IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Sequence Complexity and Work Extraction

Neri Merhav

CCIT Report #883
March 2015

Electronics
Computers
Communications

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #883 March 2015

Sequence Complexity and Work Extraction

Neri Merhav

Department of Electrical Engineering, Technion, Haifa 32000, Israel.
E–mail: merhav@ee.technion.ac.il

Abstract. We consider a simplified version of a solvable model by Mandal and Jarzynski, which constructively demonstrates the interplay between work extraction and the increase of the Shannon entropy of an information reservoir which is in contact with a physical system. We extend Mandal and Jarzynski's main findings in several directions: First, we allow sequences of correlated bits rather than just independent bits. Secondly, at least for the case of binary information, we show that, in fact, the Shannon entropy is only one measure of complexity of the information that must increase in order for work to be extracted. The extracted work can also be upper bounded in terms of the increase in other quantities that measure complexity, like the predictability of future bits from past ones. Third, we point out to a partial extension to the case of non–binary information (i.e., a larger alphabet), and finally, we extend the scope to the case where the incoming bits (before the interaction) form an individual sequence, rather than a random one. In this case, the entropy before the interaction can be replaced by the Lempel–Ziv (LZ) complexity of the incoming sequence, a fact that gives rise to an entropic meaning of the LZ complexity, not only in information theory, but also in physics.

Keywords: information exchange, second law, entropy, complexity.

1. Introduction

Information processing and the role that it plays in thermodynamics is a very well–known concept that dates back to the second half of the nineteenth century, namely, to James Clerk Maxwell and his famous gedanken experiment, known as Maxwell's demon [13]. The Maxwell demon experiment shows that an intelligent agent, with access to measurements of velocities and positions of particles in a gas, is able to separate speedy particles from the slower ones, thereby creating a temperature difference without injecting energy into the system, which is seemingly in conflict with the second law of thermodynamics. Several decades later, Leo Szilard [18] continued this line of thought, and demonstrated the conversion of heat into work, using a model of a box that contains a single particle. He showed that by measurement and control, one may be able to extract work in a closed cycle of the system, which is again, in apparent contradiction with to the second law.

This suspected violation of the second law has triggered a long–lasting controversy and many other thought–provoking gedanken experiments that have eventually furnished the basis for a rather large of volume of theoretical work concerning the role and the implications of information processing in thermodynamics. A non–exhaustive list of recent works on the modern approach of incorporating informational ingredients in physical systems includes [1], [2], [3], [5], [6], [8], [9], [10], [11], [14], [15], [17], [19], [20], and [21]. In some of these works, the informational resources are available by means of measurement and feedback control (like in the Maxwell's demon and Szilard's engine) and other works are about physical systems that include, in addition to the traditional heat reservoir, also an *information reservoir*, which interacts with the system, but without any energy exchange. The main common motive in these works is in extended versions of the second law, where the expression of the entropy increase includes an extra entropic term that is associated with the information exchange. These extended versions of the second law are, of course, intimately related to Landauer's erasure principle [12].

Unlike earlier proposed thought experiments, that were mostly described in generic terms and were not fully specified, Mandal and Jarzynski [14] were the first to propose an explicit solvable model of a concrete system that behaves in the spirit of the Maxwell demon. Specifically, they described and analyzed a relatively simple autonomous system (based on a six–state Markov jump process), that when works as an engine, it converts thermal fluctuations (heat) into mechanical work, while writing digital information onto a running tape (in the role of an information reservoir), thereby increasing its Shannon entropy. It may also act as an eraser, which implements the opposite process of losing energy while erasing information, that is, decreasing the entropy. Several variations on this model, based on similar ideas, were offered in some subsequent works, e.g., [1], [2], [3], and [15].

In this paper, we consider a simplified version‡ of Mandal and Jarzynski's model [14] and we focus on extensions of their findings in several directions.

‡ Instead of the six–state Markov process of [14], we use a two–state process, which is easier to analyze.

(i) Allowing sequences of correlated bits rather than just independent bits.

(ii) At least for the case of binary information, it is shown that, in fact, the Shannon entropy is only one measure of complexity of the information that must increase in order for work to be extracted. The extracted work can also be upper bounded in terms of the increase in other quantities that measure complexity, like the predictability of future bits from past ones.

(iii) A partial extension is offered for the case of non–binary information (i.e., digital information with a larger alphabet).

(iv) Extension of the scope to the case where the incoming bits (before the interaction) form an individual sequence, namely, a deterministic sequence rather than a random one.

In the last item above, instead of the term of information entropy before the interaction, we have the Lempel–Ziv (LZ) complexity [22] of the incoming sequence, a fact that gives rise to an entropic meaning of the LZ complexity, not only in information theory, but also in physics.

We believe that similar extensions can be offered also for the other variations of this model, that appear in [1], [2], [3], and [15], as mentioned.

2. Notation Conventions

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors, their realizations and their alphabets will be denoted, respectively, by capital letters, the corresponding lower case letters, and the corresponding calligraphic letters, all superscripted by their dimension. For example, the random vector $X^n = (X_1, \ldots, X_n)$, (n – positive integer) may take a specific vector value $x^n = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, which is the n–th order Cartesian power of $\mathcal{X}$, the alphabet of each component of this vector. The probability of an event $\mathcal{E}$ will be denoted by $P[\mathcal{E}]$. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}[\mathcal{E}]$.

The Shannon entropy of a discrete random variable X will be denoted§ by $H(X)$, that is,

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \ln P(x), \tag{1}$$

where $\{P(x),\ x \in \mathcal{X}\}$ is the probability distribution of X. When we wish to emphasize the dependence of the entropy on the underlying distribution P, we denote it by $\mathcal{H}(P)$. The binary entropy function will be defined as

$$h(p) = -p \ln p - (1-p) \ln(1-p), \qquad 0 \le p \le 1. \tag{2}$$

§ Following the customary notation conventions in information theory, $H(X)$ should not be understood as a function H of the random outcome of X, but as a functional of the probability distribution of X.

Similarly, for a discrete random vector $X^n = (X_1, \ldots, X_n)$, the joint entropy is denoted by $H(X^n)$ (or by $H(X_1, \ldots, X_n)$), and defined as

$$H(X^n) = -\sum_{x^n \in \mathcal{X}^n} P(x^n) \ln P(x^n). \tag{3}$$

The conditional entropy of a generic random variable U over a discrete alphabet $\mathcal{U}$, given another generic random variable $V \in \mathcal{V}$, is defined as

$$H(U|V) = -\sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} P(u,v) \ln P(u|v), \tag{4}$$

which should not be confused with the conditional entropy given a *specific realization* of V, i.e.,

$$H(U|V=v) = -\sum_{u \in \mathcal{U}} P(u|v) \ln P(u|v). \tag{5}$$

The mutual information between U and V is

$$\begin{aligned} I(U;V) &= H(U) - H(U|V) \\ &= H(V) - H(V|U) \\ &= H(U) + H(V) - H(U,V), \end{aligned} \tag{6}$$

where it should be kept in mind that in all three definitions, U and V can themselves be random vectors. The Kullback–Leibler divergence (a.k.a. relative entropy or cross-entropy) between two distributions P and Q on the same alphabet $\mathcal{X}$, is defined as

$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)}. \tag{7}$$

3. Setup Description, Preliminaries and Objectives

Consider the system depicted in the Fig. 1, which is a simplified version of the one in [14].

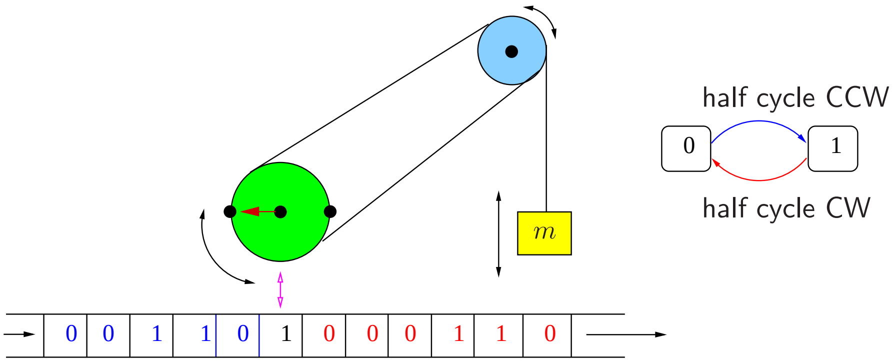

Figure 1. A system that interacts with a sequence of bits recorded on a running tape.

A device that consists of a wheel that is loaded (via a another wheel with transmission) by a mass m, interacts with a running tape that bears digital information in the form of a series of incoming bits, denoted $x_1, x_2, \ldots$, $x_i \in \{0,1\}$, $i = 1, 2, \ldots$.

The device also interacts thermally with a heat bath at temperature T (not shown in Fig. 1) in the form of heat exchange, but there is no energy exchange with the tape. During each time interval of τ seconds, $i\tau \le t < (i+1)\tau$ (i – positive integer), the device interacts with the i–th bit, x_i, in the following manner: If $x_i = 0$, then the initial state of the composite system (device plus bit) is '0' and then, due to random thermal fluctuations, the wheel may spontaneously rotate, say, half a cycle counter–clockwise (CCW) at a random time, thereby changing the state of the system to '1' and thus causing the mass to be lifted by Δ (which is half the circumference of the bigger wheel in Fig. 1). Then, at a later random time, it may rotate clockwise (CW), changing the state back to '0', and causing the mass to descend back by Δ, etc. The net change in the height of the mass, during this interval, depends, of course, only on the parity of the number of state transitions during this interval. At the end of this time interval, namely, at time $t = (i+1)\tau - 0$, the current state is recorded on the tape as the outgoing bit, denoted by y_i. Note that if $x_i = 0$ and $y_i = 1$, then the net work done by the device, during this time interval, is $\Delta W_i = mg\Delta$; otherwise $\Delta W_i = 0$. Similarly, if the incoming bit is $x_i = 1$, then the initial state is '1' and then the first state transition (if any) is associated with a CW rotation. By the same reasoning as before, at the end of the time interval, if $y_i = 0$, then the net work done by the device, during this interval, is $\Delta W_i = -mg\Delta$, otherwise, it is $\Delta W_i = 0$. Thus, in general, the work done during the i–th interval is $\Delta W_i = mg\Delta \cdot (y_i - x_i)$. Next, a new interval begins and it becomes the turn of bit x_{i+1} to interact with the device for τ seconds, and so on. It should be emphasized that this transition from the former outgoing bit y_i to a new incoming bit x_{i+1} is not accompanied by any energy exchange between the tape and the system (the wheel does not move in response to this transition). This new bit just determines which direction of rotation is enabled in which one is disabled.

The above described mechanism of back and forth transitions (with their associated rotations) within each interval is modelled as a two–state Markov jump process with transition rates $\lambda_{0\to1}$ and $\lambda_{1\to0}$, related by

$$\lambda_{0\to1} = \lambda_{1\to0} e^{-mg\Delta/kT}, \tag{8}$$

giving rise to an equilibrium (Boltzmann) distribution

$$P_{\text{eq}}[0] = \frac{1}{1+e^{-mg\Delta/kT}}; \quad P_{\text{eq}}[1] = \frac{e^{-mg\Delta/kT}}{1+e^{-mg\Delta/kT}}, \tag{9}$$

which manifests the fact that state '1' is more energetic than state '0', the energy difference being $\Delta E = mg\Delta$. At each interval, the temporal evolution of the probability of state '1' is according to the master equation:

$$\frac{\mathrm{d}P_t[1]}{\mathrm{d}t} = \lambda_{0\to1} - \lambda P_t[1] \tag{10}$$

where $\lambda \stackrel{\triangle}{=} \lambda_{0\to1} + \lambda_{1\to0}$. This simple first order differential equation is readily solved by

$$\begin{aligned} P_t[1] &= \frac{\lambda_{0\to1}}{\lambda} + \left(P_0[1] - \frac{\lambda_{0\to1}}{\lambda}\right) \cdot e^{-\lambda t} \\ &= P_{\text{eq}}[1] + (P_0[1] - P_{\text{eq}}[1]) \cdot e^{-\lambda t}, \end{aligned} \tag{11}$$

and, of course, $P_t[0]$ complements to unity. It is therefore readily seen that the mechanism that transforms the sequence of incoming bits, $x_1, x_2, \ldots$, into a sequence of outgoing bits, $y_1, y_2, \ldots$, is simply a binary–input, binary–output discrete memoryless channel‖ (DMC) $Q = [Q_{x\to y},\ x,y \in \{0,1\}]$, whose transition probabilities are given by

$$Q_{0\to 0} = 1 - Q_{0\to 1} = P_{\text{eq}}[0] + P_{\text{eq}}[1] \cdot e^{-\lambda\tau} \tag{12}$$

$$Q_{1\to 1} = 1 - Q_{1\to 0} = P_{\text{eq}}[1] + P_{\text{eq}}[0] \cdot e^{-\lambda\tau} \tag{13}$$

The expected work done by the device after n cycles is given by

$$\begin{aligned}\langle W_n\rangle &= mg\Delta \cdot \left\langle \sum_{i=1}^{n}[Y_i - X_i]\right\rangle \\ &= mg\Delta \cdot \sum_{i=1}^{n}(P[Y_i = 1] - P[X_i = 1]) \\ &= kTf \cdot \sum_{i=1}^{n}(P[Y_i = 1] - P[X_i = 1]),\end{aligned} \tag{14}$$

where $f \equiv mg\Delta/kT$. Now, from the above derived time evolution of the state distribution within an interval of duration τ, one easily finds that

$$P[Y_i = 1] = P_{\text{eq}}[1] + (P[X_i = 1] - P_{\text{eq}}[1]) \cdot e^{-\lambda\tau}, \tag{15}$$

which means a monotonic change, starting from $P[X_i = 1]$ and ending at $P_{\text{eq}}[1]$. In other words, $P[Y_i = 1]$ is always between $P(X_i = 1)$ and $P_{\text{eq}}[1]$.

We next focus on the informational (Shannon) entropy production, namely, the difference between the entropy of the outgoing bit–stream $\{Y_i\}$ and the entropy of the incoming bit–stream $\{X_i\}$. By the concavity of binary entropy function, $h(\cdot)$, it is easily seen that for every $s, t \in [0,1]$:

$$h(s) \le h(t) + (s-t) \cdot h'(t) \equiv h(t) + (s-t)\ln\frac{1-t}{t}. \tag{16}$$

Thus, setting $s = P[X_i = 1]$ and $t = P[Y_i = 1]$, we get

$$\begin{aligned}H(X_i) &\equiv h(P[X_i = 1]) \\ &\le h(P[Y_i = 1]) + (P[X_i = 1] - P[Y_i = 1])\ln\frac{1 - P[Y_i = 1]}{P[Y_i = 1]}.\end{aligned} \tag{17}$$

or equivalently,

$$(P[Y_i = 1] - P[X_i = 1])\ln\frac{1 - P[Y_i = 1]}{P[Y_i = 1]} \le H(Y_i) - H(X_i). \tag{18}$$

Now, if $P[Y_i = 1] \ge P[X_i = 1]$, then $P_{\text{eq}}[1] \ge P[Y_i = 1] \ge P[X_i = 1]$, and then

$$\begin{aligned}(P[Y_i = 1] - P[X_i = 1]) \cdot f &= (P[Y_i = 1] - P[X_i = 1])\ln\frac{1 - P_{\text{eq}}[1]}{P_{\text{eq}}[1]} \\ &\le (P[Y_i = 1] - P[X_i = 1])\ln\frac{1 - P[Y_i = 1]}{P[Y_i = 1]} \\ &\le H(Y_i) - H(X_i).\end{aligned} \tag{19}$$

‖ A memoryless channel is characterized by the assumption that the conditional probability of y^n given x^n is given by the product of conditional probabilities of y_i given x_i, $i = 1, 2, \ldots, n$.

Similarly, if $P[Y_i = 1] \leq P[X_i = 1]$, then $P_{\mathrm{eq}}[1] \leq P[Y_i = 1] \leq P[X_i = 1]$, and then again,

$$(P[Y_i = 1] - P[X_i = 1]) \cdot f \leq H(Y_i) - H(X_i) \tag{20}$$

since the terms f and $\ln\{(1-P[Y_i = 1])/P[Y_i = 1]\}$ are multiplied by $(P[Y_i = 1]-P[X_i = 1])$, which is now non–positive. Thus, in both cases, the last inequality holds, and so, as is actually shown in [14]

$$\langle \Delta W_i \rangle = kTf \cdot (P[Y_i = 1] - P[X_i = 1]) \leq kT[H(Y_i) - H(X_i)]. \tag{21}$$

Summing from $i = 1$ to n, the left–hand side of (21) gives

$$\langle W_n \rangle = kTf \cdot \sum_{i=1}^{n} (P[Y_i = 1] - P[X_i = 1]) \leq kT \sum_{i=1}^{n} [H(Y_i) - H(X_i)], \tag{22}$$

where left–hand–side (l.h.s.) is the total average total work after n cycles. The exact total average work is given by

$$\begin{aligned} \langle W_n \rangle &= kTf \cdot (1 - e^{-\lambda\tau}) \left(nP_{\mathrm{eq}}[1] - \sum_{i=1}^{n} P[X_i = 1] \right) \\ &= kTf \cdot (1 - e^{-\lambda\tau}) \left(\sum_{i=1}^{n} P[X_i = 0] - nP_{\mathrm{eq}}[0] \right), \end{aligned} \tag{23}$$

which is obviously positive if and only if $\frac{1}{n}\sum_{i=1}^{n} P[X_i = 0] > P_{\mathrm{eq}}[0]$. If $\{X_i\}$ are i.i.d. (Bernoulli), as assumed in [14] (as well as in subsequent follow–up papers mentioned earlier), then so are $\{Y_i\}$, and the right–hand side (r.h.s.) of (22) agrees with the total informational entropy production, $kT\Delta H \stackrel{\triangle}{=} kT[H(Y^n) - H(X^n)]$.

As discussed in [14], the inequality is saturated (in the sense that the ratio $f \cdot (P[Y_i = 1] - P[X_i = 1])/[H(Y_i) - H(X_i)]$ tends to unity) when $P[Y_i = 1]$ is very close to $P[X_i = 1]$ (which happens if either $\lambda\tau \ll 1$ or if $P[X_i = 1]$ is very close to $P_{\mathrm{eq}}[1]$, to begin with), but then the amount of work accumulated is very small. To approach the entropy difference limit when this difference is appreciably large, one may iterate in small steps, namely, work with $\lambda\tau \ll 1$ and feed $\{Y_i\}$ as an incoming bit–stream to another (identical, but independent) copy of the same device to generate, yet another bit–stream $\{Z_i\}$ with a further increased entropy, etc. Alternatively, one may feed $\{Y_i\}$ back to the same system. This way, with many repetitions of this process, the total work would be very close to kT times the overall growth of the Shannon entropy. This idea is in the spirit of quasi–static reversible processes in thermodynamics and statistical mechanics.

As explained in the Introduction, we extend these results in several directions:

(i) Allowing the incoming bits, $X_1, X_2 \ldots, X_n$, to be correlated rather than just independent, identically distributed (i.i.d.) bits. In this case, the sum of entropy differences, $\sum_i [H(Y_i) - H(X_i)]$, at the r.h.s. of (22) is different, in general, from the correct expression of the increase in the total Shannon entropy, $H(Y^n) - H(X^n)$, which in turn takes the correlations among the bits into account. It will be shown, nevertheless, that the correct expression associated with the entropy increase,

$kT[H(Y^n) - H(X^n)]$, is still an upper bound on the average work. This holds true for an arbitrary joint distribution of $(X_1, X_2, \ldots, X_n)$.

(ii) At least for the case of binary information, it will be shown that an inequality like (21) (even in its vector form) may hold even if the Shannon entropies on the r.h.s. are replaced by generalized entropies, which may serve as alternative measures of information complexity, such as the average probability of error in predicting the next bit X_{i+1} from the bits seen thus far $X_1, X_2, \ldots, X_i$, $i = 1, 2, \ldots, n$.

(iii) We provide a partial extension of the above to the case of non–binary information, i.e., $\{X_i\}$ and $\{Y_i\}$ take on values in a general finite alphabet, whose size may be larger than 2. The word "partial" here is due to the fact that we will not show that this holds true for an arbitrary value of τ, but only for $\tau \gg 1/\lambda$. Under the general alphabet size setting, however, item (ii) above is no longer claimed.

(iv) We extend the scope to the case where the incoming bits $x_1, x_2, \ldots, x_n$ form an individual sequence, namely, a deterministic sequence rather than a random one. In this case, in the r.h.s. of (22), the analogue of the probabilistic input entropy $H(X^n)$ will be (for large n) the Lempel–Ziv (LZ) complexity of the given sequence $x_1, x_2, \ldots, x_n$. As for the output entropy (Y^n is still a random vector), we will provide computable bounds.

4. Correlated Input Bits

Consider the case where the binary random vector $(X_1, \ldots, X_n)$, of the first n input bits, has a general joint distribution, As said, in this case, the r.h.s. of eq. (22) is no longer associated with the correct overall change in the Shannon entropy, $H(Y^n) - H(X^n)$. Nonetheless, our purpose, in this section, is to show that the latter expression (times kT) continues to be an upper bound on the expected work.

We proceed as follows. Using the fact that channel Q connecting X^n and Y^n is a DMC:

$$\begin{aligned}
H(Y^n) - H(X^n) &= \sum_{i=1}^n [H(Y_i|Y^{i-1}) - H(X_i|X^{i-1})] \\
&\geq \sum_{i=1}^n [H(Y_i|X^{i-1}, Y^{i-1}) - H(X_i|X^{i-1})] \\
&= \sum_{i=1}^n [H(Y_i|X^{i-1}) - H(X_i|X^{i-1})] \\
&= \sum_{i=1}^n \sum_{x^{i-1}} P(x^{i-1})[H(Y_i|X^{i-1} = x^{i-1}) - \\
&\quad H(X_i|X^{i-1} = x^{i-1})] \\
&= \sum_{i=1}^n \sum_{x^{i-1}} P(x^{i-1})\{h(P[Y_i = 1|X^{i-1} = x^{i-1}]) - \\
&\quad h(P[X_i = 1|X^{i-1} = x^{i-1}])\}
\end{aligned}$$

$$\begin{aligned}
&\geq f\cdot\sum_{i=1}^{n}\sum_{x^{i-1}}P(x^{i-1})(P[Y_i=1|X^{i-1}=x^{i-1}]-\\
&\quad P[X_i=1|X^{i-1}=x^{i-1}])\\
&=f\cdot\sum_{i=1}^{n}(P[Y_i=1]-P[X_i=1])\\
&=\frac{\langle W_n\rangle}{kT},
\end{aligned}\tag{24}$$

where the third line is due to the fact that Y_i is statistically independent of Y^{i-1} given X^{i-1}, and the second inequality is again due to the concavity of $h(\cdot)$.

Discussion. We have two upper bounds on the total work, $kT\sum_{i=1}^{n}[H(Y_i)-H(X_i)]$ and $kT[H(Y^n)-H(X^n)]$. As an upper bound, the former is always tighter, in other words, we argue (see Appendix A for the proof) that

$$H(Y^n)-H(X^n)\geq\sum_{i=1}^{n}[H(Y_i)-H(X_i)],\tag{25}$$

and so for the purpose of bounding the expected work, there is no point in looking at higher order entropies of the incoming and outgoing processes. However, from the physical point of view, the inequality $\langle W_n\rangle\leq kT[H(Y^n)-H(X^n)]$ remains meaningful since the difference $k[H(Y^n)-H(X^n)]-\langle W_n\rangle/T$ has the natural meaning of the total entropy production (of the combined system and its environment) for the more general case considered, i.e., where $\{X_i\}$ may be correlated. The non–negativity of this difference is then a version of the (generalized) second law of thermodynamics for systems that include information reservoirs. It follows from this discussion that if one has any control on the incoming bit sequence, then introducing correlations among them is counter–productive in the sense that it only enlarges the entropy production without enlarging the extracted work (for a given marginal probability assignment). In other words, among all input vectors with a given average marginal, $\bar{P}[x]=\frac{1}{n}\sum_{i=1}^{n}P[X_i=x]$, the best one is an i.i.d. process (i.e., a Bernoulli process) with a single–bit marginal given by $P[X_i=x]=\bar{P}[x]$ for all i. In any other case, there is an extra entropy production due to input correlations.

Note that if X^n is a codeword from a rate–R channel block code (with equiprobable messages) for reliable communication across the channel Q, namely, $H(X^n)=nR$ and $H(X^n|Y^n)$ is small by Fano's inequality [4, Section 2.10]), then

$$\begin{aligned}
H(Y^n)-H(X^n)&\approx H(Y^n|X^n)\\
&=\sum_{i=1}^{n}H(Y_i|X_i)=n[\bar{P}[0]h(Q_{0\to 0})+\bar{P}[1]h(Q_{1\to 1})].
\end{aligned}\tag{26}$$

In this case, as $H(Y^n)\approx n[R+\bar{P}[0]h(Q_{0\to 0})+\bar{P}[1]h(Q_{1\to 1})]$, one can reliably recover from Y^n both the incoming process X^n and the entire history of of (net) movements of the wheel across the various intervals, so no information is lost.

5. Other Measures of Sequence Complexity

Note that the only properties of the entropy function that were used in Section 3 were: (i) concavity, and (ii) $h'(P_{\rm eq}[1]) = f$. The second property does not pose any serious limitation because any concave function can either be scaled or added with a linear term (both without harming the concavity property), so that (ii) would hold. It follows then that the Shannon entropy is not the only measure that describes the increased complexity of information that must accompany the extracted work. In other words, there are additional measures for the amount extra randomness or the "amount of information" that must be written in order to make the system convert heat to work.

We describe a generalized entropy function that is based on a function $L_x(s)$, which is an arbitrary function of $x \in \{0, 1\}$ and a variable $s \in \mathcal{S}$, that can be thought of as a 'loss' associated with the choice of s when the observation is x. We then define a generalized entropy function as the minimum achievable average loss associated with a binary random variable X, with $P[X = 1] = 1 - P[X = 0] = p$, that is

$$\boldsymbol{h}(p) = \min_{s\in\mathcal{S}}[(1-p)\cdot L_0(s) + p\cdot L_1(s)]. \tag{27}$$

Indeed, the binary Shannon entropy $h(p)$ is obtained as a special case for $L_0(s) = -\ln(1-s)$ and $L_1(s) = -\ln s$, $\mathcal{S} = [0, 1]$, as the minimum is attained for $s^* = p$. Since $\boldsymbol{h}(p)$ is the minimum of affine functions of p, it is clearly concave. Two additional examples of entropy–like functions are the following:

(i) Let $L_x(s) = \mathcal{I}[s \neq x]$, $\mathcal{S} = \{0, 1\}$, measure the loss in (possibly erroneous) 'guessing' of x by s. In this case, $\boldsymbol{h}(p) = \min\{p, 1-p\}$.

(ii) The squared–error loss function, $L_x(s) = (x-s)^2$, $\mathcal{S} = [0, 1]$, yields $\boldsymbol{h}(p) = p(1-p)$.

The extension of (21) now asserts that the average work extraction $\langle\Delta W_i\rangle$, within a single cycle, cannot exceed

$$\frac{mg\Delta}{\boldsymbol{h}'(P_{\rm eq}[1])}\cdot\Delta\boldsymbol{h} = \frac{kTf}{\boldsymbol{h}'(P_{\rm eq}[1])}\cdot\Delta\boldsymbol{h}, \tag{28}$$

where $\Delta\boldsymbol{h} = \boldsymbol{h}(P[Y_i = 1]) - \boldsymbol{h}(P[X_i = 1])$ is the increase in the (generalized) 'complexity' in Y_i relative to X_i, and where we have assumed that $\boldsymbol{h}(\cdot)$ is differentiable at $p = P_{\rm eq}[1]$. We will comment on the non–differentiable case shortly.

Denoting $\boldsymbol{H}(X_i) = \boldsymbol{h}(P[X_i = 1])$ and $\boldsymbol{H}(Y_i) = \boldsymbol{h}(P[Y_i = 1])$, we can generalize the above discussion (including (24), provided that the first equality is considered a definition) to correlated sequences of bits, by introducing the definition

$$\boldsymbol{H}(X_i|X^{i-1}) = \sum_{x^{i-1}} P(x^{i-1})\boldsymbol{h}(P[X_i = 1|X^{i-1} = x^{i-1}]) \tag{29}$$

and similar definitions for the other generalized conditional entropies. Considering the first example above, $\boldsymbol{H}(X_i|X^{i-1})$ designates the *predictability* [7] of X_i given X^{i-1}, i.e., the minimum achievable probability of error in guessing X_i from X^{i-1}, which is certainly a reasonable measure of complexity. As for the second example above, $\boldsymbol{H}(X_i|X^{i-1})$ has the meaning of the minimum mean squared error in estimating X_i based on X^{i-1}. Here,

$\boldsymbol{h}'(P_{\text{eq}}[1])) = \tanh(f/2)$. Thus, the factor $kTf/\boldsymbol{h}'(P_{\text{eq}}[1])) = kTf/\tanh(f/2)$, which is about $kTf = mg\Delta$ at very low temperatures, and about $2kT$ at very high temperatures.

On a technical note, observe that in general $\boldsymbol{h}(\cdot)$ may not be differentiable at $P_{\text{eq}}[1]$, but due to the concavity, there are always one–sided derivatives $\boldsymbol{h}'_+(P_{\text{eq}}[1]) = \lim_{\delta\downarrow 0}[\boldsymbol{h}(P_{\text{eq}}[1]+\delta) - \boldsymbol{h}(P_{\text{eq}}[1])]/\delta$ and $\boldsymbol{h}'_-(P_{\text{eq}}[1]) = \lim_{\delta\uparrow 0}[\boldsymbol{h}(P_{\text{eq}}[1]+\delta) - \boldsymbol{h}(P_{\text{eq}}[1])]/\delta$, with $\boldsymbol{h}'_-(P_{\text{eq}}[1]) \geq \boldsymbol{h}'_+(P_{\text{eq}}[1])$. We can always use either one. In case of a strict inequality, we can choose the one that gives the tighter inequality, namely, $\boldsymbol{h}'_-(P_{\text{eq}}[1])$ if $\sum_i P[Y_i = 1] \geq \sum_i P[X_i = 1]$ and $\boldsymbol{h}'_+(P_{\text{eq}}[1])$ otherwise.

Another class of generalized entropies obey the form $\boldsymbol{H}(X) = \langle S[1/P(X)]\rangle$, where S is am arbitrary concave function (e.g., $S[u] = \ln u$ gives the Shannon entropy), which is easily seen to be concave functional of P. In the binary case considered here, this would amount to $\boldsymbol{h}(p) = pS[1/p] + (1-p)S[1/(1-p)]$. The concavity property guarantees that our earlier arguments hold for this kind of generalized entropy as well. Similar comments apply to yet another class of generalized entropies, $\boldsymbol{H}(X) = \sum_x S[P(x)]$, where S is again concave (e.g., $S[u] = -u\ln u$ gives the Shannon entropy).

This discussion sets the stage for a richer family of bounds on the extracted work, which depend on various notions of sequence complexity. Provided that $\boldsymbol{h}(\cdot)$ is differentiable at $P_{\text{eq}}[1]$, these bounds are asymptotically met in the limit of infinitesimally small differences between $P[Y_i = 1]$ and $P[X_i = 1]$, as discussed above in the context of the ordinary entropy. Nonetheless, among all generalized entropies we have discussed, only the Shannon entropy is known to be invariant under permutations, e.g., for $n = 2$, $H(X_1) + H(X_2|X_1) = H(X_2) + H(X_1|X_2)$, but in general, it not true that $\boldsymbol{H}(X_1) + \boldsymbol{H}(X_2|X_1) = \boldsymbol{H}(X_2) + \boldsymbol{H}(X_1|X_2)$. Also, it is not clear if and how any of the other entropy–like functionals continue to serve in bounding the average work when the the setup is extended to larger alphabets (see Section 6 below). These two points give rise to the special stature of the ordinary Shannon entropy, which prevails in a deeper sense and in more general situations.

6. Non–Binary Sequences

One trivial extension to the non–binary case is associated with grouping non–overlapping chunks of ℓ bits and considering them as random variables with an alphabet of size 2^ℓ. Here each input symbol, say $\boldsymbol{x}_i$, is a vector of ℓ (possibly correlated) bits $(x_{i1}, \ldots, x_{i\ell})$ and we imagine ℓ identical, independent copies of the above system, where the various bit-streams $\{x_{ij},\ i = 1, 2, \ldots\}$ are fed into the corresponding copies of the system, $j = 1, 2\ldots, \ell$, and $\{y_{ij},\ i = 1, 2, \ldots\}$ are the corresponding outgoing bit–streams. The ℓ copies operate independently during each interval. Letting $\boldsymbol{X}^n$ and $\boldsymbol{Y}^n$ denote the the collection of n input and output binary ℓ–vectors, we can now show, exactly like in (24), that $kT[H(\boldsymbol{Y}^n) - H(\boldsymbol{X}^n)]$ is an upper bound on the total work carried over the ℓ systems together, after n rounds. This is done exactly like in (24), except that now the summation is two–dimensional over $n \cdot \ell$ terms, exploiting both temporal and spatial correlations.

Another, somewhat more interesting extension to a general finite alphabet $\mathcal{X}$ of size K, concerns a Markov jump process with K states. For each state $x \in \mathcal{X}$, there is a corresponding height increment, $\Delta(x)$ (which may be positive or negative), relative to some reference state, say x_0, with $\Delta(x_0) \equiv 0$. Accordingly, each state x is associated with energy, $E(x) = mg\Delta(x)$, and the transition rates of the underlying Markov process obey detailed balance accordingly. Here we assume that τ is very large compared to all time constants so that the final distribution at each interval is nearly in equilibrium.¶ The following inequality is clearly equivalent to the inequality $D(P_0\|P_{\text{eq}}) \geq 0$:

$$\sum_{x\in\mathcal{X}}(P_{\text{eq}}[x] - P_0[x])\ln\frac{1}{P_{\text{eq}}[x]} \leq \mathcal{H}(P_{\text{eq}}) - \mathcal{H}(P_0). \tag{30}$$

Here P_0 represents the probability distribution of the incoming symbol X_i, which is also the initial distribution at each interval, and P_{eq} is the distribution of the outgoing symbol Y_i, which is the equilibrium distribution for large τ. Since

$$\ln\frac{1}{P_{\text{eq}}[x]} = \ln Z + \frac{mg\Delta(x)}{kT}, \tag{31}$$

$Z = \sum_x \exp\{-mg\Delta(x)/kT\}$ being the partition function, the l.h.s. of (30) gives the average work per cycle (in units of kT), and the r.h.s. is, of course, the entropy difference.

This discussion easily extends to the case of correlated input symbols, as in Section 4, since for large τ, the outgoing process is still i.i.d., where each symbol is distributed according to P_{eq}.

7. Individual Sequences and the LZ Complexity

Finally, we extend the scope to the case where $x_1, x_2, \ldots$ is an individual sequence, namely, an arbitrary deterministic sequence, with no assumptions concerning the mechanism that has generated it. The outgoing sequence is, of course, still random due to the randomness of the state transitions. In this setting, the LZ complexity of the incoming sequence will play a pivotal role, and therefore, before moving on to the derivation for the individual–sequence setting, we pause to provide a brief background concerning the LZ complexity, which can be thought of as an individual–sequence counterpart of entropy.

In 1978, Ziv and Lempel [22] invented their famous universal algorithm for data compression, which has been considered a major breakthrough, both from the theoretical aspects and the practical aspects of data compression. For an given (individual) infinite sequence, $x_1, x_2, \ldots$, the LZ algorithm achieves a compression ratio, which is asymptotically as good as that of the best data compression algorithm that is implementable by a finite–state machine. To the first order, the compression

¶ Here, unlike the case of a two–state Markov process, we cannot rely, in general, on the property that the state distribution at any intermediate time t, is always a convex combination of the initial distribution and the equilibrium distribution.

ratio achieved by the LZ algorithm, upon compressing the first n symbols, $x^n = (x_1, x_2, \ldots, x_n)$, i.e., the LZ complexity of x^n, is about

$$\rho(x^n) = \frac{c(x^n)\log c(x^n)}{n}, \tag{32}$$

where $c(x^n)$ is the number of distinct *phrases* of x^n obtained upon applying the so called *incremental parsing procedure*. The incremental parsing procedure works as follows. The sequence $x_1, x_2, \ldots, x_n$ is parsed sequentially (from left to right), where each parsed phrase is the shortest string that has not been encountered before as a parsed phrase, except perhaps the last phrase, which might be incomplete. For example, the sequence $x^{17} = 10001101110100010$ is parsed as $1, 0, 00, 11, 01, 110, 10, 001, 0$. The first two phrases are obviously '1 and '0 as there is no 'history' yet. The next '0' has already been seen as a phrase, but the string '00' has not yet been seen, so the next phrase is '00'. Proceeding to the next bit, '1' has already appeared as a phrase, but '11' has not, and so on. In this example then, $c(x^{17}) = 9$. The idea of the LZ algorithm is to sequentially compress the sequence phrase–by–phrase, where each phrase, say, of length r, is represented by a pointer to the location of the appearance of the first $r-1$ symbols as a previous phrase (already decoded by the de-compressor), plus an uncompressed representation of the r-th symbol of that phrase. It is shown in [22] that if the LZ algorithm is applied to a random vector X^n, that is sampled from a stationary and ergodic process, then $\rho(X^n)$ converges with probability one to the entropy rate of the process, $\bar{H} = \lim_{n\to\infty} H(X_n|X^{n-1})$. In that sense, $\rho(x^n)$ can be thought of as an analogue of entropy in the individual–sequence setting.

The general idea, in this section, is that, in the context of the entropic upper bound on the extracted work, the role of the input entropy, $H(X^n)$, of the probabilistic case, will now be played by $\rho(x^n)$, whereas $H(Y^n)$ will be upper bounded in terms of $\rho(x^n)$. Thus, the concept of LZ complexity is not only analogous to information–theoretic entropy, but in a way, it also plays an entropic role in the physical sense.

Equipped with this background, we now move on to the derivation. For simplicity, we consider the binary case, but everything can be extended to the non–binary case at least for large τ, following the considerations of Section 6. Consider then an individual binary sequence $(x_1, x_2, \ldots, x_n)$ of incoming bits. Let ℓ be a divisor of n and chop the sequence into n/ℓ non-overlapping blocks of length ℓ, $\boldsymbol{x}_i = (x_{i\ell+1}, x_{i\ell+2}, \ldots, x_{i\ell+\ell})$, $i = 0, 1, \ldots, n/\ell - 1$. Consider now the empirical distribution of ℓ–blocks

$$\hat{P}(x^\ell) = \frac{\ell}{n}\sum_{i=0}^{n/\ell-1} \mathcal{I}[\boldsymbol{x}_i = x^\ell], \qquad x^\ell \in \{0,1\}^\ell \tag{33}$$

Now, define

$$\hat{P}[X_i = 1] = \sum \hat{P}(x^\ell), \qquad i = 1, 2, \ldots, \ell \tag{34}$$

where the summation is over all binary ℓ–vectors $\{x^\ell\}$ whose i–th coordinate is 1. The average work for a given $(x_1, x_2, \ldots, x_n)$ is given by

$$\langle W_n \rangle = kTf \cdot \sum_{t=1}^{n} (\langle Y_t \rangle - x_t)$$

$$
\begin{aligned}
&= kTf \cdot \frac{n}{\ell} \cdot \sum_{i=1}^{\ell}(P[Y_i = 1] - \hat{P}[X_i = 1]) \\
&\le \frac{kTfn}{\ell} \cdot [\tilde{H}(Y^\ell) - \hat{H}(X^\ell)] \qquad (35)
\end{aligned}
$$

where $P[Y_i = 1] = \hat{P}[X_i = 1]Q_{1\to 1} + \hat{P}[X_i = 0]Q_{0\to 1}$, $\hat{H}(X^\ell)$ is the empirical entropy of ℓ–blocks associated with $x^n = (x_1, x_2, \ldots, x_n)$ and $\tilde{H}(Y^\ell)$ is the output entropy of ℓ–vectors that is induced by the input assignment $\{\hat{P}(x^\ell)\}$ and ℓ uses of the memoryless channel Q. The last inequality is simply an application of the results of Section 4 to the case where the joint distribution of X^n is $\hat{P}(\cdot)$. This already gives some meaning to the notion of entropy production in this case, where the incoming bits are deterministic. However, the choice of the parameter ℓ (among the divisors of n) appears to be somewhat arbitrary. In the following, we further obtain another bound, which is asymptotically, independent of ℓ. In this bound, $\hat{H}(X^\ell)$ will be replaced by $\rho(x^n)$. From [16, eq. (21)], we have the following lower bound on $\hat{H}(X^\ell)$ in terms of its LZ complexity (setting the alphabet size $\alpha = 2$ and passing logarithms to base e):

$$
\begin{aligned}
\frac{\hat{H}(X^\ell)}{\ell} &\ge \rho(x^n) - \frac{8\ell \ln 2}{(1-\epsilon_n)\log n} - \frac{2\ell 4^\ell \ln 2}{n} - \frac{\ln 2}{\ell} \\
&\equiv \rho(x^n) - \delta(n, \ell), \qquad (36)
\end{aligned}
$$

where $\epsilon_n \to 0$. This inequality is a result of comparing the compression ratio of a certain block code to a lower bound on the compression performance of a general finite–state machine, which is essentially $\rho(x^n)$. Of course, $\lim_{\ell\to\infty}\lim_{n\to\infty}\delta(n,\ell) = 0$. Let δ_n denote the minimum of $\delta(n,\ell)$ over all $\{\ell\}$ that are divisors of n.

It remains to deal with the entropy of Y^n. First, observe that the case of very large τ is obvious, because in this case, $H(Y^n) = n\mathcal{H}(P_{\text{eq}})$ as $\{Y_i\}$ is i.i.d. with marginal P_{eq}, regardless of x^n. Therefore, neglecting the term δ_n, the upper bound on the extracted work becomes

$$
\langle W_n \rangle \le kTn[\mathcal{H}(P_{\text{eq}}) - \rho(x^n)]. \qquad (37)
$$

Also, in this case, the second part of Section 6 is valid, and so, this derivation readily extends to the non–binary alphabet case.

For a general τ, we will remain in the binary case. Given the binary–input, binary–output DMC $Q : X \to Y$, define the single–letter function

$$
U(s) = \max\{H(Y) : \; H(X) \ge s\}. \qquad (38)
$$

The function $U(s)$ is concave and monotonically decreasing. The monotonicity is obvious. As for the concavity, indeed, let P_0 and P_1 be the achievers of $U(s_0)$ and $U(s_1)$, respectively. Then, for $0 \le \lambda \le 1$, the entropy of $P_\lambda = (1-\lambda)P_0 + \lambda P_1$ is never less than $(1-\lambda)s_0 + \lambda s_1$, and so,

$$
\begin{aligned}
U[(1-\lambda)s_0 + \lambda s_1] &\ge H(Y_\lambda) \qquad Y_\lambda \text{ being induced by } P_\lambda \text{ and } Q \\
&\ge (1-\lambda)H(Y_0) + \lambda H(Y_1) \\
&= (1-\lambda)U(s_0) + \lambda U(s_1). \qquad (39)
\end{aligned}
$$

Note that if $H(X^\ell)\ge \ell s$, then a–fortiori, $\sum_{i=1}^\ell H(X_i)\ge \ell s$, and so, for the given DMC,

$$\begin{aligned}H(Y^\ell) &\le \sum_{i=1}^\ell H(Y_i)\\ &\le \sum_{i=1}^\ell U[H(X_i)]\\ &\le \ell\cdot U\left[\frac{1}{\ell}\sum_{i=1}^\ell H(X_i)\right]\\ &\le \ell\cdot U(s). \end{aligned}\tag{40}$$

Applying this to the input distribution $\{\hat{P}(x^\ell)\}$ and the channel Q, we have, by (36):

$$\tilde{H}(Y^\ell)\le \ell\cdot U\left[\rho(x^n)-\delta_n\right],\tag{41}$$

and so, defining the function

$$V(s)\stackrel{\triangle}{=}U(s)-s,\tag{42}$$

which is concave and decreasing as well, we get the following upper bound on $\langle W_n\rangle$ in terms of the LZ complexity of x^n:

$$\langle W_n\rangle\le kTfn\cdot V\left[\rho(x^n)-\delta_n\right].\tag{43}$$

It tells us, among other things, that the more x^n is LZ–compressible, the more work extraction one can hope for.

This upper bound is tight in the sense that no other bound that depends on x^n only via its LZ compressibility $\rho(x^n)$ can be tighter, because for a given value ρ (in the range where the constraint in the maximization defining $U(\rho)$ is attained with equality) of the LZ compressibility, $\rho(x^n)$, there exist sequences with LZ compressibility ρ for which the bound $kTfnV(\rho)$ is essentially attained. This is the case, for example, for most typical sequences of the memoryless source P^* that achieves $U(\rho)$. Tighter bounds can be obtained, of course, if more detailed information is given about the empirical statistics of x^n.

The important point about the function U (and, of course, V) is that, in the jargon of information theorists, it is a single–letter function, that is, its calculation requires merely optimization in the level of marginal distributions of a single symbol, and not distributions associated with ℓ–vectors. In Appendix B, we provide an explicit expression of $U(s)$.

Acknowledgement

This research was supported by the Israel Science Foundation (ISF), grant no. 412/12.

Appendix A

Proof of Eq. (25).

The proof is by induction: For $n = 1$, this is trivially true. Assume that it is true for a given n. Then, by the memorylessness of the channel Q, $Y^n \to X^n \to X_{n+1} \to Y_{n+1}$ is a Markov chain, and so, by the data processing theorem [4, Section 2.8]

$$\begin{aligned} H(X^n) + H(X_{n+1}) - H(X^{n+1}) &= I(X_{n+1}; X^n) \\ &\geq I(Y_{n+1}; Y^n) \\ &= H(Y^n) + H(Y_{n+1}) - H(Y^{n+1}), \end{aligned} \tag{44}$$

which is equivalent to

$$H(Y^{n+1}) - H(X^{n+1}) \geq [H(Y^n) - H(X^n)] + [H(Y_{n+1}) - H(X_{n+1})], \tag{45}$$

and so,

$$H(Y^n) - H(X^n) \geq \sum_{i=1}^{n} [H(Y_i) - H(X_i)] \tag{46}$$

implies

$$H(Y^{n+1}) - H(X^{n+1}) \geq \sum_{i=1}^{n+1} [H(Y_i) - H(X_i)], \tag{47}$$

completing the proof.

Appendix B

Deriving an Explicit Expression for $U(s)$.

For the case of the binary–input, binary–output channel at hand, let us denote $\epsilon_0 = Q_{0\to 0}$ and $\epsilon_1 = Q_{1\to 0}$, and assume that $\epsilon_0 \geq \epsilon_1$ (otherwise, switch the roles of the inputs). If the input assignment is $(p, 1-p)$, then the output entropy is clearly $h(p\epsilon_0 + \bar{p}\epsilon_1)$ ($\bar{p}$ being $1 - p$). The constraint $h(p) \geq s$ is equivalent to the constraint $h^{-1}(s) \leq p \leq 1 - h^{-1}(s)$, where $h^{-1}(s)$ is the smaller of the two solutions $\{u\}$ to the equation $h(u) = s$. Denoting

$$\alpha_s = \epsilon_0 h^{-1}(s) + \epsilon_1[1 - h^{-1}(s)] \tag{48}$$

$$\beta_s = \epsilon_0[1 - h^{-1}(s)] + \epsilon_1 h^{-1}(s) \tag{49}$$

then $\epsilon_0 \geq \epsilon_1$ implies $\beta_s \geq \alpha_s$, and then

$$\begin{aligned} U(s) &= \max\{h(q): \ \alpha_s \leq q \leq \beta_s\} \\ &= \begin{cases} h(\beta_s) & \beta_s \leq \frac{1}{2} \\ \ln 2 & \alpha_s \leq \frac{1}{2} \leq \beta_s \\ h(\alpha_s) & \alpha_s \geq \frac{1}{2} \end{cases} \end{aligned} \tag{50}$$

The condition $\beta_s \leq 1/2$ is satisfied always if $\epsilon_0 \leq 1/2$. For $\epsilon_0 > 1/2 \geq \epsilon_1$, this condition is equivalent to

$$s \geq h\left(\frac{\epsilon_0 - 1/2}{\epsilon_0 - \epsilon_1}\right) \stackrel{\triangle}{=} s^* \tag{51}$$

Similarly, the condition $\alpha_s \geq 1/2$ is satisfied always if $\epsilon_1 \geq 1/2$. For $\epsilon_1 < 1/2 \leq \epsilon_0$, this condition is equivalent to

$$s \geq h\left(\frac{1/2 - \epsilon_1}{\epsilon_0 - \epsilon_1}\right) = s^* \tag{52}$$

Thus, to summarize, $U(s)$ behaves as follows:

(i) For $\epsilon_1 \geq 1/2$, $U(s) = h(\alpha_s)$ for all $s \in [0,1]$.

(ii) For $\epsilon_0 \leq 1/2$, $U(s) = h(\beta_s)$ for all $s \in [0,1]$.

(iii) For $\epsilon_1 \leq 1/2 \leq \epsilon_0$ and $\epsilon_0 + \epsilon_1 > 1$

$$U(s) = \begin{cases} \ln 2 & 0 \leq s \leq s^* \\ h(\alpha_s) & s^* < s \leq 1 \end{cases}$$

(iv) For $\epsilon_1 \leq 1/2 \leq \epsilon_0$ and $\epsilon_0 + \epsilon_1 < 1$

$$U(s) = \begin{cases} \ln 2 & 0 \leq s \leq s^* \\ h(\beta_s) & s^* < s \leq 1 \end{cases}$$

Note that for the binary symmetric channel ($\epsilon_0 + \epsilon_1 = 1$), trivially $U(s) \equiv \ln 2$ for all $s \in [0,1]$. Also, in all cases $U(1) = h[(\epsilon_0 + \epsilon_1)/2]$.

References

[1] A. C. Barato and U. Seifert, demon," *EPL – a Letters Journal Exploring the Frontiers of Physics*, 101, 60001, March 2013.

[2] A. C. Barato and U. Seifert, http://arxiv.org/pdf/1408.1224.pdf

[3] Y. Cao, Z. Gong, and H. T. Quan, http://arxiv.org/pdf/1503.015224.pdf

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second Edition, Wiley–InterScience, John Wiley & Sons, 2006.

[5] S. Deffner and C. Jarzynski, arXiv:1308.5001v1, 22 Aug 2013.

[6] M. Esposito and C. Van Den Broeck, *EPL*, 95, 40004, August 2011.

[7] M. Feder, N. Merhav, and M. Gutman, *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1258–1270, July 1992.

[8] A. Gagliardi and A. Di Carlo, http://arxiv.org/pdf/1305.2107v1.pdf

[9] J. Hoppenau and A. Angel, http://arxiv.org/pdf/1401.2270v1.pdf

[10] J. M. Horowitz and M. Esposito, http://arxiv.org/pdf/1402.3276v2.pdf

[11] J. M. Horowitz and H. Sandberg, http://arxiv.org/pdf/1409.5351.pdf

[12] R. Landauer, *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

[13] H. S. Leff and A. F. Rex, *Maxwell's demons 2*, (IOP, Bristol, 2003).

[14] D. Mandal and C. Jarzynski, *PNAS*, vol. 109, no. 29, pp. 11641–11645, July 17, 2012.

[15] D. Mandal, H. T. Quan, and C. Jarzynski, *Physical Review Letters*, 111, 030603, July 19, 2013.

[16] N. Merhav, *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1302–1310, March 2013.

[17] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, *Nature Physics*, vol. 11, pp. 131–139, February 2015.

[18] L. Szilard, Z. Phys., 53, 840 (1929).

[19] T. Sagawa and M. Ueda, http://arxiv.org/pdf/1206.2479.pdf

[20] T. Sagawa and M. Ueda, *Physical Review Letters*, 109, 180602, November 2, 2012.

[21] T. Sagawa and M. Ueda, http://arxiv.org/pdf/1307.6092.pdf

[22] J. Ziv and A. Lempel, *IEEE Trans. Inform. Theory*, vol. IT–24, no. 5, pp. 530–536, September 1978.