OPTIMIZING NON-ERGODIC FEEDBACK ENGINES*

Jordan M. Horowitz, Juan M.R. Parrondo

Departamento de Física Atómica, Nuclear y Molecular and GISC
Universidad Complutense de Madrid, 28040 Madrid, Spain

(Received May 29, 2013)

Maxwell's demon is a special case of a feedback controlled system, where the information gathered by measurement is utilized by driving a system along a thermodynamic process that depends on the measurement outcome. The demon illustrates that with feedback one can design an engine that performs work by extracting energy from a single thermal bath. Besides the fundamental questions posed by the demon — the probabilistic nature of the Second Law, the relationship between entropy and information, *etc.* — there are other practical problems related to feedback engines. One of those is the design of optimal engines, protocols that extract the maximum amount of energy given some amount of information. A refinement of the second law to feedback systems establishes a bound to the extracted energy, a bound that is met by optimal feedback engines. It is also known that optimal engines are characterized by time reversibility. As a consequence, the optimal protocol given a measurement is the one that, run in reverse, prepares the system in the post-measurement state (*preparation prescription*). In this paper, we review these results and analyze some specific features of the preparation prescription when applied to non-ergodic systems.

DOI:10.5506/APhysPolB.44.803
PACS numbers: 05.70.–a, 05.20.–y, 89.70.–a

1. Introduction

As pointed out by Maxwell in 1867 with his celebrated *demon*, a piece of information can be used to extract energy from a single thermal bath [1]. The demon is a special case of feedback control: the information about a system is gathered in a measurement, and then the system is driven along a process that depends on that measurement outcome. Subsequent examples by

* Presented at the XXV Marian Smoluchowski Symposium on Statistical Physics, "Fluctuation Relations in Nonequilibrium Regime", Kraków, Poland, September 10–13, 2012.

Szilard [1] and others (for example Refs. [2–6]) have revealed that with feedback one can design engines that perform work by extracting energy from a single thermal bath.

This connection between information and work has been made explicit by a refinement of the second law of thermodynamics in the presence of feedback [2, 7]: in a thermodynamic process with measurement and feedback, the work W done on a system is bounded by the difference between the information gained in the measurement I and the change in free energy ΔF as

$$W \geq \Delta F - kTI\,, \tag{1}$$

where T is temperature and k is Boltzmann's constant. More precisely, I is the *mutual information* between two random variables: the outcome m of the measurement and the actual value l of the quantity being measured. In an error-free measurement $m = l$, but the concept of mutual information allows us to compute the information gained in a measurement with errors. Mathematically, the mutual information reads

$$I(l;m) = H(l) + H(m) - H(l,m)\,, \tag{2}$$

where $H(X)$ is the Shannon entropy of the variable, or set of variables, X [8]. $I(l;m) = 0$ only if l and m are independent, *i.e.*, if the outcome of the measurement is completely uncorrelated with the measured magnitude l. On the other hand, if $m = l$ always, $I(l;m) = H(l) = H(m)$ is simply the Shannon entropy of l [8]. Notice also that if z is a description of the system finer than l (for instance, the microstate of the system at the instant of measurement), then $I(l;m) = I(z;m)$, provided that the conditional probability of the outcome obeys $\rho(m|l) = \rho(m|z)$.

Besides the fundamental questions posed by the demon — the probabilistic nature of the Second Law, the relationship between entropy and information, *etc.* — there are also interesting practical problems related to feedback engines. One of those is how to design optimal engines, *i.e.*, protocols that extract the maximum amount of energy given some amount of information, saturating the bound in Eq. (1) [5, 6, 9–11]. In a sequence of papers, we have shown that these optimal processes are *reversible* [11, 12]: indistinguishable from their time-reverse (constructed in a particular manner that will be described later). Building on this intuition, we proposed a method, or a recipe, for designing such optimal feedback processes which we call the *preparation prescription* [11]. Instead of looking for a protocol that extracts all the work, we turn our attention to the time-reversed process and devise a protocol that prepares the post-measurement state. In this article, we investigate how this method applies to ergodicity-breaking processes, where the phase (or state) space of the system splits into distinct

ergodic regions. The canonical example of this situation is the Szilard engine [1], where the phase space of a single ideal gas particle confined to a box is divided into to equal halves upon inserting a partition into the center of the box.

The paper is organized as follows. In Sec. 2, we briefly review the main results on the energetics of feedback control and the preparation prescription to design optimal engines. In Sec. 3, we analyze the peculiarities of the preparation prescription when applied to non-ergodic systems. In Sec. 4, we present an example of optimal design in a multi-particle Szilard engine. Finally, in Sec. 5, we summarize our results and present our main conclusions.

2. Reversible feedback and the preparation prescription

We begin with a concise review of the preparation prescription for designing reversible feedback protocols [12]. For simplicity, we only consider protocols with one feedback loop. All of our conclusions can be generalized to the case of a sequence of repeated measurements.

We have in mind a classical system whose position in phase space Γ at time t is z_t and is in thermal contact with an ideal thermal reservoir at temperature T. We drive our system away from thermodynamic equilibrium using feedback by varying the system's Hamiltonian (or energy function) $H(z, \lambda)$ through a collection of external parameters λ. From time $t = 0$ to τ, the parameters are varied according to a protocol determined by the measurement of a physical observable M at the time $t = t_{\text{meas}}$ whose outcomes m occur with conditional probability (or error) $P(m|z_{t_{\text{meas}}})$. The protocol we use, denoted $\Lambda^m = \{\lambda_t^m\}_{t=0}^{\tau}$, depends on the measurement outcome m only after time $t = t_{\text{meas}}$. During this interval, thermal fluctuations cause the system to follow a random microscopic trajectory $\gamma = \{z_t\}_{t=0}^{\tau}$. We can define a joint probability distribution $\mathcal{P}[\gamma, \Lambda^m]$ of the trajectory γ and the measurement outcome, or equivalently, the implemented protocol Λ^m. The work along this trajectory is $W[\gamma, \Lambda^m]$ and the reduction in uncertainty due to the measurement is [2, 7, 12]

$$i[\gamma, \Lambda^m] = \ln \frac{P(m|z_{t_{\text{meas}}})}{p_m} \,. \tag{3}$$

Here, the probability to measure m is $p_m = \int d\gamma \, \mathcal{P}[\gamma, \Lambda^m]$, where $d\gamma$ is a measure on the space of trajectories. Averaging over all realizations recovers the mutual information $I(z_{t_{\text{meas}}}; m) = \langle i([\gamma, \Lambda^m]) \rangle$ in Eq. (1).

With every feedback process, we can introduce a related process called the reverse process [7], which plays the role of time-reversal in the presence of feedback. We initiate the reverse process by first randomly selecting

a protocol Λ^m with probability p_m, that is *from the distribution of measurement outcomes of the feedback process*. Next, we equilibrate the system with the external parameters fixed to λ^m_τ, followed by a non-equilibrium driving according to the conjugate reverse protocol $\tilde{\Lambda}^m = \{\tilde{\lambda}_t\}_{t=0}^{\tau}$, where $\tilde{\lambda}^m_t = \lambda^m_{\tau-t}$. Time-reversal invariance guarantees that each trajectory γ of the feedback process has a conjugate twin in the reverse process $\tilde{\gamma} = \{\tilde{z}_t\}_{t=0}^{\tau}$, where $\tilde{z}_t = z^*_{\tau-t}$ and $*$ denotes momentum reversal, which is observed with probability $\tilde{\mathcal{P}}[\tilde{\gamma}, \tilde{\Lambda}^m]$.

With this setup, we have the result that the distinguishibility of the feedback process measured as the relative entropy, $D(f||g) = \int dx\, f(x) \ln(f(x)/g(x))$, between $\mathcal{P}$ and $\tilde{\mathcal{P}}$ satisfies [7, 12]

$$kTD\left(\mathcal{P}||\tilde{\mathcal{P}}\right) = W - \Delta F + kTI \geq 0 \tag{4}$$

with $I = I(z_{t_{\rm meas}}; m)$. We now see that the optimal thermodynamic process for which $W - \Delta F + kTI = 0$ occurs only when $D = 0$, which is true if and only if [8]

$$\mathcal{P}[\gamma, \Lambda^m] = \tilde{\mathcal{P}}\left[\tilde{\gamma}, \tilde{\Lambda}^m\right], \tag{5}$$

that is only when the feedback process is *indistinguishable* from its reverse [12]. This is a microscopic statement of reversibility. It is consistent with the macroscopic definition, since in a macroscopic reversible process the same sequence of states can also be traced out both forwards and backwards in time.

Equation (5) also offers insight into how to design an optimal feedback process that extracts the maximum amount of work. Instead of devising a feedback protocol implemented in response to a particular measurement, we should look for a reversible process. In particular, let us focus on the evolution at one particular time, immediately after the measurement. To this end, we integrate Eq. (5) over all trajectories passing through z at $t = t_{\rm meas}$, and divide by p_m, to deduce the equality of phase space densities conditioned on the protocol (or measurement outcome)

$$\rho_m(z, t_{\rm meas}) = \tilde{\rho}_m(\tilde{z}, \tau - t_{\rm meas})\,. \tag{6}$$

We now see that in a reversible optimal protocol the *post-measurement* state $\rho_m(z, t_{\rm meas})$ — the state prepared by the measurement — must be the same as the state *prepared* by the reverse process $\tilde{\rho}_m(\tilde{z}, \tau - t_{\rm meas})$. Our strategy to obtain reversible feedback protocols is then to design a protocol that prepares the post-measurement state [11]. As Eq. (6) suggests, by reversing this protocol, we obtain an optimal protocol to implement in the feedback process in response to that measurement outcome.

3. Preparation in non-ergodic systems

There is an apparently simple way to reversibly prepare a system in the post-measurement state $\rho_m(z, t_{\rm meas})$ from the initial state of the reverse process, $\tilde{\rho}(z, 0)$: slowly and quasi-statically vary the system Hamiltonian from its initial value $H(z, \tilde{\lambda}_0^m)$ to $H_m(z) = -kT \ln \rho_m(z, t_{\rm meas})$ so that this post-measurement state is in thermodynamic equilibrium with respect to the new Hamiltonian. This protocol has been suggested in Refs. [9, 13, 14] and, at first sight, seems to be the most general procedure for a reversible preparation, since in a reversible process the system must be in equilibrium at any time, in particular, at the beginning and end of the process.

Nevertheless, alternative and more feasible protocols can be devised if the system is not ergodic or if its dynamics presents well separated time scales, as happens in most information processing devices. Consider for instance a system whose phase space Γ at the time of measurement, $t_{\rm meas}$, can be decomposed into n distinct ergodic regions Γ_l ($\Gamma = \cup_{l=1}^n \Gamma_l$ and $\Gamma_l \cap \Gamma_k = 0$ for $l \neq k$). This partition of phase space can be the result of a rigorous ergodicity breaking in the system dynamics due to, *e.g.*, barriers higher than the total energy of the system [15] or phase transitions in the thermodynamic limit [4]. Our analysis also applies to *effective ergodicity breaking* resulting when there are slow variables (usually discrete) whose evolution is governed, for instance, by jumps over high energy barriers.

We further assume that the system is always locally in equilibrium within each ergodic region and that the measurement is merely the identification of the ergodic region where the system is located. Then, in an error-free measurement the post-measurement state will be the equilibrium distribution restricted to one of the partitions Γ_l at inverse temperature $\beta = 1/(kT)$,

$$\rho_l(z, t_{\rm meas}) = \frac{e^{-\beta H(z, \lambda_{t_{\rm meas}})}}{Z_l} \chi_l(z) \tag{7}$$

with $Z_l = \int_{\Gamma_l} e^{-\beta H}$ and $\chi_l(y)$ the characteristic function on Γ_l taking the value 1, when $y \in \Gamma_l$ and 0 otherwise.

On the other hand, a measurement with errors can be characterized by the probability that the actual value of the magnitude is l when the outcome of the measurement is m, $p(l|m)$ [1]. In this case, when the measurement

[1] The conditional probability $p(m|l)$ is a more natural way to characterize the error of a measurement device or procedure. To simplify the exposition, we use $p(l|m)$ instead. Both quantities are related by the Bayes formula: $p(l|m) = p_l p(m|l)/p_m$. Notice that in feedback control, both p_l and p_m are known (there is no need of a Beayesian *prior*). Feedback uses the information related with thermal fluctuations in a single system, but the statistical properties of such fluctuations are perfectly known.

outcome is m, the post-measurement state reads

$$\rho_m(z, t_{\rm meas}) = \sum_l p(l|m)\, \frac{e^{-\beta H(z,\lambda_{t_{\rm meas}})}}{Z_l} \chi_l(z)\,. \tag{8}$$

According to the preparation prescription, we have to design a protocol that prepares the system in this specific state $\rho_m(z, t_{\rm meas})$. To achieve this goal, it will be illuminating to discuss general features of non-ergodic systems.

In a non-ergodic system equilibration between ergodic regions Γ_l is obviously hindered. In a quasi-static process, for instance, the system is in equilibrium within a region Γ_l, like the states given by Eqs. (7) and (8), but, for a generic density $\rho(z)$, the probability to be in region Γ_l

$$p_l = \int_{\Gamma_l} dz\, \rho(z) \tag{9}$$

will, in general, differ from its equilibrium value

$$p_l^{\rm eq} = \frac{\int_{\Gamma_l} dx\; e^{-\beta H(x)}}{\int_{\Gamma} dx\, e^{-\beta H(x)}} = \frac{Z_l}{Z}\,. \tag{10}$$

In general, the actual p_l depends on the past history and/or the information that we have about the system. For example, if the system becomes non-ergodic by virtue of some symmetry breaking transition, p_l depends on the probability that the system chooses region l *at the transition point*. After the transition, the Hamiltonian can change in an arbitrary way, as far as ergodicity is not restored. The equilibrium probability $p_l^{\rm eq}$ in Eq. (10) depends on the Hamiltonian at a given time after the transition, whereas p_l depends only on the details of the transition. The probability p_l can also depend on what we know about a system: for instance, after an error-free measurement whose outcome is l, $p_l = 1$ and $p_k = 0$ for all $k \neq l$ [*cf.* Eq. (7)].

Figure 1 presents an illustration that clarifies the meaning of the non-equilibrium probability p_l. A Brownian particle at temperature T moves in a potential $V(x)$, which is modified by an external agent. The potential and the probability density $\rho(x)$ of the position x of the particle are both depicted in the figure. Initially, (a) the barrier is low enough for the particle to jump from one well to the other. Then in (b) the potential barrier is raised up to a value far above kT creating an effective ergodicity breaking for time intervals much smaller than the Kramer's mean time to cross the barrier [16]. The probability that the particle is in the left or the right region, p_l with $l = {\rm L}, {\rm R}$, is $1/2$ because the ergodicity is broken in a symmetric way. After the transition has occurred, one can lower or raise the well in an arbitrary

manner as in (c), as far as the barrier stays far above kT. The probability p_l is still $1/2$ for $l = \mathrm{L}, \mathrm{R}$, since jumps do not occur in the time scale of the process. On the other hand, the equilibrium probability p_l^{eq} in Eq. (10), obviously changes. The state depicted in (c) is in a non-equilibrium state, although the probability density equilibrates within each well. Moreover, if we measure (with no error) the position of the particle and find that it is in the left well, the post-measurement non-equilibrium state will be confined in the left well, yielding $p_{\mathrm{L}} = 1$ as depicted in (d). Hence, the non-equilibrium probability p_l depends on the history and also on our knowledge about the state of the system.

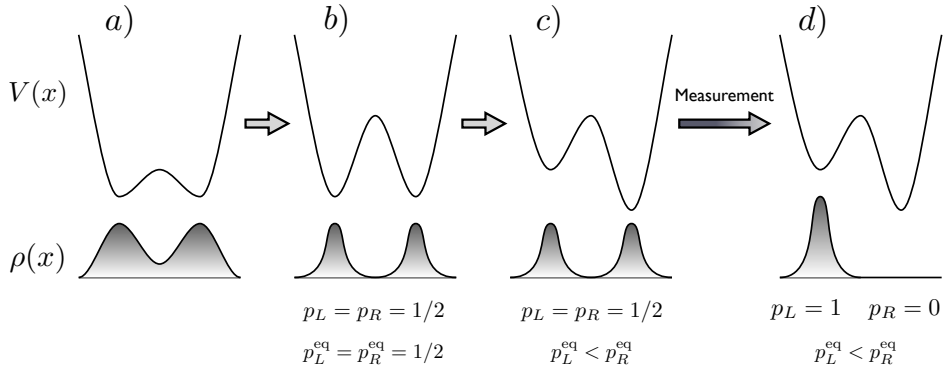

Fig. 1. An illustration of non-equilibrium states arising from ergodicity breaking and measurement. A Brownian particle at temperature T moves in a double-well potential $V(x)$ which is modified by an external agent. The potential and the probability density $\rho(x)$ of the position x of the particle are both depicted in the figure. (a) Initially the barrier is low enough for the particle to jump from one well to the other. (b) The potential barrier is raised up to some value far above kT and an effective ergodicity breaking occurs if we consider a time scale much shorter than the jump rate. The probability that the particle is in the left or the right region is $p_l = 1/2$, with $l = \mathrm{L}, \mathrm{R}$, because ergodicity is broken in a symmetric way. (c) After the transition has occurred, the left well is raised and the right one is lowered. The probability p_l remains $1/2$ for $l = \mathrm{L}, \mathrm{R}$, since jumps do not occur in the time scale of the process, whereas the equilibrium probability, p_l^{eq} in Eq. (10), changes. (d) After an error-free measurement that finds the particle in the left well, this post-measurement non-equilibrium state is now a probability density with support in the left well, yielding $p_{\mathrm{L}} = 1$.

Now we can address our main problem: how to prepare a non-ergodic system in the post-measurement state given by Eq. (8)? Since the state is non-equilibrium, we cannot apply the aforementioned preparation, consisting of a slow transition from the final Hamiltonian $H(z, \tilde{\lambda}_0^m)$ to $H_m(z) =$

$-kT \ln \rho_m(z, t_{\text{meas}})$. However, non-ergodicity provides us with a wider range of preparation strategies. The trick is to prepare any other state $\rho'_m(z)$ as long as it reversibly induces the same post-measurement distribution

$$p(l|m) = \int_{\Gamma_l} dz\, \rho'_m(z) = \int_{\Gamma_l} dz\, \rho_m(z, t_{\text{meas}}) \tag{11}$$

and is in local equilibrium. The key point is that these probabilities $p(l|m)$ depend on the critical point where the ergodicity is broken and not on the final Hamiltonian, as illustrated in Fig. 1. Once we prepare a system with the desired probabilities $p(l|m)$, one can adiabatically shift the Hamiltonian towards $H(z, \lambda_{t_{\text{meas}}})$ and complete the design of the optimal protocol.

We have applied this method in a previous paper to a multi-particle Szilard engine [11], although we did not carry out an explicit discussion of the role of non-ergodicity. This explicit analysis of the preparation prescription in non-ergodic systems allows us to consider more involved examples, like the one treated in the next section.

4. Example: two-particle Szilard engine

In this section, we highlight the utility of the preparation prescription for systems with ergodicity breaking using a two-particle Szilard engine. Previously, Kim *et al.* [5, 10] investigated the quantum multi-particle Szilard engine using a non-optimal protocol. In a subsequent article, we then showed how the preparation prescription could be used to develop an optimal feedback protocol for the classical multi-particle Szilard engine [11]. This section builds on that work to include measurement errors.

The two-particle Szilard engine consists of two ideal gas particles confined to a box of volume V connected to a thermal reservoir at temperature $kT = 1$. Furthermore, we take the particles to have a short-ranged, repulsive interaction. The engine cycle begins with the particles in equilibrium. We then quickly insert a partition dividing the box into two equal halves, breaking ergodicity. At that point, the phase space of the engine, schematically depicted in Fig. 2, is segregated into three regions that we label $l = \{\text{LL}, \text{RR}, \text{LR}\}$ for two particles in the left half, two in the right, and one in each half. We then measure l obtaining possible measurement outcomes $m = \{\text{LL}, \text{RR}, \text{LR}\}$. However, we allow for the possibility that there are errors when both particles are in the same half, but not when they are in separate halves. Specifically, when $l = \text{LL}$ (RR), we can mistakenly measure $m = \text{LR}$ instead of LL (RR) with a probability $\epsilon_{\text{LL}} \equiv p(\text{LR}|\text{LL})$ [$\epsilon_{\text{RR}} \equiv p(\text{LR}|\text{RR})$]. Then based on the measurement outcome, we extract work using an optimal, cyclic, isothermal feedback process.

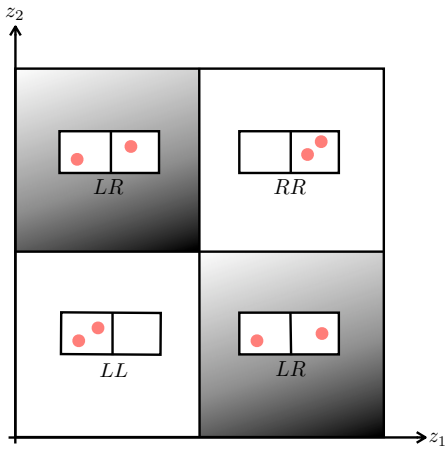

Fig. 2. Phase space schematic for the two-particle Szilard engine immediately after inserting the partition, with z_1 and z_2 the phase space positions of the two particles. Each quadrant corresponds to a ergodic region with a particular arrangement of the two particles: both in the left (LL), both in the right (RR), or in different halves (LR). The shaded squares highlight the region of phase space, where each particle is segregated into a separate half of the box.

In the light of our previous discussion on the preparation prescription (Sec. 2), the optimal protocol will prepare the engine in each ergodic region (or in a distribution over ergodic regions). When both particles are found in the same half of the box the optimal protocol is the same as in the original single-particle Szilard engine. Namely, we can prepare the engine with both particles in the left (right) half of the box by inserting the partition along the right (left) wall and then slowly shifting the partition to the center. Thus, when we find both particles in the same half of the box, we can use this protocol, in reverse, to extract the maximum amount of work.

On the other hand, it is more difficult to prepare the engine with each particle in a separate half of the box. The generic prescription requires that we reversibly prepare the equilibrium distribution for the Hamiltonian $H_m(z) = -kT \ln \rho_m(z, t_{\text{meas}})$. For error-free measurement, this Hamiltonian is infinite in the white quadrants of Fig. 2 and zero in the shaded, which requires infinite interaction energy in disjoint quadrants of phase space. In a previous article, we demonstrated that using a collection of deep potential wells we could also prepare this scenario, without recourse to such a strange Hamiltonian [11]. In the following, we build on this idea and demonstrate how we can prepare not simply both particles in separate wells, but a distribution over the regions {LL,RR,LR} corresponding measurement errors ϵ_{LL} and ϵ_{RR} that are rational numbers.

To this end, let us consider the scenario with both particles in the box at equilibrium. We then slowly lower K potential wells, n in the left half of the box and $K-n$ in the right, to a depth $E<0$ deep compared to the thermal energy $kT=1$, but shallow compared to the interaction energy, so that only one particle can occupy any well at any given time. This traps each particle in a separate well, occupying a small volume v. Next, we quickly insert the partition, followed by slowly removing the wells. As a result, the particles may be confined to either half of the box. Each particle is in a separate half ($l=\mathrm{LR}$) with probability

$$p_{\mathrm{LR}}=\frac{2n(K-n)}{K(K-1)}\,. \tag{12}$$

However, this protocol also prepares the system with both particles in the left half (LL) and the right half (RR) with probabilities

$$p_{\mathrm{LL}}=\frac{n(n-1)}{K(K-1)}\,,\qquad p_{\mathrm{RR}}=\frac{(K-n)(K-n-1)}{K(K-1)}\,, \tag{13}$$

respectively. As a consequence, it generates a distribution over the different regions, as in Eq. (11). Therefore, we can use this protocol (in reverse) as an optimal feedback protocol as long as we use a measurement with an error that results in the distribution $\{p_{\mathrm{LL}},p_{\mathrm{RR}},p_{\mathrm{LR}}\}$ over the regions of phase space given the measurement outcome $m=\mathrm{LR}$. By applying Bayes' theorem, we see this corresponds to measuring $m=\mathrm{LR}$ when $l=\mathrm{LL}$ with (conditional) probability

$$\epsilon_{\mathrm{LL}}\equiv p(\mathrm{LR}|\mathrm{LL})=\frac{n-1}{K-n}\,, \tag{14}$$

and similarly the error for both particles on the right is

$$\epsilon_{\mathrm{RR}}\equiv p(\mathrm{LR}|\mathrm{RR})=\frac{K-n-1}{n}\,. \tag{15}$$

For the special case with two wells, one in each half ($K=2$ and $n=1$), we recover error-free measurement ($\epsilon_{\mathrm{LL}}=\epsilon_{\mathrm{RR}}=0$), which was shown to be optimal in Ref. [11].

In order to verify that this protocol is, in fact, optimal for a measurement with errors ϵ_{LL} and ϵ_{RR}, we now determine the work and information conditioned on measuring the particles in separate halves. Let us return to our scenario immediately after having inserted the partition and obtained the measurement outcome $m=\mathrm{LR}$. At this point, we lower our K wells very slowly. As the wells become deeper, the depth approaches a value $E^*\sim kT=1$ at which point ergodicity begins to break, and each particle becomes trapped in a different well. The exact value of E^* will prove to

be inconsequential, but its existence is needed for the calculation. Since the process is done slowly, the average work done up to that point may be determined as an average over the ratios of the partition functions (the changes in free energy) between the initial state Z_l and the equilibrium state at the moment ergodicity breaks Z_l^* for each $l = \{\mathrm{LL}, \mathrm{RR}, \mathrm{LR}\}$ as

$$\begin{aligned} W_{\text{lower}} &= -p_{\mathrm{LL}} \ln \frac{Z^*_{\mathrm{LL}}}{Z_{\mathrm{LL}}} - p_{\mathrm{RR}} \ln \frac{Z^*_{\mathrm{RR}}}{Z_{\mathrm{RR}}} - p_{\mathrm{LR}} \ln \frac{Z^*_{\mathrm{LR}}}{Z_{\mathrm{LR}}} \\ &= -p_{\mathrm{LL}} \ln \frac{(1/2)n(n-1)v^n e^{-nE^*}}{(1/2)\,(V/2)^2} \\ &\quad -p_{\mathrm{RR}} \ln \frac{(1/2)(K-n)(K-n-1)v^{K-n} e^{-(K-n)E^*}}{(1/2)\,(V/2)^2} \\ &\quad -p_{\mathrm{LR}} \ln \frac{n(K-n)v^K e^{-KE^*}}{(V/2)^2}\,. \end{aligned} \tag{16}$$

Once the wells have passed E^*, each particle is trapped within a separate well, and the work required to lower the wells to the final value E is $w = E - E^*$. Next, we remove the partition for free. Then, we begin raising the wells with each particle trapped in a separate well doing a work $\bar{w} = E^* - E$ until we reach E^* again, and the particles begin exploring the entire box. From this point on, until the wells are completely removed, the work is

$$W_{\text{raise}} = -\ln \frac{\bar{Z}}{\bar{Z}^*} = -\ln \frac{V^2/2}{(1/2)K(K-1)v^K e^{-KE^*}}\,. \tag{17}$$

Summing these contributions, we find for the average work conditioned on measuring $m = \mathrm{LR}$

$$\begin{aligned} W &= W_{\text{lower}} + w + \bar{w} + W_{\text{raise}} \\ &= -p_{\mathrm{LL}} \ln(4p_{\mathrm{LL}}) - p_{\mathrm{RR}} \ln(4p_{\mathrm{RR}}) - p_{\mathrm{LR}} \ln(2p_{\mathrm{LR}})\,. \end{aligned} \tag{18}$$

On the other hand, the average information (reduction in uncertainty) can be determined from the formula

$$I = \sum_{l=\{\mathrm{LL},\mathrm{RR},\mathrm{LR}\}} p_l \ln \frac{\epsilon_l}{P(\mathrm{LR})} \tag{19}$$

by virtue of Eq. (3), where $P(\mathrm{LR}) = 1/(2p_{\mathrm{LR}})$ is the probability to measure LR. Thus,

$$I = p_{\mathrm{LL}} \ln(4p_{\mathrm{LL}}) + p_{\mathrm{RR}} \ln(4p_{\mathrm{RR}}) + p_{\mathrm{LR}} \ln(2p_{\mathrm{LR}})\,, \tag{20}$$

and $W + I = 0$ as desired.

5. Conclusions

In this paper, we have presented the preparation method as a recipe for designing optimal (or reversible) feedback protocols that extract the maximum amount of energy from a measurement. In many situations, our method reproduces the simplest protocol that exploits the Hamiltonian $H_m(z) = -kT \ln \rho_m(z, t_{\text{meas}})$. However, our method can generate a variety of non-trivial protocols when the system experiences some type of ergodicity breaking. In our example, the two-particle Szilard engine, we saw that the preparation led to a protocol that exploited a partitioning of phase space and avoided any non-physical Hamiltonians typical of other schemes.

This work is funded by the grants MOSAICO and ENFASIS (Spanish Government), and MODELICO (Comunidad Autonoma de Madrid). J.M.H. is supported financially by the National Science Foundation (USA) International Research Fellowship under Grant No. OISE-1059438.

REFERENCES

[1] H.S. Leff, A.F. Rex (eds.), *Maxwell's Demon: Entropy, Information, Computing*, Princeton University Press, New Jersey, 1990.
[2] T. Sagawa, M. Ueda, *Phys. Rev. Lett.* **100**, 080403 (2008).
[3] R. Kawai, J.M.R. Parrondo, C. Van den Broeck, *Phys. Rev. Lett.* **98**, 080602 (2007).
[4] J.M.R. Parrondo, *Chaos* **11**, 725 (2001).
[5] S.W. Kim, T. Sagawa, S. De Liberato, M. Ueda, *Phys. Rev. Lett.* **106**, 070401 (2011).
[6] D. Abreu, U. Seifert, *Europhys. Lett.* **94**, 10001 (2011).
[7] J.M. Horowitz, S. Vaikuntanathan, *Phys. Rev.* **E82**, 061120 (2010).
[8] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, 2006.
[9] M. Esposito, C. Van den Broeck, *Europhys. Lett.* **95**, 40004 (2011).
[10] K.H. Kim, S.W. Kim, *Phys. Rev.* **E84**, 012101 (2011).
[11] J.M. Horowitz, J.M.R. Parrondo, *New J. Phys.* **13**, 123019 (2011).
[12] J.M. Horowitz, J.M.R. Parrondo, *Europhys. Lett.* **95**, 10005 (2011).
[13] H.-H. Hasegawa, J. Ishikawa, K. Takara, D.J. Driebe, *Phys. Lett.* **A374**, 1001 (2010).
[14] K. Takara, H.-H. Hasegawa, D.J. Driebe, *Phys. Lett.* **A375**, 88 (2010).
[15] R. Marathe, J.M.R. Parrondo, *Phys. Rev. Lett.* **104**, 245704 (2010).
[16] N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed., Elsevier Ltd., New York 2007.