PAPER: CLASSICAL STATISTICAL MECHANICS, EQUILIBRIUM AND NON-EQUILIBRIUM

# Dependence of dissipation on the initial distribution over states

To cite this article: Artemy Kolchinsky and David H Wolpert *J. Stat. Mech.* (2017) 083202

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Classical statistical mechanics, equilibrium and non-equilibrium

# Dependence of dissipation on the initial distribution over states

**Artemy Kolchinsky**[1] **and David H Wolpert**[1,2,3]

[1] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, United States of America
[2] Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, United States of America
[3] Arizona State University, Tempe, AZ 85281, United States of America
E-mail: artemyk@gmail.com

Received 17 May 2017
Accepted for publication 10 July 2017
Published 16 August 2017

Online at stacks.iop.org/JSTAT/2017/083202
https://doi.org/10.1088/1742-5468/aa7ee1

**Abstract.** We analyze how the amount of work dissipated by a fixed nonequilibrium process depends on the initial distribution over states. Specifically, we compare the amount of dissipation when the process is used with some specified initial distribution to the minimal amount of dissipation possible for any initial distribution. We show that the difference between those two amounts of dissipation is given by a simple information-theoretic function that depends only on the initial and final state distributions. Crucially, this difference is independent of the details of the process relating those distributions. We then consider how dissipation depends on the initial distribution for a 'computer', i.e. a nonequilibrium process whose dynamics over coarse-grained macrostates implement some desired input-output map. We show that our results still apply when stated in terms of distributions over the computer's coarse-grained macrostates. This can be viewed as a novel thermodynamic cost of computation, reflecting changes in the distribution over inputs rather than the logical dynamics of the computation.

**Keywords:** dynamical processes

## Contents



## 1. Introduction

The past few decades have seen great advances in nonequilibrium statistical physics [1–16], resulting in many novel predictions and experiments [17–19]. Some of the most important results of this research have been powerful new tools for analyzing the *dissipated work* (or 'dissipation' for short) in nonequilibrium processes. Dissipation is the amount of work done on an evolving system that exceeds the theoretical minimal amount needed to drive such a system from its initial to its final distribution [9, 11, 20–22]. Equivalently, it is proportional to the (irreversible) entropy production during the course of the process, i.e. the total change in entropy of the system minus the amount of entropy that flows from the heat bath to the system in the form of heat.

Several expressions for the amount of dissipation in any given process have been derived by exploiting the detailed fluctuation theorems (DFTs) [5, 6, 23–25], typically under the assumption of dynamics that obeys local detailed balance. These results express the dissipation in terms of the Kullback–Leibler (KL) divergence [26, 27] between the probability density over state trajectories occurring in the original process and the probability density under a special 'time-reversed' version of the process. However these results are impractical for quantifying dissipation in many cases of

https://doi.org/10.1088/1742-5468/aa7ee1


J. Stat. Mech. (2017) 083202

interest, since computing the KL divergence requires integration of a probability density over all possible trajectories.

Related research has investigated lower bounds on dissipation by studying *optimal processes*. These are processes that achieve minimal dissipation subject to some specified set of constraints [11, 20, 21]. For example, optimal processes have been identified for transforming some desired initial Hamiltonian and state distribution into a different Hamiltonian and state distribution under a finite-time constraint [28–30], or while obeying a constraint on allowable work fluctuations [31]. Some authors have also considered how changes to the initial distribution affect the work and dissipation *if the process is changed to be optimal for the new distribution* [2, 3, 20]. Such research is concerned with processes that minimize dissipation, and more generally with how dissipation varies with changes to the process.

Here we consider a complementary problem, which to our knowledge has never been previously analyzed. We suppose that there is a fixed process $\mathcal{P}$, coupled to a heat bath that is at a constant temperature. We then consider a very common real-world scenario, in which this same process can be run with different initial distributions over states. We ask, how does the the amount of work dissipated by $\mathcal{P}$ vary with changes to the initial distribution? What is the maximal cost in extra dissipation that can arise by using one initial distribution rather than another? How do these answers depend on the details of the process $\mathcal{P}$?

Surprisingly, we find that the dependence of dissipation on the initial distribuiton has a simple information-theoretic form. Let $q_0$ be an initial distribution over the states for which $\mathcal{P}$ dissipates the minimal amount of work. We prove that the dissipation arising from using some arbitrary initial distribution $r_0$ is the dissipation arising from using $q_0$, plus the reduction of the Kullback–Leibler (KL) divergence between $r_0$ and $q_0$ from the beginning to the end of $\mathcal{P}$. The additional dissipation incurred when $\mathcal{P}$ is initialized with $r_0 \neq q_0$ is independent of all intermediate details of *how* $\mathcal{P}$ changes the initial distribution into the final one.

Our analysis provides a useful and novel tool for calculating dissipated work for a given thermodynamic process run on a given initial distribution. For example, suppose we design a process to be dissipationless (i.e. thermodynamically reversible) when run with some initial state distribution. Our analysis can be used to calculate exactly how much work would be dissipated if that process were run with some other state distribution. As a demonstration appendix C, we consider a published model of Maxwell's demon [32], a device that extracts work from an incoming stream of bits, and compute dissipation as a function of the distribution of bits.

More generally, consider a fixed process connected to a heat bath that is at a constant temperature, which dissipates least work when prepared with some particular initial distribution. For example, this might be a process in which a volume of gas expands while pushing against a piston and lifting a weight. There will be some 'optimal' initial distribution of the states of the gas which minimizes dissipated work. Our results state how much more work will be dissipated when the gas is prepared with some other initial distribution.

After deriving this result, we extend it to analyze dissipation in a physical computer. More precisely, we suppose that there is a coarse-graining of the states of our system into a set of macrostates. These macrostates are identified with logical values and the

J. Stat. Mech. (2017) 083202

dynamics over the macrostates identified with the (possibly noisy) computation. The initial distribution over the macrostates may reflect how a user of the computer initializes its logical values. As before, we consider how the additional dissipation incurred by a computer, above and beyond the minimum, depends on the initial macrostate distribution. We show that the additional dissipation is still given by the drop in KL divergence, only now stated in terms of distributions over the macrostates.

To illustrate the implications of this result for thermodynamics of computation, suppose we construct a process that performs a given computation, and that achieves zero dissipation for *some* initial distribution over its states (e.g. when employed by one particular user of that computer). Our results quantify how much computer will dissipate if it is instead initialized according to a different distribution (e.g. if the computer is employed by some other user).

We emphasize that these results are equalities, not just bounds. Furthermore, our results give dissipation in terms of a difference in two KL divergences, concerning initial and final *state distributions*. Thus they differ fundamentally from previously derived DFTs, which give dissipation in terms of a single KL divergence, concerning forward and time-reversed *trajectory distributions*. Moreover, in contrast to such DFTs, our results do not assume local detailed balance.

Our analysis of dissipated work should also be distinguished from earlier analyses of *reversible work*, in particular analyses expressing reversible work as a reduction of KL divergence between nonequilibrium and equilibrium distributions at initial and final times plus the difference of equilibrium free energies [21]. Reversible work is the work required to perform a given transformation using an optimal process, and can be thermodynamically recovered by reversing the process. Dissipated work, on the other hand, is work that is irreversibly lost as entropy production. Furthermore, in general the KL divergences that arise in our analysis do not necessarily involve equilibrium distributions.

There is one previously-known result that is a special case of our analysis: if a system is prepared with some nonequilibrium distribution $r_0$ and then undergoes a nondriven process in which it fully relaxes to equilibrium, the dissipated work is equal to the KL divergence between $r_0$ and the equilibrium distribution [33]. Our analysis generalizes this earlier result significantly, allowing for processes that do not relax fully to equilibrium. It also applies to processes that are driven by an external work reservoir, in which the equilibrium changes over time, during which the system can remain arbitrarily far from equilibrium at all times.

## 2. Formal background

We consider a physical system with a countable set of microstates $X$ that evolves across a countable set of times $t \in \{0, \Delta\tau, 2\Delta\tau, \ldots, 1\}$, while in contact with a heat bath at temperature $T$. We use $x_{0..1} := (x_0, x_{\Delta\tau}, \ldots, x_1)$ to indicate a particular trajectory through the system's state space. The system may also be connected to a work reservoir throughout its evolution, which causes the system's Hamiltonian to change with time. We indicate the trajectory through the space of Hamiltonians as $H_{0..1} := (H_0, H_{\Delta\tau}, \ldots, H_1)$.



J. Stat. Mech. (2017) 083202

Note that the units of time are arbitrary and $\Delta\tau$ can be arbitrarily small (though non-zero). Accordingly our results hold exactly no matter how long the process takes, and in particular even in the quasi-static limit. The choice of countable state space and discretized time is used to simplify analysis, in line with much of the literature [5, 14, 34, 35]. However, our approach should extend to continuous state space and continuous time.

Write the distribution of the system's state at $t$ as $p_t(x)$, or equivalently $p(x_t)$. Due to thermal fluctuations and driving by the work reservoir, the system undergoes a stochastic dynamics, represented by a conditional distribution of trajectories given initial states, $p(x_{0..1}|x_0)$ (we make no assumptions about whether this dynamics is first-order Markovian or not). The conditional distribution over trajectories in turn induces a conditional distribution of final states given initial states, $p(x_1|x_0) = \sum_{x'_{0..1}} \delta_{x'_1,x_1} p(x'_{0..1}|x_0)$, which we sometimes refer to as a map that takes initial states $x_0$ to final states $x_1$.

We refer to a given pair of $H_{0..1}$ and $p(x_{0..1}|x_0)$ as a (thermodynamic) process operating on the system, indicated generically as $\mathcal{P}$. Note that any process $\mathcal{P}$ can be prepared with different initial distributions $p_0$, giving different trajectory probabilities $p(x_{0..1}) := p(x_{0..1}|x_0)\, p_0(x_0)$.

Given a sequence of Hamiltonians $H_{0..1}$, the total work done on the system if it follows state trajectory $x_{0..1}$ is

$$W(x_{0..1}) = \sum_{t\in\{0,\Delta\tau,..,1\}} H_{t+\Delta\tau}(x_t) - H_t(x_t). \tag{1}$$

For an initial distribution $p_0$, the expected work across all trajectories is

$$\langle W\rangle_{p_0} = \sum_{x_{0..1}} p_0(x_0)\, p(x_{0..1}|x_0) W(x_{0..1}).$$

Suppose we seek to drive the system from some particular (possibly non-equilibrium) initial distribution $p_0$ to some final distribution $p_1$, while changing the Hamiltonian from $H_0$ to $H_1$. Define the non-equilibrium free energy [11] of a system with Hamiltonian $H_t$ and distribution $p_t(x)$ as

$$\mathcal{F}(H_t, p_t) := \langle H_t\rangle_{p_t} - kT \cdot S(p_t),$$

where $S(p) := -\sum_x p(x) \ln p(x)$ indicates Shannon entropy (in nats). (Note that $\mathcal{F}$ is equal to the equilibrium free energy when $p_t$ is the Boltzmann distribution for Hamiltonian $H_t$.) For any process $\mathcal{P}$ that transforms $(p_0, H_0) \to (p_1, H_1)$, expected work is lower bounded by

$$\langle W\rangle_{p_0} \geqslant \mathcal{F}(H_1, p_1) - \mathcal{F}(H_0, p_0)\,. \tag{2}$$

This inequality reflects the modern understanding of the second law [9, 11, 21, 22].

The difference of non-equilibrium free energies is called the reversible work. Reversible work is the portion of expected work that could be recovered from the heat bath and system after the process finishes, by transforming the system from $H_1, p_1$ back to $H_0, p_0$ in a thermodynamically reversible manner (in this way completing a thermodynamic cycle). Reversible work can be either positive or negative, depending on $H_0, p_0, H_1$ and $p_1$.

J. Stat. Mech. (2017) 083202

Dissipated work or simply dissipation is the portion of expected work that cannot be thermodynamically recovered [11, 21, 33]. It is written as

$$W_d(p_0) := \langle W \rangle_{p_0} - [\mathcal{F}(H_1, p_1) - \mathcal{F}(H_0, p_0)]\,. \tag{3}$$

The dissipation associated with a process is always non-negative, and it is zero iff the process is thermodynamically reversible. (Dissipation should not to be confused with the dissipated heat, which is the total energy transferred to the heat bath, nor with expected total work minus the change in *equilibrium* free energies, which is also sometimes called dissipated work [33, 36, 37].)

Define $\mathcal{Q}(x_0)$ as the expected total heat transferred from the bath to the system during the process if the system starts in $x_0$ [6]. By conservation of energy we can write this as

$$\mathcal{Q}(x_0) := \sum_{x'_{0..1}} p(x'_{0..1}|x_0)(H_1(x'_1) - H_0(x_0) - W(x'_{0..1})),$$

so that the total expected heat transferred is $\langle \mathcal{Q}(X_0) \rangle_{p_0} = \sum_{x_0} p(x_0)\mathcal{Q}(x_0)$. This allows us to rewrite dissipation as

$$W_d(p_0) = kT[S(p_1) - S(p_0)] - \langle \mathcal{Q}(X_0) \rangle_{p_0}, \tag{4}$$

where $p_1(x') = \sum_x p(x_1|x_0)\, p_0(x_0)$ is the final state distribution when the process is initialized with $p_0$. Thus, dissipation is proportional to the entropy change that does not correspond to heat exchanged with the heat bath, which is called the (irreversible) entropy production [9, 11, 38].

In the remainder of this paper we choose units so that $kT = 1$.

## 3. Dissipation due to incorrect priors

Let $q_0$ be an initial distribution that achieves minimum dissipation for a given $\mathcal{P}$,

$$q_0 := \arg\min_{p_0} W_d(p_0). \tag{5}$$

We call $q_0$ the prior distribution for $\mathcal{P}$ (for reasons made clear below). We do not assume that the prior distribution is unique.

While $q_0$ is an initial distribution that results in minimal dissipation, in general $\mathcal{P}$ may be prepared with some initial distribution $r_0$, which we call the environment distribution that need not equal $q_0$. By definition,

$$W_d(r_0) - W_d(q_0) \geqslant 0.$$

We call this extra dissipation when using $r_0$ rather than $q_0$ the incorrect prior dissipation. Notice that if $\mathcal{P}$ achieves zero dissipation for some initial distribution, then $W_d(q_0) = 0$ and dissipation and incorrect prior dissipation are equivalent.

Several papers have shown that it is possible to design a process that implements any given stochastic map $p(x_1|x_0)$ with zero dissipation for any given initial distribution $p_0$ [39–41]. Incorrect prior dissipation first appeared in these analyses: it was shown that a *particular type of process* that implements a given stochastic map and achieves



J. Stat. Mech. (2017) 083202

zero dissipation for a particular $q_0$ will dissipate work when prepared with a different initial distribution $r_0 \neq q_0$. Here we generalize these previous analyses; the main result of our paper is a simple expression for incorrect prior dissipation that applies to *any* thermodynamic process.

To derive our main result, note that by definition, the prior $q_0$ minimizes the differentiable function $W_d$ over the set of all valid probability distributions. We assume that $q_0$ has full support, i.e. it is in the interior of the unit simplex. (This assumption will often hold; appendix D presents one particular sufficient condition concerning $p(x_1|x_0)$.) Then, for any initial state distribution $r_0$, the directional derivative at $q_0$ must obey

$$(r_0 - q_0) \cdot \nabla W_d(q_0) = 0, \tag{6}$$

where $\cdot$ indicates the dot product.

Next we use equation (4) to write the $|X|$ components of $\nabla W_d(p_0)$,

$$\begin{aligned}\frac{\partial W_d}{\partial p(x_0)}(p_0) &= \Big[-\sum_{x_1} p(x_1|x_0) \ln\Big(\sum_{x_0'} p_0(x_0')\, p(x_1|x_0')\Big) - 1\Big] + \Big[\ln p(x_0) + 1\Big] - \mathcal{Q}(x_0) \\ &= -\sum_{x_1} p(x_1|x_0) \ln p_1(x_1) + \ln p_0(x_0) - \mathcal{Q}(x_0).\end{aligned} \tag{7}$$

Combining equations (7) and (4) lets us express the inner products as

$$q_0 \cdot \nabla W_d(q_0) = S(q_1) - S(q_0) - \langle Q \rangle_{q_0} = W_d(q_0) \tag{8}$$

$$\begin{aligned} r_0 \cdot \nabla W_d(q_0) &= C(r_1\|q_1) - C(r_0\|q_0) - \langle Q \rangle_{r_0} \\ &= D(r_1\|q_1) - D(r_0\|q_0) + W_d(r_0). \end{aligned} \tag{9}$$

where $C(p\|q) := -\sum_x p(x) \ln q(x)$ is the cross-entropy function and $D(p\|q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} = C(p\|q) - S(p)$ is the Kullback–Leibler (KL) divergence [26].

Combining equations (6), (8) and (9) leads to our main result: incorrect prior dissipation for any distribution $r_0$ is

$$W_d(r_0) - W_d(q_0) = D(r_0\|q_0) - D(r_1\|q_1). \tag{10}$$

(See appendix A for an extension of this result for the case where all distributions are restricted to a convex subset of the unit simplex.)

Recall that the KL divergence $D(r\|q)$ is an information-theoretic measure of the distinguishability of distributions $r$ and $q$ [26]. Thus, our main result states that 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 $p(x_1|x_0)$ [42, 43]. It is non-negative due to the KL data processing inequality [44, lemma 3.11]. (This is consistent with our main result, since incorrect prior dissipation measures extra dissipation relative to the minimum possible.)

Interestingly, the contraction of KL divergence reflects the logical reversibility of the map $p(x_1|x_0)$. If $p(x_1|x_0)$ specifies a logically-reversible map from $x_0$ to $x_1$ (i.e. a permutation over $X$), then incorrect prior dissipation is 0 for all $r_0$. At the other extreme, if



J. Stat. Mech. (2017) 083202

$p(x_1|x_0)$ is an input-independent map, where changing $x_0$ has no effect on the resultant distribution over $x_1$, then $D(r_1\|q_1) = 0$ and incorrect prior dissipation reaches its maximum value of $D(r_0\|q_0)$. In addition, in this case the prior distribution that minimizes $W_d(.)$ is unique, since $W_d(r_0) = D(r_0\|q_0) = 0$ iff $r_0 = q_0$. More generally, in appendix E we prove that if and only if $p\,(x_1|x_0)$ is not a logically reversible map, then there must exist an $r_0$ such that $W_d(r_0) - W_d(q_0) > 0$.

For another perspective on equation (10), note that by the chain rule for KL divergence [26, equation (2.67)],

$$\begin{aligned} D(r(X_0, X_1)\|q(X_0, X_1)) &= D(r_0\|q_0) + D(r(X_1|X_0)\|q(X_1|X_0)) \\ &= D(r_1\|q_1) + D(r(X_0|X_1)\|q(X_0|X_1)). \end{aligned} \tag{11}$$

However, since $r(x_1|x_0) = q(x_1|x_0) = p(x_1|x_0)$, $D(r(X_1|X_0)\|q(X_1|X_0)) = 0$. Thus, equation (10) is equivalent to

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

(See also [41].) 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 equation (10) is why we refer to the initial distribution $q_0$ as a 'prior'.)

In appendix A, we show that if $q_0$ is *not* assumed to have full support, then the RHS of (10) becomes a lower bound (rather than an equality) on the incorrect prior dissipation.

## 4. Discussion of incorrect prior dissipation

In this section, we present some important implications and generalizations of our main result, as well as some caveats that are important to keep in mind.

Note that a thermodynamic process $\mathcal{P}$ is specified by a large set of real numbers: the values of the Hamiltonian $H_{0..1}$ and the conditional distribution $p(x_{0..1}|x_0)$. (In fact, in the $\Delta\tau \to 0$ limit this set is infinite.) However, by equation (4), the dissipation function $W_d(\cdot)$ can be specified using only $|X|^2$ real numbers: the $|X|$ values of $\mathcal{Q}(x_0)$ and the $|X|(|X| - 1)$ values of $p(x_1|x_0)$. Unfortunately, the values $\mathcal{Q}(x_0)$ may be impractical to compute for a given $\mathcal{P}$, since they involve expectations over a very large set of trajectories. Indeed, the distribution over trajectories may not even be fully specified if some details of the process are unknown.

Equation (10) shows that $W_d(\cdot)$ can alternatively be parameterized by the $|X|^2$ numbers: the value of $W_d(q_0)$, the $|X| - 1$ values of $q_0$, and the $|X|(|X| - 1)$ values of $p(x_1|x_0)$. This also means that, perhaps surprisingly, calculating the amount of dissipation above the minimum possible only requires knowledge of the stochastic map $p(x_1|x_0)$ and a minimizer $q_0$, and does not depend on any specifics of the intermediate process. Given some initial distribution $r_0$, all physical details of how $\mathcal{P}$ manages to transform $q_0 \to q_1$ and $r_0 \to r_1$ are irrelevant for evaluating incorrect prior dissipation.

It is important to emphasize that our analysis above does not specify how to find the minimizer $q_0$. In some cases, it may be possible to find $q_0$ via numerical minimization of



J. Stat. Mech. (2017) 083202

the convex function $W_d(p_0)$ over a $|X|^2$-dimensional space. (See appendix B for a proof that $W_d$ is convex.) In others, such minimization may be achievable via analytical techniques, or it may be possible to analytically find an initial distribution that achieves zero dissipation (which must then be a minimizer). Some previous studies have used these kinds of techniques to find priors $q_0$ and our results can provide additional insight into those studies. For example, one published model of Maxwell's demon used numerical methods to derive an inequality for dissipated work [32, equation (10)]. As we show in appendix C, our results can be used in a straightforward manner to derive this inequality analytically—and in fact provide an exact expression for dissipated work.

It is also important to emphasize that our main result concerns only one contributor to the total dissipated work (namely the amount in addition to the minimum amount possible). Moreover, dissipated work itself is just one contributor to expected total work. Thus, for instance, the fact that incorrect prior dissipation is related to the logical irreversibility of the map $p(x_1|x_0)$ does not necessarily have implications for the relationship between logical irreversibility and the *total* dissipation and/or *total* work involved in carrying out the map[4]. In addition, note that the prior distribution $q_0$, which minimizes dissipation, will not generally be the initial distribution that minimizes expected total work. Indeed, since total expect work is linear in the initial distribution over states, the distribution that minimizes expected total work is a delta function about $x_0^\star = \arg\min_{x_0} \sum_{x_{0..1}} p(x_{0..1}|x_0)W(x_{0..1})$. In general, that delta function distribution will not minimize dissipation.

There are some conditions, however, when incorrect prior dissipation *can* be related to expected total work. Consider the case when the process $\mathcal{P}$ is thermodynamically-reversible for some initial distribution, meaning that $W_d(q_0) = 0$. Then, the expected work when $\mathcal{P}$ is prepared with initial distribution $r_0$ is

$$\begin{aligned}\langle W\rangle_{r_0} &= D(r_0\|q_0) - D(r_1\|q_1) + \mathcal{F}(r_1, H_1) - \mathcal{F}(r_0, H_0)\\ &= \langle H_1\rangle_{r_1} - \langle H_0\rangle_{r_0} + C(r_0\|q_0) - C(r_1\|q_1).\end{aligned} \tag{12}$$

If, furthermore, both the initial and final Hamiltonians $H_0$ and $H_1$ are uniform over the space of allowed states, then expected work for initial distribution $r_0$ is $C(r_0||q_0) - C(r_1||q_1)$. (See [40] for an example of a physical system where this is the case.)

Finally, it is possible to generalize our main result in two important ways, as shown in appendix A. First, when the minimizer $q_0$ does *not* have full support, incorrect prior dissipation is lower-bounded by (rather than equal to) the contraction of KL divergence. In addition, our main result can be generalized to the case when $q_0$ is not the minimizer of $W_d$ over all possible initial distributions, but only within some convex subspace of distributions. Then, the result holds for any other initial distribution $r_0$ within the same subspace. The latter generalization is used in the next section to derive a coarse-grained version of equation (10).

[4] Indeed, though logical reversibility and thermodynamic 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 [45], assuming that the distribution over the states of the bit is exactly known to the designer.



J. Stat. Mech. (2017) 083202

## 5. Thermodynamics of computation

We now extend our main result, to apply to the thermodynamics of computation. Formally, this means that we analyze the implications of our main result for physical systems that perform information-processing operations over some coarse-grained degrees of freedom.

Recent advances in nonequilibrium statistical physics [11, 45] have extended and clarified the pioneering analysis of of Landauer, Bennett and others [46–49] regarding the fundamental thermodynamics cost of information processing. In this section, we consider the implications of incorrect prior dissipation for thermodynamics of computation.

In keeping with previous 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 $v \in V$ (the set of CMs are equivalent to what are called the 'information bearing degrees of freedom' in [50], and the 'information states' in [51]). $\mathcal{P}$ induces a stochastic dynamics over $X$, and the (possibly non-deterministic) computation is identified with the associated dynamics over CMs. We use $\pi(v_1|v_0)$ to indicate this dynamical process over CMs, i.e. to indicate a single iteration of the computation that maps inputs $v_0$ to outputs $v_1$. The canonical example of this kind of computation is a single iteration of a laptop, modifying the bit pattern in its memory (i.e. its CM) [50]. In practice, computers are usually designed to perform the same operation over their CMs from one iteration to the next. Formally, this means that their dynamics are first-order Markovian and time-homogeneous.

In previous work [41], we showed that for any given $\pi$ and input distribution, a computer can be designed that implements $\pi$ with zero dissipation for that input distribution. Here, we instead consider how the amount of dissipation for a fixed, given computer depends on the choice of input distribution. We recover a coarse-grained version of equation (10), expressing incorrect prior dissipation for a distribution over input CMs. Thus, the exact same equations that determine how dissipation varies with the initial distribution over *micro*states also determine how dissipation in a computer varies with the initial distribution over computational *macro*states.

Formally, let $g : X \to V$ be the coarse-graining function that maps the microstates of a computer to its CMs. Let $s(x|v)$ be a fixed distribution over the microstates corresponding to the specified macrostate $v$, and so obeys $s(x|v) = 0$ if $v \neq g(x)$. We use the random variables $V_0$ and $V_1$ to indicate the CM at the beginning and end of the process, respectively. To avoid confusion between distributions over CMs and those over microstates, distributions over CM are superscripted with a $V$. Thus, we write $p_0^V(.)$ and $p_1^V(.)$ to indicate the distribution over CMs at $t = 0$ and at $t = 1$, respectively, and similarly for $q_0^V, q_1^V, r_0^V$ and $r_1^V$.

When combined with the conditional update distribution $\pi(v_1|v_0)$, any initial distribution $p_0^V$ over CMs induces a final distribution over states of $V$ at $t = 1$ in the obvious way. Such a $p_0^V$ also induces a $t = 0$ mixture distribution over microstates, given by averaging the distributions $s(x|v)$ over all possible $v$. It will be useful to write this mixture with the shorthand

$$[\Phi(p^V)](x) = \sum_v s(x|v)\, p^V(v) = s(x|g(x))\, p^V(g(x)). \tag{13}$$

Thus, $\Phi(.)$ is a map that takes distributions over $V$ to distributions over $X$. The image of $\Phi$, $\mathcal{T}$, is a convex subset of the set of distributions over $X$, containing all possible mixtures of $s(x|v)$ induced by distributions over $V$.

We make two assumptions in our analysis of computers, which capture some physical properties of what is commonly meant by 'computers', both in the real world and in the literature on thermodynamics of computation.

First, we assume the initial distribution over microstates is determined by specifying the initial distribution over CMs. This assumption reflects the fact that in current real-world computers, the input is set by selecting some computational macrostate (e.g. setting the pattern of logical bits in memory). It is *not* selected by the user selecting a particular microstate of the system (which would occur, for example, if the user set the positions and momenta of all atoms and electrons in the computer). Formally, this assumption means that any allowed initial distribution $p_0$ must be an element of $\mathcal{T}$, and will satisfy $p_0(x_0) = [\Phi(p_0^V)](x_0)$ for some initial distribution over CMs $p_0^V$. We call such an initial distribution over CMs an input distribution.

Second, we assume that the distribution of microstates, conditioned on the respective macrostate, is the same at the beginning and end of the thermodynamic process. This assumption guarantees that the dynamics of a computer's logical state is first-order Markovian and time-homogeneous; in other words, the computer can be run for multiple iterations, and it is guaranteed to obey the same logical rules in each iteration. We formalize this assumption by requiring that the dynamics obey

$$p(x_1|v_0, v_1) = \frac{p(x_1, v_1|v_0)}{p(v_1|v_0)} = \frac{\sum_{x_0} \delta_{v_1, g(x_1)} p(x_1|x_0) s(x_0|v_0)}{p(v_1|v_0)} = s(x_1|v_1)$$

for all $x_1, v_0, v_1$. In words, this states 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 that provided by the fact that $x_1$ belongs to CM $v_1$). This assumption also means that as long as the initial microstate distribution $p_0$ is induced by some distribution over CMs, then the output microstate distribution $p_1$ is also induced by some distribution over CMs (i.e. that $p_1 \in \mathcal{T}$ so long as $p_0 \in \mathcal{T}$). We refer to any process that obeys this condition as computationally cyclic.

When the computer is run with the microstate distribution $\Phi(p_0^V)$, the amount of dissipated work is $W_d(\Phi(p_0^V))$. Accordingly we refer to $W_d(\Phi(p_0^V))$ as the *dissipation of the (macrostate) input distribution* $p_0^V$, and when clear from context, write it simply as $W_d(p_0^V)$.

In analogy with the case of dynamics over $X$, we say that an input distribution $q_0^V$ is a prior for the computer if it achieves minimum dissipation among all input distributions. As we did in our analysis of priors over microstates, we assume that the prior $q_0^V$ has full support. (More formally, see appendix D for a sufficient condition on $\pi$ under which this assumption will hold.)

Let $q_0 := \Phi(q_0^V) \in \mathcal{T}$ indicate the microstate distribution induced by $q_0^V$. By definition of $q_0^V$, $q_0$ has minimum dissipation within the convex set $\mathcal{T}$. Furthermore, by our assumption that $q_0^V$ has full support, $\Phi(q_0^V)$ will be in the relative interior of $\mathcal{T}$.

Now consider any other input distribution $r_0^V$, as well as its associated microstate distribution $r_0 := \Phi(r_0) \in \mathcal{T}$. Using the general statement of dissipation due to incorrect priors derived in appendix A,

$$\begin{aligned} W_d(\Phi(r_0^V)) - W_d(\Phi(q_0^V)) &= D(r_0\|q_0) - D(r_1\|q_1) \\ &= D(r(V_0,X_0)\|q(V_0,X_0)) - D(r(V_1,X_1)\|q(V_1,X_1)) \\ &= D(r_0^V\|q_0^V) - D(r_1^V\|q_1^V) \\ &\quad + D(r(X_0|V_0)\|q(X_0\|V_0)) - D(r(X_1|V_1)\|q(X_1|V_1)), \end{aligned}$$

where the second line follows because $v_0$ and $v_1$ are deterministic functions of $x_0$ and $x_1$, and the third line follows from the chain rule for KL divergence. Next, note that by definition $r(x_0|v_0) = s(x_0|v_0) = q(x_0|v_0)$. So $D(r(X_0|V_0)\|q(X_0|V_0)) = 0$. In addition, by the cyclic condition, $r(x_1|v_1) = \sum_{v_0} r(v_0|v_1)\, p(x_1|v_0,v_1) = s(x_1|v_1)$ and similarly for $q(x_1|v_1)$. So $D(r(X_1|V_1)\|q(X_1|V_1)) = 0$.

Combining leads to a coarse-grained version of our main result: for dynamics over CMs, incorrect prior dissipation for any input distribution $r_0^V$ is

$$W_d(r_0^V) - W_d(q_0^V) = D(r_0^V\|q_0^V) - D(r_1^V\|q_1^V).$$

The obvious analog of equation (12) (and the associated discussion) holds for 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 [41]. However the analysis here holds for any computer, no matter how it operates.

As before, if $q_0^V$ does not have full support, we recover an inequality rather than an equality appendix A.

## 6. Conclusion

For a fixed nonequilibrium process, we have quantified the additional dissipation arising from using some arbitrary initial distribution, relative to the dissipation incurred when using the initial distribution that achieves minimal dissipation. This additional dissipation has a simple, information-theoretic form, being equal to the the contraction of KL divergence between the actual and optimal initial distributions over the course of the process.

We also considered computers, i.e. processes that implement some stochastic map over a set of coarse-grained variables. We showed that our main result applies to distributions over coarse-grained states of a system, so long as the fine-grained dynamics obey several conditions. Landauer and co-workers pioneered analysis of the thermodynamic cost of computation; in its modern formulation, Landauer's bound considers the minimal (dissipation-free) total work needed to perform a given computation [9, 11]. Our result extends these analyses to include the dissipation cost of computation, and in particular its dependence on the initial distribution of the computer's states.

Our results are derived with few assumptions. They do not require that the dynamics obey local detailed balance, nor that they are Markovian. In addition, they hold for both quasi-static and finite time processes, and regardless of how far the process is from equilibrium.



J. Stat. Mech. (2017) 083202

## 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. TWCF0079/AB47 from the Templeton World Charity Foundation, Grant No. FQXi-RHl3-1349 from the FQXi foundation, and Grant No. CHE-1648973 from the US 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.

## Appendix A. Dissipation due to incorrect priors over convex spaces

Let $\Delta$ be a convex subset of the set of all distributions over state space $X$. For a given process $\mathcal{P}$, define the prior distribution in $\Delta$ as

$$q_0 := \arg\min_{p_0 \in \Delta} W_d(\mathcal{P}, p_0).$$

**Proposition 1.** *For all initial distributions $r_0 \in \Delta$,*

$$W_d(r_0) - W_d(q_0) \geqslant D(r_0 \| q_0) - D(r_1 \| q_1), \tag{A.1}$$

*where $D(\cdot\|\cdot)$ is the KL divergence. If the prior distribution $q_0$ is in the relative interior of $\Delta$, the inequality is tight:*

$$W_d(r_0) - W_d(q_0) = D(r_0 \| q_0) - D(r_1 \| q_1). \tag{A.2}$$

**Proof.** First use equation (4) in the main text to write the $|X|$ components of $\nabla W_d(\,p_0)$ as

$$\begin{aligned}\frac{\partial W_d}{\partial p(x_0)}(\,p_0) &= \Big[ -\sum_{x_1} p(x_1|x_0) \ln \sum_{x_0'} p_0(x_0')\, p(x_1|x_0') - 1 \Big] + \Big[ \ln p(x_0) + 1 \Big] - \mathcal{Q}(x_0) \\ &= -\sum_{x_1} p(x_1|x_0) \ln p_1(x_1) + \ln p_0(x_0) - \mathcal{Q}(x_0).\end{aligned} \tag{A.3}$$

Note that by equations (7) and (4) in the main text, even though $W_d(\,p_0)$ is not a linear function of $p_0$, it is still true that for any $p_0$,

$$W_d(\,p_0) = \sum_{x_0} p(x_0) \frac{\partial W_d}{\partial p(x_0)}(\,p_0) = p_0 \cdot \nabla W_d(\,p_0). \tag{A.4}$$

The prior $q_0$ minimizes $W_d$ in the convex space $\Delta$. Then, the directional derivative at $q_0$ toward $r_0 \in \Delta$, written as $(r_0 - q_0) \cdot \nabla W_d(q_0)$, must be non-negative, since otherwise $W_d$ could be decreased by slightly perturbing $q_0$ toward $r_0$. Thus,

$$(r_0 - q_0) \cdot \nabla W_d(q_0) \geqslant 0.$$

By equation (A.4),

$$r_0 \cdot \nabla W_d(q_0) \geqslant W_d(q_0).$$

Plugging in equation (A.3), we rewrite,

$$\begin{aligned} r_0 \cdot \nabla W_d(q_0) &= -\langle \mathcal{Q} \rangle_{r_0} - C(r_0 \| q_0) + C(r_1 \| q_1) \\ &= W_d(r_0) - \left[ D(r_0 \| q_0) - D(r_1 \| q_1) \right], \end{aligned} \tag{A.5}$$

where $C(\cdot\|\cdot)$ is the cross-entropy. Combining establishes the inequality proposition A.1.

If $q_0$ is in the relative interior of $\Delta$, the directional derivatives at $q_0$ must be positive toward *and* away-from $r_0$. Thus,

$$(r_0 - q_0) \cdot \nabla W_d(q_0) \geqslant 0 \text{ and } (q_0 - r_0) \cdot \nabla W_d(q_0) \geqslant 0,$$

leading to

$$(r_0 - q_0) \cdot \nabla W_d(q_0) = 0. \tag{A.6}$$

Combining equations (A.6), (A.4) and (A.5) establishes the equality proposition A.2. □

## Appendix B. Dissipated work is convex

First, consider two initial distributions, specified by the conditional probability distribution $w(X_0 = x_0|C = 0)$ and $w(X_0 = x|C = 1)$, as well as the mixture $w(X_0 = x) = \sum_c p(C = c)\, w(X = x|C = c)$. At the end of the process, these distributions are mapped to $w(X_1 = x|C = 0)$, $w(X_1 = x|C = 1)$, and $w(X_1 = x) = \sum_c p_C(C = c)\, w(X_1 = x|C = c)$.

To demonstrate that $W_d$ is convex, we will show that

$$p(C = 0)W_d(w(X_0|C = 0)) + p(C = 1)W_d(w(X_0|C = 1)) \geqslant W_d(w(X_0)).$$

First, we subtract the RHS from the LHS, while using the expression for dissipated work equation (4). The linear terms drop out, leaving the entropy terms:

$$\begin{aligned} & p(C = 0)W_d(w(X_0|C = 0)) + p(C = 1)W_d(w(X_0|C = 1)) - W_d(w(X_0)) \\ =\, & p(C = 0)\left[S(w(X_1|C = 0)) - S(w(X_0|C = 1))\right] \\ & + p(C = 1)\left[S(w(X_1|C = 1)) - S(w(X_0|C = 1))\right] \\ & - \left[S(w(X_1)) - S(w(X_0))\right] \\ =\, & MI(X_0; C) - MI(X_1; C) \\ \geqslant\, & 0. \end{aligned}$$

The last line follows from the data processing inequality for mutual information [26].

## Appendix C. Analysis of 'Maxwell's demon' model of Mandal and Jarzynski

We consider the work dissipated in the thermodynamic process corresponding to one 'interaction interval' of the information-processing 'demon' described in [32]. Let $X = \{A0, B0, C0, A1, B1, C1\}$ represent the state space of the model. Also let $V$ be a coarse-graining of $X$ into a binary state (corresponding to the state of the bit on the tape), where $V = 0$ corresponds to $\{A0, B0, C0\}$ and $V = 1$ corresponds to $\{A1, B1, C1\}$.

As in our main text, it will useful to distinguish distributions over $V$ from those over $X$ with a superscript, e.g. writing $r_0^V$ rather than $r_0$.

The model is parameterized by:

(i) $\tau$: the amount of time the demon interacts with each incoming bit, i.e. the length of a single interaction interval. In our framework, this means that $t \in [0,1]$ maps to a duration of physical time $\tau$.

(ii) $\delta$: set the 'excess' of 0s in the incoming bit distribution (i.e. the distribution of $V$ at the beginning of the interaction interval), via $\delta = r_0^V(V=0) - r_0^V(V=1)$.

(iii) $\epsilon$: set the 'excess' of 0 in the equilibrium distribution of outgoing bits (i.e. the distribution of $V$ at the end of an interaction interval as $\tau \to \infty$), via $\epsilon = p_{\text{eq}}^V(V=0) - p_{\text{eq}}^V(V=1)$ [32, equation (6b)].

The parameter $\epsilon$ is used to define a continuous-time $|X| \times |X|$ rate matrix $\mathscr{R}$ [32, equation (S1)] specifying system dynamics. This rate matrix is then used to define a transition matrix $\Pi^\tau := e^{\tau \mathscr{R}}$ for interactions of duration $\tau$.

In [32], it is noted that dissipation is 0 when $\delta = \epsilon$. In their Supporting Information, the authors provide a complex derivation showing that dissipation is non-negative for other cases. This is shown analytically for the quasi-static limit of interval lengths ($\tau \to \infty$, i.e. when each interaction interval takes an infinite amount of time), but only numerically for finite interval lengths [32, equation (S20)]. Here we show how to use the results in our paper to prove *strict* positivity simply, and analytically, for all time scales.

Note that $\mathscr{R}$ is irreducible, and hence has a unique stationary distribution, which we call $p_{\text{eq}}^X(x)$ ($p_{\text{eq}}^V$ is a marginalization of this stationary distribution onto the $V$ subspace). When the 6-state system is prepared with initial distribution $p_{\text{eq}}^X$, no work gets done [32] and the nonequilibrium free energy doesn't change, hence $W_d\left(p_{\text{eq}}^X\right)$ is 0. Therefore, in the language of our main text, $p_{eq}$ is a prior distribution for this thermodynamic process, since no other initial distribution can achieve lower dissipation.

Using our main result, we write dissipation when the system is prepared with initial distribution $r_0^X$ and allowed to interact for duration $\tau$ as

$$\begin{aligned} W_d\left(r_0^X\right) - W_d\left(p_{\text{eq}}^X\right) &= W_d\left(r_0^X\right) \\ &= D\left(r_0^X \| p_{\text{eq}}^X\right) - D\left(\Pi^\tau r_0^X \| \Pi^\tau p_{\text{eq}}^X\right) \\ &= D\left(r_0^X \| p_{\text{eq}}^X\right) - D\left(\Pi^\tau r_0^X \| p_{\text{eq}}^X\right) \\ &> 0 \quad \text{whenever } r_0^X \neq p_{\text{eq}}^X \end{aligned} \tag{C.1}$$

where the inequality arises from the fact that irreducible rate matrices have strict convergence to equilibrium [52, section 3.5].

Note that $\epsilon = \delta$ means that $r_0^V(v) = p_{\text{eq}}^V(v)$. Thus $\epsilon = \delta$ is a necessary condition for $r_0^X(x) = p_{\text{eq}}^X(x)$ (though not sufficient, since we would also need $r_0(x|v) = p_{\text{eq}}(x|v)$). We have thus shown that $\epsilon = \delta$ is a necessary condition for dissipation to be 0, and that when $\epsilon \neq \delta$, dissipation is guaranteed to be strictly positive.



J. Stat. Mech. (2017) 083202

Observe also that for any specific values $\tau, \epsilon, \delta$, and $r_0\,(x|v)$, we can use equation (C.1) above to compute dissipation exactly.

## Appendix D. Sufficient conditions for prior 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 equation (7) in the main text to write

$$\frac{\partial W_d}{\partial p(x_0)}(\,p_0) = -\mathcal{Q}(x_0) - \sum_{x_1} p(x_1|x_0) \ln\left[\frac{\sum_{x_0'} p(x_1|x_0')\,p(x_0')}{p(x_0)}\right] \tag{D.1}$$

for any distribution $p_0$ over $X$.

Define $q_0 := \arg\min_{p_0 \in \Delta} W_d(\,p_0)$, where $\Delta$ is the $|X|$-dimensional unit simplex. To show that $q_0$ has full support, hypothesize that there exists some $x_0^\star$ such that $q(x_0^\star) = 0$. Now consider the one-sided derivative $\frac{\partial W_d}{\partial p(x_0^\star)}(q_0)$. By the assumption that $p\,(x_1|x_0) > 0$ for all $x_0, x_1$, the numerator inside the logarithm in equation (D.1) is nonzero, while by hypothesis the denominator is 0. Thus, the argument of the logarithm is positive infinite and (since $\mathcal{Q}(x_0^\star)$ is finite, by assumption) $\frac{\partial W_d}{\partial p(x_0^\star)}(q_0)$ is negative infinite. Moreover, for any $x_0'$ where $q(x_0') > 0$, $\frac{\partial W_d}{\partial p_0(x_0')}(q_0)$ is finite. This means that $W_d(q_0)$ can be reduced by increasing $q(x_0^\star)$ 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 $q_0^V := \arg\min_{p_0^V \in \Delta^V} W_d(\Phi(\,p_0^V))$, where $\Delta^V$ 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, $\frac{\partial W_d(\Phi(p_V))}{\partial p_V(v_0)}$. Let $p_0 := \Phi(\,p_0^V)$, and then use the chain rule, equations (13), (D.1), and then (13) in the main text again to write

$$\begin{aligned}
\frac{\partial W_d(\Phi(p_V))}{\partial p_V(v_0)}\left(\,p_0^V\right) &= \sum_{x_0} \frac{\partial W_d}{\partial p(x_0)}(\,p_0)\,\frac{\partial[\Phi(p_V)](x_0)}{\partial p_V(v_0)}\left(\,p_0^V\right) \\
&= \sum_{x_0} \frac{\partial W_d}{\partial p(x_0)}(\,p_0)\;s\,(x_0|v_0) \\
&= \sum_{x_0}\left[-\mathcal{Q}(x_0) - \sum_{x_1} p(x_1|x_0) \ln \frac{\sum_{x_0'} p(x_1|x_0')\,p(x_0')}{p(x_0)}\right] s\,(x_0|v_0) \\
&= \sum_{x_0}\left[-\mathcal{Q}(x_0) - \sum_{x_1} p(x_1|x_0) \ln \frac{\sum_{x_0'} p(x_1|x_0')\,p(x_0')}{s\,(x_0|v_0)\;p_V(v_0)}\right] s\,(x_0|v_0)
\end{aligned} \tag{D.2}$$

for any distribution $p_V^0$ over $V$.

Proceeding as before, hypothesize that there exists some $v_0^\star$ such that $q_V(v_0^\star) = 0$, and use equation (D.2) to evaluate $\frac{\partial W_d(\Phi(p_V))}{\partial p(v_0^\star)}(q_0^V)$. By our hypothesis, the associated

value of the denominator in the logarithm in equation (D.2) 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 $\frac{\partial W_d(\Phi(p_V))}{\partial p(v_0^\star)}(q_0^V)$ is negative infinite. At the same time, $\frac{\partial W_d(\Phi(p_V))}{\partial p(v_0')}(q_0^V)$ is finite for any $v_0'$ where $q_V(v_0') > 0$. Thus $W_d(\Phi(q_0^V))$ can be reduced by increasing $q_V(v_0^\star)$ 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.

## Appendix E. Proof of strictly positive dissipation for non-invertible maps

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 $W_d(r_0) - W_d(q_0) > 0$ iff $p\,(x_1|x_0)$ is not an invertible map.*

**Proof.** If $q_0$ has full support, then $suppr_0 \subseteq suppq_0$ for all $r_0$. Then equation (10) states that if initial distribution $r_0$ is used, extra dissipation is equal to

$$\begin{aligned} W_d(r_0) - W_d(q_0) &= D(r_0\|q_0) - D(r_1\|q_1) \\ &= D(r(X_0|X_1)\|q(X_0|X_1)) \end{aligned} \tag{E.1}$$

KL divergence is invariant under invertible transformations. Therefore, if $p\,(x_1|x_0)$ is an invertible map, then $D(r_0\|q_0) = D(r_1\|q_1) \implies W_d(r_0) - W_d(q_0) = 0\ \forall\ r_0$.

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 $\mathbf{1}^T = (1, \ldots, 1)^T$ with eigenvalue 1.

Furthermore, it is known that if $M$ is not an invertible map, i.e. permutation matrix, then $|\det M| < 1$ [53]. 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 $\mathbf{s}$ represent the non-zero left eigenvector corresponding to $\lambda$. Note that due to biorthgonality of eigenvectors, $\mathbf{s}\mathbf{1}^T = 0$. We use $s(x)$ to refer to elements of $\mathbf{s}$ indexed by $x \in X$. Without loss of generality, assume $\mathbf{s}$ 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 $\mathbf{s}$ and because $\mathbf{s}\mathbf{1}^T = 0$, $r_0$ is a valid probability distribution.

We use the notation $s\,(x_1) := \sum_{x_0} s\,(x_0)\ p\,(x_1|x_0)$ and $r\,(x_1) := \sum_{x_0} r\,(x_0)\ p\,(x_1|x_0) = q\,(x_1) + s\,(x_1)$. We also use the notation $\mathcal{C} := suppr_1$. The fact that $q_0$ has full support also means that $\mathcal{C} \subseteq suppq_1$.



J. Stat. Mech. (2017) 083202

The proof proceeds by contradiction. Assume that $W_d(r_0) - W_d(q_0) = 0$. Using equation (E.1) and due to properties of KL divergence, this means that for each $x_0 \in X$ and $x_1 \in \mathcal{C}$,

$$\begin{aligned}
q\left(x_0|x_1\right) &= r\left(x_0|x_1\right) \\
\frac{q\left(x_0\right)\, p\left(x_1|x_0\right)}{q\left(x_1\right)} &= \frac{r\left(x_0\right)\, p\left(x_1|x_0\right)}{r\left(x_1\right)} \\
\frac{r\left(x_1\right)}{q\left(x_1\right)} p\left(x_1|x_0\right) &= \frac{r\left(x_0\right)}{q\left(x_0\right)} p\left(x_1|x_0\right) \\
\frac{q\left(x_1\right) + s\left(x_1\right)}{q\left(x_1\right)} p\left(x_1|x_0\right) &= \frac{q\left(x_0\right) + s\left(x_0\right)}{q\left(x_0\right)} p\left(x_1|x_0\right) \\
\frac{s\left(x_1\right)}{q\left(x_1\right)} p\left(x_1|x_0\right) &= \frac{s\left(x_0\right)}{q\left(x_0\right)} p\left(x_1|x_0\right) \\
s\left(x_1\right) q\left(x_0|x_1\right) &= s\left(x_0\right)\, p\left(x_1|x_0\right)
\end{aligned}$$

Taking absolute value of both sides gives

$$\left|s\left(x_1\right)\right| q\left(x_0|x_1\right) = \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right)$$

Summing over $x_0 \in X$ and $x_1 \in \mathcal{C}$,

$$\begin{aligned}
\sum_{x_1 \in \mathcal{C}} \sum_{x_0 \in X} \left|s\left(x_1\right)\right| q\left(x_0|x_1\right) &= \sum_{x_1 \in \mathcal{C}} \sum_{x_0 \in X} \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right) \\
\sum_{x_1 \in X} \left|s\left(x_1\right)\right| - \sum_{x_1 \notin \mathcal{C}} \left|s\left(x_1\right)\right| &= \sum_{x_1 \in X} \sum_{x_0 \in X} \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right) \\
&\quad - \sum_{x_1 \notin \mathcal{C}} \sum_{x_0 \in X} \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right)
\end{aligned} \tag{E.2}$$

Note that for all $x_1 \notin \mathcal{C}$, $r\left(x_1\right) = 0$, meaning that $s\left(x_1\right) = -q\left(x_1\right)$. Thus,

$$\sum_{x_1 \notin \mathcal{C}} \left|s\left(x_1\right)\right| = \sum_{x_1 \notin \mathcal{C}} q\left(x_1\right)$$

Furthermore, for all $x_1 \notin \mathcal{C}$, $r\left(x_1\right) = \sum_{x_0} r\left(x_0\right)\, p\left(x_1|x_0\right) = 0$. Thus, for all $x_0 \in X$ where $p\left(x_1|x_0\right) > 0$ for some $x_1 \notin \mathcal{C}$, $r\left(x_0\right) = 0$, meaning $s\left(x_0\right) = -q\left(x_0\right)$. This allows us to rewrite the last term in equation (E.2) as

$$\begin{aligned}
&\sum_{x_1 \notin \mathcal{C}} \sum_{x_0 \in X} \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right) \\
&= \sum_{x_1 \notin \mathcal{C}} \sum_{x_0 : p(x_1|x_0) > 0} \left|s\left(x_0\right)\right|\, p\left(x_1|x_0\right) \\
&= \sum_{x_1 \notin \mathcal{C}} \sum_{x_0 : p(x_1|x_0) > 0} q\left(x_0\right)\, p\left(x_1|x_0\right) \\
&= \sum_{x_1 \notin \mathcal{C}} q\left(x_1\right)
\end{aligned}$$



J. Stat. Mech. (2017) 083202

Cancelling terms that equal $\sum_{x_1 \notin \mathcal{C}} q\left(x_1\right)$ from both sides of equation (E.2), we rewrite

$$\sum_{x_1} \left|s\left(x_1\right)\right| = \sum_{x_1} \sum_{x_0} \left|s\left(x_0\right)\right| \, p\left(x_1|x_0\right) = \sum_{x_0} \left|s\left(x_0\right)\right| \tag{E.3}$$

In matrix notation, equation (E.3) states that

$$\left\|\mathbf{s}M\right\|_1 = \left\|\mathbf{s}\right\|_1 \tag{E.4}$$

where $\left\|\cdot\right\|_1$ indicates the vector $\ell_1$ norm. However, by definition $\mathbf{s}M = \lambda\mathbf{s}$. Hence,

$$\left\|\mathbf{s}M\right\|_1 = \left\|\lambda\mathbf{s}\right\|_1 = \left|\lambda\right| \left\|\mathbf{s}\right\|_1 < \left\|\mathbf{s}\right\|_1$$

meaning that equation (E.4) 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 $W_d(r_0) - W_d(q_0) > 0$. □

## References


[1] Touchette H and Lloyd S 2004 *Physica* A **331** 140–72
[2] Sagawa T and Ueda M 2009 *Phys. Rev. Lett.* **102** 250602
[3] Dillenschneider R and Lutz E 2010 *Phys. Rev. Lett.* **104** 198903
[4] Sagawa T and Ueda M 2012 *Phys. Rev. Lett.* **109** 180602
[5] Crooks G E 1999 *Phys. Rev.* E **60** 2721
[6] Crooks G E 1998 *J. Stat. Phys.* **90** 1481–7
[7] Chejne Janna F, Moukalled F and Gómez C A 2013 *Int. J. Thermodyn.* **16** 97–101
[8] Jarzynski C 1997 *Phys. Rev. Lett.* **78** 2690
[9] Esposito M and Van den Broeck C 2011 *Europhys. Lett.* **95** 40004
[10] Esposito M and Van den Broeck C 2010 *Phys. Rev.* E **82** 011143
[11] Parrondo J M, Horowitz J M and Sagawa T 2015 *Nat. Phys.* **11** 131–9
[12] Pollard B S 2016 *Open Syst. Inf. Dyn.* **23** 1650006
[13] Wiesner K, Gu M, Rieper E and Vedral V 2012 *Proc. R. Soc.* A **468** 4058–66
[14] Still S, Sivak D A, Bell A J and Crooks G E 2012 *Phys. Rev. Lett.* **109** 120604
[15] Prokopenko M, Lizier J T and Price D C 2013 *Entropy* **15** 524–43
[16] Prokopenko M and Lizier J T 2014 *Sci. Rep.* **4** 5394
[17] Dunkel J 2014 *Nat. Phys.* **10** 409–10
[18] Roldán É, Martinez I A, Parrondo J M and Petrov D 2014 *Nat. Phys.* **10** 457–461
[19] Bérut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R and Lutz E 2012 *Nature* **483** 187–9
[20] Hasegawa H H, Ishikawa J, Takara K and Driebe D 2010 *Phys. Lett.* A **374** 1001–4
[21] Takara K, Hasegawa H H and Driebe D 2010 *Phys. Lett.* A **375** 88–92
[22] Deffner S and Lutz E 2012 arXiv:1201.3888
[23] Seifert U 2005 *Phys. Rev. Lett.* **95** 040602
[24] Jarzynski C 2006 *Phys. Rev.* E **73** 046105
[25] Seifert U 2012 *Rep. Prog. Phys.* **75** 126001
[26] Cover T M and Thomas J A 2012 *Elements of Information Theory* (New York: Wiley)
[27] Mackay D 2003 *Information Theory, Inference, and Learning Algorithms* (Cambridge: Cambridge University Press)
[28] Schmiedl T and Seifert U 2007 *Phys. Rev. Lett.* **98** 108301
[29] Sivak D A and Crooks G E 2012 *Phys. Rev. Lett.* **108** 190602
[30] Aurell E, Gawędzki K, Mejía-Monasterio C, Mohayaee R and Muratore-Ginanneschi P 2012 *J. Stat. Phys.* **147** 487–505
[31] Funo K, Shitara T and Ueda M 2016 *Phys. Rev.* E **94** 062112
[32] Mandal D and Jarzynski C 2012 *Proc. Natl Acad. Sci.* **109** 11641–5
[33] Kawai R, Parrondo J M R and Van den Broeck C 2007 *Phys. Rev. Lett.* **98** 080602
[34] Hatano T 1999 *Phys. Rev.* E **60** R5017




J. Stat. Mech. (2017) 083202

[35] Chernyak V Y, Chertkov M and Jarzynski C 2006 *J. Stat. Mech.* P08001
[36] Gomez-Marin A, Parrondo J and Van den Broeck C 2008 *Europhys. Lett.* **82** 50002
[37] Parrondo J M, Van den Broeck C and Kawai R 2009 *New J. Phys.* **11** 073008
[38] Deffner S and Lutz E 2011 *Phys. Rev. Lett.* **107** 1079–7114
[39] Maroney O 2009 *Phys. Rev.* E **79** 031105
[40] Wolpert D H 2016 *Entropy* **18** 138
[41] Wolpert D H 2016 Extending landauer's bound from bit erasure to arbitrary computation (arXiv:1508.05319)
[42] Ahlswede R and Gács P 1976 *Ann. Probab.* **4** 925–39
[43] Cohen J E, Iwasa Y, Rautu G, Beth Ruskai M, Seneta E and Zbaganu G 1993 *Linear Algebra Appl.* **179** 211–35
[44] Csiszar I and Körner J 2011 *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Cambridge: Cambridge University Press)
[45] Sagawa T 2014 *J. Stat. Mech.* P03025
[46] Landauer R 1961 *IBM J. Res. Dev.* **5** 183–91
[47] Bennett C H 1982 *Int. J. Theor. Phys.* **21** 905–40
[48] Zurek W H 1989 *Nature* **341** 119–24
[49] Zurek W H 1989 *Phys. Rev.* A **40** 4731–51
[50] Bennett C H 2003 *Stud. Hist. Phil. Sci.* B **34** 501–10
[51] Deffner S and Jarzynski C 2013 *Phys. Rev.* X **3** 041003
[52] Greven A, Keller G and Warnecke G 2003 *Entropy* (Princeton, NJ: Princeton University Press)
[53] Goldberg K 1966 *J. Res. Natl Bur. Stand.* B **70B** 157–8

J. Stat. Mech. (2017) 083202