Relations Between Work and Entropy Production for General Information–Driven, Finite–State Engines

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering, Technion, Haifa 32000, Israel.

E–mail: merhav@ee.technion.ac.il

Abstract. We consider a system model of a general finite–state machine (ratchet) that simultaneously interacts with three kinds of reservoirs: a heat reservoir, a work reservoir, and an information reservoir, the latter being taken to be a running digital tape whose symbols interact sequentially with the machine. As has been shown in earlier work, this finite–state machine can act as a demon (with memory), which creates a net flow of energy from the heat reservoir into the work reservoir (thus extracting useful work) at the price of increasing the entropy of the information reservoir. Under very few assumptions, we propose a simple derivation of a family of inequalities that relate the work extraction with the entropy production. These inequalities can be seen as either upper bounds on the extractable work or as lower bounds on the entropy production, depending on the point of view. Many of these bounds are relatively easy to calculate and they are tight in the sense that equality can be approached arbitrarily closely. In their basic forms, these inequalities are applicable to any finite number of cycles (and not only asymptotically), and for a general input information sequence (possibly correlated), which is not necessarily assumed even stationary. Several known results are obtained as special cases.

Keywords: information exchange, second law, entropy production, Maxwell demon, work extraction, finite–state machine.

1. Introduction

The fact that information processing plays a very interesting role in thermodynamics, has already been recognized in the second half of the nineteenth century, namely, when Maxwell proposed his celebrated gedanken experiment, known as Maxwell's demon [20]. According to the Maxwell demon experiment, a demon with access to information on momenta and positions of particles in a gas, at every given time, is cable of separating between fast–moving particles and slower ones, thus forming a temperature difference without supplying external energy, which sounds in contradiction to the second law of thermodynamics. A few decades later, Szilard [27] pointed out that it is possible to convert heat into work, when considering a box with a single particle. In particular, using a certain protocol of measurement and control, one may be able to produce work in each cycle of the system, which is again, in apparent contradiction with to the second law, since no external energy is injected.

These intriguing observations have created a considerable dispute and controversy in the scientific community. Several additional thought–provoking gedanken experiments have ultimately formed the basis for a vast amount of theoretical work associated with the role of informational ingredients in thermodynamics. An incomplete list of modern articles along these lines, include [1], [2], [3], [4], [5], [6], [7], [8], [9], [11], [12], [13], [14], [15], [16], [17], [18], [21], [22], [23], [24], [25], and [26]. These articles can be basically divided into two main categories. In the first category, the informational ingredient is in the form of measurement and feedback control (just like in the Maxwell's demon and Szilard's engine) and the second category is about physical systems that include, beyond the traditional heat reservoir (heat bath), also a work reservoir and an *information reservoir*, which interacts with the system entropically, but with no energy exchange. The information reservoir, which is a relatively new concept in physics [3], [4], [12], may be, for instance, a large memory register or a digital tape carrying a long sequence of bits, which interact sequentially with the system and may change during this interaction. Basically, the main results, in all these articles, are generalized forms of the second law of thermodynamics, where the entropy increase consists of an extra term that is concerned with information exchange, such as mutual information (for systems with measurement and feedback control) or Shannon entropy increase (for systems with a information reservoir).

In contrast to the early proposed thought experiments, that were typically described in general terms of an "intelligent agent" and were not quite described in full detail, Mandal and Jarzynski [21] were the first to devise a concrete model of a system that behaves basically like a demon. Specifically, they described and analyzed a simple autonomous system, based on a finite–state Markov process, that when operates as an engine, it converts heat into mechanical work, and, at the same time, it writes bits serially on a tape, which plays the role of an information reservoir. Here, the word "writes" refers to a situation where the entropy of the output bits recorded on the tape (after the interaction), is larger than the entropy of the input bits (before the interaction). It can also act as an eraser, which performs the reversed process of losing energy while "deleting" information, that is, decreasing the entropy. Several variants on this physical model, which are based on quite similar ideas, were offered in some later articles. These include: [2] – where the running tape can move both back and forth, [3] – where the interaction time with each bit is a random variable rather than fixed parameter, [4] – with three different points of view on information–driven systems, [5] – with the upper energy level being time–varying, [9] – with a model based on enzyme kinetics, [11] – with a quantum model, [16] – with a thermal tape, and [22], which concerns an information–driven refrigerator, where instead of the work, heat is transferred from a cold reservoir into a hotter one.

In a recent series of interesting papers, [6], [7], [8], Boyd, Mandal and Crutchfield considered a system model of a demon (ratchet) that is implemented by a general finite–state machine (FSM) that simultaneously interacts with a heat reservoir (heat bath at fixed temperature), a work reservoir (i.e., a given mass that may be lifted by the machine), and an information reservoir (a digital tape, as described above). The state variable of the FSM, which manifests the memory of the ratchet to past input and output information, interacts with the current bit of the information reservoir during one unit of time, a.k.a. the interaction interval (or cycle), and then the machine produces the next state and the output bit, before it turns to process the next input bit, etc. The operation of the ratchet during one cycle is then characterized by the joint probability distribution of the next state and the output bit given the current state and the input bit. Perhaps the most important result in [6], [7] and [8], is that for a stationary input process (i.e., the incoming sequence of tape bits), the work extraction per cycle is asymptotically upper bounded by kT times the difference between the Shannon entropy rate of the tape output process and that of the input process (both in units of nats‡ per cycle), i.e., eq. (5) of [6] (here k is the Boltzmann constant and T is the temperature). In addition to this general result, various conclusions are drawn in those papers. For example, the uselessness of ratchet memory when the input process is memoryless (i.i.d.), as well as its usefulness (for maximizing work extraction) when the input process is correlated, are both discussed in depth, and several interesting examples are demonstrated. While the above mentioned upper bound on the work extraction, [6, eq. (5)], seems reasonable and interesting, some concerns arise upon reading its derivation in [6, Appendix A], and these concerns are discussed in some detail in the Appendix.

In this paper, we consider a similar setup, but we focus is on the derivation of a family of alternative inequalities that relate work extraction to entropy production. The new proposed inequalities have the following advantages.

(i) The approach taken and the derivation are very simple.

(ii) The underlying assumptions about the input process, the ratchet, and the other parts of the system, are rather mild.

(iii) The inequalities apply to any finite number of cycles.

‡ 1 nat = $\log_2 e$ bits. Entropy defined using the natural base logarithm has units of nats.

(iv) For a stationary input process, the inequalities are simple and the resulting bounds are relatively easy to calculate.

(v) The inequalities are tight in the sense that equality can be approached arbitrarily closely.

(vi) Some known results are obtained as special cases.

The remaining part of the paper is organized as follows. In Section 2, we establish some notation conventions. In Section 3, we describe the physical system model. In Section 4, we derive our basic work/entropy–production inequality. In Section 5, we discuss this inequality and explore it from various points of view. Finally, in Section 6, we derive a more general family of inequalities, which have the flavor of fluctuation theorems.

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

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

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

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

§ 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.

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. Similarly, the conditional mutual information between U and V given W is

$$\begin{aligned} I(U;V|W) &= H(U|W) - H(U|V,W) \\ &= H(V|W) - H(V|U,W) \\ &= H(U|W) + H(V|W) - H(U,V|W). \end{aligned} \tag{7}$$

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

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\ln\frac{P(x)}{Q(x)}. \tag{8}$$

3. System Model Description

As in the previous articles on models of physical systems with an information reservoir, our system consists of the following ingredients: a heat bath at temperature T, a work reservoir, here designated by a wheel loaded by a mass m, an information reservoir in the form of a digital input tape, a corresponding output tape, and a certain device, which is the demon, or ratchet, in the terminology of [6], [7], [8]. The ratchet interacts (separately) with each one of the other parts of the system (see Fig. 1).

The input tape consists of a sequence of symbols, $x_1, x_2, \ldots$, from a finite alphabet $\mathcal{X}$ (say, binary symbols where $\mathcal{X} = \{0,1\}$), that are serially fed into the ratchet, which in turn processes these symbols sequentially, while going through a sequence of internal states, $s_1, s_2, \ldots$, taking values in a finite set $\mathcal{S}$. The ratchet outputs another sequence of symbols, $y_1, y_2, \ldots$, which are elements of the same alphabet, $\mathcal{X}$, as the input symbols. The state of the ratchet is an internal variable that encodes the memory that the ratchet has with regard to its history. In the n–th cycle of the process ($n = 1, 2, \ldots$), while the ratchet is at state s_n, it is fed by the input symbol x_n and it produces the pair (y_n, s_{n+1}) in stochastic manner, according to a given conditional distribution, $P(y_n, s_{n+1}|x_n, s_n)$, where y_n is the output symbol at the n–th cycle and s_{n+1} is the next state.

We now describe the mechanism that dictates this conditional distribution, along with the concurrent interactions among the ratchet, the heat bath and the work reservoir. The n–th cycle of the process occurs during the time interval, $(n-1)\tau \leq$

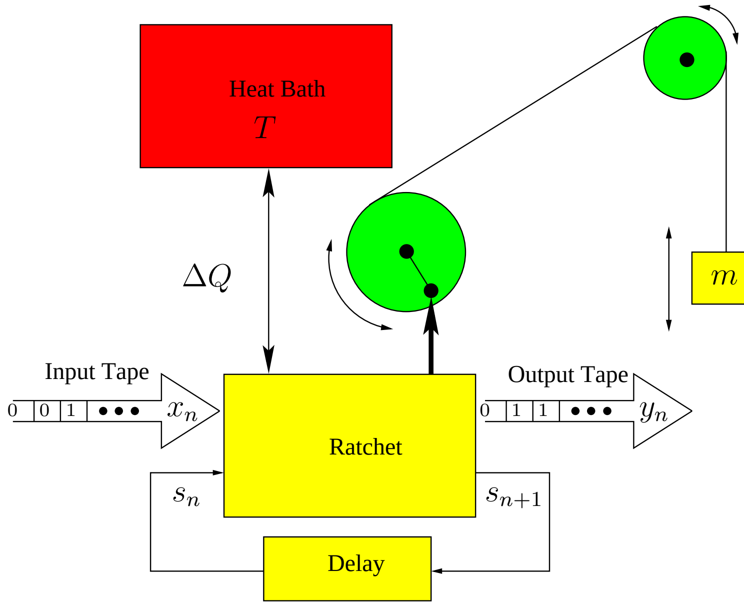

Figure 1. The physical system model.

$t < n\tau$, in other words, the duration of each cycle is τ seconds, where $\tau > 0$ is a given parameter. During each such interval, the symbol and the state form together a Markov jump process, (ξ_t, σ_t), whose state‖ set is the product set $\mathcal{X} \times \mathcal{S}$ and whose matrix of Markov–state transition rates is $M[(\xi, \sigma) \to (\xi', \sigma')]$, $\xi, \xi' \in \mathcal{X}$, $\sigma, \sigma' \in \mathcal{S}$. The random Markov–state transitions of this process are caused by spontaneous thermal fluctuations that result from the interaction with the heat bath. The Markov process is initialized at time $t = (n-1)\tau$ according to $(\xi_{(n-1)\tau}, \sigma_{(n-1)\tau}) = (x_n, s_n)$. At the end of this interaction interval, i.e., at time $t = n\tau - 0$, when the process is its final state $(\xi_{n\tau-0}, \sigma_{n\tau-0})$, the ratchet records the output symbol as $y_n = \xi_{n\tau-0}$ and the next ratchet state becomes $s_{n+1} = \sigma_{n\tau-0}$, and then the $(n+1)$–st cycle begins in the same manner, etc.

Denoting by $\Pi_t(\xi, \sigma)$ the probability of finding the Markov process in state (ξ, σ) at time t, it is clear from the above description, that the conditional distribution $P(y_n, s_{n+1}|x_n, s_n)$, that was mentioned before, is the solution $\{\Pi_{n\tau-0}(y, s)\}$ to the master equations (see, e.g., [28, Chap. 5]),

$$\frac{\mathrm{d}\Pi_t(\xi, \sigma)}{\mathrm{d}t} = \sum_{\xi', \sigma'} \{\Pi_t(\xi', \sigma') M[(\xi', \sigma') \to (\xi, \sigma)] - \Pi_t(\xi, \sigma) M[(\xi, \sigma) \to (\xi', \sigma')]\},$$

when the initial condition is $\Pi_{(n-1)\tau}(\xi, \sigma) = \mathcal{I}\{(\xi, \sigma) = (x_n, s_n)\}$.

Associated with each state, (ξ, σ), of the Markov process, there is a given energy $E(\xi, \sigma) = mg \cdot \Delta(\xi, \sigma)$, $\Delta(\xi, \sigma)$ being the height level of the mass m (relative to some reference height associated with an arbitrary Markov state). As the Markov process

‖ Note that from this point and onward, there are two different notions of "state", one of which is the state of ratchet, which is just s_n (or σ_t), and the other one is the state of the Markov process, which is the pair (x_n, s_n) (or (ξ_t, σ_t)). To avoid confusion, we will use the terms "ratchet state" and "Markov state" correspondingly, whenever there is room for ambiguity.

jumps from (ξ, σ) to (ξ', σ'), the ratchet lifts the mass by $\Delta(\xi', \sigma') - \Delta(\xi, \sigma)$, thus performing an amount of work given by $E(\xi', \sigma') - E(\xi, \sigma)$, whose origin is heat extracted from the heat bath (of course, the direction of the flow of energy between the heat bath and the work reservoir is reversed when these energy differences change their sign). It should be pointed out that the input tape does not supply energy to the ratchet, in other words, at the switching times, $t = n\tau$, although the state of the Markov process changes from $(\xi_{n\tau-0}, \sigma_{n\tau-0}) = (y_n, s_{n+1})$ to $(\xi_{n\tau}, \sigma_{n\tau}) = (x_{n+1}, s_{n+1})$, this switching is not assumed to be accompanied by a change in energy (the mass is neither raised nor lowered). In other words, the various energy levels, $E($
$xi, \sigma)$, have only a relative meaning, and so, after N cycles, the total amount of work carried out by the ratchet is given by

$$W_N = \sum_{n=1}^{N} [E(y_n, s_{n+1}) - E(x_n, s_n)]. \tag{9}$$

It will be assumed that the sequence of input symbols is governed by a stochastic process, which is designated by $X_1, X_2, \ldots$, and which obeys a given probability law P, that is,

$$\Pr\{X_1 = x_1,\ X_2 = x_2, \ldots, X_n = x_n\} = P(x_1, x_2, \ldots, x_n), \tag{10}$$

for every positive integer n and every $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, where $P(x_1, x_2, \ldots, x_n)$ is the probability distribution function. No special assumptions will be made concerning the process (not even stationarity) unless this will be specified explicitly. Following the notation conventions described in Section 2, the notation of the input sequence using capital X emphasizes that this is a random process. By the same token, when we wish to emphasize the induced randomness of the ratchet state sequence and the output sequence, we denote them by $\{S_n\}$ and $\{Y_n\}$, respectively.

To summarize, our model consists of two sets of stochastic processes in two different levels: one level lies in the larger time scale which is discrete (indexed by the integer n), and this is where the processes $\{X_n\}$, $\{Y_n\}$ and $\{S_n\}$ take place. The probability distributions of these processes are denoted by the letter P. The other level is in the smaller time scale, which is continuous, and this is where the Markov–jump pair process $\{(\xi_t, \sigma_t)\}$ takes place during each interaction interval of length τ. The joint probability distribution of (ξ_t, σ_t) is denoted by Π_t. The connection between the two kinds of processes is that at times $t = (n-1)\tau$, $n = 1, 2, \ldots$, (ξ_t, σ_t) is set to (X_n, S_n), and at times $t = n\tau - 0$, (Y_n, S_{n+1}) is set to (ξ_t, σ_t).

4. The Basic Work/Entropy–Production Inequality

As said, we are assuming that within each interaction interval, $(n-1)\tau \le t < n\tau$, the pair (ξ_t, σ_t) is a Markov jump process. For convenience of the exposition, let us temporarily shift the origin and redefine this time interval to be $0 \le t < \tau$. Since each Markov state (ξ, σ), is associated with energy level $E(\xi, \sigma)$, the equilibrium distribution is the canonical distribution,

$$\Pi_{\text{eq}}(\xi,\sigma) = \frac{e^{-\beta E(\xi,\sigma)}}{Z(\beta)}, \tag{11}$$

where $\beta = \frac{1}{kT}$ is the inverse temperature and

$$Z(\beta) = \sum_{(\xi,\sigma)\in\mathcal{X}\times\mathcal{S}} e^{-\beta E(\xi,\sigma)}. \tag{12}$$

The Markovity of the process implies that $D(\Pi_t\|\Pi_{\text{eq}})$ is monotonically non–increasing in t (see, e.g., [28, Chap. V.5], [19, Theorem 1.6], [10, Section 4.4]), and so,

$$D(\Pi_\tau\|\Pi_{\text{eq}}) \le D(\Pi_0\|\Pi_{\text{eq}}), \tag{13}$$

which is clearly equivalent to

$$\sum_{(\xi,\sigma)\in\mathcal{X}\times\mathcal{S}} [\Pi_\tau(\xi,\sigma) - \Pi_0(\xi,\sigma)] \cdot \ln\frac{1}{\Pi_{\text{eq}}(\xi,\sigma)} \le \mathcal{H}(\Pi_\tau) - \mathcal{H}(\Pi_0). \tag{14}$$

Since

$$\ln\frac{1}{\Pi_{\text{eq}}(\xi,\sigma)} = \ln Z(\beta) + \beta E(\xi,\sigma) \equiv \ln Z(\beta) + \beta mg\Delta(\xi,\sigma), \tag{15}$$

the left–hand side (l.h.s.) of (14) gives the average work per cycle (in units of kT), and the right–hand side (r.h.s.) is the difference between the entropy of the final Markov state within the cycle, (ξ_τ,σ_τ), and the entropy of the initial Markov state, (ξ_0,σ_0). Returning to the notation of the discrete time processes (indexed by n), we have then just shown that

$$\langle\Delta W_n\rangle \equiv \langle E(Y_n,S_{n+1})\rangle - \langle E(X_n,S_n)\rangle \le kT\cdot[H(Y_n,S_{n+1}) - H(X_n,S_n)],$$

and so, the total average work after N cycles is upper bounded by

$$\langle W_N\rangle \equiv \sum_{n=1}^N \langle\Delta W_n\rangle \le kT\cdot\sum_{n=1}^N [H(Y_n,S_{n+1}) - H(X_n,S_n)]. \tag{16}$$

Eq. (16) serves as our basic work/entropy–production inequality.

A slightly different form is the following:

$$\begin{aligned}\frac{\langle W_N\rangle}{kT} &\le \sum_{n=1}^N [H(Y_n|S_{n+1}) - H(X_n|S_n)] + \sum_{n=1}^N [H(S_{n+1}) - H(S_n)] \\ &= \sum_{n=1}^N [H(Y_n|S_{n+1}) - H(X_n|S_n)] + H(S_{N+1}) - H(S_1).\end{aligned} \tag{17}$$

The first sum in the last expression is the (conditional) entropy production associated with the input–output relation of the system, whereas the term $H(S_{N+1}) - H(S_1)$ can be understood as the contribution of the ratchet state to the net entropy production throughout the entire process of N cycles. If the ratchet has many states and N is not too large, the latter contribution might be significant, but if the number of ratchet states, $|\mathcal{S}|$, is fixed, then the relative contribution of ratchet–state entropy production term, which cannot exceed $\ln|\mathcal{S}|$, becomes negligible compared to the input–output entropy production term for large N. In particular, if we divide both sides of the inequality by N, then as $N \to \infty$, the term $\frac{\ln|\mathcal{S}|}{N}$ tends to zero, and so, the the average work per cycle is asymptotically upper bounded by $\frac{kT}{N}\sum_{n=1}^{N}[H(Y_n|S_{n+1}) - H(X_n|S_n)]$. This expression is different from the general upper bound of [6], [7], [8], where it was argued that $\langle W_N \rangle / NkT$ is asymptotically upper bounded by

$$\frac{1}{N}[H(Y^N) - H(X^N)] = \frac{1}{N}\sum_{n=1}^{N}[H(Y_n|Y^{n-1}) - H(X_n|X^{n-1})]. \tag{18}$$

While both the first term in (17) and (18) involve sums of differences between conditional output and input entropies, the conditionings being used in the two bounds are substantially different. Our bound suggests that the relevant information "memorized" by both the input process and the output process, is simply the ratchet state that is coupled to it, rather than its own past, as in (18). These conditionings on the states can be understood to be the residual input–output entropy production that is *not* part of the entropy production of the ratchet state (which is in general, correlated to the input and output). Moreover, the last line of (17) is typically easier to calculate than (18), as will be discussed and demonstrated in the sequel.

Yet another variant of (16) is obtained when the chain rule of the entropy is applied in the opposite manner, i.e.,

$$\frac{\langle W_N \rangle}{kT} \le \sum_{n=1}^{N}[H(Y_n) - H(X_n)] + \sum_{n=1}^{N}[H(S_{n+1}|Y_n) - H(S_n|X_n)]. \tag{19}$$

Here the first term is the input–output entropy production and the second term is the conditional entropy production of the ratchet state. However, this form is less useful than (17).

5. Discussion on the Bounds and Their Variants

In this section, we discuss eqs. (16) and (17) as well as several additional variants of these inequalities.

5.1. Tightness and Achievability

The first important point concerning inequality (16) is that it is potentially tight in the sense that the ratio between the two sides of eq. (16) may approach unity arbitrarily closely. To see this, consider first the case where $\Pi_0(\xi,\sigma)$ is close to $\Pi_{\rm eq}(\xi,\sigma)$ in the sense that

$$\Pi_0(\xi,\sigma) = \Pi_{\rm eq}(\xi,\sigma)\cdot[1+\epsilon(\xi,\sigma)], \quad (\xi,\sigma)\in\mathcal{X}\times\mathcal{S} \tag{20}$$

where $\epsilon \equiv \max_{\xi,\sigma}|\epsilon(\xi,\sigma)| \ll 1$ and obviously,

$$\sum_{\xi,\sigma}\Pi_{\rm eq}(\xi,\sigma)\epsilon(\xi,\sigma) = 0 \tag{21}$$

since $\{\Pi_0(\xi,\sigma)\}$ must sum up to unity. Assume also that $\Pi_\tau(\xi,\sigma)$ is even much closer to $\Pi_{\mathrm{eq}}(\xi,\sigma)$ in the sense that the ratio $\Pi_\tau(\xi,\sigma)/\Pi_{\mathrm{eq}}(\xi,\sigma)$ is between $1-\epsilon^2$ and $1+\epsilon^2$. Now, the work per cycle is given by

$$\begin{aligned}\langle\Delta W\rangle &= \sum_{\xi,\sigma}\Pi_\tau(\xi,\sigma)E(\xi,\sigma)-\sum_{\xi,\sigma}\Pi_0(\xi,\sigma)E(\xi,\sigma)\\ &= \sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)E(\xi,\sigma)+O(\epsilon^2)-\sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)[1+\epsilon(\xi,\sigma)]E(\xi,\sigma)\\ &= -\sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)\epsilon(\xi,\sigma)E(\xi,\sigma)+O(\epsilon^2). \end{aligned}\tag{22}$$

On the other hand, the entropy production per cycle is given by

$$\Delta\mathcal{H}\equiv\mathcal{H}(\Pi_\tau)-\mathcal{H}(\Pi_0)\tag{23}$$

$$= \sum_{\xi,\sigma}\Pi_0(\xi,\sigma)\ln\Pi_0(\xi,\sigma)-\sum_{\xi,\sigma}\Pi_\tau(\xi,\sigma)\ln\Pi_\tau(\xi,\sigma)\tag{24}$$

$$\begin{aligned}&= \sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)[1+\epsilon(\xi,\sigma)]\ln\{\Pi_{\mathrm{eq}}(\xi,\sigma)[1+\epsilon(\xi,\sigma)]\}-\\ &\quad\sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)\ln\Pi_{\mathrm{eq}}(\xi,\sigma)+O(\epsilon^2)\end{aligned}\tag{25}$$

$$= \sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)\epsilon(\xi,\sigma)\ln\Pi_{\mathrm{eq}}(\xi,\sigma)+O(\epsilon^2),\tag{26}$$

where the last line is obtained using (21). Now, the difference $kT\Delta\mathcal{H}-\langle\Delta W\rangle$ is given by $kT\cdot[D(\Pi_0\|\Pi_{\mathrm{eq}})-D(\Pi_\tau\|\Pi_{\mathrm{eq}})]$. But,

$$D(\Pi_0\|\Pi_{\mathrm{eq}}) = \sum_{\xi,\sigma}\Pi_0(\xi,\sigma)\ln[1+\epsilon(\xi,\sigma)]\tag{27}$$

$$= \sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)[1+\epsilon(\xi,\sigma)]\ln[1+\epsilon(\xi,\sigma)]\tag{28}$$

$$= \frac{1}{2}\sum_{\xi,\sigma}\Pi_{\mathrm{eq}}(\xi,\sigma)\epsilon^2(\xi,\sigma)+o(\epsilon^2)\tag{29}$$

$$= O(\epsilon^2)\tag{30}$$

and similarly, $D(\Pi_\tau\|\Pi_{\mathrm{eq}})=O(\epsilon^4)$. We have seen then that while both $kT\Delta\mathcal{H}$ and $\langle\Delta W\rangle$ scale linearly with $\{\epsilon(\xi,\sigma)\}$ (for small $\epsilon(\xi,\sigma)$), the difference between them scales with $\{\epsilon^2(\xi,\sigma)\}$. Thus, if both $\langle\Delta W\rangle$ and $kT\Delta\mathcal{H}$ are positive, the ratio between them may be arbitrarily close to unity, provided that $\{\epsilon(\xi,\sigma)\}$ are sufficiently small.

Even if Π_0 and Π_{eq} differ considerably, it is still possible to approach the entropy production bound, but this may require many small steps (in the spirit of quasi–static processes in classical thermodynamics), i.e., a chain of many systems of the type of Fig. 1, where the output bit–stream of each one of them serves as the input bit–stream to the next one. This approach was hinted already in [23] and later also in [8]. If we think of Π_0 as the canonical distribution with respect to some Hamiltonian $E_0(\xi,\sigma)$ (which is always possible, say, by defining $E_0(\xi,\sigma)=-kT\ln\Pi_0(\xi,\sigma)$), then we can design a long sequence of distributions, $\Pi^{(1)},\Pi^{(2)},\ldots,\Pi^{(L)}=\Pi_{\mathrm{eq}}$ (L – large positive integer), such that $\Pi^{(i)}$ has "Hamiltonian" $(1-i/L)E_0(\xi,\sigma)+(i/L)E(\xi,\sigma)$, $i=1,2,\ldots,L$, so that the distance between every two consecutive distributions (in the above sense) is of the order of $\epsilon = 1/L$ and hence the gap between the entropy production and the incremental work, pertaining to the passage from $\Pi^{(i)}$ to $\Pi^{(i+1)}$, is of the order of $\epsilon^2 = 1/L^2$, so that even if we sum up all these gaps, the total cumulative gap is of the order of L steps times $1/L^2$, which is $1/L$, and hence can still be made arbitrarily small by selecting L large enough.

5.2. Memoryless and Markov Input Processes

Most of the earlier works on systems with information reservoirs assumed that the input process $\{X_n\}$ is memoryless, i.e., that $P(x_1, \ldots, x_N)$ admits a product form for all N. In this case, S_n, which is generated by $X_1, \ldots, X_{n-1}$, must be statistically independent of X_n, and so, in eq. (17), $H(X_n|S_n) = H(X_n)$. We therefore obtain from (17), the following:

$$\frac{\langle W_N \rangle}{kT} \leq \sum_{n=1}^{N} [H(Y_n|S_{n+1}) - H(X_n)] + H(S_{N+1}) - H(S_1) \tag{31}$$

$$= \sum_{n=1}^{N} [H(Y_n) - H(X_n] - \sum_{n=1}^{N} I(S_{n+1}; Y_n) + H(S_{N+1}) - H(S_1). \tag{32}$$

As already mentioned in the context of (17), if we divide both sides by N and take the limit $N \to \infty$, the term $\frac{1}{N}[H(S_{N+1}) - H(S_1)] \leq \frac{1}{N} \ln |\mathcal{S}|$ vanishes as $N \to \infty$, and if we also drop the negative contribution of the mutual information terms, we further enlarge the expression to obtain the familiar bound that the asymptotic work per cycle cannot exceed the limit of $kT \cdot \frac{1}{N} \sum_{n=1}^{N} [H(Y_n) - H(X_n]$. As discussed also in [6], [7], [8], this bound is valid (and can be approached, following the discussion in the previous subsection) also by a memoryless ratchet, namely, a ratchet with one internal state only. Moreover, it is not only that there is nothing to lose from using a memoryless ratchet, but on the contrary – there is, in fact, a lot to lose if the ratchet uses memory in a non–trivial manner: this loss is expressed in the negative term $-\sum_{n=1}^{N} I(S_{n+1}; Y_n)$. The loss can, of course, be avoided if we make sure that at the end of each cycle, the two components of the Markov state, namely, S_{n+1} and Y_n, are statistically independent, and so, $I(S_{n+1}; Y_n) = 0$ for all n. If τ is large enough so that Π_{eq} is approached, and if $E(\xi, \sigma)$ is additive (namely, $E(\xi, \sigma) = E_1(\xi) + E_2(\sigma)$), then $\Pi_{\text{eq}}(\xi, \sigma) = \Pi_{\text{eq}}(\xi)\Pi_{\text{eq}}(\sigma)$, and this is the case. Indeed, in [21], for example, this is the case, as there are six Markov states ($|\mathcal{X}| = 2$ times $|\mathcal{S}| = 3$) and $\Pi_{\text{eq}}(\xi, \sigma) = e^{-\beta mgh\xi}/[3(1 + e^{-\beta mgh})]$, $\xi \in \{0, 1\}, \sigma \in \{A, B, C\}$.

Example. Consider a binary memoryless source with $\Pr\{X_n = 1\} = 1 - \Pr\{X_n = 0\} = p$, and a two–state ratchet, with a state set $\mathcal{S} = \{A, B\}$. The joint process $\{(X_n, S_n)\}$ (as well as $\{(\xi_t, \sigma_t)\}$ within each interaction interval) is therefore a four–state process with state set $\{A0, B0, A1, B1\}$. Let the energy levels be $E(A0) = 0$, $E(B0) = \epsilon$, $E(A1) = 2\epsilon$ and $E(B1) = 3\epsilon$, where $\epsilon > 0$ is a given energy quantum. The Markov jump process $\{(\xi_t, \sigma_t)\}$ has transition rates, $M[A0 \to B0] = M[B0 \to A1] = M[A1 \to B1] = e^{-\beta\epsilon}$, $M[B1 \to A1] = M[A1 \to B0] = M[B0 \to A0] = 1$ (in some units of frequency) and all other transition rates are zero (see Fig. 2). This process obeys detailed balance and its equilibrium distribution is given by $\Pi_{\text{eq}}[A0] = 1/Z$, $\Pi_{\text{eq}}[B0] = e^{-\beta\epsilon}/Z$, $\Pi_{\text{eq}}[A1] = e^{-2\beta\epsilon}/Z$, and $\Pi_{\text{eq}}[B1] = e^{-3\beta\epsilon}/Z$, where $Z = 1 + e^{-\beta\epsilon} + e^{-2\beta\epsilon} + e^{-3\beta\epsilon}$.

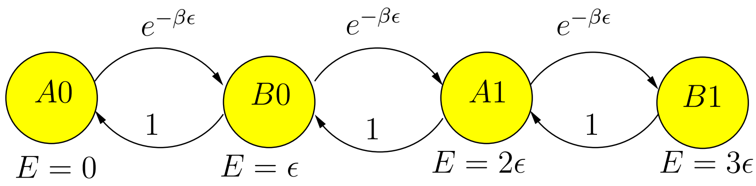

Figure 2. Example of the Markov jump process.

Suppose that τ is very large compared to the time constants of the process, so that $\Pi_\tau(\xi, \sigma)$ can be well approximated by the equilibrium distribution. Then, it is straightforward to see that

$$P(Y_n = 0|S_{n+1} = A) = \frac{\Pi_{\text{eq}}[A0]}{\Pi_{\text{eq}}[A0] + \Pi_{\text{eq}}[A1]} = \frac{1}{1 + e^{-2\beta\epsilon}} \tag{33}$$

and similarly for $P[Y_n = 0|S_{n+1} = B]$. Therefore,

$$H(Y_n|S_{n+1}) = h\left(\frac{1}{1 + e^{-2\beta\epsilon}}\right), \tag{34}$$

where $h(\cdot)$ is the binary entropy function, defined in Section 2. As for the input entropy, we have $H(X_n|S_n) = H(X_n) = h(p)$. Therefore, the upper bound on the work per cycle is

$$\langle \Delta W_n \rangle \le h\left(\frac{1}{1 + e^{-2\beta\epsilon}}\right) - h(p). \tag{35}$$

It follows that a necessary condition for the ratchet to operate as an engine (rather than as an eraser) is $p < 1/(1 + e^{2\beta\epsilon})$ or $p > 1/(1 + e^{-2\beta\epsilon})$. Using similar considerations, the exact work extraction is also easy to calculate in this example, but we will not delve into it any further. This concludes the example.

Consider next the case where the input process is a stationary first order Markov process, i.e.,

$$P(x^N) = P(x_1) \prod_{n=1}^{N-1} P(x_{n+1}|x_n). \tag{36}$$

As described above, in the discrete time scale, the ratchet is characterized by the input–output transition probability distribution $P(y, s'|x, s) = \Pr\{Y_n = y, S_{n+1} = s'|X_n = x, S_n = s\}$. Consider the corresponding marginal conditional distribution

$$P(s'|x, s) = \sum_{y \in \mathcal{X}} P(y, s'|x, s). \tag{37}$$

Then, assuming that the initial ratchet state, S_1, is independent of the initial input symbol, X_1, we have

$$P(x^N, s^N) = P(x_1)P(s_1) \prod_{n=1}^{N-1} [P(x_{n+1}|x_n)P(s_{n+1}|x_n, s_n)], \tag{38}$$

which means that the pair process $\{(X_n, S_n)\}$ is a first order Markov process as well. Let us assume that the transition matrix of this Markov pair process is such that there exists a unique stationary distribution $P(x, s) = \Pr\{X_n = x, S_n = s\}$. Once the stationary distribution $P(x, s)$ is found, the input–output–state joint distribution is dictated by the ratchet input–output transition probability distribution $\{P(y, s'|x, s)\}$, according to

$$P(x, s, y, s') = P(x, s)P(y, s'|x, s), \tag{39}$$

which is the joint distribution of the quadruple (X_n, S_n, Y_n, S_{n+1}) in the stationary regime. Once this joint distribution is found, one can (relatively) easily compute the stationary average work extraction per cycle, $\langle \Delta W_n \rangle = \langle E(Y_n, S_{n+1}) \rangle - \langle E(X_n, S_n) \rangle$, as well as the stationary joint entropies $H(X_n, S_n)$ and $H(Y_n, S_{n+1})$ (or $H(X_n|S_n)$ and $H(Y_n|S_{n+1})$) in order to calculate the entropy–production bound. This should be contrasted with the bound in [6] (see also [7], [8]), where, as mentioned earlier, $\langle W_N/NkT \rangle$ is asymptotically upper bounded by $\lim_{N\to\infty} \frac{1}{N}[H(Y^N) - H(X^N)]$, whose calculation is not trivial, as Y^N is a hidden Markov process, for which there is no closed–form expression for the entropy rate.

A good design of a ratchet would be in the quest of finding the transition distribution $\{P(y, s'|x, s)\}$ that maximizes the work extraction (or its entropy production bound) for the given Markov input process. This is an optimization problem with a finite (and fixed) number of parameters. If, in addition, one has the freedom to control the parameters of the Markov input process, say, by transducing a given source of randomness, e.g., a random bit–stream, then of course, the optimization will include also the induced joint distribution $\{P(x, s)\}$. If such a transducer is a one–to–one mapping, then its operation does not consume energy. For example, if the raw input stream is a sequence of independent fair coin tosses (i.e., a purely random bit–stream), this transducer can be chosen to be the decoder of an optimal lossless data compression scheme for the desired input process P.

5.3. Conditional Entropy Bounds

We now return to the case of a general input process. For a given $n = 1, 2, \ldots$, let us denote $u_n = (x^{n-1}, y^{n-1}, s^n)$, which is the full input–output–state history available at time n, and define $v_n = f_n(u_n)$, where f_n is an arbitrary function. If f_n is a many–to–one function, then v_n designates some partial history information, for example, $v_n = x^{n-1}$, or $v_n = y^{n-1}$. Once again, when we wish to emphasize the randomness of all these variables, we use capital letters: $U_n = (X^{n-1}, Y^{n-1}, S^n)$, $V_n = f_n(U_n)$, etc. Now consider the application of the H–theorem (eq. (13)) with $\Pi_0(\xi, \sigma) = P(X_n = \xi, S_n = \sigma|V_n = v_n)$, instead of the unconditional distribution as before. Then, using the Markovity of the dynamics within each interaction interval, the same derivation as in Section 3 would now yield

$$\begin{aligned}\langle \Delta W_n|V_n = v_n\rangle &\equiv \langle E(Y_n, S_{n+1})|V_n = v_n\rangle - \langle E(X_n, S_n)|V_n = v_n\rangle \\ &\le kT[H(Y_n, S_{n+1}|V_n = v_n) - H(X_n, S_n|V_n = v_n)],\end{aligned} \tag{40}$$

where the notation $\langle \cdot | V_n = v_n \rangle$ designates conditional expectation given $V_n = v_n$. Averaging both sides with respect to (w.r.t.) the randomness of V_n, we get

$$\begin{aligned}\langle \Delta W_n \rangle &\equiv \langle E(Y_n, S_{n+1}) \rangle - \langle E(X_n, S_n) \rangle \\ &\leq kT[H(Y_n, S_{n+1}|V_n) - H(X_n, S_n|V_n)],\end{aligned} \tag{41}$$

and summing all inequalities from $n = 1$ to $n = N$, we obtain the family of bounds,

$$\begin{aligned}\langle W_N \rangle &\equiv \sum_{n=1}^{N} [\langle E(Y_n, S_{n+1}) \rangle - \langle E(X_n, S_n) \rangle] \\ &\leq kT \sum_{n=1}^{N} [H(Y_n, S_{n+1}|V_n) - H(X_n, S_n|V_n)],\end{aligned} \tag{42}$$

with a freedom in the choice of V_n (or, equivalently, the choice of the function f_n). Now, one may wonder what is the best choice that would yield the tightest bound in this family. Conditioning reduces entropy, but it reduces both the entropy of (Y_n, S_{n+1}) and that of (X_n, S_n), so it may not be immediately clear what happens to the difference. A little thought, however, shows that the best choice of V_n is null, namely, the unconditional entropy bound of Section 3 is no worse than any bound of the form (42). To see why this is true, observe that

$$\begin{aligned}& H(Y_n, S_{n+1}|V_n) - H(X_n, S_n|V_n) \\ &= H(Y_n, S_{n+1}) - H(X_n, S_n) + I(V_n; X_n, S_n) - I(V_n; Y_n, S_{n+1}) \qquad (43) \\ &\geq H(Y_n, S_{n+1}) - H(X_n, S_n), \qquad (44)\end{aligned}$$

where the inequality follows from the data processing inequality [10, Sect. 2.8], as V_n and (Y_n, S_{n+1}) are statistically independent given (X_n, S_n), owing to the Markov property of the process $\{(\xi_t, \sigma_t)\}$. Consequently, $I(V_n; X_n, S_n) \geq I(V_n; Y_n, S_{n+1})$, and the inequality is achieved when V_n is degenerate. Thus, for the purpose of upper bounding the work, the conditioning on any partial history V_n turns out to be completely useless.

However, the family of inequalities (42) may be more interesting when we consider them as lower bounds on entropy production rather than upper bounds on extractable work. Specifically, consider the case $V_n = (X^{n-1}, Y^{n-1})$. Then, the work/entropy-production inequality reads

$$\begin{aligned}\frac{\langle W_N \rangle}{kT} &\leq \sum_{n=1}^{N} [H(Y_n, S_{n+1}|X^{n-1}, Y^{n-1}) - H(X_n, S_n|X^{n-1}, Y^{n-1})] \\ &\leq \sum_{n=1}^{N} [H(Y_n, S_{n+1}|Y^{n-1}) - H(X_n|X^{n-1}, Y^{n-1}) - \\ &\quad H(S_n|X^n, Y^{n-1})] \\ &= \sum_{n=1}^{N} [H(Y_n|Y^{n-1}) + H(S_{n+1}|Y^n) - \\ &\quad H(X_n|X^{n-1}) - H(S_n|X^n, Y^{n-1})] \\ &= H(Y^N) - H(X^N) + \sum_{n=1}^{N} [H(S_{n+1}|Y^n) - H(S_n|X^n, Y^{n-1})],\end{aligned} \tag{45}$$

where the first equality is since Y^{n-1} is independent of X_n given X^{n-1}. Now, the second term in the last line of eq. (45) is equivalent to

$$H(S_{N+1}|Y^N) - H(S_1|X_1) + \sum_{n=2}^{N}[H(S_n|Y^{n-1}) - H(S_n|X^n, Y^{n-1})]$$
$$= H(S_{N+1}|Y^N) - H(S_1|X_1) + \sum_{n=2}^{N} I(S_n; X^n|Y^{n-1}). \tag{46}$$

In general, this expression can always be upper bounded by $N \ln |\mathcal{S}|$, and so, we obtain the following lower bound on the output entropy

$$H(Y^N) \geq H(X^N) + \frac{\langle W_N \rangle}{kT} - N \ln |\mathcal{S}|. \tag{47}$$

Suppose now that $\{X_n\}$ is a memoryless process, or even a Markov process. Then, as mentioned earlier (see also [6, 7, 8]), $\{Y_n\}$ is a hidden Markov process, and as already explained before, the joint entropy of Y^N is difficult to compute and it does not have a simple closed–form expression. On the other hand, the above lower bound on $H(Y^N)$ is relatively easy to calculate, as $P(x^N)$ has a simple product form and $\langle W_N \rangle$ depends only on the marginals of (X_n, S_n) and (Y_n, S_{n+1}), which can be calculated recursively from the transition probabilities $\{P(y_n, s_{n+1}|x_n, s_n)\}$, for $n = 1, 2, \ldots, N$, and if in addition, $\{X_n\}$ is stationary, then (X_n, Y_n) and (Y_n, S_{n+1}) have stationary distributions too, as described before. While one may suspect that $N \ln |\mathcal{S}|$ might be a loose bound for the second term on the right–most side of (45), there are, nevertheless, situations where it is quite a reasonable bound, especially when $\ln |\mathcal{S}|$ is small (compared to $H(X^N)/N + \langle W_N \rangle /NkT$). Moreover, if the marginal entropy of S_n is known to be upper bounded by some constant $H_0 < \ln |\mathcal{S}|$, then $\ln |\mathcal{S}|$ can be replaced by H_0 in the above lower bound.

6. More General Inequalities

An equivalent form of the basic result of Section 3 is the following:

$$\mathcal{H}(\Pi_0) - \beta \langle E(\xi_0, \sigma_0) \rangle \leq \mathcal{H}(\Pi_\tau) - \beta \langle E(\xi_\tau, \sigma_\tau) \rangle, \tag{48}$$

The l.h.s. can be thought of as the negative free energy of the Markov state at time $t = 0$ (multiplied by a factor of β), and the r.h.s. is the same quantity at time $t = \tau$. In other words, if we define the random variable

$$\phi_t(\xi, \sigma) = -\ln \Pi_t(\xi, \sigma) - \beta E(\xi, \sigma), \tag{49}$$

then what we have seen in Section 3 is that

$$\langle \phi_0(\xi, \sigma) \rangle_0 \leq \langle \phi_t(\xi, \sigma) \rangle_t, \tag{50}$$

where $\langle \cdot \rangle_t$ denotes expectation w.r.t. Π_t. Equivalently, if we denote $\phi(X_n, S_n) = -\ln P(X_n, S_n) - \beta E(X_n, S_n)$, $\phi(Y_n, S_{n+1}) = -\ln P(Y_n, S_{n+1}) - \beta E(Y_n, S_{n+1})$, and we take $t = \tau$, this becomes

$$\langle \phi(X_n, S_n) \rangle \leq \langle \phi(Y_n, S_{n+1}) \rangle, \tag{51}$$

where the expectations at both sides are w.r.t. the randomness of the relevant random variables.

In this section, we show that this form of the inequality relation extends to more general moments of the random variables $\phi(X_n, S_n)$ and $\phi(Y_n, S_{n+1})$. As is well known, the H–theorem applies to generalized divergence functionals and not only to the Kullback–Leibler divergence $D(\Pi_t \| \Pi_{\text{eq}})$, see [19, Theorem 1.6], [28, Chap. V.5]. Let Q be any convex function and suppose that $\Pi_{\text{eq}}(x,s) > 0$ for every (x,s). Then according to the generalized H–theorem,

$$D_Q(\Pi_t \| \Pi_{\text{eq}}) = \sum_{x,s} \Pi_{\text{eq}}(x,s) Q\left(\frac{\Pi_t(x,s)}{\Pi_{\text{eq}}(x,s)}\right) \tag{52}$$

decreases monotonically as a function of t, and so,

$$D_Q(\Pi_\tau \| \Pi_{\text{eq}}) \le D_Q(\Pi_0 \| \Pi_{\text{eq}}). \tag{53}$$

Now,

$$\begin{aligned} D_Q(\Pi_t \| \Pi_{\text{eq}}) &= \left\langle \frac{\Pi_{\text{eq}}(\xi,\sigma)}{\Pi_t(\xi,\sigma)} \cdot Q\left(\frac{\Pi_\tau(\xi,\sigma)}{P_{\text{eq}}(\xi,\sigma)}\right)\right\rangle_t \\ &= \frac{1}{Z} \cdot \left\langle e^{\phi_t(\xi,\sigma)} \cdot Q\left(Z \cdot e^{-\phi_t(\xi,\sigma)}\right)\right\rangle_t \end{aligned} \tag{54}$$

In the corresponding inequality between $D_Q(\Pi_\tau \| \Pi_{\text{eq}})$ and $D_Q(\Pi_0 \| \Pi_{\text{eq}})$, the external factor of $1/Z$, obviously cancels out. Also, since $Q(u)$ is convex iff $Q(Z \cdot u)$ (Z – constant) is convex, we can re–define the latter as our convex function Q to begin with, and so, by the generalized H–theorem¶

$$\Lambda(t) \equiv \left\langle e^{\phi_t(\xi,\sigma)} \cdot Q\left(e^{-\phi_t(\xi,\sigma)}\right)\right\rangle_t \tag{55}$$

is monotonically decreasing for any convex function Q. It now follows that

$$\left\langle e^{\phi(X_n,S_n)} \cdot Q\left(e^{-\phi(X_n,S_n)}\right)\right\rangle \ge \left\langle e^{\phi(Y_n,S_{n+1})} \cdot Q\left(e^{-\phi(Y_n,S_{n+1})}\right)\right\rangle. \tag{56}$$

This class of inequalities has the flavor of fluctuation theorems concerning $\phi(X_n, S_n)$ and $\phi(Y_n, S_{n+1})$. We observe that unlike the classical H–theorem, which makes a claim only about the the first moments of $\phi(X_n, S_n)$ and $\phi(Y_n, S_{n+1})$, here we have a more general statement concerning the monotonicity of moments of a considerably wide family of functions of these random variables. For example, choosing $Q(u) = -\ln u$ gives

$$\left\langle \phi(X_n, S_n) e^{\phi(X_n,S_n)}\right\rangle \ge \left\langle \phi(Y_n, S_{n+1}) e^{\phi(Y_n,S_{n+1})}\right\rangle, \tag{57}$$

which is somewhat counter–intuitive, in view of (51), as the function $f(u) = ue^u$ is monotonically increasing.

An interesting family of functions $\{Q\}$ is the family of power functions, defined as $Q_z(u) = u^{1-z}$ for $z \le 0$ and $z \ge 1$ and $Q_z(u) = -u^{1-z}$ for $z \in [0,1]$. Here we obtain that

$$\langle \exp\{z\phi(X_n, S_n)\}\rangle \le \langle \exp\{z\phi(Y_n, S_{n+1})\}\rangle \qquad \text{for } \; z \in [0,1] \tag{58}$$

¶ Note that the classical H–theorem is obtained as a special case by the choice $Q(u) = u \ln u$.

and

$$\langle \exp\{z\phi(X_n, S_n)\}\rangle \geq \langle \exp\{z\phi(Y_n, S_{n+1})\}\rangle \qquad \text{for } z \notin [0,1] \tag{59}$$

Note that for $z > 1$, $P(X_n = x, S_n = s)$ must be strictly positive for all (x, s) with $E(x, s) < \infty$, for otherwise, there is a singularity. We have therefore obtained inequalities that involve the characteristic functions of $\phi(X_n, S_n)$ and $\phi(Y_n, S_{n+1})$. It is interesting to observe that the direction of the inequality is reversed when the parameter z crosses both the values $z = 0$ and $z = 1$.

Appendix A

Some Concerns About the Derivation of Eq. (5) of [6].

In Appendix A of [6], eq. (5) of that paper is derived, namely, the inequality that upper bounds the work extraction per cycle by kT times the difference between the Shannon entropy rate of the output process and that of the input process, as mentioned in the Introduction. The derivation in [6, Appendix A] begins from the second law of thermodynamics, and on the basis of the second law, it states that the joint *Shannon entropy* of the entire system, consisting of the ratchet state, the input tape, the output tape, and the heat bath, must not decrease with time (eq. (A2) in [6]).

The first concern is that while the second law is an assertion about the increase of the thermodynamic entropy (which is, strictly speaking, defined for equilibrium), some more care should be exercised when addressing the increase of the Shannon entropy. To be specific, we are familiar with two situations (in classical statistical physics) where the Shannon entropy is known to be non–decreasing. The first is associated with Hamiltonian dynamics, where the total Shannon entropy simply remains fixed, due to the Liouville theorem, as argued, for example, in [12, Section III], and indeed, ref. [12] is cited in [6] (in the context of eq. (2) therein), but there is no assumption in [6] about Hamiltonian dynamics, and it is not even clear that Hamiltonian dynamics can be assumed in this model setting, in the first place, due to the discrete nature of the input and output information streams, as well as the ratchet state. Two additional assumptions made in [12], but not in [6], are that the system is initially prepared in a product state (i.e., the states of the different parts of the system are statistically independent) [12, eq. (27)] and that the heat bath is initially in equilibrium [12, eq. (28)]. By contrast, the only assumption made in [6] is that the ratchet has a finite number of states (see first sentence in [6, Appendix A]).

The second situation where the Shannon entropy is known to be non–decreasing is when the state of the system is a Markov process, which has a uniform stationary state distribution, owing to the H–Theorem (see, for example, [28, Chap. V, Sect. 5]). However, it is not clear that the total system under discussion obeys Markov dynamics with a stationary distribution (let alone, the uniform distribution), because the tape moves in one direction only, so states accessible at a given time instant are no longer accessible at later times (after n cycles, the machine has converted n input bits to output bits, so the position of the tape relative to the ratchet, indexed by n, should be part of the Markovian state).

Another concern is that in Appendix A of [6], it is argued that the state of the heat bath is independent of the states of the ratchet and the tape at all times, with the somewhat vague explanation that "they have no memory of the environment" (see the text immediately after eq. (A4) of [6]). While this independence argument may make sense with regard to the initial preparation (at time $t = 0$) of the system (again, as assumed also in [12]), it is less clear why this remains true also at later times, after the systems have interacted for a while. Note that indeed, in [12], the various components of the system are not assumed independent at positive times.

To summarize, there seems to be some room for concern that more assumptions may be needed in [6] beyond the assumption on a finite number of ratchet states.

References

[1] A. Allahverdyan, D. Janzing, and G. Mahler, *J. Stat. Mech.*, P09011 2009.
[2] A. C. Barato and U. Seifert, *EPL*, **101** (2013) 60001.
[3] A. C. Barato and U. Seifert, *Phys. Rev. E* 90, 042150, October 31, 2014.
[4] A. C. Barato and U. Seifert, *Phys. Rev. Lett.*, 112 090601 2014.
[5] M. Bauer, A. C. Barato, and U. Seifert, *J. Stat. Mech.*, P09010 2014.
[6] A. B. Boyd, D. Mandal, and J. P. Crutchfield, *New J. Phys.* **18** 023049, 2016.
[7] A. B. Boyd, D. Mandal, and J. P. Crutchfield, https://arxiv.org/abs/1606.08506
[8] A. B. Boyd, D. Mandal, and J. P. Crutchfield, https://arxiv.org/pdf/1609.05353.pdf
[9] Y. Cao, Z. Gong, and H. T. Quan, http://arxiv.org/pdf/1503.01522.pdf
[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second Edition, Wiley–InterScience, John Wiley & Sons, 2006.
[11] S. Deffner, *Phys. Rev. E*, **88** 062128 (2013).
[12] S. Deffner and C. Jarzynski, *Phys. Rev. X* **3**, 041003, 2013.
[13] M. Esposito and C. Van Den Broeck, *EPL*, **95** (2011) 40004.
[14] A. Gagliardi and A. Di Carlo, http://arxiv.org/pdf/1305.2107v1.pdf
[15] D. Hartich, A. C. Barato, and U. Seifert, *J. Stat. Mech.*, P02016 2014.
[16] J. Hoppenau and A. Angel, *EPL*, **105** (2014) 50002.
[17] J. M. Horowitz and M. Esposito, *Phys. Rev. X*, **4** 031015, July 26, 2014.
[18] J. M. Horowitz and H. Sandberg, *New J. of Physics*, **16** (2014) 125007.
[19] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley & Sons, 1979.
[20] H. S. Leff and A. F. Rex, *Maxwell's Demon 2*, CRC Press, Taylor & Francis Group, 2002.
[21] D. Mandal and C. Jarzynski, *PNAS*, vol. 109, no. 29, pp. 11641–11645, July 17, 2012.
[22] D. Mandal, H. T. Quan, and C. Jarzynski, *Phys. Rev. Lett.*, 111, 030602, July 19, 2013.
[23] N. Merhav, J. Stat. Mech., P06037, 2015.
[24] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, *Nature Physics*, vol. 11, pp. 131–139, February 2015.
[25] T. Sagawa and M. Ueda, *Phys. Rev. Lett.*, **109**, 180602 (2012).
[26] T. Sagawa and M. Ueda, *New J. Phys.*, **15** 125012, December 2013.
[27] L. Szilard, Z. Phys., 53, 840 (1929).
[28] N. G. van Kampen, *Stochastic Processes in Physics and Chemsitry*, Elesevier Science, Amsterdam, The Nederlands, 2003.