

Optimal work extraction and mutual information in a generalized Szilárd engine

Juyong Song,¹ Susanne Still,² Rafael Díaz Hernández Rojas,^{3,*} Isaac Pérez Castillo,⁴ and Matteo Marsili⁵

¹*Samsung Research, Samsung Electronics Co., Ltd., Seoul, 06765, Korea*

²*Department of Information and Computer Sciences, and Department of Physics and Astronomy, University of Hawai'i at Mānoa, Honolulu, 96822, Hawaii, USA*

³*Dipartimento di Fisica, Università di Roma "La Sapienza", P.le A. Moro 5, I-00185 Rome, Italy*

⁴*Departamento de Física, Universidad Autónoma Metropolitana-Iztapalapa, San Rafael Atlirco 186, Ciudad de México 09340, Mexico*

⁵*The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, 34151, Italy*

A 1929 Gedankenexperiment proposed by Szilárd, often referred to as “Szilárd’s engine”, has served as a foundation for computing fundamental thermodynamic bounds to information processing. While Szilárd’s original box could be partitioned into two halves and contains one gas molecule, we calculate here the maximal average work that can be extracted in a system with N particles and q partitions, given an observer which counts the molecules in each partition, and given a work extraction mechanism that is limited to pressure equalization. We find that the average extracted work is proportional to the mutual information between the one-particle position and the vector containing the counts of how many particles are in each partition. We optimize this quantity over the initial locations of the dividing walls, and find that there exists a critical number of particles $N^*(q)$ below which the extracted work is maximized by a symmetric configuration of the q partitions, and above which the optimal partitioning is asymmetric. Overall, the average extracted work is maximized for a number of particles $\tilde{N}(q) < N^*(q)$, with a symmetric partition. We calculate asymptotic values for $N \rightarrow \infty$.

I. INTRODUCTION

The thought experiment known as “Maxwell’s Demon” [1] addressed the issue that the Second Law of thermodynamics is statistical in nature. An ideal gas at temperature T is enclosed in an isolated container divided into two equal parts by a fixed wall with a trap door operated by some sentient being, later called a “demon”, who will open the door to incoming particles, sorting them by velocity. This process would result in a temperature gradient that could be used to obtain work from the system. Maxwell’s idea started an ongoing debate, to which Szilárd contributed significantly with a model that circumvents the necessity of a sentient being, replacing it by a simple mechanism which, importantly, retained the main feature of the “demon”, namely that of having a memory.

Szilárd’s engine consists of a single particle gas within a container divided into two equal partitions separated by a movable, frictionless wall [2]. When the container is put into contact with a single thermal reservoir at temperature T , the movable wall may then be used to extract work (*e.g.* by lifting a weight), when moved towards the empty side of the box, as the particle transfers kinetic energy in successive elastic collisions. Being able to do this requires knowledge of which side is empty at the beginning of work extraction, to remain present throughout, *i.e.*, it requires a memory. When operated cyclically, the average extractable work is compensated by the average amount of work that has to be done to run the memory. An adiabatic, isothermal volume expansion yields work

$W_{\text{ext}} = k_B T \ln \left(\frac{V}{V/2} \right) = k_B T \ln 2$, which corresponds to $k_B T$ times the mutual information captured about the coarse grained particle location. This idea has served as a foundation not only for computing fundamental thermodynamic bounds for information processing (*e.g.* [3–7]), but also for concretely demonstrating how information can be turned into work, and vice versa. Recently, interest in these issues has spiked with increased experimental capabilities [8–23]. Different variations of the standard Szilárd engine have been discussed, including generalisations to N -particles systems [24, 25] and non-ideal (classical or quantum) particles [26, 27].

Here, we study how much work can be extracted, on average, when a Szilárd engine, operated quasi-statically, contains an ideal gas with N particles, and when q partitions can be created in the box. We assume that the observer counts and memorizes how many particles fall into each partition, and then exploit the isothermal expansion of the ideal gas in the different compartments to extract work, as in the original Szilárd box [2]. We show in Sec. II that the average extracted work is proportional to the mutual information retained in memory about the location of a single particle, not of the location vector of the ensemble. The latter information controls the minimal cost for memorizing the counts, whereby their difference controls a lower bound on dissipation of the engine (when run cyclically). We calculate how much average work can maximally be extracted when the choice of where to place the movable walls before the measurement is optimized, for fixed N and q . To build intuition, Sec. III A treats the case with only one movable wall, $q = 2$, and confirms agreement with previous work. The general case is then treated in Sec. III B.

* rafael.diazhernandezrojas@uniroma1.it

II. MUTUAL INFORMATION AND WORK

We consider a Szilárd engine, generalized to N particles inside a container of longitudinal size L and transverse unit area. The gas in the container is coupled to a thermal reservoir, and is in thermal equilibrium at the beginning of each cycle. The mass of the walls is assumed to be much larger than the mass of the particles (which is set to unity). The engine is run cyclically as follows:

1. Preparation step (assumed not to require work): Insert walls which divide the container along the longitudinal axis into q partitions with lengths $\ell = (\ell_1, \dots, \ell_q)$, see Fig. 1a.
2. Measurement and data representation step: the observer has *a priori* knowledge of the experimental setup, including ℓ , and is provided with a snapshot of the N particles' x-positions, denoted by $\mathbf{x} = (x_1, \dots, x_N)$. Using this measurement, the observer commits the counts of how many particles reside in each partition to memory, denoted by $\mathbf{k} = (k_1, \dots, k_N)$.
3. Work extraction step: using the information committed to memory, weights are attached in such a way that work is extracted while the walls move quasi-statically until the pressure is equalized across the container. The observer knows the number density in each partition, k_i/ℓ_i , and thus the local pressure, $P_i = k_B T k_i/\ell_i$ [28]. This enables the observer to determine how much each partition will be moved by the expansion of the gas, and in which direction. In equilibrium, the pressure is $P_i^{(\text{eq})} = N k_B T/L$ in all partitions, which implies that the lengths of the partitions after work extraction are $\ell_i^{(\text{eq})} = k_i L/N$.
4. Return to beginning: Walls are pulled out (assumed not to require work).

We know that the total information captured by the observer's memory, about the available data determines the energetic cost of the memory, while only the relevant part of this information determines the potential energetic gain [7]. To assess the engine's overall efficiency, we need to calculate both aspects.

In the setup we consider here, available data are the x-positions of the ensemble, \mathbf{x} , and work extraction can happen only via movement of the walls towards pressure equalization. Since the particles of an ideal gas are non-interacting, only information about the locations of the N individual particles matters, none of the correlations between particle positions make any difference to calculating the density, and thus the pressure, in the different partitions.

But the counts \mathbf{k} do contain information about the ensemble, because knowledge of k_i constrains the possible counts in the other partitions, $k_{j \neq i}$. Counting is thus intrinsically wasteful in this situation, because it captures

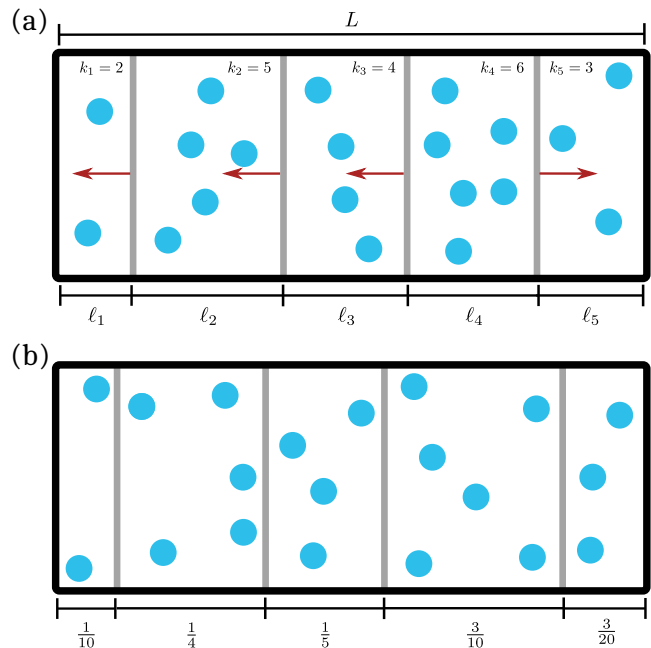


Figure 1. Schematic representation of a Szilard engine with $q = 5$ partitions and $N = 20$ particles. Panel (a) depicts the initial state, where the length of each partition is given by ℓ_i and the red arrows indicate in which direction each wall is going to move as the engine performs work. The final state, which is assumed to be reached quasi-statically is shown in panel (b), where the length of each partition is given by $\ell_i^{(\text{eq})} = k_i L/N$ and thus the pressure is the same in all of them. The values indicated in the lower panel corresponds to the counts obtained by setting $L = 1$.

some irrelevant information. Overall engine dissipation is proportional to irrelevant information [7] when the observer is viewed as a part of the information engine and the observer's energetic costs are taken into account.

We thus have a quantitative expectation of the costs and gains associated with the observer's memory, given the physical constraints imposed by our specific work extraction protocol. To make that precise, let \mathbf{K} denote a random variable with realizations $\mathbf{k} \in \mathcal{K}$. This, together with $p(\mathbf{k}|\mathbf{x})$, characterizes the memory we have chosen here: ensemble values of all N x -positions are mapped onto memory states, which are specified by the count vector \mathbf{k} . We are using the following standard notational shortcuts: vectors are bold face symbols, entropy functionals are written as $H[X] = -\langle \ln p(x) \rangle_{p(x)}$, and $H[\mathbf{X}] = -\langle \ln p(\mathbf{x}) \rangle_{p(\mathbf{x})} = -\langle \ln p(x_1, \dots, x_N) \rangle_{p(x_1, \dots, x_N)}$, where $\langle \cdot \rangle_p$ denotes the average. Conditional entropy and information are written accordingly [29]. For readers unfamiliar with information theory, all terms and calculations pertaining to this section are written out in detail in Appendix A.

While the counts committed to memory capture information in the amount of $I[\mathbf{X}, \mathbf{K}]$, the given work extraction protocol should not allow the observer to re-

cover that much work. We expect that the observer can use only the total information captured about individual particle locations, which is $\sum_{i=1}^N I[X_i, \mathbf{K}] = NI[X, \mathbf{K}]$, because the particles are identical.

A. Extractable work and relevant information

The work extracted [30] when all partitions change from ℓ_i to $k_i L/N$, is

$$\begin{aligned} W(\mathbf{k}) &= \sum_{i=1}^q \int_{\ell_i}^{L \frac{k_i}{N}} dV_i P_i = k_B T \sum_{i=1}^q \int_{\ell_i}^{L \frac{k_i}{N}} dV_i k_i / V_i \\ &= k_B T \sum_{i=1}^q k_i \ln \frac{k_i L}{N \ell_i}, \end{aligned} \quad (1)$$

The expected extracted work [31], $\langle W \rangle$, then results from averaging $W(\mathbf{k})$ over the distribution $P(\mathbf{k})$ of measurement vectors given by the multinomial distribution

$$P(\mathbf{k}) = \frac{N!}{k_1! \cdots k_q!} p_1^{k_1} \cdots p_q^{k_q}, \quad (2)$$

where $p_i = \ell_i/L$ is the probability of finding any one particle in partition i .

We now calculate the relevant information captured by the counts:

$$I[X, \mathbf{K}] = \sum_{\mathbf{k}} P(\mathbf{k}) \int dx P(x|\mathbf{k}) \ln \frac{P(x|\mathbf{k})}{P(x)}. \quad (3)$$

First, note that the probability of finding a single particle in any location along the x-axis is $P(x) = 1/L$. Second, within each partition, j , the overall probability of finding a particle, given the count vector \mathbf{k} , is k_j/N . The probability of finding a single particle in position x within partition j is uniform over the length of the partition, ℓ_j . Therefore, the probability of finding a particle in position x , given the counts \mathbf{k} is:

$$P(x|\mathbf{k}) = \frac{k_i}{N \ell_i}, \quad \sum_{j=0}^{i-1} \ell_j \leq x < \sum_{j=1}^i \ell_j, \quad (4)$$

for $i = 1, \dots, q$, using the convention $\ell_0 = 0$.

Putting everything together lets us compute how much information the counts contain about the location of a single particle:

$$I[X, \mathbf{K}] = \left\langle \sum_{i=1}^q \frac{k_i}{N} \ln \frac{k_i L}{N \ell_i} \right\rangle_{P(\mathbf{k})}. \quad (5)$$

We thus arrive at our main result: combining Eqs. (1) and (5), tells us that the average extracted work is proportional to N times the single particle location information captured by the counts,

$$\langle W \rangle = k_B T N I[X, \mathbf{K}]. \quad (6)$$

B. Irrelevant information retained in memory

To physically run the memory that keeps the counts of particles in each partition, heat in the amount of at least $k_B T I[\mathbf{X}, \mathbf{K}]$ joules has to be dissipated per cycle, on average (e.g. [6, 7]). But the memory can be used to extract, on average, work up to only $k_B T N I[X, \mathbf{K}]$ joules per cycle. The lower limit on overall average dissipation per cycle is the difference, which is proportional to the irrelevant information retained in memory [7]. Here, the counts keep irrelevant information in the amount of

$$I_{\text{irrel}} = I[\mathbf{X}, \mathbf{K}] - NI[X, \mathbf{K}] \quad (7)$$

$$= I[\mathbf{X}|\mathbf{K}] \geq 0, \quad (8)$$

equal to the conditional multi information,

$$I[\mathbf{X}|\mathbf{K}] = \left\langle \ln \left[\frac{P(x_1, \dots, x_N | k_1, \dots, k_q)}{\prod_i^N P(x_i | k_1, \dots, k_q)} \right] \right\rangle_{P(\mathbf{x}, \mathbf{k})}, \quad (9)$$

a non-negative quantity. To get from line (7) to line (8), note that $I[\mathbf{X}, \mathbf{K}] - \sum_{i=1}^N I[X_i, \mathbf{K}] = I[\mathbf{X}|\mathbf{K}] - I[\mathbf{X}]$. But the multi information $I[\mathbf{X}] = 0$ is zero here, because the particles are non-interacting.

The mapping from particle locations to the vector of counts characterizes the memory accessible to the information engine. Counting is a deterministic mapping, which means that the conditional entropy $H[\mathbf{K}|\mathbf{X}]$ is zero, since the counts are completely determined by the locations (unless measurement errors are taken into account). Therefore, we have $I[\mathbf{X}, \mathbf{K}] = H[\mathbf{K}]$. The irrelevant information retained in the counts is thus $I_{\text{irrel}}^{\text{count}} =$

$$H[\mathbf{K}] - NI[X, \mathbf{K}] = \ln \left[\frac{N^N}{N!} \right] - \sum_{i=1}^q \left\langle \ln \left[\frac{k_i^{k_i}}{k_i!} \right] \right\rangle_{p(k_i)} \quad (10)$$

We use Stirling's approximation, $n! \cong n^n e^{-n} \sqrt{2\pi n}$, to estimate for very large N , and for equidistant partitions:

$$I_{\text{irrel}}^{\text{count}} \cong \frac{1}{2} \left((q-1) \ln(2\pi N) - q \ln(q) \right) \quad (11)$$

C. Discussion

Better memories could perhaps do better, as could different work extraction mechanisms. The problem of finding an optimal memory was discussed for the case of generalized, partially observable information engines in [7]. The case where particles have a weak, short range, repulsive potential is studied in [26]. Furthermore, it is obvious that if the particles can initially be prepared in a non-equilibrium state, then a larger amount of work can be extracted from a cycle of the Szilárd engine, a point made in [32]. However, none of these variations on the original Szilárd scheme are the focus of this paper.

Instead, here we explore the consequences of the choices made in our specific setup: a naive, counting, observer, and a work extraction mechanism based solely on

pressure equalization. These choices constitute a *direct* extension of the Szilárd engine to N particles and q partitions, without the introduction of additional variations or optimizations. In this straightforward extension of the Szilárd engine to N particles and q partitions there is then only one remaining choice, namely the initial placement of the dividers. Conceptually, we could ascribe the task of inserting the partitions to the observer, which would also justify the fact that the observer knows ℓ .

We will now compute those partition locations ℓ that maximize relevant information, and hence maximize extractable work (Sec. II A). This tells us how to design the best naive N particle q partition classical extension of Szilárd's engine.

III. MAXIMIZING WORK EXTRACTION BY POSITIONING DIVIDER PARTITIONS

We now ask: given a system of N particles and q partitions, does there exist a partitioning that is optimal in the sense that it maximizes the average extractable work? That means, we seek to find a location vector $\hat{\ell}$ that maximizes $I[X, \mathbf{K}]$.

A. Maximum work extraction for a single movable wall

To build intuition, let us start with the case of one movable wall, i.e., two partitions. Let p denote the probability of finding a particle in the left partition of longitudinal size ℓ , which is $p = \ell/L = \langle k \rangle / N$. The probability of observing k particles out of the N possible in the left partition is then:

$$P(k) = \binom{N}{k} p^k (1-p)^{N-k}. \quad (12)$$

The conditional probability $P(x|k)$ is:

$$P(x|k) = \begin{cases} k/N\ell, & 0 \leq x \leq \ell \\ (N-k)/N(L-\ell), & \ell < x \leq L \end{cases}, \quad (13)$$

and the marginal probability is $P(x) = 1/L$. This yields the mutual information between the measurement k and the position of any single particle:

$$I[X, K] = \sum_k P(k) \left[\frac{k}{N} \ln \frac{kL}{N\ell} + \frac{N-k}{N} \ln \frac{(N-k)L}{N(L-\ell)} \right], \quad (14)$$

which is in agreement with earlier work [24, 25].

For one particle, a quick and intuitive calculation shows that the maximal value of $I[X, K]$ is attained by placing the wall in the middle. But this is not always the case for any number of particles, N . To illustrate this,

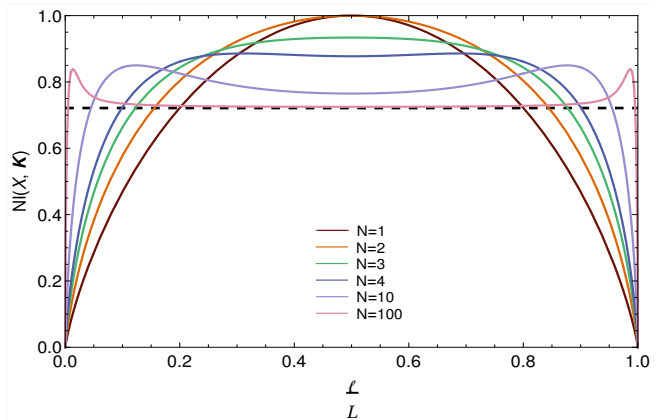


Figure 2. The mutual information $NI[X, \mathbf{K}]$ as a function of the position of the wall $p = \ell/L$, for a Szilárd's engine containing N particles. The different curves correspond to different values of N , while the dotted line is the asymptotic limit $\frac{1}{2 \ln(2)} \simeq 0.7213$ bits .

we plot $NI[X, K]$ against $p = \ell/L$ in Fig. 2, for various values of N . Only for $N \leq 3$ does the optimal position of the movable wall, $\hat{\ell}$, correspond to halving the volume ($\hat{\ell} = L/2$). For larger numbers of particles, the optimal partition is given by an asymmetric configuration of the partitions: the symmetric solution with the wall in the middle becomes a local minimum, and two maxima appear at $\hat{\ell}$ and $1 - \hat{\ell}$. This agrees with what was reported in [33]. To understand the mechanism behind the appearance of this asymmetric solution, we can perform an asymptotic expansion of $NI[X, K]$ for large N . Introducing the average count $n := \langle k \rangle = N \ell/L$, and the random variable Δ , such that $k = n + \Delta$, we first write the mutual information as:

$$\begin{aligned} NI[X, K] &= \sum_k P(k) \left[k \ln \frac{kL}{\ell N} + (N-k) \ln \frac{(N-k)L}{N(L-\ell)} \right] \\ &= \sum_k P(k) \left[(n + \Delta) \ln \left(1 + \frac{\Delta}{n} \right) \right. \\ &\quad \left. + (N - n - \Delta) \ln \left(1 - \frac{\Delta}{N - n} \right) \right], \end{aligned} \quad (15)$$

and then perform a Taylor expansion for small values of Δ . Recalling the following statistical properties of the binomial distribution, $\langle \Delta \rangle = 0$, $\langle \Delta^2 \rangle = Np(1-p)$, $\langle \Delta^3 \rangle = Np(1-p)(1-2p)$ and $\langle \Delta^4 \rangle = 3N(N-2)p^2(1-p)^2 + Np(1-p)$, we obtain

$$NI[X, K] = \frac{1}{2} + \frac{1}{4N} + \frac{(1-2p)^2}{12Np(1-p)} + \mathcal{O}(N^{-2}). \quad (16)$$

Apart from the first term, all orders go to zero in the limit of an infinite number of particles. For finite N , however, the second term gives the leading order on the decaying

value for the symmetric partition, while the third one is responsible of making the symmetric partition become a local minimum, because, as p deviates from $1/2$, this term increases. The same asymptotic expansion can be carried out for the general case of q partitions, which we will calculate later.

For a more quantitative analysis, we use an integral representation of the natural logarithm,

$$\ln z = \int_0^\infty \frac{du}{u} (e^{-u} - e^{-zu}), \quad (17)$$

to rewrite Eq. (5) as

$$NI[X, \mathbf{K}] = \sum_{i=1}^q \sum_{k_i} P(k_i) k_i \ln \left(\frac{k_i}{\langle k_i \rangle} \right) \quad (18)$$

$$= \sum_{i=1}^q \int_0^\infty \frac{du}{u} \left[\langle k_i (e^{-u} - e^{-k_i u}) \rangle - \langle k_i \rangle (e^{-u} - e^{-(k_i)u}) \right]. \quad (19)$$

Noticing that

$$\begin{aligned} \langle e^{-k_i u} \rangle &= \sum_{k_i=0}^N \binom{N}{k_i} (p_i e^{-u})^{k_i} (1-p_i)^{N-k_i} \\ &= (1-p_i(1-e^{-u}))^N, \end{aligned} \quad (20)$$

we obtain

$$NI[X, \mathbf{K}] = \sum_{i=1}^q \mathcal{F}_N(n_i), \quad (21)$$

with $n_i := \langle k_i \rangle$, and the function $\mathcal{F}_N(x)$ is given by

$$\mathcal{F}_N(x) = \int_0^\infty \frac{du}{u^2} \left\{ \left[1 - \frac{x}{N} (1 - e^{-u}) \right]^N - e^{-xu} \right\}. \quad (22)$$

We can hence write Eq. (14) as

$$NI[X, K] = \mathcal{F}_N(n) + \mathcal{F}_N(N-n), \quad (23)$$

where $n = \langle k \rangle$ is the average count.

Thus, maximizing $NI[X, K]$ with respect to the position ℓ of the wall for a fixed number of particles is equivalent to maximizing $\mathcal{F}_N(n) + \mathcal{F}_N(N-n)$ with respect to n . A plot of the function $\mathcal{F}_N(n)/\ln(2)$ (scaled to be in units of bits) can be found in Fig. 3 for various values of N . The optimal value $\hat{n}(N)$ which maximizes the mutual information must obey $\mathcal{F}'_N(n) = \mathcal{F}'_N(N-n)$. For $N \leq 3$ one finds that $\hat{n} = \frac{N}{2}$, and for larger numbers of particles we have that the optimal solution is given by an asymmetric partition (consistent with Fig. 2). For $N \rightarrow \infty$, either $\mathcal{F}_N(n)$ or $\mathcal{F}_N(N-n)$ becomes the dominant term in Eq. (23), and one can write that

$$\begin{aligned} \lim_{N \rightarrow \infty} NI[X, K] &= \mathcal{F}_\infty(n) \\ &= \int_0^\infty \frac{du}{u^2} (e^{-n(1-e^{-u})} - e^{-nu}). \end{aligned} \quad (24)$$

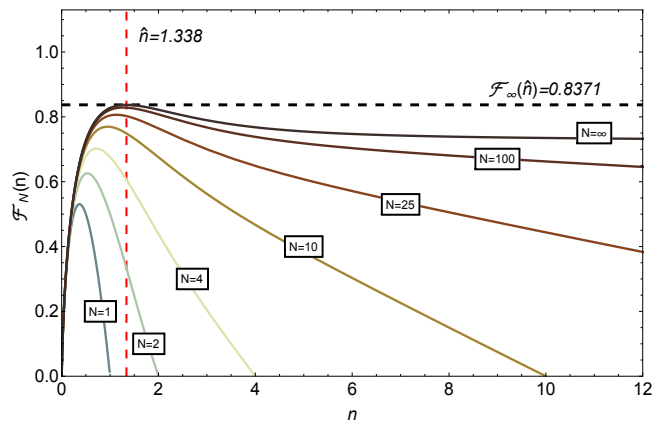


Figure 3. $\mathcal{F}_N(n)/\ln(2)$ [in bits] as a function of n , plotted for various values of N . In the limit of an infinite number of particles the maximum of $\mathcal{F}_N(n)$ is achieved at $\hat{n} \simeq 1.338$ and has a value of $\mathcal{F}_\infty(\hat{n}) \simeq 0.8371$ bits.

The maximal value of the mutual information occurs at $\hat{n}(\infty) \simeq 1.338$ and is $\mathcal{F}_\infty(\hat{n}) \simeq 0.8371$ bits. This is to be contrasted with a symmetric partition, which would give a lower value for the mutual information of $\frac{1}{2} \ln(2) \simeq 0.7213$ bits (see Fig. 2).

The larger the number of particles, N , the closer to the edge of the box we have to insert the movable wall to maximize average work extraction, while the average work extracted by a wall in the middle goes to zero. This explains not only why, with a container filled with some regular gas (N roughly between 10^{22} and 10^{23}), zero work can be extracted, on average, by putting a wall in the middle, but also why there is no chance to extract macroscopic work by implementing the optimal partitioning, because the necessary distances become much too extreme to realize.

Optimizing the average extracted work (for $q = 2$) also with respect to the number of particles, N , gives as the best choice either one or two particles, (with the wall in the middle), as can be appreciated from Figs. 2 and 5.

B. Optimal work extraction with q partitions

For one particle, we can trivially insert as many partitions as our experimental setup allows, and measure to the same resolution, in order to get more work out of the information engine, but we equally have to spend more energy to run the memory. We have, for one particle, that the cost and the potential benefit of the memory are precisely equal, because $I[\mathbf{X}, \mathbf{K}] = I[X, \mathbf{K}]$ for $N = 1$. Therefore, the overall bound on the engine's dissipation is unaffected. With one particle, a cyclically run Szilárd engine can, in principle, achieve zero dissipation.

How much work can be extracted from a Szilárd box with N particles and q partitions? We can use Eq. (21) to analyze this general case similarly to the $q = 2$ case discussed above. Finding the optimal partition that max-

imizes the mutual information for fixed N and q , *i.e.*, $\hat{\boldsymbol{\ell}} = (\hat{\ell}_1, \dots, \hat{\ell}_q) := \arg \max_{\boldsymbol{\ell}} (NI[X, \mathbf{X}])$, together with the maximal value, $\hat{\mathcal{I}}_q(N) = \max_{\boldsymbol{\ell}} (NI[X, \mathbf{K}])$, then reduces to finding the number vector $\mathbf{n} = (n_1, \dots, n_q)$ that maximizes Eq. (21). The number vector has to be normalized, $\sum_{i=1}^q n_i = N$, and to carry out the optimization, we introduce

$$\mathcal{I}_q^{(N)}(\mathbf{n}, \lambda) = \sum_{i=1}^q \mathcal{F}_N(n_i) - \lambda \left(\sum_{i=1}^q n_i - N \right), \quad (25)$$

which must be maximized with respect to $\{\mathbf{n}, \lambda\}$, where λ is a Lagrange multiplier. Then the optimal $\hat{\mathbf{n}}$ must obey

$$\begin{aligned} 0 &= \left. \frac{\partial \mathcal{I}_q^{(N)}(\mathbf{n}, \lambda)}{\partial n_i} \right|_{\mathbf{n}=\hat{\mathbf{n}}}, & i = 1, \dots, q, \\ 0 &= \left. \frac{\partial \mathcal{I}_q^{(N)}(\mathbf{n}, \lambda)}{\partial \lambda} \right|_{\mathbf{n}=\hat{\mathbf{n}}}, \end{aligned} \quad (26)$$

yielding:

$$\begin{aligned} \lambda &= \mathcal{F}'_N(\hat{n}_i), & i = 1, \dots, q, \\ N &= \sum_{i=1}^q \hat{n}_i. \end{aligned} \quad (27)$$

We must, moreover, be sure that

$$\sum_{i,j=1}^q \left. \frac{\partial^2 \mathcal{I}_q^{(N)}(\mathbf{n}, \lambda)}{\partial n_i \partial n_j} \right|_{\mathbf{n}=\hat{\mathbf{n}}} \delta n_i \delta n_j < 0, \quad (28)$$

with $\delta n_i = n_i - \hat{n}_i$, or equivalently,

$$\sum_{i=1}^q \mathcal{F}''_N(\hat{n}_i) (\delta n_i)^2 < 0. \quad (29)$$

with $\sum_{i=1}^q \delta n_i = 0$.

To understand the solutions to this system of equations, we plot $\mathcal{F}'_N(n)$ and $\mathcal{F}''_N(n)$ for different values of N and q in Fig. 4. The symmetric solution $n_i = N/q$ is stable as long as $\mathcal{F}''_N(N/q) < 0$ (red curve), but beyond that point unstable, and we need to investigate asymmetric solutions. For a given value of N and q , the equation $\lambda = \mathcal{F}'_N(\hat{n}_i)$ has two solutions $\hat{n}_i = n^-$ and n^+ with $n^- < n^+$. Since the Lagrange multiplier imposes a global constraint, this implies that, regardless of the order of the indices labeling the partitions, a general solution may correspond to having q^- partitions with solution n^- and q^+ partitions with n^+ , such that $q = q^- + q^+$ and $N = q^- n^- + q^+ n^+$. Importantly, the solution n^- is such that $\mathcal{F}''_N(n^-) \leq 0$, while for n^+ we have instead that $\mathcal{F}''_N(n^+) \geq 0$ (see solid red line in Fig. 4). Therefore, the stability condition (29) reads

$$|\mathcal{F}''_N(n^+)| \sum_{\ell=1}^{q^+} (\delta n_{\ell}^+)^2 - |\mathcal{F}''_N(n^-)| \sum_{\ell=1}^{q^-} (\delta n_{\ell}^-)^2 < 0, \quad (30)$$

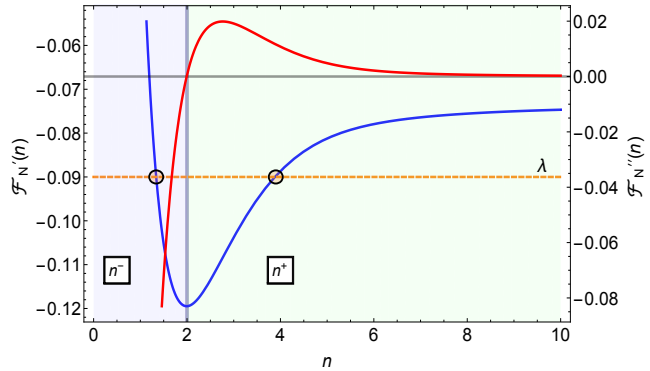


Figure 4. Plot of $\mathcal{F}'_N(n)$ (solid blue line) and $\mathcal{F}''_N(n)$ (solid red line) as a function of n , for $N = 10$. As we can see here, for a given value of λ (dashed orange line), there are two solutions of $\mathcal{F}'_N(n) = \lambda$ (represented here with black circles). These, denoted as n^- and n^+ are such that $\mathcal{F}''_N(n^-) < 0$ and $\mathcal{F}''_N(n^+) > 0$.

with

$$\sum_{\ell=1}^{q^+} \delta n_{\ell}^+ + \sum_{\ell=1}^{q^-} \delta n_{\ell}^- = 0. \quad (31)$$

This automatically implies that $q^+ \leq 1$, because if q^+ was 2 or larger, the inequality (30) could be violated by the choice $\delta n_{\ell}^- = 0$ for all $\ell = 1, \dots, q^-$. This means that the optimal asymmetric partition corresponds to having one large partition and $q - 1$ small and equal partitions.

This, in turn, allows us to write down a general expression for $\hat{\mathcal{I}}_q(N)$, namely:

$$\hat{\mathcal{I}}_q(N) = \mathcal{F}_N(n^+) + (q - 1)\mathcal{F}_N(n^-), \quad (32)$$

with the constraint $n^+ + (q - 1)n^- = N$. Expression (32) also contains the symmetric solution, corresponding to $n^+ = n^- = N/q$. Thus, the optimal value of the mutual information can always be written as

$$\mathcal{I}_q(N) = \max_{0 \leq n^+ \leq N} \left[\mathcal{F}_N(n^+) + (q - 1)\mathcal{F}_N\left(\frac{N - n^+}{q - 1}\right) \right], \quad (33)$$

and is plotted in Fig. 5 (brown triangle markers), and compared to the mutual information achieved for the symmetric partition, corresponding to $n_i = N/q$ for $i = 1, \dots, q$ (orange rhomboid markers). We see that, for a fixed q , there exists a critical value of particles $N^*(q)$ such that for $N \leq N^*(q)$ the optimal partition is always the symmetric one, while for $N > N^*(q)$, the optimal partition is asymmetric. We also found that for $q \gg 1$, this critical number of particles is given by $N^*(q) \simeq 2.1803q$. Additionally, Fig. 5 shows that there exists a value $\hat{N}(q) < N^*(q)$ for which $\mathcal{I}_q(N)$ is maximal at fixed q .

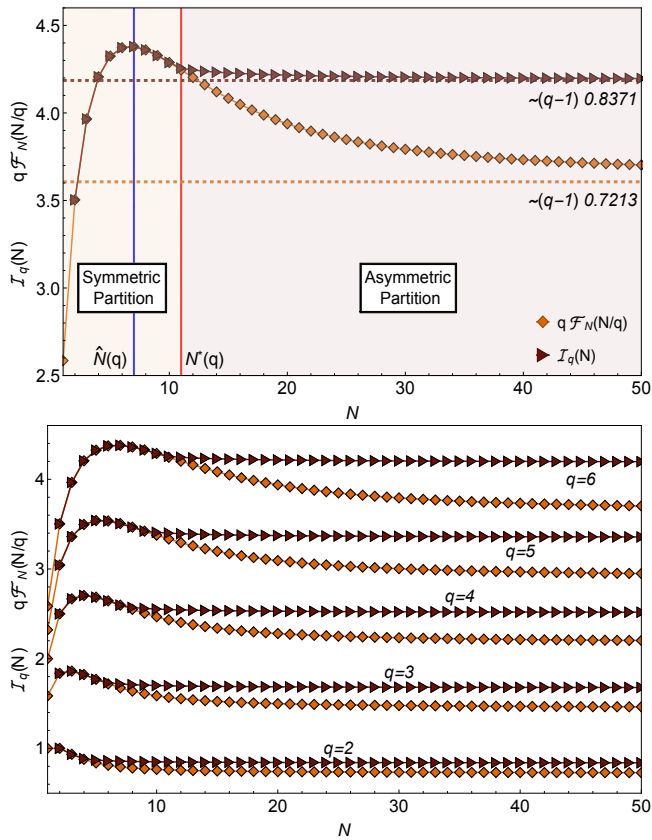


Figure 5. Plot of the mutual information for the symmetric partition $q\mathcal{F}_N(N/q)$ (orange rhomboid markers) and the optimal mutual information $\mathcal{I}_q(N)$ (brown triangle markers) as a function of N , in units of nats. Top panel: Plots for a particular value on the number of partitions ($q = 6$). As we can see, there exists a critical value $N^*(q)$ (vertical solid red line) on the number of particles, below which the optimal partition is the symmetric one, while above it the optimal partition corresponds to an asymmetric partition. Moreover the optimal mutual information becomes maximal for a particular number of particles $\hat{N}(q) < N^*(q)$ ($\hat{N}(s)$ is shown by the vertical solid blue line). Bottom panel: plot for various values of q , shown the same general features.

We obtain an asymptotic value of $\hat{N}(q)$ by noticing that, since this maximum is achieved for a symmetric partition we must have that $\hat{N} = nq$. The optimal mutual information is then given by $\hat{N}I[X, \mathbf{K}] = q\mathcal{F}_{nq}(n)$, so that for a large number of partitions, and using Eq. (24) we obtain $\hat{N}(q) \sim 1.338q$. Note also that as N goes to infinity, the value of $\mathcal{I}_q(N)$ given by Eq. (33) will be dominated by the $q - 1$ smaller partitions, yielding the asymptotic value of $\mathcal{I}_q(N) \sim (q - 1)0.8371$, as shown in the top panel of Fig. 5.

IV. CONCLUSIONS

In a Szilárd engine that can use an ideal gas with N particles, and for which the box can be partitioned into q partitions with walls that can move to extract work via a quasi-static process, the average extracted work is proportional to the information retained in the counts of how many particles fall into each partition about the single particle locations. The cost of running a memory that contains counts of how many particles are in each partition is proportional to the information retained about the ensemble locations. Run cyclically, the engine's efficiency is thus limited by the difference—the information retained in memory that is not relevant with respect to work extraction. This provides a non-negative lower bound on engine dissipation.

We calculated the maximum average extracted work, by optimizing over the initial wall locations, for given N and q . For fixed q , there is a critical value $N^*(q)$ below which the optimal partitioning is symmetric (partitions of equal volume), and above which an asymmetric solution is preferable. The maximum value of the average extractable work occurs at $\hat{N}(q) < N^*(q)$. Asymptotically, as $N \rightarrow \infty$, $\hat{N}(q)$ is linear in q , as is the maximal extractable average work.

The extension of Szilárd's engine we have explored here is basic in the sense that it allows for work extraction only via movement of the partitions along the x -axis, without building in additional work extraction mechanisms, such as used e.g. in [26], and in that it uses a naive, counting observer, without requiring an optimal observer, as discussed in [7].

Appendix A: Notation and Calculations

Vectors are bold face symbols, e.g., $\mathbf{x} = (x_1, \dots, x_N)$. Entropy is written as $H[X] = -\langle \ln p(x) \rangle_{p(x)}$, and $H[\mathbf{X}] = -\langle \ln p(\mathbf{x}) \rangle_{p(\mathbf{x})} = -\langle \ln p(x_1, \dots, x_N) \rangle_{p(x_1, \dots, x_N)}$, where $\langle \cdot \rangle_p$ denotes the average over the probability distribution p . In the main text, we sometimes use capital letters to denote probability distributions, in this Appendix we use small letters.

Conditional entropy is written as

$$H[X|Y] = -\langle \ln p(x|y) \rangle_{p(x,y)}, \quad (\text{A1})$$

$$H[\mathbf{X}|Y] = -\langle \ln p(\mathbf{x}|y) \rangle_{p(\mathbf{x},y)} \quad (\text{A2})$$

$$= -\langle \ln p(x_1, \dots, x_N|y) \rangle_{p(x_1, \dots, x_N, y)}, \quad (\text{A3})$$

$$H[X|\mathbf{Y}] = -\langle \ln p(x|\mathbf{y}) \rangle_{p(x,\mathbf{y})} \quad (\text{A4})$$

$$= -\langle \ln p(x|y_1, \dots, y_M) \rangle_{p(x, y_1, \dots, y_M)}, \quad (\text{A5})$$

$$H[\mathbf{X}|\mathbf{Y}] = -\langle \ln p(\mathbf{x}|\mathbf{y}) \rangle_{p(\mathbf{x},\mathbf{y})} \quad (\text{A6})$$

$$= -\langle \ln p(x_1, \dots, x_N|y_1, \dots, y_M) \rangle_{p(x_1, \dots, x_N, y_1, \dots, y_M)}, \quad (\text{A7})$$

respectively, and mutual information is defined the usual way [29], written as

$$I[X, Y] = H[X] - H[X|Y], \quad (\text{A8})$$

which can be written as the relative entropy between the joint distribution $p(x, y)$ and the product of the marginals, or between the conditional $p(x|y)$ and the marginal $p(x)$:

$$I[X, Y] = \left\langle \ln \left[\frac{p(x, y)}{p(x)p(y)} \right] \right\rangle_{p(x, y)} = \left\langle \ln \left[\frac{p(x|y)}{p(x)} \right] \right\rangle_{p(x, y)} \quad (\text{A9})$$

While mutual information is defined for two random variables, total correlation, also called multi information, is defined as the relative entropy between the joint distribution of N random variables $p(X_1, \dots, X_N)$ and the product of the marginals, $\prod_{i=1}^N p(X_i)$. We write it as

$$I[\mathbf{X}] = I[X_1, \dots, X_N] \quad (\text{A10})$$

$$= \left\langle \ln \left[\frac{p(x_1, \dots, x_N)}{\prod_{i=1}^N p(x_i)} \right] \right\rangle_{p(x_1, \dots, x_N)} \quad (\text{A11})$$

$$= \sum_{i=1}^N H[X_i] - H[X_1, \dots, X_N] \quad (\text{A12})$$

Conditional total correlation, also called conditional multi information, is defined similarly, with all distributions conditional on the outcome of another random variable, say Y :

$$I[\mathbf{X}, Y] = I[X_1, \dots, X_N|Y] \quad (\text{A13})$$

$$= \left\langle \ln \left[\frac{p(x_1, \dots, x_N|y)}{\prod_{i=1}^N p(x_i|y)} \right] \right\rangle_{p(x_1, \dots, x_N, y)} \quad (\text{A14})$$

$$= \sum_{i=1}^N H[X_i|Y] - H[X_1, \dots, X_N|Y]. \quad (\text{A15})$$

Similarly for $\mathbf{y} = (y_1, \dots, y_M)$, we have

$$I[\mathbf{X}, \mathbf{Y}] = H[\mathbf{X}] - H[\mathbf{X}|\mathbf{Y}] \quad (\text{A16})$$

$$= H[\mathbf{Y}] - H[\mathbf{Y}|\mathbf{X}]. \quad (\text{A17})$$

Using these definitions, it is easy to see that

$$\begin{aligned} I[\mathbf{X}, \mathbf{K}] &= \sum_{i=1}^N I[X_i, \mathbf{K}] \\ &= H[\mathbf{X}] - H[\mathbf{X}|\mathbf{K}] - \sum_{i=1}^N H[X_i] + \sum_{i=1}^N H[X_i|\mathbf{K}] \quad (\text{A18}) \\ &= I[\mathbf{X}|\mathbf{K}] - I[\mathbf{X}]. \quad (\text{A19}) \end{aligned}$$

Since the particles are identical, we have

$$NI[X, \mathbf{K}] = \sum_{i=1}^N I[X_i, \mathbf{K}], \quad (\text{A20})$$

and because they are non-interacting, the variables X_i are independently distributed before the measurement, i.e. $P(\mathbf{X}) = \prod_{i=1}^N P(X_i)$, whereby $I[\mathbf{X}] = 0$.

Eq. (8) in the main text follows directly from these observations:

$$I[\mathbf{X}, \mathbf{K}] - NI[X, \mathbf{K}] = I[\mathbf{X}|\mathbf{K}]. \quad (\text{A21})$$

Appendix B: Validity of our results in the quasi-static limit

In response to a request by one of the reviewers, this short appendix recalls elementary definitions in thermodynamics [30] that are used throughout this paper. Specifically, we assume: i) that the system is at all times in contact with a heat bath at constant temperature T , and ii) that the engine is operated quasi-statically during the complete work extraction process. Both of these assumptions are important because they allow us to model the engine as a set of coupled ideal gases instead of using a stochastic description for the collisions between the particles and the walls. This means that our results are not restricted to the thermodynamic limit, $N \rightarrow \infty$, but hold for any number of particles.

Assumption i) implies that for long enough times, our system will be Markovian, while ii) guarantees that dynamical processes occur over long time scales.

Wall i separates two chambers, of lengths ℓ_i and ℓ_{i+1} , respectively. Each time a particle hits the i -th wall the latter changes its position by a small amount ϵ , and thus, after a fixed time t , its displacement is given by $t(m_{i+1} - m_i)$, where m_i is the number of collisions impinging on the wall coming from the i -th partition. It is clear that m_i is a random variable, but given that particles do not interact with each other, and given that the system has a constant temperature, the number of collisions from either side can be described as a Poission process with constant rate. Hence, $\langle m_i \rangle \propto tk_i/\ell_i$ and similarly its variance fulfills that $\text{Var}[m_i] \propto tk_i/\ell_i$. By taking the $\epsilon \rightarrow 0$ limit it is clear that the time needed for any wall to move by a finite amount diverges as $1/\epsilon$, thus recovering the quasi-static limit. Note also that in this regime, the fluctuations of the position of the walls vanish as $\sqrt{\epsilon}$, and thus can be safely neglected. Therefore, when the engine is operated within this regime, the equilibrium condition is simply determined by the equality of pressure in all the partitions, as we used in the main text, but no further assumption on the value of N is needed.

ACKNOWLEDGMENTS

J.S., S.S. and R.D.H.R. thank The Abdus Salam International Centre for Theoretical Physics (ICTP) for hospitality and support during a visit in 2016 when the first part of this work was carried out. J.S. was partially supported by ICTP through the OEA-AC-98. S.S. is grateful for funding from the Foundational Questions Institute (together with the Fetzer Franklin Fund), Grant No. FQXi-RFP-1820. I.P.C. acknowledges financial support from the project UNAM-DGAPA-PAPIIT IA103417 and support from the London Mathematical Laboratory, where he is an external fellow.

-
- [1] J. Maxwell, *Theory of Heat*, Text-books of science (Longmans, 1871).
- [2] L. Szilard, *Zeitschrift für Physik* **53**, 840 (1929).
- [3] R. Landauer, *IBM Journal of Research and Development* **5**, 183 (1961).
- [4] C. H. Bennett, *Int. J. Theor. Phys.* **21**, 905 (1982).
- [5] T. Sagawa, *Prog. Theor. Phys.* **127**, 1 (2012).
- [6] J. M. Parrondo, J. M. Horowitz, and T. Sagawa, *Nature physics* **11**, 131 (2015).
- [7] S. Still, *Phys. Rev. Lett.* **124**, 050601 (2020).
- [8] T. Sagawa and M. Ueda, *Phys. Rev. Lett.* **102**, 250602 (2009).
- [9] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, *Nature* **483**, 187 (2012).
- [10] D. Mandal and C. Jarzynski, *PNAS* **109**, 11641 (2012).
- [11] T. Sagawa and M. Ueda, *Phys. Rev. E* **85**, 021104 (2012).
- [12] Z. Lu, D. Mandal, and C. Jarzynski, *Physics Today* **67**, 60 (2014).
- [13] J. V. Koski, V. F. Maisi, T. Sagawa, and J. P. Pekola, *Phys. Rev. Lett.* **113**, 030601 (2014).
- [14] Y. Jun, M. Gavrilov, and J. Bechhoefer, *Phys. Rev. Lett.* **113**, 190601 (2014).
- [15] I. A. Martínez, E. Roldán, L. Dinis, D. Petrov, J. M. R. Parrondo, and R. A. Rica, *Nature physics* **12**, 67 (2016).
- [16] J. Hong, B. Lambson, S. Dhuey, and J. Bokor, *Science advances* **2**, e1501492 (2016).
- [17] M. c. v. Gavrilov and J. Bechhoefer, *Phys. Rev. Lett.* **117**, 200601 (2016).
- [18] M. Gavrilov, R. Chétrite, and J. Bechhoefer, *PNAS* **114**, 11097 (2017).
- [19] E. Lathouwers and J. Bechhoefer, *Phys. Rev. E* **95**, 062144 (2017).
- [20] A. Kumar and J. Bechhoefer, *Appl. Phys. Lett.* **113**, 183702 (2018).
- [21] G. Paneru, D. Y. Lee, T. Tlusty, and H. K. Pak, *Phys. Rev. Lett.* **120**, 020601 (2018).
- [22] T. Admon, S. Rahav, and Y. Roichman, *Phys. Rev. Lett.* **121**, 180601 (2018).
- [23] D. H. Wolpert, C. Kempes, P. F. Stadler, and J. A. Grochow, eds., *The energetics of computing in life and machines* (SFI Press, 2019).
- [24] S. W. Kim, T. Sagawa, S. De Liberato, and M. Ueda, *Phys. Rev. Lett.* **106**, 070401 (2011).
- [25] K.-H. Kim and S. W. Kim, *Phys. Rev. E* **84**, 012101 (2011).
- [26] J. M. Horowitz and J. M. R. Parrondo, *New Journal of Physics* **13**, 123019 (2011).
- [27] J. Bengtsson, M. N. Tengstrand, A. Wacker, P. Samuelsson, M. Ueda, H. Linke, and S. M. Reimann, *Phys. Rev. Lett.* **120**, 100601 (2018).
- [28] The volume of partition i is ℓ_i , because the container has transverse unit area.
- [29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, 2006).
- [30] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, 1985).
- [31] Averages are denoted by $\langle \cdot \rangle$.
- [32] L. Touzo, M. Marsili, N. Merhav, and É. Roldán, *Journal of Statistical Mechanics: Theory and Experiment* **2020**, 093403 (2020).
- [33] P. Pal and A. Jayannavar, *arXiv preprint arXiv:1612.07007* (2016).