then due to Assumption (2) the condition in Prop. 1 is met. So for any $r_0 \in \Delta$, $W_d(r_0) = W_d(q_0) + D(r_0//q_0) - D(r_1//q_1)$. From now on we leave specification of $S$ implicit when the meaning is clear.

We refer to $W_d(r_0) - W_d(q_0)$ as “incorrect prior dissipation". Recall that the KL divergence $D(r//q)$ is an information-theoretic measure of the distinguishability of distributions $r$ and $q$. So incorrect prior dissipation measures the decrease in our ability to distinguish whether the initial distribution was $q_0$ or $r_0$ as the system evolves from $t = 0$ to $t = 1$. Formally, this drop reflects the “contraction of KL divergence" under the action of the map $p(x_1|x_0)$ [39, 40].

By the KL data processing inequality [41, Lemma 3.11], incorrect prior dissipation is non-negative. It achieves its minimum value of 0 if the dynamics $p(x_1|x_0)$ form an invertible map, regardless of the prior $q_0$. At the other extreme, if $p(x_1|x_0)$ is an input-independent map, $D(r_1//q_1) = 0$ and incorrect prior dissipation reaches it’s maximum value. More generally, in the SM [37] we establish the following:

**Proposition 2.** There exists $r_0$ with $W_d(r_0) - W_d(q_0) > 0$ iff $p(x_1|X_0)$ is not an invertible map.

For another perspective on Prop. 1, note that by the chain rule for KL divergence [42, Eq. 2.67],

$$D(r(X_0, X_1)//q(X_0, X_1)) = D(r_0//q_0) + D(r(X_1|X_0)//q(X_1|X_0))$$

$$= D(r_1//q_1) + D(r(X_0|X_1)//q(X_0|X_1))$$

However since $r(x_1|x_0) = q(x_1|x_0) = p(x_1|x_0)$,

$$D(r(X_1|X_0)//q(X_1|X_0)) = 0.$$ So Prop. 1 is equivalent to

$$W_d(r_0) - W_d(q_0) = D(r(X_0|X_1)//q(X_0|X_1))$$

(See also [24].) In this expression $r(x_0|x_1)$ and $q(x_0|x_1)$ are Bayesian posterior probabilities of the initial state conditioned on the final state, for the assumed priors $r_0$ and $q_0$ respectively, and the shared likelihood function $p(x_1|x_0)$. (This Bayesian formulation of Prop. 1 is why we refer to the initial distribution $q_0$ as a “prior".)

Recall that a thermodynamic process $\mathcal{P}$ is specified by a continuous-time sequence of Hamiltonians and the conditional distribution over state trajectories. The set of such processes is infinite-dimensional. However Eq. (3) shows that the associated dissipation function $W_d(\cdot)$ can be parameterized by only $|X|^2$ real numbers: the $|X|$ values of $Q(x_0)$ plus the $|X|(|X| - 1)$ values of $p(x_1|x_0)$. No features of the thermodynamic process beyond those numbers are necessary to calculate the dissipation for an arbitrary $p_0$. Similarly, Prop. 1 shows that the incorrect prior dissipation from using $r_0$ rather than $q_0$ depends only on $r_0, q_0, r_1$ and $q_1$; all physical details of
how \( \mathcal{D} \) manages to transform \( q_0 \rightarrow q_1 \) and \( r_0 \rightarrow r_1 \) are irrelevant. As a result, we can reparameterize the set of all dissipation functions \( W_d(\cdot) \) as the \(|X|^2 \) numbers \( (q_0, W_d(q_0), p(x_1|x_0)) \).

Prop. 1 concerns only one component of expected total work, namely the dissipated work. However this is related to the expected total work via Eq. (2). Plugging Prop. 1 into Eq. (2) establishes the following:

**Corollary 3.** If \( W_d(q_0) = 0 \), then the expected total work expended by \( \mathcal{D} \) when it starts with distribution \( r_0 \) is \( (H_1)_{r_1} - (H_0)_{r_0} + C(r_0||q_0) - C(r_1||q_1) \).

In particular, suppose we have a process that is thermodynamically reversible for some initial distribution \( q_0 \) and that both \( H_0 \) and \( H_1 \) are uniform over the space of allowed states. Then if that process is run with some initial distribution \( r_0 \), the expected value of the total work expended is \( C(r_0||q_0) - C(r_1||q_1) \). (See [23] for an example of a physical system where this is the case.)

Finally, it is important to note that in general, \( q_0 \), the initial distribution that achieves minimal dissipation, will not be the initial distribution that achieves minimal expected total work. Let \( x_0^* = \arg \min_{x_0} \sum_{x_0} p(x_0|x_0) W(x_0|x_1) \) and define \( r_0(x_0) = \delta_{x_0, x_0^*} \). Then \( r_0 \) will achieve minimal expected total work, but will not in general be equal to \( q_0 \). (Indeed, as we establish in the SM, under generic conditions \( q_0 \) will lie in the relative interior of the simplex, and so cannot be the same as this work-minimizing \( r_0 \).)

**A new thermodynamic cost of computation.**—We now consider a process whose coarse-grained dynamics over computational macrostates (CMs) implements some desired input-output map.

Formally, let \( V \) indicate the set of computational macrostates (CMs), and \( \pi(v_1|v_0) \) the desired computation, i.e., the stochastic map from inputs \( v_0 \) to outputs \( v_1 \). Let \( g : X \rightarrow V \) be a coarse-graining function of a computer that maps its microstates to its CMs, and let \( s(x|v) \) be a fixed distribution of microstates given macrostates, with the condition that \( s(x|v) = 0 \) if \( v \neq g(x) \). Any distribution \( p_V(v) \) over CMs induces a mixture distribution over microstates, given by forming an average of distributions \( s(x|v) \) over all possible \( v \). Recalling that \( \Delta \) and \( \Delta^V \) indicate the \(|X|\) and \(|V|\)-dimensional unit simplices respectively, define the function \( \Phi : \Delta^V \rightarrow \Delta \) as

\[
[\Phi(p_V)](x) = \sum_v s(x|v)p_V(v) = s(x|g(x))p_V(g(x))
\]

Let \( S \) indicate the image of \( \Phi \). \( S \) is a convex subset of \( \Delta \) containing all distributions over \( X \) which are mixtures of \( s(x|v) \) induced by some distribution over \( V \).

Our analysis of thermodynamics of computation requires three conditions:

1) \( p(x_0) \in S \). Thus, initial conditions are determined up to an initial distribution over CMs, which we call the input distribution, \( p_V(v_0) \) (abbreviated as \( p^V_0 \)).

2) The thermodynamic process implements \( \pi \); \( \forall v_0, v_1, p(v_1|v_0) = \sum_{x_0,x_1} \delta_{v_1,g(x_1)}p(x_1|x_0) s(x_0|v_0) = \pi(v_1|v_0) \).

3) The thermodynamic process is thermodynamically cyclic with respect to \( V \), i.e., we require that \( \forall x_1, v_0, v_1, p(x_1|v_0, v_1) = \sum_{x_0} \delta_{v_1,g(x_1)}p(x_1|x_0)s(x_0|v_0) p(v_1|v_0) = s(x_1|v_1) \).

It is now appreciated that to analyze the thermodynamics of Maxwell’s demon, one should consider a complete thermodynamic cycle of the demon [32]. Similarly, to analyze the thermodynamics of an iteration of a computer that will run for multiple iterations, one must consider a full thermodynamic cycle of the computer. However, we do not impose this requirement in its strongest form, i.e., that the distribution over microstates returns to its initial distribution after each iteration of the computer. This strong form would imply that the distribution over CMs must be the same at the beginning and end of every iteration, i.e., that the computation must be the identity map.

Instead, in the context of computers, the appropriate formulation of the requirement that the process carry out a full thermodynamic cycle is stated by condition (3). Specifically, this condition guarantees two important properties. First, it means that \( p(x_1) \in S \), so long as \( p(x_0) \in S \). Thus, by condition (2), if a second iteration of the thermodynamic process is run starting from the \( t = 1 \) microstate distribution, then the distribution of CMs will again evolve according to \( \pi \); therefore, the dynamics over CMs will be time-translation invariant. Second, the cyclic condition guarantees that \( v_0 \) is conditionally independent of \( x_1 \) given \( v_1 \), i.e., there is no information about \( v_0 \) “hidden” in the microstate \( x_1 \), beyond the information that \( x_1 \) lies in \( v_1 \). This means that when the computer is run for multiple iterations, the dynamics of the CMs will be first-order Markovian, a common desideratum for coarse-grainings of dynamical systems [43].

When the computer is run with the microstate distribution \( \Phi(p^V_0) \), the amount of dissipated work is specified by \( W_d(\Phi(p^V_0)) \). Accordingly we refer to \( W_d(\Phi(p^V_0)) \) as the dissipation of the (macrostate) input distribution \( p^V_0 \), and when clear from context, write it simply as \( W_d(p^V_0) \).

In analogy with the case of dynamics over \( X \), we say that an input distribution \( q^S_0 \) is a prior for the computer if it achieves minimum dissipation among all input distributions. For instance, in previous work [24], we showed that for any given \( \pi \) and \( p^V_0 \), a computer can be designed that implements \( \pi \) with zero dissipation for input distribution \( p^V_0 \); in this case, \( p^V_0 \) would be a prior distribution.
We now make the additional assumption that \( q_0^V \) has full support \([38]\). The following result is the analog of Prop. 1 for dynamics over CMs:

**Corollary 4.** If a computer is cyclic and its prior \( q_0^V \) has full support, then for any distribution \( r_0^V \)

\[
W_d(r_0^V) - W_d(q_0^V) = D(r_0^V || q_0^V) - D(r_1^V || q_1^V)
\]

**Proof.** Let \( q_0 = \Phi(q_0^V) \), \( q(x_0, v_0) = q(x_0)\delta_{v_0, g(x_0)} = q_0^V(v_0)s(x_0|v_0) \), and similarly for \( q_1, r_0, \) and \( r_1 \).

By definition, \( q_0^V := \arg\min_{p \in \Delta^V} W_d(\Phi(p_0^V)) \), meaning that \( \Phi(q_0^V) = \arg\min_{p_0 \in S} W_d(p_0) \), where \( S \) is the convex subset of \( \Delta \) that is the image of \( \Phi \). Because \( q_0^V \) has full support, \( \Phi(q_0^V) \) is in the relative interior of \( S \). We proceed by applying Prop. 1 and expanding

\[
W_d(\Phi(r_0^V)) - W_d(\Phi(q_0^V)) \\
= D(r_0^V || q_0^V) - D(r_1^V || q_1^V) \\
= D(r_0^V || q_0^V) - D(r_1^V || q_1^V) \\
+ D(r_0^V || q_0^V) - D(r_0^V || q_1^V) \\
+ D(r_0^V || q_1^V) - D(r_1^V || q_1^V) \\
\]

where the third line follows because \( v_0 \) and \( v_1 \) are deterministic functions of \( x_0 \) and \( x_1 \), and the fourth line from the chain rule for KL divergence.

By definition, \( r(x_0|v_0) = s(x_0|v_0) = q(x_0|v_0) \), and so

\[
D(r_0^V || q_0^V) = D(r_1^V || q_1^V) = D(r_0^V || q_0^V) \
\]

The obvious analog of Corollary 3 (and the associated discussion) holds for cyclic computers, if we replace distributions over microstates by distributions over CMs. These results agree with the analysis for a specific model of a computer in [24]. However the analysis here holds for any computer, no matter how it operates.

**Discussion.**—We have identified a new kind of thermodynamic cost present in all nonequilibrium thermodynamic processes. This new cost arises due to varying the distribution over initial states of a physical system away from the optimal initial distribution, rather than due to varying the subsequent thermodynamic process away from the optimal one. This distinguishes this new cost from most of the costs considered in the literature. Indeed, in contrast to many analyses of other thermodynamic costs, our analysis of this new cost does not invoke the second law, and in contrast to most formulations of the second law, our results are equalities rather than inequalities.

We then showed that the same kind of thermodynamic cost arises in (cyclic) computers. This thermodynamic cost of computation arises whenever the distribution over initial states of a computer differs from the one that minimizes the dissipation caused by running that computer. It is a new kind of thermodynamic computation cost, in addition to the computation costs considered in analyses of the Landauer bound \([9, 11]\). However it is similar to the costs considered in those analyses, in that it equals zero for logically invertible computations implemented with cyclic computers.

**Acknowledgments** —We would like to thank the Santa Fe Institute for helping to support this research. This paper was made possible through the support of Grant No. TWC00079/AB47 from the Templeton World Charity Foundation, Grant No. FQXi-RHI3-1349 from the FQXi foundation, and Grant No. CHE-1648973 from the U.S. National Science Foundation. The opinions expressed in this paper are those of the authors and do not necessarily reflect the view of Templeton World Charity Foundation.

---

[1] H. Touchette and S. Lloyd, Physica A: Statistical Mechanics and its Applications 331, 140 (2004).
[2] T. Sagawa and M. Ueda, Physical review letters 102, 250602 (2009).
[3] R. Dillenschneider and E. Lutz, Physical review letters 104, 189903 (2010).
[4] T. Sagawa and M. Ueda, Physical review letters 109, 180602 (2012).
[5] G. E. Crooks, Physical Review E 60, 2721 (1999).
[6] G. E. Crooks, Journal of Statistical Physics 90, 1481 (1998).
[7] F. Clejne Janna, F. Moukalled, and C. A. Gómez, International Journal of Thermodynamics 16, 97 (2013).
[8] C. Jarzynski, Physical Review Letters 78, 2690 (1997).
[9] M. Esposito and C. Van den Broeck, EPL (Europhysics Letters) 95, 40004 (2011).
[10] M. Esposito and C. Van den Broeck, Physical Review E 82, 011143 (2010).
[11] J. M. Parrondo, J. M. Horowitz, and T. Sagawa, Nature Physics 11, 131 (2015).
[12] B. S. Pollard, Open Systems & Information Dynamics 23, 1650006 (2016).
[13] K. Wiesner, M. Gu, E. Rieper, and V. Vedral, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 468, 4058 (2012).
[14] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Physical review letters 109, 120604 (2012).
[15] M. Prokopenko, J. T. Lizier, and D. C. Price, Entropy 15, 524 (2013).
[16] M. Prokopenko and J. T. Lizier, Nature Scientific reports 4 (2014).
[17] J. Dunkel, Nature Physics 10, 409 (2014).
[18] E. Roldán, I. A. Martínez, J. M. Parrondo, and D. Petrov, Nature Physics (2014).
[19] A. Béret, A. Arakelyan, A. Petrosysan, S. Ciliberto, R. Dillenschneider, and E. Lutz, Nature 483, 187 (2012).
[20] K. Takara, H.-H. Hasegawa, and D. Driebe, Physics Letters A 375, 88 (2010).
[21] S. Deffner and E. Lutz, arXiv preprint arXiv:1201.3888 (2012).
[22] K. Takara, H.-H. Hasegawa, and D. Driebe, Physics Letters A 375, 88 (2010).
[23] D. H. Wolpert, Entropy 18, 138 (2016).
Indeed, though logical reversibility and thermal reversibility were associated in early work on the thermodynamics of computation, it is now understood that they are independent. For instance, one can design a process to erase a bit in a thermodynamically reversible manner even though bit-erasure is logically-irreversible [30], assuming that the distribution over the states of the bit is exactly known to the designer.

See Supplemental Material at [URL will be inserted by publisher].

In the SM we prove that Assumption (2) holds if all components of $p(x_1|x_0)$ are nonzero. In the case of a computer, we also show that this condition is sufficient for the prior input distribution to have full support.

Acknowledgments...
SUPPLEMENTARY MATERIAL

A. Sufficient conditions for \( q_0 \) to have full support

In this section of the SM we assume that all components of \( p(x_1 | x_0) \) are nonzero and that \( \mathcal{Q}(x_0) \) is finite for all \( x_0 \), and show that this means that both minimizers \( q(x_0) \) and \( q(v_0) \) have full support.

To begin, expand Eq. (4) to write

\[
\frac{\partial W_d}{\partial p(x_0)}(p_0) = -\mathcal{Q}(x_0) - \sum_{x_1} p(x_1 | x_0) \ln \left[ \frac{\sum_{x_0'} p(x_1 | x_0') p(x_0')} {p(x_0)} \right] \tag{8}
\]

for any distribution \( p_0 \) over \( X \).

Define \( q_0 := \arg \min_{p_0 \in \Delta} W_d(p_0) \), recalling that \( \Delta \) is the \(|X|\)-dimensional unit simplex. To show that \( q_0 \) has full support, hypothesize that there exists some \( x_0^* \) such that \( q(x_0^*) = 0 \). Now consider the one-sided derivative \( \frac{\partial W_d}{\partial p(x_0)}(q_0) \).

By the assumption that \( p(x_1 | x_0) > 0 \) for all \( x_0, x_1 \), the numerator inside the logarithm in Eq. (8) is nonzero, while by hypothesis the denominator is 0. Thus, the argument of the logarithm is positive infinite and (since \( \mathcal{Q}(x_0^*) \) is finite, by assumption) \( \frac{\partial W_d}{\partial p(x_0)}(q_0) \) is negative infinite. Moreover, for any \( x_0^* \) where \( q(x_0^*) > 0 \), \( \frac{\partial W_d}{\partial p(x_0)}(q_0) \) is finite. This means that \( W_d(q_0) \) can be reduced by increasing \( q(x_0^*) \) and (to maintain normalization) reducing \( q(x_0^*) \), contrary to the definition of \( q_0 \) as a minimizer. Therefore our hypothesis must be wrong.

Next, consider the prior input distribution \( q_0^V := \arg \min_{p_0' \in \Delta^V} W_d(\Phi(p_0^V)) \), recalling that \( \Delta^V \) is the \(|V|\)-dimensional unit simplex. To show that \( q_0^V \) has full support under the above assumptions, consider the partial derivative of dissipation wrt to each entry of the input probability distribution, \( \frac{\partial W_d(\Phi(p_0^V))}{\partial p_v(v_0)} \). Let \( p_0 := \Phi(p_0^V) \), and then use the chain rule, Eq. (7), Eq. (8), and then Eq. (7) again to write

\[
\frac{\partial W_d(\Phi(p_0^V))}{\partial p_v(v_0)}(p_0^V) = \sum_{x_0} \frac{\partial W_d}{\partial p(x_0)}(p_0) \frac{\partial \Phi(p_0^V)(x_0)}{\partial p_v(v_0)}(p_0^V)
\]

\[
= \sum_{x_0} \frac{\partial W_d}{\partial p(x_0)}(p_0) \frac{s(x_0 | v_0)}{p(x_0)}
\]

\[
= \sum_{x_0} \left[ -\mathcal{Q}(x_0) - \sum_{x_1} p(x_1 | x_0) \ln \frac{\sum_{x_0'} p(x_1 | x_0') p(x_0')}{p(x_0)} \right] s(x_0 | v_0)
\]

\[
= \sum_{x_0} \left[ -\mathcal{Q}(x_0) - \sum_{x_1} p(x_1 | x_0) \ln \frac{\sum_{x_0'} p(x_1 | x_0') s(x_0 | v_0) p(x_0')}{p(x_0)} \right] s(x_0 | v_0) \tag{9}
\]

for any distribution \( p_0^V \) over \( V \).

Proceeding as before, hypothesize that there exists some \( v_0^* \) such that \( q_v(v_0^*) = 0 \), and use Eq. (9) to evaluate \( \frac{\partial W_d(\Phi(p_0^V))}{\partial p_v(v_0^*)}(q_0^V) \). By our hypothesis, the associated value of the denominator in the logarithm in Eq. (9) is zero. Since \( p(x_1 | x_0) \) is always nonzero by assumption, this means the sum over \( x_1 \) is positive infinite. Since by assumption \( \mathcal{Q}(x_0) \) is bounded, this means that \( \frac{\partial W_d(\Phi(p_0^V))}{\partial p_v(v_0^*)}(q_0^V) \) is negative infinite. At the same time, \( \frac{\partial W_d(\Phi(p_0^V))}{\partial p_v(v_0^*)}(q_0^V) \) is finite for any \( v_0^* \) where \( q_v(v_0^*) > 0 \). Thus \( W_d(\Phi(q_0^V)) \) can be reduced by increasing \( q_v(v_0^*) \) and (to maintain normalization) reducing \( q_v(v_0^*) \), contrary to the definition of \( q_0^V \) as a minimizer. Therefore our hypothesis must be wrong.
Suppose the driven dynamics \( p(X_1|X_0) \) is a stochastic map from \( X \to X \) that results in minimal dissipation for some prior distribution \( q_0 \).

**Theorem 1.** Suppose that \( q_0 \) has full support. Then, there exists \( r_0 \) with incorrect prior dissipation \( W_d(r_0) - W_d(q_0) > 0 \) if \( p(X_1|X_0) \) is not an invertible map.

**Proof.** If \( q_0 \) has full support, then \( \text{supp} r_0 \subseteq \text{supp} q_0 \) for all \( r_0 \). Then Prop. 1 in the main text states that if initial distribution \( r_0 \) is used, extra dissipation is equal to:

\[
W_d(r_0) - W_d(q_0) = D(r_0||q_0) - D(r_1||q_1) = D(r(X_0|X_1)||q(X_0|X_1))
\]  

(10)

KL divergence is invariant under invertible transformations. Therefore, if \( p(X_1|X_0) \) is an invertible map, then \( D(r_0||q_0) = D(r_1||q_1) \implies W_d(r_0) - W_d(q_0) = 0 \) for all \( r_0 \).

We now prove that if \( p(X_1|X_0) \) is not an invertible map, then there exists \( r_0 \) such that \( W_d(r_0) - W_d(q_0) > 0 \). For simplicity, write the dynamics \( p(X_1|X_0) \) as the right stochastic matrix \( M \). Because \( M \) is a right stochastic matrix, it has a right (column) eigenvector \( 1^T = (1, \ldots , 1)^T \) with eigenvalue 1.

Furthermore, it is known that if \( M \) is not an invertible map, i.e. permutation matrix, then \( |\det M| < 1 \). Since the determinant is the product of the eigenvalues and the magnitude of any eigenvalue of a stochastic matrix is upper bounded by 1, \( M \) must have at least one eigenvalue \( \lambda \) with \( |\lambda| < 1 \). Let \( s \) represent the non-zero left eigenvector corresponding to \( \lambda \). Note that due to biorthgonality of eigenvectors, \( s1^T = 0 \). We use \( s(x) \) to refer to elements of \( s \) indexed by \( x \in X \). Without loss of generality, assume \( s \) is scaled such that \( \max_x |s(x)| = \min_{x_0} q(x_0) \) (which is greater than 0, by assumption that \( q_0 \) has full support).

We now define \( r_0 \) as:

\[
r(x_0) := q(x_0) + s(x_0)
\]

Due to the scaling of \( s \) and because \( s1^T = 0 \), \( r_0 \) is a valid probability distribution.

We use the notation \( s(x) := \sum_{x_0} s(x_0) p(x| x_0) \) and \( r(x) := \sum_{x_0} r(x_0) p(x| x_0) = q(x) + s(x) \). We also use the notation \( \mathcal{C} := \text{supp} r_1 \). The fact that \( q_0 \) has full support also means that \( \mathcal{C} \subseteq \text{supp} q_1 \).

The proof proceeds by contradiction. Assume that \( W_d(r_0) - W_d(q_0) = 0 \). Using Eq. 10 and due to properties of KL divergence, this means that for each \( x_0 \in X \) and \( x_1 \in \mathcal{C} \):

\[
\frac{q(x_0|x_1)}{q(x_1)} = \frac{r(x_0)p(x_1|x_0)}{r(x_1)}
\]

\[
\frac{q(x_1)}{q(x_0)}p(x_1|x_0) = \frac{r(x_1)p(x_1|x_0)}{r(x_0)}
\]

\[
\frac{q(x_1) + s(x_1)}{q(x_1)}p(x_1|x_0) = \frac{q(x_0) + s(x_0)}{q(x_0)}p(x_1|x_0)
\]

Taking absolute value of both sides gives:

\[
|s(x_1)|q(x_0|x_1) = |s(x_0)|p(x_1|x_0)
\]

Summing over \( x_0 \in X \) and \( x_1 \in \mathcal{C} \):

\[
\sum_{x_1 \in \mathcal{C}} \sum_{x_0 \in X} |s(x_1)|q(x_0|x_1) = \sum_{x_1 \in \mathcal{C}} \sum_{x_0 \in X} |s(x_0)|p(x_1|x_0)
\]

(11)

\[
\sum_{x_1 \in X} |s(x_1)| - \sum_{x_1 \notin \mathcal{C}} |s(x_1)| = \sum_{x_1 \in X} \sum_{x_0 \in X} |s(x_0)|p(x_1|x_0) - \sum_{x_1 \notin \mathcal{C}} \sum_{x_0 \in X} |s(x_0)|p(x_1|x_0)
\]
Note that for all \( x_1 \notin C \), \( r(x_1) = 0 \), meaning that \( s(x_1) = -q(x_1) \). Thus,

\[
\sum_{x_1 \notin C} |s(x_1)| = \sum_{x_1 \notin C} q(x_1)
\]

Furthermore, for all \( x_1 \notin C \), \( r(x_1) = \sum_{x_0 \in X} r(x_0) p(x_1|x_0) = 0 \). Thus, for all \( x_0 \in X \) where \( p(x_1|x_0) > 0 \) for some \( x_1 \notin C \), \( r(x_0) = 0 \), meaning \( s(x_0) = -q(x_0) \). This allows us to rewrite the last term in Eq. 11 as:

\[
\sum_{x_1 \notin C} \sum_{x_0 \in X} |s(x_0)| p(x_1|x_0) = \sum_{x_1 \notin C \cap x_0, p(x_1|x_0) > 0} |s(x_0)| p(x_1|x_0) = \sum_{x_1 \notin C} \sum_{x_0 \in X, p(x_1|x_0) > 0} q(x_0) p(x_1|x_0) = \sum_{x_1 \notin C} q(x_1)
\]

Cancelling terms that equal \( \sum_{x_1 \notin C} q(x_1) \) from both sides of Eq. 11, we rewrite:

\[
\sum_{x_1} |s(x_1)| = \sum_{x_1} \sum_{x_0} |s(x_0)| p(x_1|x_0) = \sum_{x_0} |s(x_0)|
\]  

(12)

In matrix notation, Eq. 12 states that:

\[
\|sM\|_1 = \|s\|_1
\]

(13)

where \( \|\cdot\|_1 \) indicates the vector \( \ell_1 \) norm. However, by definition \( sM = \lambda s \). Hence:

\[
\|sM\|_1 = \|\lambda s\|_1 = |\lambda| \|s\|_1 < \|s\|_1
\]

meaning that Eq. 13 cannot be true and the original assumption \( W_d(r_0) - W_d(q_0) = 0 \) is incorrect. We have shown that for non-invertible maps, there always exists an \( r_0 \) for which \( W_d(r_0) - W_d(q_0) > 0 \).