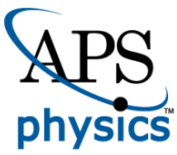


Modeling Maxwell's demon with a microcanonical Szilard engine

Suriyanarayanan Vaikuntanathan and Christopher Jarzynski

Phys. Rev. E **83**, 061120 — Published 15 June 2011

DOI: 10.1103/PhysRevE.83.061120

Modeling Maxwell's demon with a microcanonical Szilard engine

Suriyanarayanan Vaikuntanathan[1] and Christopher Jarzynski[1,2]

[1] *Chemical Physics Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742*

[2] *Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742*

Following recent work by Marathe and Parrondo [PRL, **104**, 245704 (2010)], we construct a classical Hamiltonian system whose energy is reduced during the adiabatic cycling of external parameters, when initial conditions are sampled microcanonically. Combining our system with a device that measures its energy, we propose a cyclic procedure during which energy is extracted from a heat bath and converted to work, in apparent violation of the second law of thermodynamics. This paradox is resolved by deriving an explicit relationship between the average work delivered during one cycle of operation, and the average information gained when measuring the system's energy.

I. INTRODUCTION

The Kelvin-Planck statement of the second law of thermodynamics asserts that no process is possible whose sole result is the extraction of energy from a heat bath, and the conversion of that energy into work. [1] Because this statement is formulated in terms of energy rather than entropy, it provides an attractive starting point for exploring the microscopic foundations of the second law. This is particularly true when we consider an immediate corollary of the Kelvin-Planck statement: when a thermally isolated system, initially in equilibrium, evolves under a cyclic variation of external parameters, its internal energy cannot decrease.[1] Since an isolated system exchanges no heat with its surroundings, and is governed by familiar equations of motion – Hamiltonian dynamics in the classical case, or the Schrödinger equation for a non-relativistic quantum system – relatively few theoretical tools are needed to embark on an investigation of this statement.

Let us formulate the problem as follows. A finite, classical system is described by a Hamiltonian $H(\mathbf{z};\vec{\lambda})$, where $\mathbf{z} = (q,p)$ denotes a point in $2D$-dimensional phase space, and $\vec{\lambda} = (\lambda_1, \cdots, \lambda_n)$ is a set of externally controlled parameters. At time $t = 0$ the system's initial conditions are sampled from an equilibrium distribution $p^{\rm eq}(\mathbf{z})$, and then for $0 \le t \le \tau$ the system evolves under Hamilton's equations as the parameters are made to trace out a closed loop in $\vec{\lambda}$-space. We will use the notation $\vec{\lambda}_c(t)$ to denote such a cyclic protocol for varying the parameters, beginning and ending at $\vec{\lambda}^A \equiv \vec{\lambda}_c(0) = \vec{\lambda}_c(\tau)$. The work performed on the system during this process is the net change in the value of the Hamiltonian,

$$W = H(\mathbf{z}_\tau;\vec{\lambda}^A) - H(\mathbf{z}_0;\vec{\lambda}^A), \tag{1}$$

where the trajectory $\mathbf{z}_t$ describes the system's evolution from $t = 0$ to $t = \tau$. Since Hamiltonian dynamics are deterministic, the value of W is fully determined by the initial conditions: $W = W(\mathbf{z}_0)$. The Kelvin-Planck statement, viewed as a statistical prediction about averages, then implies the inequality,

$$\langle W\rangle \equiv \int {\rm d}\mathbf{z}_0\, p^{\rm eq}(\mathbf{z}_0)\, W(\mathbf{z}_0) \ge 0. \tag{2}$$

We now ask, for what choices of the equilibrium distribution $p^{\rm eq}(\mathbf{z})$ can this result be established rigorously?

When initial conditions are sampled from a canonical distribution

$$p^{\rm eq}_{can}(\mathbf{z}) \propto \exp\left[-\beta H_A(\mathbf{z})\right] \qquad , \qquad H_A(\mathbf{z}) \equiv H(\mathbf{z};\vec{\lambda}^A), \tag{3}$$

[1] If its energy were to decrease, then at the end of the process the system could be returned to its initial state by equilibrating it with a heat bath at temperature T, resulting in the net conversion of heat to work.

Eq. 2 follows directly from the properties of Hamilton's equations [2–4]. In fact, this result extends to any distribution of initial conditions that is a decreasing function of energy [3, 4]. Somewhat surprisingly, however, Eq. 2 is not universally valid when initial conditions are sampled from a microcanonical distribution,

$$p^{\rm eq}_{\mu can}(\mathbf{z}) \propto \delta\left[E_i - H_A(\mathbf{z})\right] \tag{4}$$

This has been discussed by Allahverdyan and Nieuwenhuizen [3], but to the best of our knowledge it was Sato [5] who first constructed a counter-example, involving a perturbed, one-dimensional harmonic oscillator. For microcanonically sampled initial conditions, Sato described a cyclic variation of the Hamiltonian that results in a negative value of average work, $\langle W \rangle < 0$. More recently, Marathe and Parrondo [6] have developed another counterexample to Eq. 2, involving a particle inside a box with hard walls and an insertable barrier. For a given initial energy, Marathe and Parrondo describe a cyclic manipulation of the walls and the barrier, whose net effect is to reduce the energy of the system. Ultimately, the particle can be brought arbitrarily close to zero kinetic energy by a succession of such cycles, with a different protocol for each cycle.

Inspired by Ref. [6], in the present paper we introduce and analyze another model system that violates Eq. 2. We consider a classical particle moving in a one-dimensional potential well, described by a pair of external parameters $\vec{\lambda} = (\lambda_L, \lambda_R)$ (see Eq. 5 and Fig. 1). We will discuss the design of protocols for varying these parameters cyclically with time, $\vec{\lambda}_c(t)$, in a manner that lowers the energy of the system. In particular, for any choice of initial particle energy E_i, we will construct a protocol (which depends on the value of E_i) that reduces the particle's kinetic energy arbitrarily close to zero in a single cycle, bringing the system to a final state in which the particle sits nearly motionless at the bottom of the potential well. In effect, the system is cooled near to "absolute zero" temperature.

Our model, like those of Refs. [5, 6], suggests that a perpetual-motion device of the second kind could be constructed, operating by the following steps.

1. The system is brought into contact and allowed to equilibrate with a thermal reservoir at temperature T. The reservoir is then removed.

2. The energy of the now-isolated system is measured.

3. The system is subjected to a cyclic protocol that reduces its kinetic energy close to zero (as discussed above).

By repeatedly performing this sequence of steps, we obtain a scenario in which energy is systematically extracted from the reservoir (step 1) and delivered as work to the agent that carries out the cyclic protocol (step 3). This is reminiscent of Maxwell's demon [7–9], only here the demon's role is to implement a cyclic protocol $\vec{\lambda}_c(t)$ based on the measured energy of the system, instead of opening or closing a trapdoor based on the observed motion of nearby particles. The key to exorcising the demon – that is, to reconciling this scenario with the second law of thermodynamics – is to recognize that the repeated measurements of energy in step 2 result in the accumulation of information. In order for the device to satisfy the "sole result" stipulation of the Kelvin-Planck statement (see above), this information must eventually be erased. As famously discussed by Landauer [10], and by Bennett [11] in the context of Szilard's engine [12] – another incarnation of Maxwell's demon – the erasure of information carries an unavoidable thermodynamic cost of $k_B T \ln 2$ per bit. We will show by explicit calculation that this cost ultimately wipes out any gains made by our device: in the process of erasing the accumulated information, all of the work harvested by the device is returned as heat to the thermal reservoir.

In Sec. II we introduce our model and discuss protocols $\vec{\lambda}_c(t)$ that reduce the energy of the system. In Sec. III we discuss the average amount of work that is extracted per cycle, when carrying out the three-step procedure discussed above; this amount depends on the precision with which the initial energy is measured in step 2. Using Landauer's principle for the work that must eventually be expended to erase the accumulated information ($k_B T \ln 2$ per bit), we will show that this is no less than the work extracted in step 3, regardless of the precision with which the initial energy is measured. Thus in the final accounting, after all the bits of information are reset to zero, the device is unable to deliver work and the second law is rescued from the demon.

II. MODEL AND PROTOCOLS

Consider a classical particle of unit mass moving in one dimension, governed by a Hamiltonian

$$H(\mathbf{z};\vec{\lambda}) = \frac{p^2}{2} + U(q;\vec{\lambda}) \equiv \frac{p^2}{2} + q^4 - \begin{cases} \lambda_L\, q^2 & \text{if } q \le 0 \\ \lambda_R\, q^2 & \text{if } q \ge 0 \end{cases} \tag{5}$$

where $\mathbf{z} = (q,p)$ is a point in the phase space of the particle, and $\vec{\lambda} = (\lambda_L, \lambda_R)$ is a point in two-dimensional parameter space, with $\lambda_L, \lambda_R \ge 0$. The parameter λ_L modulates the shape of the potential energy function in the region $q < 0$: when $\lambda_L > 0$, there is a local minimum at $q_L^{\min} = -\sqrt{\lambda_L/2}$, as illustrated in Fig. 1. Similarly, the value of λ_R specifies a minimum at

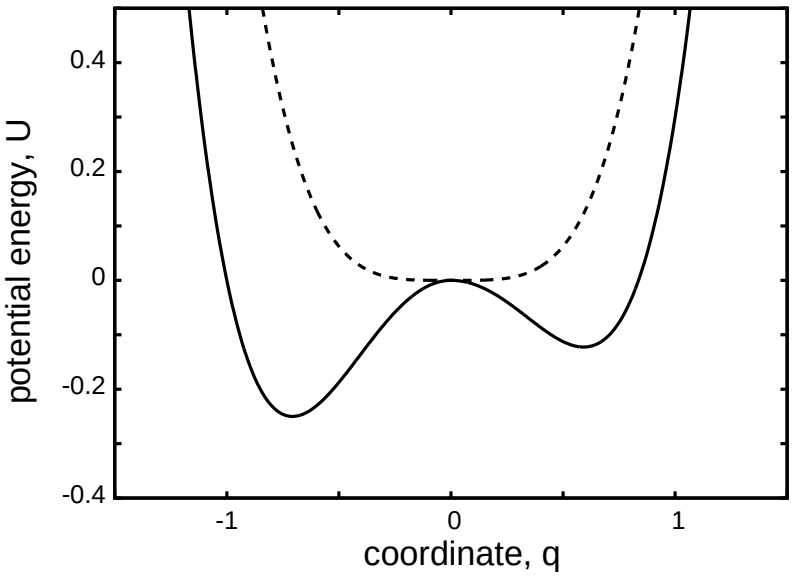
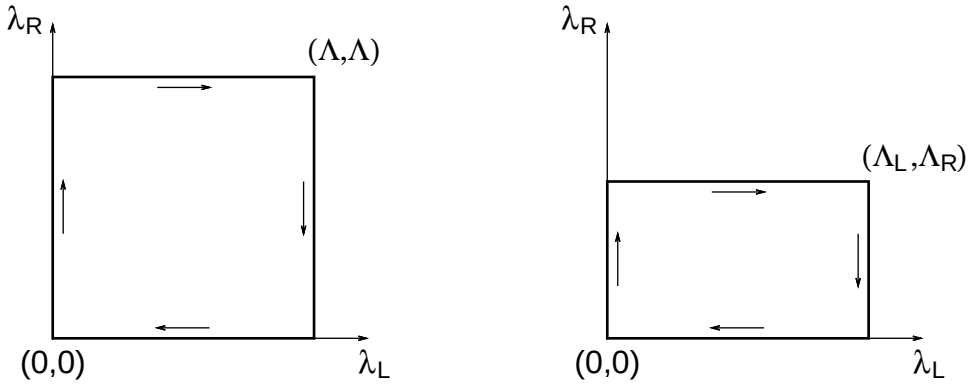


FIG. 1. The solid curve depicts the potential $U(q; 1.0, 0.7)$, with local minima at $q_L^{\min} = -\sqrt{0.5}$ and $q_R^{\min} = +\sqrt{0.35}$ (see text). The dashed curve is the unperturbed, quartic potential $U(q; 0, 0)$.



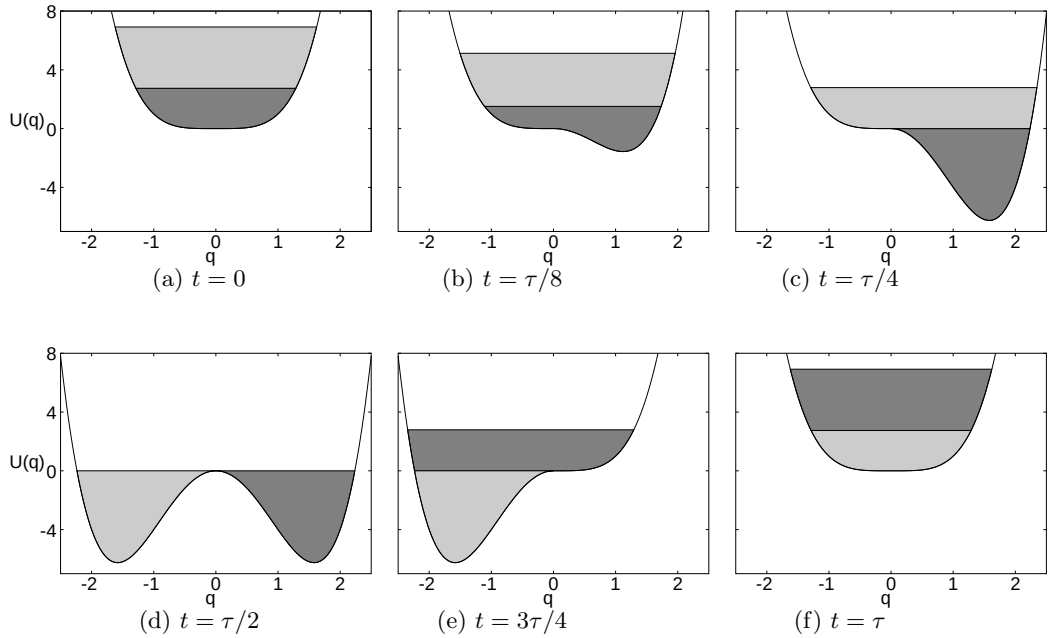
(a) Symmetric protocol.

(b) Asymmetric protocol.

FIG. 2. The cyclic protocols $\vec{\lambda}_c(t)$, depicted here, proceed clockwise from the origin.

$q_R^{\min} = +\sqrt{\lambda_R/2}$. We will refer to these regions as the left well and the right well. When $\vec{\lambda} = (0, 0)$, the particle moves in a quartic potential, which we call the unperturbed system.

Now imagine a protocol $\vec{\lambda}_c(t)$ whereby the parameters are made to trace out the perimeter of the square shown in Fig. 2(a), starting and ending at $\vec{\lambda} = (0, 0)$. For simplicity we assume a constant speed, $|\mathrm{d}\vec{\lambda}/\mathrm{d}t| = 4\Lambda/\tau$. The deformation of the potential during this protocol can be pictured as follows. Starting from the unperturbed quartic potential, the right well gradually drops down,



(a) $t = 0$ (b) $t = \tau/8$ (c) $t = \tau/4$

(d) $t = \tau/2$ (e) $t = 3\tau/4$ (f) $t = \tau$

FIG. 3. Snapshots of the potential energy function as $\vec{\lambda}$ is varied according to the protocol shown in Fig. 2(a), with $\Lambda = 5.0$ (hence $E_1 = 2.744$ and $E_2 = 6.914$, see Eq. 6). The shaded regions illustrate the evolution of sets I and II, in the quasi-static limit $\tau \to \infty$.

forming a local minimum that moves from the origin to $\sqrt{\Lambda/2}$ (see Fig. 3(a) - 3(c)) as λ_R increases from 0 to Λ. Next, as λ_L increases from 0 to Λ the left well drops down, forming a local minimum that comes to rest at $-\sqrt{\Lambda/2}$, with a local maximum at the origin (Fig. 3(d)). These two stages are then undone (Figs. 3(e), 3(f)). The net effect is a piston-like pumping of the right and left wells. For this protocol, let $\mathbf{z}_t$ denote a trajectory evolving under the time-dependent Hamiltonian $H(\mathbf{z}; \lambda_c(t))$.

For a given choice of Λ, let us define two energy values,

$$E_1(\Lambda) = \left(\frac{1}{3I_0}\right)^{4/3} \Lambda^2 \qquad , \qquad E_2(\Lambda) = \left(\frac{2}{3I_0}\right)^{4/3} \Lambda^2, \tag{6}$$

where

$$I_0 = \int_{-1}^{+1} \mathrm{d}y \, \sqrt{1 - y^4} = \frac{\sqrt{\pi}\,\Gamma(5/4)}{\Gamma(7/4)} \approx 1.74804. \tag{7}$$

These in turn define three regions of phase space, I, II, and III, according to the value of the

unperturbed Hamiltonian $H_0(\mathbf{z}) \equiv p^2/2 + q^4$:

$$\begin{aligned} I &: 0 < H_0(\mathbf{z}) < E_1 \\ II &: E_1 < H_0(\mathbf{z}) < E_2 \\ III &: E_2 < H_0(\mathbf{z}) \end{aligned} \tag{8}$$

We now claim that when the protocol $\vec{\lambda}_c(t)$ shown in Fig. 2(a) is implemented quasi-statically, $\tau \to \infty$, then the net effect is to swap regions I and II. That is, trajectories with initial conditions $\mathbf{z}_0$ in region I end with final conditions $\mathbf{z}_\tau$ in region II, and vice-versa. (See, however, the discussion of subtleties associated with this limit, in Sec. IV.) Fig. 3 and the following paragraphs convey how this swap proceeds. For convenience, we will use the terms *set I* and *set II* to refer to trajectories with initial conditions in regions I and II of phase space, respectively. The shaded regions in Fig. 3 depict the evolution of these sets of trajectories, as a sequence of snapshots from $t = 0$ to $t = \tau$.

By Hamilton's equations we have

$$\frac{\mathrm{d}}{\mathrm{d}t} H(\mathbf{z}_t; \vec{\lambda}_t) = \frac{\mathrm{d}\vec{\lambda}}{\mathrm{d}t} \cdot \frac{\partial H}{\partial \vec{\lambda}}(\mathbf{z}_t; \vec{\lambda}_t) = -q_t^2 \left[\dot{\lambda}_L \, \theta(-q_t) + \dot{\lambda}_R \, \theta(+q_t) \right] \tag{9}$$

where $\theta(\cdot)$ is the unit step function. During the first stage of the process, $0 < t < \tau/4$, we have $\dot{\lambda}_L = 0$ and $\dot{\lambda}_R > 0$, therefore as the right well drops down the value of $H(\mathbf{z}_t; \vec{\lambda}_t)$ decreases whenever $q_t > 0$. As a result, some trajectories acquire negative energies ($H < 0$) and become trapped in the right well. As shown in Fig. 3(c) – and as justified quantitatively by Eqs. 10 - 15 below – at the end of this stage the trajectories belonging to set I are trapped.

During the second stage, $\tau/4 < t < \tau/2$, the left well drops down, trapping the trajectories in set II. As this occurs, the trajectories in set I remain trapped in the right well.

From $\tau/2 < t < 3\tau/4$, as the right well rises and ultimately disappears, the trajectories in set I gain energy (Fig. 3(e)), and during the fourth and final stage, $3\tau/4 < t < \tau$, all trajectories gain energy as the left well gradually rises until it disappears. The situation at $t = \tau$, shown in Fig. 3(f), reflects the swap that has occurred between sets I and II, relative to Fig. 3(a).

Due to adiabatic averaging, the energy-ordering of the trajectories within each set remains fixed in the quasi-static limit: if we were to subdivide the lightly shaded region II in Fig. 3(a) into a stack of narrow horizontal bands, then the vertical ordering of these bands would remain unchanged throughout the process.

A proper analysis of this process involves the theory of adiabatic invariants, with careful attention paid to the phase space separatrix that is present during the interval $\tau/4 < t < 3\tau/4$, when $U(q)$ has a local maximum at $q = 0$ [13, 14]. However, the essence of what occurs should be

intuitively clear from the above discussion. A useful analogy is provided by imagining a container initially filled with three layers of a viscous, incompressible fluid, labeled I, II and III in vertically ascending order. Two syringes are attached to the bottom of the container. First one syringe extracts the lowest layer I of the fluid, bringing layer II to the bottom of the container. Next, the other syringe extracts layer II. Then the fluid layers are re-injected in the same order in which they were removed, resulting in the rearrangement of these layers.

The incompressibility of the fluid in this analogy corresponds to Liouville's theorem: phase space volume is preserved under Hamiltonian dynamics. To justify quantitatively our assertion that the protocol $\vec{\lambda}_c(t)$ swaps regions I and II, we must show that the phase space volumes corresponding to the darkly shaded regions in Figs. 3(a) and Figs. 3(d) are equal (in other words, it is precisely the trajectories in set I that get trapped in the right well), and similarly that the phase space volumes of the lightly shaded regions in Figs. 3(a) and Figs. 3(d) are equal.

Let $\Omega(E;\vec{\lambda})$ denote the volume of phase space enclosed by the surface $H(\mathbf{z};\vec{\lambda}) = E$:

$$\begin{aligned}\Omega(E;\vec{\lambda}) &= \int \mathrm{d}\mathbf{z}\ \theta\left[E - H(\mathbf{z};\vec{\lambda})\right] \\ &= \int_{E>U} \mathrm{d}q\ \sqrt{8\left[E - U(q;\vec{\lambda})\right]}\end{aligned} \tag{10}$$

where we have integrated over momentum to get to the second line. When either $E = 0$ or $\vec{\lambda} = \vec{0}$ the remaining integral can be evaluated analytically:

$$\Omega(E;\vec{0}) = \int_{-E^{1/4}}^{+E^{1/4}} \mathrm{d}q\ \sqrt{8(E-q^4)} = \sqrt{8}E^{3/4}I_0 \tag{11a}$$

$$\Omega(0;\vec{\lambda}) = \int_{-\sqrt{\lambda_L}}^{\sqrt{\lambda_R}} \mathrm{d}q\ \sqrt{-8U(q;\vec{\lambda})} = \sqrt{\frac{8}{9}}\left(\lambda_L^{3/2} + \lambda_R^{3/2}\right) = \Omega_L + \Omega_R \tag{11b}$$

with I_0 given by Eq. 7. The quantity

$$\Omega_L(\lambda_L) \equiv \sqrt{\frac{8}{9}}\,\lambda_L^{3/2} \tag{12}$$

is the volume of phase space for which $H < 0$ and $q < 0$, and $\Omega_R(\lambda_R)$ is defined similarly for $H < 0$ and $q > 0$.

Using Eq. 11a, the phase space volumes of regions I and II, defined by Eq. 8, are

$$\Omega_I = \sqrt{8}E_1^{3/4}I_0 \qquad , \qquad \Omega_{II} = \sqrt{8}\left(E_2^{3/4} - E_1^{3/4}\right)I_0 \tag{13}$$

In Fig. 3(d) the lightly and darkly shaded regions correspond to phase space volumes $\Omega_L(\Lambda)$ and $\Omega_R(\Lambda)$, respectively, which are equal in value:

$$\Omega_L(\Lambda) = \Omega_R(\Lambda) = \sqrt{\frac{8}{9}}\,\Lambda^{3/2} \tag{14}$$

Combining these results with Eq. 6 we find that

$$\Omega_I = \Omega_R(\Lambda) \qquad , \qquad \Omega_{II} = \Omega_L(\Lambda) \tag{15}$$

This establishes that our qualitative description of what occurs during this process, as illustrated in Fig. 3, is indeed consistent with the preservation of phase space volume, as mandated by Liouville's theorem.

The picture developed in the preceding paragraphs suggests the following relationship between the initial (E_i) and final (E_f) energy of the system, in the limit $\tau \to \infty$:

$$\Omega(E_f;\vec{0}) = \Omega(E_i;\vec{0}) + \Omega_{II} \qquad \text{if} \quad 0 < E_i < E_1 \tag{16a}$$

$$\Omega(E_f;\vec{0}) = \Omega(E_i;\vec{0}) - \Omega_I \qquad \text{if} \quad E_1 < E_i < E_2 \tag{16b}$$

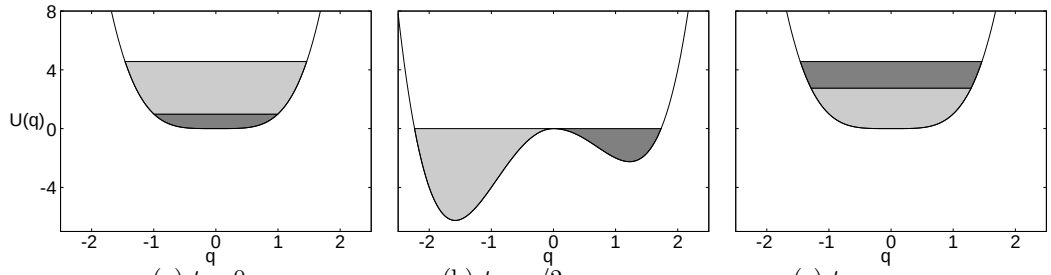
$$\Omega(E_f;\vec{0}) = \Omega(E_i;\vec{0}) \qquad \text{if} \quad E_2 < E_i \tag{16c}$$

with E_1 and $E_2 = 2^{4/3}E_1$ determined by the value of Λ (Eq. 6). Combining these results with Eq. 13 (note that $\Omega_I = \Omega_{II}$) we obtain

$$E_f = \begin{cases} \left(E_i^{3/4} + E_1^{3/4}\right)^{4/3} & \text{if} \quad 0 < E_i < E_1 \\ \left(E_i^{3/4} - E_1^{3/4}\right)^{4/3} & \text{if} \quad E_1 < E_i < E_2 \\ E_i & \text{if} \quad E_2 < E_i \end{cases} \tag{17}$$

As a test of Eq. 17, we sampled 10^5 initial conditions $\mathbf{z}_0 = (q_0, p_0)$ from a microcanonical ensemble at energy $E_i = H_0(\mathbf{z}_0) = 2.8$, near the bottom of region II (see Fig. 3). For each initial condition $\mathbf{z}_0$ we generated a trajectory $\mathbf{z}_t$ by integrating Hamilton's equations as the parameters were varied as in Fig. 2(a), with $\tau = 12000$. The resulting distribution of final energies $E_f = H_0(\mathbf{z}_\tau)$, spanning a range from $E_{f,\min} = 0.0030$ to $E_{f,\max} = 0.0150$, was characterized by a mean value $\overline{E_f} = 0.0106$ and a standard deviation $\sigma_{E_f} = 0.0014$, in excellent agreement with the value $E_f = 0.0104$ predicted by Eq. 17. (The small discrepancies reflect the fact that the duration $\tau = 12000$ is finite.) While these numerical results support the analysis leading to Eq. 17, some caveats are in order. In particular, Liouville's theorem itself rules out the possibility that *all* initial conditions with energy $E_i = 2.8$ lead to a net decrease of energy, $E_f < E_i$. We defer a discussion of this issue to Sec. IV.

To this point we have considered a symmetric protocol, Fig. 2(a), in which each well reaches the same maximal depth, determined by the value of Λ (Fig. 3(d)). However, the analysis is easily generalized to the asymmetric protocol shown in Fig. 2(b), in which the parameters are varied



(a) $t = 0$ (b) $t = \tau/2$ (c) $t = \tau$

FIG. 4. Similar to Fig. 3, but for an asymmetric protocol, Fig. 2(b), with $\Lambda_L = 5.0$ and $\Lambda_R = 3.0$. The phase space volume of set I (the darkly shaded region) remains constant, as does the volume of set II (lightly shaded), but the two volumes differ: $\Omega_I \neq \Omega_{II}$.

around a rectangle with corners at $(0, 0)$ and (Λ_L, Λ_R). Regions I, II and III are defined as in Eq. 8, but now the energies E_1 and E_2 are defined by

$$E_1(\vec{\Lambda}) = \left(\frac{\Lambda_R^{3/2}}{3I_0}\right)^{4/3} \quad , \quad E_2(\vec{\Lambda}) = \left(\frac{\Lambda_R^{3/2} + \Lambda_L^{3/2}}{3I_0}\right)^{4/3} \tag{18}$$

When the protocol is implemented quasi-statically, the net result is a rearrangement of sets I and II, as depicted in Fig. 4. Eq. 16 now leads to the result

$$E_f = \begin{cases} \left(E_i^{3/4} + E_2^{3/4} - E_1^{3/4}\right)^{4/3} & \text{if } 0 < E_i < E_1 \\ \left(E_i^{3/4} - E_1^{3/4}\right)^{4/3} & \text{if } E_1 < E_i < E_2 \\ E_i & \text{if } E_2 < E_i \end{cases} \tag{19}$$

The viscous fluid analogy also applies to this situation, only now the syringes remove different quantities of fluid, $\Omega_I \neq \Omega_{II}$. Alternatively, the processes illustrated in Figs. 3 and 4 are analogous to a simple shuffle of a deck of cards, in which a stack of adjacent cards (region II) is removed from the middle of the deck and transferred to the bottom.

It should now be clear how to design a quasi-static protocol that lowers the energy of the system almost to zero, for a given initial energy $E_i = H_0(\mathbf{z}_0)$. Namely, we choose Λ_R such that E_i is slightly above E_1, thus locating the initial conditions near the bottom of region II. If we then implement the protocol shown in Fig. 2, in either its symmetric ($\Lambda_L = \Lambda_R = \Lambda$) or asymmetric ($\Lambda_L \neq \Lambda_R$) version, the system will be trapped near the bottom of the left well at $t = \tau/2$, and will end the process with $E_f \approx 0$. This outcome is independent of the value of Λ_L, which simply determines the width (in energy) of region II.

III. EXORCISING MAXWELL'S DEMON

Let us now return to the perpetual-motion device of the second kind proposed in the Introduction: after equilibrating the system with a thermal reservoir at temperature T (step 1), we measure the initial energy E_i (step 2), then choose a protocol that reduces the energy near to zero (step 3). The amount of work we extract during this cycle – equivalently, *minus* the amount of work we perform on the system – is given by

$$W_{\text{extracted}} = -W = E_i - E_f < E_i \tag{20}$$

If we repeat this process many times, then the average work extracted per cycle satisfies

$$\langle W_{\text{extracted}} \rangle < \langle E_i \rangle = \int \mathrm{d}\mathbf{z}_0 \, p^{\text{eq}}(\mathbf{z}_0) \, H_0(\mathbf{z}_0) = \frac{3}{4} \, \beta^{-1} \qquad (\beta^{-1} \equiv k_B T) \tag{21}$$

where the canonical distribution $p^{\text{eq}} \propto \exp(-\beta H_0)$ reflects initial equilibration with the reservoir.[2] To approach this upper bound of $(3/4)k_BT$ per cycle, in which the thermal energy of the system is entirely converted to work ($E_f = 0$), the initial energy must be measured with high precision, allowing us to choose a protocol for which $E_i - E_1(\vec{\Lambda})$ is tiny but positive (Eqs. 17, 19). However, as mentioned in the Introduction, these measurements generate information that must ultimately be erased, at a cost of $\beta^{-1} \ln 2$ per bit. There is a competition at play here: increased precision brings us closer to the maximal extracted work, but carries the penalty of increased accumulation of information.

To address this issue, imagine a measurement apparatus that reports the initial energy of the system with finite precision. Specifically, given the initial microstate $\mathbf{z}_0$, the apparatus outputs one of K values associated with specified energy intervals A, B, C, $\cdots$. Taking $K = 4$ for purpose of illustration, the apparatus outputs A, B, C, or D according to

$$\begin{aligned} A &: 0 < H_0(\mathbf{z}_0) < E_A \\ B &: E_A < H_0(\mathbf{z}_0) < E_B \\ C &: E_B < H_0(\mathbf{z}_0) < E_C \\ D &: E_C < H_0(\mathbf{z}_0) \end{aligned} \tag{22}$$

where the values E_A, E_B, and E_C are fixed properties of the apparatus.

Now consider the following strategy for choosing a cyclic protocol, based on the output of the measurement apparatus.

---

[2] In Eq. 21 we have used the identity $\langle E \rangle = -(\partial/\partial\beta) \ln Z$, with $Z \equiv \int \mathrm{d}\mathbf{z} \exp(-\beta H_0) = \sqrt{8\pi}\Gamma(5/4)\beta^{-3/4}$.

• Output = A: Do nothing to the system, as it is already in the lowest-energy interval.

• Output = B: Using Eq. 18, set $E_1(\vec{\Lambda}) = E_A$ and $E_2(\vec{\Lambda}) = E_B$, that is choose (Λ_L, Λ_R) so that interval B in Eq. 22 corresponds to region II in Eq. 8. Next, implement the asymmetric protocol of Fig. 2(b), under which initial conditions from this region are transferred to the bottom of the potential well, as in Fig. 4.

• Output = C: Set $E_1(\vec{\Lambda}) = E_B$ and $E_2(\vec{\Lambda}) = E_C$, then implement the asymmetric protocol. Again, the energy interval containing the initial conditions – interval C, in this case – is shuffled to the bottom of the potential.

• Output = D: Set $E_1(\vec{\Lambda}) = E_C$ and $E_2(\vec{\Lambda}) = E^*$, where $E^* > E_C$ is an arbitrary cutoff energy, then implement the asymmetric protocol. In this case, initial conditions from the region between E_C and E^* are transferred to the bottom of the potential, whereas if $H(\mathbf{z}_0) > E^*$ the protocol produces no net change in the energy of the system.

This strategy takes advantage of the limited knowledge provided by the measurement of the initial energy. When it is implemented, the energy of the system decreases (that is, $E_f < E_i$) if $E_A < E_i < E^*$, and remains unchanged otherwise. Thus, on average per cycle, work is extracted from the system,

$$\langle W_{\text{extracted}} \rangle > 0 \tag{23}$$

and ultimately from the reservoir that replenishes the system's energy.

Over $N \gg 1$ repetitions of the process, the measurement apparatus generates a symbolic string of length N, of the form $BDCCADA\cdots$. Letting P_X denote the probability of outcome $X \in \{A, B, C, D\}$ in a given measurement, the number of bits required to encode this string is given by

$$N_{\text{bits}} = N\mathcal{H}/\ln 2, \tag{24}$$

where

$$\mathcal{H} = -\sum_X P_X \ln P_X \tag{25}$$

is the Shannon entropy of the measurement [15]. Now, both $\langle W_{\text{extracted}} \rangle$ and $\mathcal{H}$ depend on E_A, E_B and E_C, and the former also depends on E^*. In the following section we establish that, no matter what values these parameters take, the inequality

$$\langle W_{\text{extracted}} \rangle \leq \beta^{-1}\mathcal{H} \tag{26}$$

is satisfied. The extraction of work thus comes at the cost of the accumulation of information: on average, at least one bit is written per $\beta^{-1}\ln 2$ of extracted work. [3]

We now turn our attention to the eventual cost of erasing this information. By Landauer's principle, the average work required to erase one bit of information is no less than $\beta^{-1}\ln 2$. Therefore, since the number of bits generated per cycle is $\mathcal{H}/\ln 2$ (Eq. 24), the average work required to erase the information accumulated in one cycle of operation satisfies

$$\langle W_{\text{erasure}}\rangle \geq \beta^{-1}\mathcal{H} \tag{27}$$

Combining Eqs. 26 and 27, we find that the work required to erase the accumulated information exceeds – or at best, matches – the work extracted during the cycle:

$$\langle W_{\text{extracted}}\rangle \leq \beta^{-1}\mathcal{H} \leq \langle W_{\text{erasure}}\rangle \tag{28}$$

Thus our model obeys the Kelvin-Planck statement of the second law, as it had better do! Eq. 28 highlights the two logically distinct steps we take in reconciling our model with the second law. Although the second half of this inequality chain (that is, Landauer's principle) is derived by appeal to the second law itself [10], the first half (Eq. 26) is obtained without assuming the second law: in Sec. III A we do not infer Eq. 26 by arguing that the second law demands it, rather we will derive this inequality directly.

Eq. 26 is a special case of an inequality recently derived by Sagawa and Ueda (see Eq. 3 of Ref. [16] or, in the quantum setting, Eq. 14 of Ref. [17]), which generalizes the second law of thermodynamics to processes with feedback, such as the one considered in this paper. This inequality also follows readily from recent generalizations [16, 18, 19] of the nonequilibrium work relation [2] and Crooks's fluctuation theorem [20] to nonequilibrium processes with feedback. In the following derivation, we do not directly invoke these results, instead we provide a self-contained analysis that is pertinent to our particular model.

A. Bound on work

Consider a cyclic process with the measurement apparatus described by Eq. 22 above. For initial conditions $\mathbf{z}_0$, let $\mathbf{z}_\tau^X(\mathbf{z}_0)$ denote the final conditions, after implementation of the cyclic protocol

[3] In the original Szilard engine, which involves a single particle in a chamber, this relationship is straightforward: the determination whether the particle is in the left or right half of the chamber produces exactly one bit of information, $\mathcal{H} = \ln 2$, and standard thermodynamics gives the amount of work extracted during the subsequent isothermal expansion, $W_{\text{extracted}} = \beta^{-1}\ln 2$.

corresponding to measurement outcome $X \in \{A, B, C, D\}$. The work performed on the system as it evolves from $\mathbf{z}_0$ to $\mathbf{z}_\tau^X(\mathbf{z}_0)$ is given by

$$W = H_0(\mathbf{z}_\tau^X(\mathbf{z}_0)) - H_0(\mathbf{z}_0) \tag{29}$$

Over many repetitions of the process, with the protocol X determined by the measurement of initial energy, the average work performed on the system is

$$\langle W \rangle = \sum_X^{A,B,C,D} \int_{\mathbf{z}_0 \in X} \mathrm{d}\mathbf{z}_0 \, p^{\mathrm{eq}}(\mathbf{z}_0) \, \left(H_0(\mathbf{z}_\tau^X(\mathbf{z}_0)) - H_0(\mathbf{z}_0) \right) \tag{30}$$

where $p^{\mathrm{eq}} \propto \exp(-\beta H_0)$, and $\int_{\mathbf{z}_0 \in X}$ indicates integration over all microstates $\mathbf{z}_0$ that result in the measurement outcome X. Eq. 30 can be rewritten as

$$\langle W \rangle = \beta^{-1} \sum_X \int_{\mathbf{z} \in X} \mathrm{d}\mathbf{z} \, p^{\mathrm{eq}}(\mathbf{z}) \ln \frac{p^{\mathrm{eq}}(\mathbf{z})}{p^{\mathrm{eq}}(\mathbf{z}_\tau^X(\mathbf{z}))} \tag{31}$$

(dropping the subscript 0). Let us now define two functions

$$f_X(\mathbf{z}) \equiv \begin{cases} p^{\mathrm{eq}}(\mathbf{z})/P_X & \text{if } \mathbf{z} \in X \\ 0 & \text{if } \mathbf{z} \notin X \end{cases} \tag{32}$$

$$g_X(\mathbf{z}) \equiv p^{\mathrm{eq}}(\mathbf{z}_\tau^X(\mathbf{z})) \tag{33}$$

where $P_X \equiv \int_{\mathbf{z} \in X} p^{\mathrm{eq}}(\mathbf{z})$ is the probability that the outcome of the measurement is X. We can interpret $f_X(\mathbf{z})$ as the probability distribution of initial microstates, *conditioned* on the outcome X. Moreover, $\int \mathrm{d}\mathbf{z} \, g_X(\mathbf{z}) = \int \mathrm{d}\mathbf{z}_\tau^X \, p^{\mathrm{eq}}(\mathbf{z}_\tau^X) = 1$ (since phase volume is preserved, $\mathrm{d}\mathbf{z} = \mathrm{d}\mathbf{z}_\tau^X(\mathbf{z})$, by Liouville's theorem), therefore $g_X(\mathbf{z})$ can also be interpreted as a probability distribution on phase space.

With these definitions, Eq. 31 becomes

$$\langle W \rangle = \beta^{-1} \sum_X \int \mathrm{d}\mathbf{z} \, P_X f_X(\mathbf{z}) \ln \frac{P_X f_X(\mathbf{z})}{g_X(\mathbf{z})} \tag{34}$$

$$= \beta^{-1} \sum_X P_X \int \mathrm{d}\mathbf{z} \, f_X(\mathbf{z}) \ln \frac{f_X(\mathbf{z})}{g_X(\mathbf{z})} + \beta^{-1} \sum_X P_X \ln P_X \tag{35}$$

The integral appearing in Eq. 35 is the relative entropy or Kullback-Leibler divergence between the distributions $f_X(\mathbf{z})$ and $g_X(\mathbf{z})$; this quantity is equal to zero if the two distributions are identical and is positive otherwise [15]:

$$\int f_X \ln \frac{f_X}{g_X} = D[f_X || g_X] \geq 0 \tag{36}$$

Thus the first sum on the right side of Eq. 35 is non-negative, hence

$$\langle W \rangle \geq \beta^{-1} \sum_X P_X \ln P_X = -\beta^{-1} \mathcal{H} \tag{37}$$

which is equivalent to Eq. 26, the bound we set out to establish. [4]

The above derivation hinges on the non-negativity of relative entropy. A similar approach has recently been taken to obtain inequalities related to the second law of thermodynamics [21–24], in situations when the system of interest does not necessarily begin (or end) in states of thermal equilibrium. (See also Ref. [25] for an alternative derivation of such inequalities.)

While the calculation presented here assumes a measurement apparatus with four possible outcomes, it should be clear that the analysis generalizes to any finite number of energy intervals. In fact, we can even drop the assumption that the measurement is strictly correlated with energy. That is, suppose phase space is divided into N regions (not necessarily corresponding to energy intervals) and suppose that when the system is in microstate $\mathbf{z}$, the measurement apparatus returns a value X that identifies the region of phase space to which that microstate belongs. Finally, a cyclic protocol is assigned to each possible outcome. It can be verified by the reader that the steps leading to Eq. 37 (equivalently Eq. 26) remain valid.

Moreover, to this point we have considered a measurement apparatus that is error-free: if the initial microstate $\mathbf{z}_0$ belongs in region X, then the measurement outcome is necessarily X. Let us now consider a more general situation in which $P(X|\mathbf{z}_0)$ represents the probability that the apparatus outputs the value X, when a measurement is performed on a system in microstate $\mathbf{z}_0$. In the Appendix we analyze this scenario and derive the bound

$$\langle W_{\text{extracted}} \rangle \leq \beta^{-1} \mathcal{I} \tag{38}$$

where $\mathcal{I}$ is the *mutual information* [15] between the variable $\mathbf{z}_0$ and X. For error-free measurements (e.g. Eq. 22), $\mathcal{I} = \mathcal{H}$ and Eq. 38 reduces to Eq. 37. When the apparatus is capable of making errors, then $\mathcal{I} < \mathcal{H}$ [15], which conforms nicely to the intuition that an error-prone measuring device degrades our ability to extract work from the system. In either case Eq. 26 remains valid.

Finally, we note that the results derived in this section can be generalized to systems evolving according to stochastic equations of motion [26].

---

[4] In fact, as long as our measurement apparatus has more than one possible outcome X, this result will be a strict inequality, since $f_X(\mathbf{z}) = 0 \neq g_X(\mathbf{z})$ for any $\mathbf{z} \notin X$, hence $D[f_X||g_X] > 0$.

IV. DISCUSSION AND CONCLUSIONS

The past few years have seen considerable interest in the thermodynamics of small systems and in the applicability of the second law to various nanoscale scenarios (see Ref. [27] for a recent review), including those involving feedback. Motivated by the recent work of Marathe and Parrondo [6], we have studied a model single-particle system that is "cooled" under the quasi-static cycling of external parameters, when initial conditions are sampled microcanonically. We have used this model to construct a procedure for systematically harvesting energy from a thermal reservoir and converting that energy to work, in seeming violation of the Kelvin-Planck statement of the second law. This procedure, however, involves the repeated measurement of the energy of the system. Modeling the measurement apparatus in Sec. III, we have shown by explicit calculation that the average work delivered per operating cycle does not exceed the average work that must eventually be expended (in accordance with Landauer's principle) to erase the information acquired in the act of measuring the initial energy. Thus on balance the Kelvin-Planck statement remains satisifed.

Our model illustrates the idea – which traces back to Maxwell and Szilard – that knowledge about the microscopic state of a system can be exploited to circumvent the second law of thermodynamics, loosely speaking [8]. In this setting, Eq. 37 places a bound on the work that can be extracted during a cyclic process, following a measurement that provides information about the initial state of the system. As already mentioned, similar bounds have been obtained and studied in the past few years, both for quantum systems [17, 28–30] and for systems evolving according to stochastic equations of motion [16, 18, 19, 31–38]. We also note that Eq. 35, a precursor to Eq. 37, generalizes the relative entropy work relation of Kawai, Parrondo and Van den Broeck [39] to processes with feedback.

Let us now return to a point mentioned in Sec. II: the apparent incompatibility of Eq. 17 with Liouville's theorem. Consider a single *energy shell*, that is the set of all points $\mathbf{z}_0$ with a particular value of energy $E_i = H_0(\mathbf{z}_0)$. This set, which we denote $\mathcal{S}_i$, has the topology of a simple closed loop in phase space. Let us assume that this energy shell is located in region II, hence $E_1 < E_i < E_2$. If we evolve trajectories from initial conditions in $\mathcal{S}_i$, using the protocol in Fig. 2(a), we arrive at a set of final conditions, $\mathcal{S}_f$, which also has the topology of a simple closed loop:

$$\mathcal{S}_i = \{\mathbf{z}_0 \,|\, H_0(\mathbf{z}_0) = E_i\} \quad \rightarrow \quad \mathcal{S}_f = \{\mathbf{z}_\tau(\mathbf{z}_0) \,|\, H_0(\mathbf{z}_0) = E_i\} \tag{39}$$

By Liouville's theorem, these loops enclose equal volumes of phase space: $\Omega[\mathcal{S}_f] = \Omega[\mathcal{S}_i]$. This, however, is incompatible with a literal interpretation of Eq. 17, which seems to assert that *every*

initial condition with energy E_i leads to a net decrease of energy, $E_f < E_i$, in other words that $\mathcal{S}_f$ is contained entirely in the interior of $\mathcal{S}_i$. To address this apparent contradiction, we sketch a more careful interpretation of Eq. 17.

For any finite duration τ, there exist *some* initial conditions $\mathbf{z}_0 \in \mathcal{S}_i$ that yield trajectories for which the system's energy increases: $H_0(\mathbf{z}_\tau(\mathbf{z}_0)) > E_i$. We will refer to these trajectories as "bad actors", as they spoil the picture shown in Fig. 3. [5] While bad actors exist for any finite τ, the probability to generate one of these trajectories generally decreases with increasing τ, for initial conditions sampled microcanonically from $\mathcal{S}_i$. We have observed this trend in numerical simulations over a range from $\tau = 1200$ to 2000 (data not shown); and as mentioned in Sec. II, for $E_i = 2.8$ and $\tau = 12000$ no bad actors were observed among 10^5 trajectories. Thus for large but finite τ, we expect $\mathcal{S}_f$ to be a highly convoluted, closed loop – necessarily enclosing the same volume of phase space as $\mathcal{S}_i$ – with much of the loop concentrated at low energies near the value predicted by Eq. 17, but with tendrils reaching into the region of energies higher than than E_i. We believe this issue deserves a more careful treatment, but this is beyond the scope of the present paper. We end with a conjecture regarding the quasi-static limit:

$$\lim_{\tau\to\infty} P\left[|H_0(\mathbf{z}_\tau) - E_f| < \frac{\epsilon}{2}\right] = 1 \quad \text{for any } \epsilon > 0 \tag{40}$$

where the quantity inside the limit is the probability to generate a trajectory whose final energy falls within an interval of width ϵ around the value predicted by Eq. 17, and microcanonical sampling at energy E_i is assumed. We believe this conjecture represents the proper way to understand the validity of Eq. 17 and Fig. 3. Similar comments apply to Eq. 19 and Fig. 4.

Our results suggest several avenues for future research.

First, it would be interesting to explore a quantum-mechanical version of our model system. Here, the possibility of tunneling between the left and right wells introduces a new aspect to the problem, possibly spoiling the picture developed in Sec. II by preventing particles from getting trapped.

Because the protocols discussed in Sec. II involve the quasi-static cycling of external parameters, it is natural to wonder whether the swapping of regions I and II (illustrated in Fig. 3) can be described in terms of a geometric phase.

Finally, we have not explicitly modeled the "demon" in Sec. III. Instead, we have assumed the existence of some mechanism by which a particular outcome of the measurement leads to the

[5] In simulations, we have observed bad actors that begin near the bottom of region II, but get trapped in the right well at the end of the first stage of the process, e.g. just before $t = \tau/4$ in Fig. 3. As a result, they do not get drawn into the left well during the second stage. They subsequently "float" on top of the darkly shaded set I in Fig. 3, and end the process with $H_0(\mathbf{z}_\tau(\mathbf{z}_0)) \approx E_2$.

implementation of the corresponding protocol. It would be interesting, however, to model this mechanism explicitly within a Hamiltonian framework, either by introducing additional degrees of freedom to model the demon or by specifying coupling terms between the measurement device and the system. In this case, we anticipate that the bound on extracted work will be given in terms of the correlation between the state of the system and the state of the measuring device and/or demon [28, 29, 31, 33].

ACKNOWLEDGMENTS

We gratefully acknowledge useful discussions and correspondence with Eric Heller, Jordan Horowitz, Daniel Lathrop, Rahul Marathe, Juan Parrondo and Wojciech Zurek, as well as financial support from the National Science Foundation (USA) under grants CHE-0841557 and DMR-0906601, and the University of Maryland, College Park.

Appendix A: Analysis of error-prone measurement devices

Consider a measurement apparatus with a discrete set of possible outputs, $X = A, B, C, \cdots$, and let $P(X|\mathbf{z}_0)$ denote the probability to obtain outcome X, when the measurement is performed on a system in microstate $\mathbf{z}_0$. We assume that every measurement produces some outcome, hence $\sum_X P(X|\mathbf{z}_0) = 1$ for any $\mathbf{z}_0$. As before, a cyclic protocol is chosen based on the outcome of the measurement. For initial conditions $\mathbf{z}_0$, let $\mathbf{z}_\tau^X(\mathbf{z}_0)$ denote the final conditions, after implementation of the protocol corresponding to outcome X. The work performed on the system is given by Eq. 29, and averaging over many repetitions of the process gives us

$$\langle W \rangle = \int \mathrm{d}\mathbf{z}_0 \, p^{\mathrm{eq}}(\mathbf{z}_0) \sum_X P(X|\mathbf{z}_0) \, \left(H_0(\mathbf{z}_\tau^X(\mathbf{z}_0)) - H_0(\mathbf{z}_0) \right) \tag{A1}$$

$$= \beta^{-1} \sum_X \int \mathrm{d}\mathbf{z}_0 \, P(\mathbf{z}_0, X) \, \ln \frac{p^{\mathrm{eq}}(\mathbf{z}_0)}{p^{\mathrm{eq}}(\mathbf{z}_\tau^X(\mathbf{z}_0))} \tag{A2}$$

where $P(\mathbf{z}_0, X)$ is the joint probability that the system is initially in microstate $\mathbf{z}_0$ and the measurement outcome is X. Dropping the subscript 0, we now introduce two probability distributions (compare with Eqs. 32, 33)

$$f_X(\mathbf{z}) \equiv P(\mathbf{z}|X) = P(\mathbf{z}, X)/P_X \tag{A3}$$

$$g_X(\mathbf{z}) \equiv p^{\mathrm{eq}}(\mathbf{z}_\tau^X(\mathbf{z})) \tag{A4}$$

where $P_X = \int \mathrm{d}\mathbf{z}\, P(\mathbf{z}, X)$ is the net probability to generate the outcome X, and $P(\mathbf{z}|X)$ denotes the conditional probability distribution that the initial microstate is $\mathbf{z}$, given the measurement outcome X. In terms of these distributions we now have

$$\langle W \rangle = \beta^{-1} \sum_X \int \mathrm{d}\mathbf{z}\, P(\mathbf{z}, X)\, \ln \left[\frac{f_X(\mathbf{z})}{g_X(\mathbf{z})} \cdot \frac{P_X\, p^{\mathrm{eq}}(\mathbf{z})}{P_X\, f_X(\mathbf{z})} \right] \tag{A5}$$

$$= \beta^{-1} \sum_X P_X \int \mathrm{d}\mathbf{z}\, f_X(\mathbf{z}) \ln \frac{f_X(\mathbf{z})}{g_X(\mathbf{z})} - \beta^{-1} \sum_X \int \mathrm{d}\mathbf{z}\, P(\mathbf{z}, X) \ln \frac{P(\mathbf{z}, X)}{p^{\mathrm{eq}}(\mathbf{z}) P_X} \tag{A6}$$

On the last line, the first term is a relative entropy, and therefore non-negative; while the second term (apart from the factor β^{-1}) is the mutual information between $\mathbf{z}$ and X. We thus arrive at

$$\langle W \rangle \geq -\beta^{-1} \mathcal{I} \tag{A7}$$

equivalently Eq. 38.

---

[1] C. B. P. Finn, *Thermal Physics*, 2nd ed. (Chapman and Hall, 1993).

[2] C. Jarzynski, Physical Review Letters **78**, 2690 (1997).

[3] A. Allahverdyan and T. Nieuwenhuizen, Physica A **305**, 542 (2002).

[4] M. Campisi, Studies in History and Philosophy of Modern Physics **39**, 181 (2008).

[5] K. Sato, Journal of the Physical Society of Japan **71**, 1065 (April 2002).

[6] R. Marathe and J. M. R. Parrondo, Physical Review Letters **104**, 245704 (June 2010).

[7] J. C. Maxwell, *Theory of Heat* (Longmans, Green and Co., London, 1871).

[8] *Maxwell's Demon 2: Enropy, Information, Computing*, edited by H. S. Leff and A. F. Rex (Institute of Physics Publishing, Bristol and Philadelphia, 2003).

[9] K. Maruyama, F. Nori, and V. Vedral, Rev. Mod. Phys. **81**, 1 (January - March 2009).

[10] R. Landauer, IBM Journal of Research and Development **5**, 183 (1961).

[11] C. H. Bennett, International Journal of Theoretical Physics **21**, 905 (1982).

[12] L. Szilard, Zeitschrift für Physik **53**, 840 (1929).

[13] J. L. Tennyson, J. R. Cary, and D. F. Escande, Physical Review Letters **56**, 2117 (May 1986).

[14] J. R. Cary, D. F. Escande, and J. L. Tennyson, Physical Review A **34**, 4256 (November 1986).

[15] T. M. Cover. and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, 2006).

[16] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (Mar 2010).

[17] T. Sagawa and M. Ueda, Phys. Rev. Lett. **100**, 080403 (Feb 2008).

[18] M. Ponmurugan, Phys. Rev. E **82**, 031129 (Sep 2010).

[19] J. M. Horowitz and S. Vaikuntanathan, Phys. Rev. E **82**, 061120 (Dec 2010).

[20] G. E. Crooks, Phys. Rev. E **60**, 2721 (Sep 1999).

[21] M. Esposito, K. Lindenberg, and C. Van den Broeck, New J. Phys. **12**, 013013 (2010).

[22] H.-H. Hasegawa, J. Ishikawa, K. Takara, and D. J. Driebe, Physics Letters A **374**, 1001 (2010).

[23] K. Takara, H.-H. Hasegawa, and D. J. Driebe, Physics Letters A **375**, 88 (2010).

[24] M. Esposito and C. Van den Broeck, "Second law and landauer principle far from equilibrium," (2011), arXiv:1104.5165v1.

[25] C. Jarzynski, J. Stat. Phys. **96**, 415 (1999).

[26] S. Vaikuntanathan, unpublished.

[27] C. Jarzynski, Annu. Rev. Cond. Matt. Phys. **2**, 329 (2011).

[28] H. W. Zurek, "Maxwell's demon, szilard's engine and quantum measurements," (2003), arXiv:quant-ph/0301076v1.

[29] K. Jacobs, Phys. Rev. A **80**, 012322 (Jul 2009).

[30] S. W. Kim, T. Sagawa, S. De Liberato, and M. Ueda, Phys. Rev. Lett. **106**, 070401 (February 2011).

[31] H. Touchette and S. Lloyd, Phys. Rev. Lett. **84**, 1156 (Feb 2000).

[32] H. K. Kim and H. Qian, Phys. Rev. E **75**, 022102 (Feb 2007).

[33] F. J. Cao and M. Feito, Phys. Rev. E **79**, 041118 (Apr 2009).

[34] H. Suzuki and Y. Fujitani, Journal of the Physical Society of Japan **78**, 074007 (Jul 2009).

[35] Y. Fujitani and H. Suzuki, Journal of the Physical Society of Japan **79**, 104003 (Oct 2010).

[36] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, Nature Physics **6**, 988 (December 2010).

[37] D. Abreu and U. Seifert, EPL (Europhysics Letters) **94**, 10001 (2011).

[38] J. M. Horowitz and J. M. R. Parrondo, "Thermodynamic reversibility in feedback processes," (2011), arXiv:1104.0332v1.

[39] R. Kawai, J. M. R. Parrondo, and C. V. den Broeck, Phys. Rev. Lett. **98**, 080602 (Feb 2007).