Data Processing Theorems and the Second Law of Thermodynamics

Neri Merhav[*]

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

Abstract

We draw relationships between the generalized data processing theorems of Zakai and Ziv (1973 and 1975) and the dynamical version of the second law of thermodynamics, a.k.a. the Boltzmann H–Theorem, which asserts that the Shannon entropy, $H(X_t)$, pertaining to a finite–state Markov process $\{X_t\}$, is monotonically non–decreasing as a function of time t, provided that the steady–state distribution of this process is uniform across the state space (which is the case when the process designates an isolated system). It turns out that both the generalized data processing theorems and the Boltzmann H–Theorem can be viewed as special cases of a more general principle concerning the monotonicity (in time) of a certain generalized information measure applied to a Markov process. This gives rise to a new look at the generalized data processing theorem, which suggests to exploit certain degrees of freedom that may lead to better bounds, for a given choice of the convex function that defines the generalized mutual information. Indeed, we demonstrate an example of a certain setup of joint source–channel coding, where this idea yields an improved lower bound on the distortion, relative to both the 1973 Ziv–Zakai lower bound and the lower bound obtained from the ordinary data processing theorem.

Index Terms: Data processing inequality, convexity, perspective function, H–Theorem, thermodynamics, detailed balance.

---

[*] This work was supported by the Israel Science Foundation (ISF) grant no. 208/08.

1 Introduction

In [10] (see also [8]), Csiszár considered a generalized notion of the divergence between two probability distributions, a.k.a. the *f–divergence*, by replacing the negative logarithm function, of the classical divergence,

$$D(P_1\|P_2) = \int \mathrm{d}x \cdot P_1(x)\left[-\log\frac{P_2(x)}{P_1(x)}\right], \tag{1}$$

with a general convex function[1] Q, which satisfies certain regularity conditions, i.e.,[2]

$$D_Q(P_1\|P_2) = \int \mathrm{d}x \cdot P_1(x) \cdot Q\left(\frac{P_2(x)}{P_1(x)}\right). \tag{2}$$

The reader is referred also to [19], [1], and [9] for related work and independent derivations, as well as to [11, Section 4] for a tutorial on the subject.

When the f–divergence was applied to the joint distribution (in the role of P_1) and the product of marginals (in the role of P_2) of two random variables X and Y, it yielded a generalized notion of mutual information,

$$I_Q(X;Y) = \int \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y) \cdot Q\left(\frac{P_X(x)P_Y(y)}{P_{XY}(x,y)}\right) = \int \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y)\cdot Q\left(\frac{P_Y(y)}{P_{Y|X}(y|x)}\right), \tag{3}$$

which was shown in [10] to obey a data processing inequality, thus extending the well known data processing inequality of the ordinary mutual information (see, e.g., [7, Section 2.8]).

The same ideas were introduced independently by Ziv and Zakai [24], with the primary motivation of using it to obtain sharper distortion bounds for classes of simple codes for joint source–channel coding (e.g., of block length 1), as well as certain situations of signal detection and estimation (see also [2]). The idea was the following: given a communication system symbolized by a Markov chain $U \to X \to Y \to V$, where U designates the source, X represents the channel input, Y corresponds to the channel output and V stands for the decoder output, one uses the data processing inequality pertaining to I_Q:

$$I_Q(U;V) \le I_Q(X;Y), \tag{4}$$

---

[1] Originally, this function was denoted by f in [10], hence the name f–divergence. Here, we denote the convex function by Q, following the notation in [24] and [25], which are central references of this paper.

[2] In [10] and later references, the notation pertaining to the expression on the r.h.s. of eq. (2) is actually $D(P_2\|P_1)$ (or $D(P_2, P_1)$). Here, we prefer the notation $D(P_1\|P_2)$ to keep consistency with the notation of the ordinary divergence (1), which will also play a role throughout the sequel. Also, the somewhat more general definition in [10] allows integration w.r.t. an arbitrary measure $\lambda(\cdot)$, not necessarily the Lebesgue measure.

and then defines both a “rate–distortion function,” $R_Q(d) = \min\{I_Q(U;V):\ E\rho(U,V) \le d\}$ ($\rho(\cdot,\cdot)$ being a distortion measure) and a “channel capacity,” $C_Q = \max I_Q(X;Y)$, to derive a lower bound on the distortion d from the implied inequality

$$R_Q(d) \le C_Q. \tag{5}$$

In the sequel, this will be referred to as the 1973 version of the generalized data processing theorem.

In a somewhat less well known work [25], Zakai and Ziv have substantially further generalized their data processing theorems, so as to apply to even more general information measures, and this will be referred to as the 1975 version. This generalized information measure was in the form

$$\begin{aligned} I_Q(X;Y) &= \int \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y) \cdot Q\left(\frac{\mu^1_{XY}(x,y)}{P_{XY}(x,y)}, \ldots, \frac{\mu^k_{XY}(x,y)}{P_{XY}(x,y)}\right) \\ &= \int \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y) \cdot Q\left(\frac{\mu^1_{Y|X}(y|x)}{P_{Y|X}(y|x)}, \ldots, \frac{\mu^k_{Y|X}(y|x)}{P_{Y|X}(y|x)}\right), \end{aligned} \tag{6}$$

where Q is now an arbitrary convex function of k variables and $\{\mu^i_{XY}(x,y)\}$ are arbitrary positive measures (not necessarily probability measures) that are defined consistently with the Markov conditions and where $\mu^i_{Y|X}(y|x) = \mu^i_{XY}(x,y)/P_X(x)$. It was shown in [25, Theorem 7.1] that the distortion bounds obtained from (6) are tight in the sense that there always exist a convex function Q and measures $\{\mu^i_{XY}\}$ that would yield the exact distortion pertaining to the optimum communication system, and so, there is no room for improvement of this class of bounds.[3]

By setting $\mu^i_{Y|X}(y|x) = P_{Y|X}(y|x_i)$, $i = 1, 2, \ldots, k-1$, where $\{x_i\}$ are $k-1$ particular letters in the alphabet of X, and $\mu^k_{Y|X}(y|x) = P_Y(y)$, they defined yet another generalized information measure that satisfies the data processing theorem as

$$\boldsymbol{E}\left\{Q\left(\frac{P_{Y|X}(Y|X_1)}{P_{Y|X}(Y|X)}, \ldots, \frac{P_{Y|X}(Y|X_{k-1})}{P_{Y|X}(Y|X)}, \frac{P_Y(Y)}{P_{Y|X}(Y|X)}\right)\right\}, \tag{7}$$

where the expectation is taken w.r.t. the joint distribution

$$P(x_1, \ldots, x_{k-1}, x, y) = P_X(x)P_{Y|X}(y|x)P_X(x_1)P_X(x_2)\cdots P_X(x_{k-1}).$$

In both [24] and [25], there are many examples how these data processing inequalities can be used to improve on earlier distortion bounds.

---

[3] This result is non–constructive, however, in the sense that this choice of Q and $\{\mu^i_{XY}\}$ depends on the optimum encoder and decoder.

The data processing theorems of Csiszár and Zakai and Ziv form one aspect of this work. The other aspect, which may seem unrelated at first glance (but will nevertheless be shown here to be strongly related) is the second law of thermodynamics, or more precisely, *Boltzmann's H–theorem.* The second law of thermodynamics tells that in an isolated physical system (i.e., when no energy flows in or out), the entropy increases. Since one of the basic postulates of statistical physics tells that all states of the system, which have the same energy, also have the same probability in equilibrium, it follows that the stationary (equilibrium) distribution of these states must be uniform, because all accessible states must have the same energy when the system is isolated. Indeed, if the state of this system is designated by a Markov process, $\{X_t\}$ with a uniform stationary state distribution, the Boltzmann H–theorem tells that the Shannon entropy (or the Gibbs' entropy in the language of physicists) of X_t, $H(X_t)$, cannot decrease with t, which is has the spirit of the second law, but as will be explained later, it is not quite exactly equivalent to the second law.

We show, in this paper, that the generalized data processing theorems of [10], [24], and [25] on the one hand, and the Boltzmann H–theorem, on the other hand, are all special cases of a more general principle, which asserts that a certain generalized information measure, applied to the underlying Markov process must be a monotonic function of time. This unified framework provides a new perspective on the generalized data processing theorem. Beyond the fact that this new perspective may be interesting on its own right, it naturally suggests to exploit certain degrees of freedom of the Ziv–Zakai generalized mutual information that may lead to better bounds, for a given choice of the convex function that defines this generalized mutual information. These additional degrees of freedom may be important, because the variety of convex functions $\{Q\}$ which are convenient to work with, is rather limited. The fact that better bounds may indeed be obtained is demonstrated by an example.

Finally, two technical comments are in order: The first comment is that, for reasons of simplicity of the exposition, throughout the rest of the paper, all random variables will be assumed to take on values in finite alphabets. Extensions of data processing inequalities to countable and continuous alphabets are available and can be found, for example, in [17] and in [18], as well as in [24]. The second comment concerns the notation. As can be understood from the previous paragraphs, this paper is concerned with two different setups having two different corresponding sets of random variables. One is associated with various versions of the data processing theorem that applies to a set of random variables denoted by different letters, like U, V, X, and Y, and the other is concerned with a Markov process denoted $\{X_t\}$. In the former, to avoid confusion between the various probability distributions and measures, we will use the standard convention of substricting the probability distributions and measures by the names of the random variables (as was done above). In the latter, on the other hand, since there is no apparent room for ambiguity, no subscripts of this type will be used, but merely the subscript t that designates time.

The outline of the remaining part of this paper is as follows. In Section 2, we provide some background on Markov processes with a slight physical flavor, which will include the notion of detailed balance, global balance, as well as known results like the Boltzmann H–theorem, and its generalizations to information measures other than the entropy. In Section 3, we relate the generalized version of the Boltzmann H–theorem and the generalized data processing theorems and we formalize the unified framework that supports both. This is done, first for the 1973 version [24] of the Ziv–Zakai data processing theorem (along with an example), and then for the 1975 version by Zakai and Ziv [25]. Finally, in Section 4, we summarize and conclude.

2 Background

Many dynamical models of physical systems describe their microscopic states (or microstates, for short) as obeying certain differential equations (e.g., the Langevin equation) that gives rise to their statistical description as Markov processes, like a random walk, a Brownian motion, or any other diffusion process. This model of a Markov process may be defined either in discrete time or in continuous time.

In this section, we provide some background on Markov processes, first in the general level, with emphasis on concepts like detailed balance and time reversal symmetry, both for isolated systems and for non–isolated systems (Subsection 2.1), and then the focus will shift to the evolution of information measures (associated with Markov processes), like the entropy, the divergence, the mutual information and their generalized versions (Subsection 2.2).

2.1 Detailed Balance and Global Balance

We begin with an isolated system in continuous time, which is not necessarily assumed to have reached yet its stationary distribution pertaining to equilibrium. Let us suppose that the state X_t may take on values in a finite set $\mathcal{X}$. For $x, x' \in \mathcal{X}$, let us define the state transition rates

$$W_{xx'} = \lim_{\delta \to 0} \frac{\Pr\{X_{t+\delta} = x' | X_t = x\}}{\delta} \qquad x' \neq x \tag{8}$$

which means, in other words,

$$\Pr\{X_{t+\delta} = x' | X_t = x\} = W_{xx'} \cdot \delta + o(\delta). \tag{9}$$

Denoting

$$P_t(x) = \Pr\{X_t = x\}, \tag{10}$$

it is easy to see that for small δ,

$$P_{t+\delta}(x) = \sum_{x' \neq x} P_t(x') W_{x'x}\delta + P_t(x)\left(1 - \sum_{x' \neq x} W_{xx'}\delta\right) + o(\delta), \tag{11}$$

where the first sum describes the probabilities of all possible transitions from other states to state x and the second term describes the probability of not leaving state x. Subtracting $P_t(x)$ from both sides, dividing by δ, and taking the limit of $\delta \to 0$, we immediately obtain the following set of differential equations:

$$\frac{\mathrm{d}P_t(x)}{\mathrm{d}t} = \sum_{x'} [P_t(x') W_{x'x} - P_t(x) W_{xx'}], \quad x \in \mathcal{X}, \tag{12}$$

where W_{xx} is defined in an arbitrary manner, e.g., $W_{xx} \equiv 0$ for all $x \in \mathcal{X}$. In the physics terminology (see, e.g., [16],[22]), these equations are called the *master equations*.[4] When the process reaches stationarity, i.e., for all $x \in \mathcal{X}$, $P_t(x)$ converge to some $P(x)$ that is time–invariant, then

$$\sum_{x'} [P(x') W_{x'x} - P(x) W_{xx'}] = 0, \quad \forall\, x \in \mathcal{X}. \tag{13}$$

This situation is called *global balance* or *steady state*. When the physical system under discussion is isolated, namely, no energy flows into the system or out, the steady state distribution must be

---

[4] Note that the master equations apply in discrete time too, provided that the derivative on the l.h.s. is replaced by a simple difference, $P_{t+1}(x) - P_t(x)$, and $\{W_{xx'}\}$ are replaced by one–step state transition probabilities.

uniform across all states, because all accessible states must be of the same energy and the equilibrium probability of each state depends solely on its energy. Thus, in the case of an isolated system, $P(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$. From quantum mechanical considerations, as well as considerations pertaining to time reversibility in the microscopic level,[5] it is customary to assume $W_{xx'} = W_{x'x}$ for all pairs $\{x, x'\}$. We then observe that, not only do $\sum_{x'}[P(x')W_{x'x} - P(x)W_{xx'}]$ all vanish, but moreover, each individual term in this sum vanishes, as

$$P(x')W_{x'x} - P(x)W_{xx'} = 0, \qquad \forall\ x, x'. \tag{14}$$

This property, where $P(x')W_{x'x} - P(x)W_{xx'} = 0$ for all x and x', is called *detailed balance*, and as said, when $P(x)$ is constant (the uniform distribution across $\mathcal{X}$), it is equivalent to $W_{x'x} = W_{xx'}$ for all x and x'. The detailed balance property is stronger than global balance, and it means equilibrium, which is stronger than steady state. While both steady–state and equilibrium refer to situations of time–invariant state probabilities $\{P(x)\}$, a steady–state still allows cyclic "flows of probability." For example, a Markov process with cyclic deterministic transitions $1 \to 2 \to 3 \to 1 \to 2 \to 3 \to \cdots$ is in steady state, provided that the probability distribution of the initial state is uniform $(1/3, 1/3, 1/3)$, however, the cyclic flow among the states is in one direction. On the other hand, in detailed balance ($W_{xx'} = W_{x'x}$ for an isolated system), which is equilibrium, there is no net flow in any cycle of states. All the net cyclic probability fluxes vanish, and therefore, time reversal would not change the probability law, that is, $\{X_{-t}\}$ has the same probability law as $\{X_t\}$ (see [15, Sect. 1.2]). For example, if $\{Y_t\}$ is a binary symmetric source (i.e., each Y_t takes on values equiprobably and independently in $\{-1, +1\}$), then X_t, defined recursively by

$$X_{t+1} = (X_t + Y_t) \bmod K \tag{15}$$

(K being a positive integer), has a symmetric state–transition probability matrix W, a uniform stationary state distribution, and it satisfies detailed balance.

---

[5] Consider, for example, an isolated system of moving particles of mass m and position vectors $\{\boldsymbol{r}_i(t)\}$, obeying the differential equations $m\mathrm{d}^2\boldsymbol{r}_i(t)/\mathrm{d}t^2 = \sum_{j\neq i} F(\boldsymbol{r}_j(t) - \boldsymbol{r}_i(t))$, $i = 1, 2, \ldots, n$, ($F(\boldsymbol{r}_j(t) - \boldsymbol{r}_i(t))$ being mutual interaction forces), which remain valid if the time variable t is replaced by $-t$ since $\mathrm{d}^2\boldsymbol{r}_i(t)/\mathrm{d}t^2 = \mathrm{d}^2\boldsymbol{r}_i(-t)/\mathrm{d}(-t)^2$.

2.2 Monotonicity of Information Measures

Returning to the case where the process $\{X_t\}$ pertaining to our isolated system has not necessarily reached equilibrium, let us take a look at the entropy of the state

$$H(X_t) = -\sum_{x\in\mathcal{X}} P_t(x)\log P_t(x). \tag{16}$$

The Boltzmann H–theorem (see, e.g., [4, Chap. 7], [14, Sect. 3.5], [16, pp. 171–173] [22, pp. 624–626]) asserts that $H(X_t)$ is monotonically non–decreasing. At first glance, this result may seem as a restatement of the second law of thermodynamics, which tells that the entropy of an isolated system cannot decrease with time. It is important to stress, however, that it is not quite precise, because $H(X_t)$ is not really the physical entropy of the system outside the regime of equilibrium (see, e.g., [21]). The second law simply states that if a system is thermally isolated, then for any process that begins in equilibrium and ends in (possibly, another) equilibrium, the entropy of the *final* state is never smaller than the entropy of the *initial* state, but there is no statement concerning monotonic evolution of the entropy (whatever its definition may be) along the process itself, when the system is out of equilibrium. For more discussion about the differences between the H–theorem and the second law, the reader is referred to [6] and [13].

To see why $H(X_t)$ is non–decreasing, we next show that detailed balance implies

$$\frac{\mathrm{d}H(X_t)}{\mathrm{d}t} \geq 0, \tag{17}$$

where for convenience, we denote $\mathrm{d}P_t(x)/\mathrm{d}t$ by $\dot{P}_t(x)$. Now,

$$\begin{aligned}
\frac{\mathrm{d}H(X_t)}{\mathrm{d}t} &= -\sum_x [\dot{P}_t(x)\log P_t(x) + \dot{P}_t(x)] \\
&= -\sum_x \dot{P}_t(x)\log P_t(x) \\
&= -\sum_x\sum_{x'} W_{x'x}[P_t(x') - P_t(x)]\log P_t(x) \\
&= -\frac{1}{2}\sum_{x,x'} W_{x'x}[P_t(x') - P_t(x)]\log P_t(x) - \\
&\quad \frac{1}{2}\sum_{x,x'} W_{x'x}[P_t(x) - P_t(x')]\log P_t(x') \\
&= \frac{1}{2}\sum_{x,x'} W_{x'x}[P_t(x') - P_t(x)]\cdot[\log P_t(x') - \log P_t(x)]
\end{aligned}$$

$$\geq 0, \tag{18}$$

where in the second line we used the fact that $\sum_x \dot{P}_t(x) = 0$, in the third line we used detailed balance ($W_{xx'} = W_{x'x}$), and the last inequality is due to the increasing monotonicity of the logarithmic function: the product $[P_t(x') - P_t(x)] \cdot [\log P_t(x') - \log P_t(x)]$ cannot be negative for any pair (x, x'), as the two factors of this product are either both negative, both zero, or both positive. Thus, $H(X_t)$ cannot decrease with time.

The H–theorem has a discrete–time analogue: If a finite–state Markov process has a symmetric transition probability matrix (which is the discrete–time counterpart of the above detailed balance property), which means that the stationary state distribution is uniform, then $H(X_t)$ is a monotonically non–decreasing sequence.

A well–known paradox, in this context, is associated with the notion of the *arrow of time.* On the one hand, we are talking about time–reversible processes, obeying detailed balance, but on the other hand, the increase of entropy suggests that there is an asymmetry between the two possible directions at which the time axis can be swept, the forward direction and the backward direction. If we go back in time, the entropy would decrease. So is there an arrow of time? This paradox was resolved by Boltzmann himself, once he made the clear distinction between equilibrium and non–equilibrium situations: The notion of time reversibility is associated with equilibrium, where the process $\{X_t\}$ is stationary. On the other hand, the increase of entropy is a result that belongs to the non–stationary regime, where the process is on its way to stationarity and equilibrium. In the latter case, the system has been initially prepared in a non–equilibrium situation. Of course, when the process is stationary, $H(X_t)$ is fixed and there is no contradiction.

So far we discussed the property of detailed balance only for an isolated system, where the stationary state distribution is the uniform distribution. How is the property of detailed balance defined when the stationary distribution is non–uniform? For a general Markov process, whose steady state–distribution is not necessarily uniform, the condition of detailed balance, which means time–reversibility [15], reads

$$P(x)W_{xx'} = P(x')W_{x'x}, \tag{19}$$

in the continuous–time case. In the discrete–time case (where t takes on positive integer values only), it is defined by a similar equation, except that $W_{xx'}$ and $W_{x'x}$ are replaced by the corresponding one–step state transition probabilities, i.e.,

$$P(x)P(x'|x) = P(x')P(x|x'), \tag{20}$$

where

$$P(x'|x) \stackrel{\triangle}{=} \Pr\{X_{t+1} = x'|X_t = x\}. \tag{21}$$

The physical interpretation is that now our system is (a small) part of a much larger isolated system, where the large system obeys detailed balance w.r.t. the uniform equilibrium distribution, as before.

A well known example of a process that obeys detailed balance, in its more general form, is the M/M/1 queue with an arrival rate λ and service rate μ ($\lambda < \mu$). Here, since all states are arranged along a line, with bidirectional transitions between neighboring states only (see Fig. 1), there cannot be any cyclic probability flux. The steady–state distribution is well–known to be geometric

$$P(x) = \left(1 - \frac{\lambda}{\mu}\right) \cdot \left(\frac{\lambda}{\mu}\right)^x, \qquad x = 0, 1, 2, \ldots, \tag{22}$$

which indeed satisfies the detailed balance $\lambda P(x) = \mu P(x+1)$ for all x. Thus, the Markov process $\{X_t\}$, designating the number of customers in the queue at time t, is time–reversible.

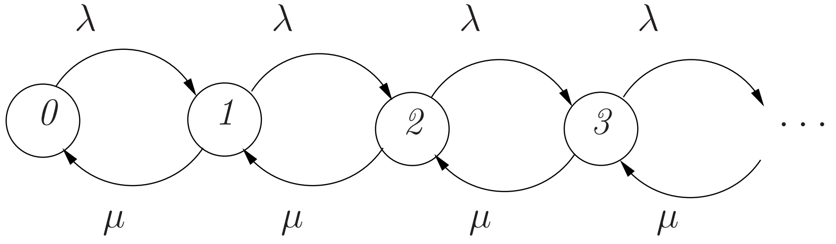

Figure 1: State transition diagram of an M/M/1 queue.

For the sake of simplicity, from this point onward, our discussion will focus almost exclusively on discrete–time Markov processes, but the results to be stated, will hold for continuous–time Markov processes as well. We will continue to denote by $P_t(x)$ the probability of $X_t = x$, except that now t will be limited to take on integer values only. The one–step state transition probabilities will be denoted by $\{P(x'|x)\}$, as mentioned earlier.

How does the H–theorem extend to situations where the stationary state distribution is not uniform? In [7, p. 82], it is shown (among other things) that the divergence,

$$D(P_t\|P) = \sum_{x\in\mathcal{X}} P_t(x)\log\frac{P_t(x)}{P(x)}, \tag{23}$$

where $P = \{P(x),\ x \in \mathcal{X}\}$ is a stationary state distribution, is a monotonically non–increasing function of t. Does this result have a physical interpretation, like the H–theorem and its connotation with the second law of thermodynamics? When it comes to non–isolated systems, where the steady state distribution is non–uniform, the extension of the second law of thermodynamics, replaces the principle of increase of entropy by the principle of decrease of free energy, or equivalently, the decrease of the difference between the free energy at time t and the free energy in equilibrium. The information–theoretic counterpart of this free energy difference is the divergence considered to be $D(P_t\|P)$ (see, e.g., [3]). Thus, the monotonic decrease of $D(P_t\|P)$ seems to have a simple physical interpretation in the spirit of of free energy decrease, which is the natural extension of the entropy increase.[6] Indeed, particularizing this to the case where P is the uniform distribution (as in an isolated system), then

$$D(P_t\|P) = \log|\mathcal{X}| - H(X_t), \tag{24}$$

which means that the decrease of the divergence is equivalent to the increase of entropy, as before. However, here the result is more general than the H–theorem from an additional aspect: It does not require detailed balance. It only requires the existence of a stationary state distribution. Note that even in the earlier case of an isolated system, detailed balance, which means symmetry of the state transition probability matrix ($P(x'|x) = P(x|x')$), is a stronger requirement than uniformity of the stationary state distribution, as the latter requires merely that the matrix $\{P(x'|x)\}$ would be doubly stochastic, i.e., $\sum_x P(x|x') = \sum_x P(x'|x) = 1$ for all $x' \in \mathcal{X}$, which is weaker than symmetry of the matrix itself. The results shown in [7] are, in fact, somewhat more general: Let $P_t = \{P_t(x)\}$ and $P_t' = \{P_t'(x)\}$ be two time–varying state–distributions pertaining to the same Markov chain, but induced by two different initial state distributions, $\{P_0(x)\}$ and $\{P_0'(x)\}$, respectively. Then $D(P_t\|P_t')$ is monotonically non–increasing. This is easily seen as follows:

$$D(P_t\|P_t') = \sum_x P_t(x)\log\frac{P_t(x)}{P_t'(x)}$$

---

[6] Once again, we reiterate that a similar digression as before applies here too: The free energy is defined only in equilibrium, thus it is not clear that $D(P_t\|P)$ is really the free energy out of equilibrium.

$$
\begin{aligned}
&= \sum_{x,x'} P_t(x)P(x'|x)\log\frac{P_t(x)P(x'|x)}{P_t'(x)P(x'|x)} \\
&= \sum_{x,x'} P(X_t=x,\ X_{t+1}=x')\log\frac{P(X_t=x,\ X_{t+1}=x')}{P'(X_t=x,\ X_{t+1}=x')} \\
&\ge D(P_{t+1}\|P'_{t+1}) \qquad (25)
\end{aligned}
$$

where the last inequality follows from the data processing theorem of the divergence: the divergence between two joint distributions of (X_t, X_{t+1}) is never smaller than the divergence between corresponding marginal distributions of X_{t+1}. Another interesting special case of this result is obtained if we now take the first argument of the divergence to be the a stationary state distribution: This will mean that $D(P\|P_t)$ is also monotonically non–increasing.

In [15, Theorem 1.6], there is a further extension of all the above monotonicity results, where the ordinary divergence is actually replaced by the f–divergence (though the term "f–divergence" is not mentioned in [15]): If $\{X_t\}$ is a Markov process with a given state transition probability matrix $\{P(x'|x)\}$, then the function

$$U(t) = D_Q(P\|P_t) = \sum_{x\in\mathcal{X}} P(x)\cdot Q\left(\frac{P_t(x)}{P(x)}\right) \qquad (26)$$

is monotonically non–increasing, provided that Q is a convex function. To see why this is true, define the backward transition probability matrix by

$$\tilde{P}(x|x') = \frac{P(x)P(x'|x)}{P(x')}, \qquad (27)$$

where again, $P(x)$ and $P(x')$ pertain to a stationary state distribution. Obviously,

$$\sum_x \tilde{P}(x|x') = 1 \qquad (28)$$

for all $x' \in \mathcal{X}$, and so,

$$\frac{P_{t+1}(x)}{P(x)} = \sum_{x'} \frac{P_t(x')P(x|x')}{P(x)} = \sum_{x'} \frac{\tilde{P}(x'|x)P_t(x')}{P(x')}. \qquad (29)$$

By the convexity of Q:

$$
\begin{aligned}
U(t+1) &= \sum_x P(x)\cdot Q\left(\frac{P_{t+1}(x)}{P(x)}\right) \\
&= \sum_x P(x)\cdot Q\left(\sum_{x'} \tilde{P}(x'|x)\frac{P_t(x')}{P(x')}\right)
\end{aligned}
$$

$$\begin{aligned}
&\le \sum_x \sum_{x'} P(x)\tilde{P}(x'|x) \cdot Q\left(\frac{P_t(x')}{P(x')}\right) \\
&= \sum_x \sum_{x'} P(x')P(x|x') \cdot Q\left(\frac{P_t(x')}{P(x')}\right) \\
&= \sum_{x'} P(x') \cdot Q\left(\frac{P_t(x')}{P(x')}\right) = U(t). \qquad (30)
\end{aligned}$$

Moreover, it is shown in [15, Theorem 1.6] that if Q is strictly convex and if P_0 is not the uniform distribution, then $U(t+1) < U(t)$ for all t. Now, a few interesting choices of the function Q may be considered: As proposed in [15, p. 19], for $Q(u) = u \ln u$, we have $U(t) = D(P_t \| P)$, and we are back to the aforementioned result in [7]. Another interesting choice is $Q(u) = -\ln u$, which gives $U(t) = D(P \| P_t)$. Thus, the monotonicity of $D(P \| P_t)$ is also obtained as a special case.[7] Yet another choice is $Q(u) = -u^s$, where $s \in [0,1]$ is a parameter. This would yield the increasing monotonicity of $\sum_x P^{1-s}(x)P_t^s(x)$, a divergence that plays a role in the theory of asymptotic exponents of error probabilities pertaining to the optimum likelihood ratio test between two probability distributions [23, Chapter 3]. In particular, the choice $s = 1/2$ yields balance between the Chernoff divergences of the two kinds of error and it is intimately related to the Bhattacharyya distance.

In the case of detailed balance, there is another physical interpretation of the approach to equilibrium and the temporal decrease of $U(t)$ [15, p. 20]: Returning, for a moment, to the realm of continuous–time Markov processes, we can write the master equations as follows:

$$\frac{\mathrm{d}P_t(x)}{\mathrm{d}t} = \sum_{x'} \frac{1}{R_{xx'}} \left[\frac{P_t(x')}{P(x')} - \frac{P_t(x)}{P(x)}\right] \qquad (31)$$

where $R_{xx'} = [P(x')W_{x'x}]^{-1} = [P(x)W_{xx'}]^{-1}$. Imagine now an electrical circuit where the indices $\{x\}$ designate the various nodes. Nodes x and x' are connected by a wire with resistance $R_{xx'}$ and every node x is grounded via a capacitor with capacitance $P(x)$ (see Fig. 2). If $P_t(x)$ is the charge at node x at time t, then the master equations are the Kirchoff equations of the currents at each node in the circuit. Thus, the way in which probability spreads across the states is analogous to the way charge spreads across the circuit and probability fluxes are now analogous to electrical

---

[7] We are not yet in a position to obtain the monotonicity of $D(P_t \| P_t')$ as a special case of the monotonicity of $D_Q(P \| P_t)$. This will require a slight further extension of this information measure, to be carried out later on.

currents. If we now choose $Q(u) = \frac{1}{2}u^2$, then

$$U(t) = \frac{1}{2}\sum_x \frac{P_t^2(x)}{P(x)}, \tag{32}$$

which means that the energy stored in the capacitors dissipates as heat in the wires until the system reaches equilibrium, where all nodes have the same potential, $P_t(x)/P(x) = \text{const.}$, and hence detailed balance corresponds to the situation where all individual currents vanish (not only their algebraic sum). This constant potential must, of course, be 1 because both $P_t(x)$ and $P(x)$ are probability distributions.

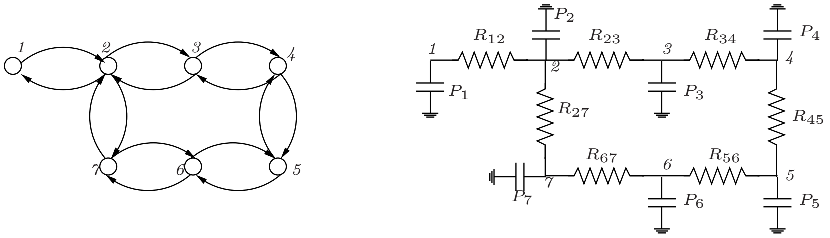

Figure 2: State transition diagram of a Markov chain (left part) and the electric circuit that emulates the dynamics of $\{P_t(x)\}$ (right part).

We have seen, in the above examples, that various choices of the function Q yield various f–divergences between $\{P(x))\}$ and $\{P_t(x)\}$, which are both marginal distributions of a single symbol x. What about joint distributions of two or more symbols? Consider, for example, the function

$$J(t) = \sum_{x,x'} P(X_0 = x, X_t = x') \cdot Q\left(\frac{P(X_0 = x)P(X_t = x')}{P(X_0 = x, X_t = x')}\right), \tag{33}$$

where Q is convex as before. Here, by the same token, $J(t)$ is the f–divergence between the joint probability distribution $\{P(X_0 = x, X_t = x')\}$ and the product of marginals $\{P(X_0 = x)P(X_t = x')\}$, namely, it is the generalized mutual information of [10],[24], and [25], as mentioned in the Introduction.[8] Now, using a similar chain of inequalities as before, we get the non–increasing monotonicity of $J(t)$ as follows:

$$\begin{aligned} J(t) = \sum_{x,x',x''} & P(X_0 = x, X_t = x', X_{t+1} = x'')\times \\ & Q\left(\frac{P(X_0 = x)P(X_t = x')}{P(X_0 = x, X_t = x')} \cdot \frac{P(X_{t+1} = x''|X_t = x')}{P(X_{t+1} = x''|X_t = x')}\right) \end{aligned}$$

---

[8] It should be noted, however, that in both [10] and [24] a technical condition was imposed on Q, that is, $\lim_{x\to 0} xQ(1/x) = 0$, but it was not required in [25], where the focus was on the finite alphabet case.

$$
\begin{aligned}
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \sum_{x'} P(X_t = x' | X_0 = x, X_{t+1} = x'') \times \\
&Q\left(\frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_t = x', X_{t+1} = x'')}\right) \\
&\geq \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \cdot Q\left(\sum_{x'} P(X_t = x' | X_0 = x, X_{t+1} = x'') \times \right. \\
&\left. \frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_t = x', X_{t+1} = x'')}\right) \\
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') Q\left(\sum_{x'} \frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_{t+1} = x'')}\right) \\
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \cdot Q\left(\frac{P(X_0 = x) P(X_{t+1} = x'')}{P(X_0 = x, X_{t+1} = x'')}\right) \\
&= J(t+1). \qquad (34)
\end{aligned}
$$

This time, we assumed only the Markov property of (X_0, X_t, X_{t+1}) (not even homogeneity). This is, in fact, nothing but the 1973 version of the generalized data processing theorem of Ziv and Zakai [24], which was mentioned in the Introduction.

3 A Unified Framework

In this section, we relate the generalized version of the Boltzmann H–theorem and the generalized data processing theorems and formalize the uniform framework that supports both. This is done, first for the 1973 version [24] of the Ziv–Zakai data processing theorem (along with an example), and then for the 1975 version by Zakai and Ziv [25].

In spite of the general resemblance (via the notion of the f–divergence), the last monotonicity result, concerning $J(t)$, and the monotonicity of $D(P_t \| P'_t)$, do not seem, on the face of it, to fall in the framework of the monotonicity of the f–divergence $D_Q(P \| P_t)$. This is because in the latter, there is an additional dependence on a stationary state distribution that appears neither in $D(P_t \| P'_t)$ nor in $J(t)$. However, two simple observations can put them both in the framework of the monotonicity of $D_Q(P \| P_t)$.

The first observation is that the monotonicity of $U(t) = D_Q(P \| P_t)$ continues to hold (with a straightforward extension of the proof) if $P_t(x)$ is extended to be a vector of time varying state distributions $(P_t^1(x), P_t^2(x), \ldots, P_t^k(x))$, and Q is taken to be a convex function of k variables.

Moreover, in fact, each component $P_t^i(x)$ does not have to be necessarily a probability distribution. It can be any function $\mu_t^i(x)$ that satisfies the recursion

$$\mu_{t+1}^i(x) = \sum_{x'} \mu_t^i(x')P(x|x'), \qquad 1 \le i \le k. \tag{35}$$

This means that $\sum_x \mu_t^i(x)$ can be any constant, that may depend on i, but not on t (moreover, this constant may be ∞ when the alphabet size is infinite). Let us then denote $\boldsymbol{\mu}_t(x) = (\mu_t^1(x), \mu_t^2(x), \ldots, \mu_t^k(x))$ and assume that Q is jointly convex in all its k arguments. Then the redefined function

$$\begin{aligned} U(t) &= \sum_{x\in\mathcal{X}} P(x) \cdot Q\left(\frac{\boldsymbol{\mu}_t(x)}{P(x)}\right) \\ &= \sum_{x\in\mathcal{X}} P(x) \cdot Q\left(\frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right) \end{aligned} \tag{36}$$

is monotonically non–increasing with t.

The second observation is rooted in convex analysis, and it is related to the notion of the perspective of a convex function and its convexity property [5]. Here, a few words of background are in order. Let $Q(\boldsymbol{u})$ be a convex function of the vector $\boldsymbol{u} = (u_1, \ldots, u_k)$ and let $v > 0$ be an additional variable. Then, the function

$$\tilde{Q}(v, u_1, u_2, \ldots, u_k) \triangleq v \cdot Q\left(\frac{u_1}{v}, \frac{u_2}{v}, \ldots, \frac{u_k}{v}\right) \tag{37}$$

is called the *perspective function* of Q. A well–known property of the perspective operation is conservation of convexity, in other words, if Q is convex in $\boldsymbol{u}$, then $\tilde{Q}$ is convex in $(v, \boldsymbol{u})$. The proof of this fact, which is straightforward, can be found, for example, in [5, p. 89, Subsection 3.2.6] (see also [12]) and it is brought here for the sake of completeness: Letting λ_1 and λ_2 be two non–negative numbers summing to unity and letting $(v_1, \boldsymbol{u}_1)$ and $(v_2, \boldsymbol{u}_2)$ be given, then

$$\begin{aligned} \tilde{Q}(\lambda_1(v_1, \boldsymbol{u}_1) + \lambda_2(v_2, \boldsymbol{u}_2)) &= (\lambda_1 v_1 + \lambda_2 v_2) \cdot Q\left(\frac{\lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2}{\lambda_1 v_1 + \lambda_2 v_2}\right) \\ &= (\lambda_1 v_1 + \lambda_2 v_2) \cdot Q\left(\frac{\lambda_1 v_1}{\lambda_1 v_1 + \lambda_2 v_2} \cdot \frac{\boldsymbol{u}_1}{v_1} + \frac{\lambda_2 v_2}{\lambda_1 v_1 + \lambda_2 v_2} \cdot \frac{\boldsymbol{u}_2}{v_2}\right) \\ &\le \lambda_1 v_1 Q\left(\frac{\boldsymbol{u}_1}{v_1}\right) + \lambda_2 v_2 Q\left(\frac{\boldsymbol{u}_2}{v_2}\right) \\ &= \lambda_1 \tilde{Q}(v_1, \boldsymbol{u}_1) + \lambda_2 \tilde{Q}(v_2, \boldsymbol{u}_2). \end{aligned} \tag{38}$$

Putting these two observations together, we can now state the following result:

Theorem 1 *Let*

$$V(t) = \sum_{x:\ P(x)>0} \mu_t^0(x) Q\left(\frac{\mu_t^1(x)}{\mu_t^0(x)}, \frac{\mu_t^2(x)}{\mu_t^0(x)}, \ldots, \frac{\mu_t^k(x)}{\mu_t^0(x)}\right), \tag{39}$$

where Q is a convex function of k variables and $\{\mu_t^i(x)\}_{i=0}^k$ are arbitrary functions that satisfy the recursion

$$\mu_{t+1}^i(x) = \sum_{x'} \mu_t^i(x') P(x|x'), \qquad i = 0, 1, 2, \ldots, k, \tag{40}$$

and where $\mu_t^0(x)$ is moreover strictly positive. Assume further that $\tilde{Q}$, the perspective of Q, satisfies the condition $\lim_{\alpha\to 0} \alpha\tilde{Q}(v/\alpha, u_1/\alpha, \ldots u_k/\alpha) = 0$ for every $(v, u_1, \ldots, u_k)$. Then, $V(t)$ is a monotonically non–increasing function of t.

Proof. Using the above mentioned observations, the proof of Theorem 1 is straightforward: Letting P be a stationary state distribution of $\{X_t\}$, we have:

$$\begin{aligned}
V(t) &= \sum_{x:\ P(x)>0} \mu_t^0(x) Q\left(\frac{\mu_t^1(x)}{\mu_t^0(x)}, \frac{\mu_t^2(x)}{\mu_t^0(x)}, \ldots, \frac{\mu_t^k(x)}{\mu_t^0(x)}\right) \\
&= \sum_{x:\ P(x)>0} P(x) \cdot \frac{\mu_t^0(x)}{P(x)} Q\left(\frac{\mu_t^1(x)/P(x)}{\mu_t^0(x)/P(x)}, \ldots, \frac{\mu_t^k(x)/P(x)}{\mu_t^0(x)/P(x)}\right) \\
&= \sum_{x:\ P(x)>0} P(x) \tilde{Q}\left(\frac{\mu_t^0(x)}{P(x)}, \frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right) && (41) \\
&= \sum_x P(x) \tilde{Q}\left(\frac{\mu_t^0(x)}{P(x)}, \frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right), && (42)
\end{aligned}$$

where the last line holds by the regularity condition concerning $\tilde{Q}$, as $0 \cdot \tilde{Q}(v/0, u_1/0, \ldots, u_k/0)$ is considered zero. Since $\tilde{Q}$ is the perspective of the convex function Q, then it is convex as well, and so, the monotonicity of $V(t)$ follows from the first observation above. This completes the proof of Theorem 1.

It is now readily seen that both $D(P_t \| P_t')$ and $J(t)$ are special cases of $V(t)$, at least when $P(x) > 0$ for all x, and hence we have essentially covered all special cases seen thus far under the umbrella of the more general information functional $V(t)$.[9] Note also that the restriction of the

---

[9] The limitation $P(x) > 0$ means that x is a recurrent state of the Markov process. This limitation becomes immaterial if we simply eliminate all transient states and possibly focus on one (or more) of the closed subsets of communicating states. Such a removal of transient states does not change significantly the probability law since the transient states cease to occur after finite time anyway.

summation over x only to symbols with positive steady state probability is immaterial if Q satisfies a similar technical condition as required for $\tilde{Q}$ in Theorem 1, namely, $\lim_{\alpha\to 0}\alpha Q(u_1/\alpha,\ldots u_k/\alpha)=0$ for every $(u_1,\ldots,u_k)$, and so, $0\cdot Q(u_1/0,\ldots,u_k/0)$ is considered zero. In such a case, the summation can obviously be extended to cover the entirety of $\mathcal{X}$.

It is important to observe that the same idea exactly can be applied, first of all, to the 1973 version of the Ziv–Zakai data processing theorem (regardless of the above described monotonicity results concerning Markov processes): Analogously to $V(t)$, consider the generalized mutual information functional

$$J_Q(X;Y)\triangleq\sum_{(x,y):\ P_{XY}(x,y)>0}\mu^0_{XY}(x,y)Q\left(\frac{\mu^1_{XY}(x,y)}{\mu^0_{XY}(x,y)}\right), \tag{43}$$

where $\mu^0_{XY}(x,y)>0$ and $\mu^1_{XY}(x,y)$ are arbitrary functions that are consistent with the Markov conditions, i.e., for any Markov chain $X\to Y\to Z$, these functions satisfy

$$\mu^i_{XZ}(x,z)=\sum_y\mu^i_{XY}(x,y)P_{Z|Y}(z|y)=\sum_y\mu^i_{YZ}(y,z)P_{X|Y}(x|y),\qquad i=0,1. \tag{44}$$

Then, $J_Q(X;Y)$ satisfies the data processing inequalities,

$$J_Q(X;Y)\ge J_Q(X;Z);\quad J_Q(Y;Z)\ge J_Q(X;Z) \tag{45}$$

because, similarly as in the proof of Theorem 1,

$$\begin{aligned}J_Q(X;Y)&=\sum_{(x,y):\ P_{XY}(x,y)>0}P_{XY}(x,y)\cdot\frac{\mu^0_{XY}(x,y)}{P_{XY}(x,y)}\cdot Q\left(\frac{\mu^1_{XY}(x,y)/P_{XY}(x,y)}{\mu^0_{XY}(x,y)/P_{XY}(x,y)}\right)\\&=\sum_{x,y}P_{XY}(x,y)\cdot\tilde{Q}\left(\frac{\mu^0_{XY}(x,y)}{P_{XY}(x,y)},\frac{\mu^1_{XY}(x,y)}{P_{XY}(x,y)}\right),\end{aligned} \tag{46}$$

which is a Zakai–Ziv information functional of the 1975 version [25] and hence it satisfies a data processing inequality. Once again, we are assuming that $\tilde{Q}$ satisfies a technical condition similar to the one in Theorem 1.

It is natural to ask, at this point, whether the summation over (x,y) in the above definition of $J_Q(X;Y)$ can also be extended to cover all $\mathcal{X}\times\mathcal{Y}$, without the restriction of $P_{XY}(x,y)>0$, and still the resulting information measure would satisfy a data processing inequality. The answer turns out to be affirmative, and the simplest way to prove it is the direct one (see Appendix A). Thus,

$J_Q(X;Y)$ as defined in (43), both with and without the positivity limitation, are valid information measures in the sense of admitting data processing inequalities. Throughout the sequel, we will adopt the definition without the positivity limitation, namely, $J_Q(X;Y)$ will be re–defined as:

$$J_Q(X;Y) = \sum_{x,y} \mu_{XY}^0(x,y) Q\left(\frac{\mu_{XY}^1(x,y)}{\mu_{XY}^0(x,y)}\right). \tag{47}$$

By the same token, in the realm of Markov processes, Theorem 1 can be modified analogously: One may re–define $V(t)$ as the summation over the entire alphabet, not just those letters with $P(x) > 0$, and then the technical condition concerning the perspective $\tilde{Q}$ can be removed (see the ending sentence of Appendix A for a hint on that). Thus, both with $J_Q(X;Y)$ and $V(t)$, we have the freedom to choose whether the summation would be on the entire alphabet or only over those symbols (or combinations of symbols) with positive probability. Both are valid information measures with the desired properties.

Returning to data processing inequalities, what functions, $\mu_{XY}^0(x,y)$ and $\mu_{XY}^1(x,y)$, can be consistent with the Markov conditions? Two such functions are, of course, $\mu_{XY}^0(x,y) = P_{XY}(x,y)$ and $\mu_{XY}^1(x,y) = P_X(x)P_Y(y)$, which bring us back to the 1973 Ziv–Zakai information measure. We can, of course, swap their roles and obtain a generalized version of the lautum information [20], which is also known to satisfy a data processing inequality. For additional options, let us consider a communication system, operating on single symbols (block length 1), where the source symbol u is mapped into a channel input $x = f(u)$, by a deterministic encoder f, which is then fed into the channel $P_{Y|X}(y|x)$, and the channel output y is in turn mapped into the reconstruction symbol $v = g(y)$. As is argued in [25], the function $\mu_{UY}(u,y) = P_U(u)P_{Y|U}(y|u_0)$ is consistent with the Markov conditions for any given source symbol u_0. Indeed, since the encoder is assumed deterministic, $P_{Y|U}(y|u_0) = P_{Y|X}(y|f(u_0)) = P_{Y|X}(y|x_0)$, and it is easily seen that

$$\mu_{UV}(u,v) = P_U(u)P_{V|U}(v|u_0) = \sum_y P_U(u)P_{Y|U}(y|u_0)P_{V|Y}(v|y) = \sum_y \mu_{UY}(u,y)P_{V|Y}(v|y) \tag{48}$$

and

$$\begin{aligned}\mu_{UY}(u,y) &= P_U(u)P_{Y|U}(y|u_0) \\ &= \sum_x P_{U|X}(u|x)P_X(x)P_{Y|U}(y|u_0) \\ &= \sum_x P_{U|X}(u|x)P_X(x)P_{Y|X}(y|x_0) = \sum_x P_{U|X}(u|x)\mu_{XY}(x,y).\end{aligned} \tag{49}$$

Of course, every linear combination of all these functions is also consistent with the Markov conditions. Thus, we can take

$$\mu^0_{XY}(x,y) = s_0 P_{XY}(x,y) + \sum_{x_i\in\mathcal{X}} s_i P_X(x) P_{Y|X}(y|x_i) \tag{50}$$

and

$$\mu^1_{XY}(x,y) = t_0 P_{XY}(x,y) + \sum_{x_i\in\mathcal{X}} t_i P_X(x) P_{Y|X}(y|x_i), \tag{51}$$

where $\{s_i\}$ and $\{t_i\}$ are the (arbitrary) coefficients of these linear combinations (with the limitation that $s_i \geq 0$ for all i, with at least one $s_i > 0$). Thus, we may define

$$J_Q(X;Y) = \sum_{x,y}\left[s_0 P_{XY}(x,y) + \sum_{x_i\in\mathcal{X}} s_i P_X(x) P_{Y|X}(y|x_i)\right]\cdot Q\left(\frac{t_0 P_{XY}(x,y) + \sum_{x_i\in\mathcal{X}} t_i P_X(x) P_{Y|X}(y|x_i)}{s_0 P_{XY}(x,y) + \sum_{x_i\in\mathcal{X}} s_i P_X(x) P_{Y|X}(y|x_i)}\right), \tag{52}$$

or, equivalently,

$$J_Q(X;Y) = \sum_{x,y} P_X(x)\left[s_0 P_{Y|X}(y|x) + \sum_{x_i\in\mathcal{X}} s_i P_{Y|X}(y|x_i)\right]\cdot Q\left(\frac{t_0 P_{Y|X}(y|x) + \sum_{x_i\in\mathcal{X}} t_i P_{Y|X}(y|x_i)}{s_0 P_{Y|X}(y|x) + \sum_{x_i\in\mathcal{X}} s_i P_{Y|X}(y|x_i)}\right). \tag{53}$$

Moreover, to eliminate the dependence on the specific encoder, we can think of $\{x_i\}$ as independent random variables $\{X_i\}$, take the expectation w.r.t. their randomness (in the same spirit as in [25]), and obtain the following information measure

$$\boldsymbol{E}\left\{\sum_{x,y} P_X(x)\left[s_0 P_{Y|X}(y|x) + \sum_i s_i P_{Y|X}(y|X_i)\right]\cdot Q\left(\frac{t_0 P_{Y|X}(y|x) + \sum_i t_i P_{Y|X}(y|X_i)}{s_0 P_{Y|X}(y|x) + \sum_i s_i P_{Y|X}(y|X_i)}\right)\right\}, \tag{54}$$

where the expectation is w.r.t. the product measure of $\{X_i\}$, $P_{X_1X_2,\ldots}(x_1,x_2,\ldots) = \prod_i P_X(x_i)$. These are essentially the most general information measures, that obey a data processing inequality, that we can get with a univariate convex function Q.

As a simple example, returning to eq. (53) and taking $s_0 = 1$, $t_0 = 0$, $s_i = sP_X(x_i)$ ($s \geq 0$, a parameter), and $t_i = P_X(x_i)$, $x_i \in \mathcal{X}$, we have $\mu^0_{XY}(x,y) = P_{XY}(x,y) + sP_X(x)P_Y(y)$, and $\mu^1_{XY}(x,y) = P_X(x)P_Y(y)$, and the resulting generalized mutual information reads

$$J_Q(X;Y) = \sum_{x,y} P_X(x)[P_{Y|X}(y|x) + sP_Y(y)]\cdot Q\left(\frac{P_Y(y)}{P_{Y|X}(y|x) + sP_Y(y)}\right). \tag{55}$$

The interesting point concerning these generalized mutual information measures is that even if we remain in the framework of the 1973 version of the Ziv–Zakai data processing theorem of a univariate function Q (as opposed to the 1975 version where it is multivariate), we have added an extra degrees of freedom (in the above example, the parameter s), which may be used in order to improve the obtained bounds. If the inequality $R_Q(d) \leq C_Q$ can be transformed into an inequality on the distortion d, where the lower bound depends on s, then this bound can be maximized w.r.t. the parameter s. If the optimum $s > 0$ yields a distortion bound which is larger than that of $s = 0$, then we have improved on [24] for the given choice of the convex function Q. Sometimes this optimization may not be a trivial task, but even if we can just identify one positive value of s (including the limit $s \to \infty$) that is better than $s = 0$, then we have improved on the generalized data processing bound of [24], which corresponds to $s = 0$. This additional degree of freedom may be important, because, as mentioned in the Introduction, the variety of convex functions $\{Q\}$ which are convenient to work with, is somewhat limited (most notably, the functions $Q(z) = z^2$, $Q(z) = 1/z$, $Q(z) = -\sqrt{z}$ and some piecewise linear functions [24],[25]). The next example demonstrates this point.

Example. Consider the information functional (55) with the convex function $Q(z) = -\sqrt{z}$. Then, the corresponding generalized mutual information is

$$\begin{aligned}
J_Q(X;Y) &= -\sum_{x,v} P_X(x)[P_{Y|X}(y|x) + sP_X(x)] \cdot \sqrt{\frac{P_X(x)}{P_{Y|X}(y|x) + sP_Y(y)}} \\
&= -\sum_{x,y} P_X(x)\sqrt{P_Y(y)[P_{Y|X}(y|x) + sP_Y(y)]} \\
&= -\sum_{x,y} P_X(x)P_Y(y)\sqrt{s + \frac{P_{Y|X}(y|x)}{P_Y(y)}}. \qquad (56)
\end{aligned}$$

Consider now the above–described problem of joint source–channel coding, for the following source and channel: The source is designated by a random variable U, which is uniformly distributed over the alphabet $\mathcal{U} = \{0, 1, \ldots, K-1\}$. The reproduction variable, V, takes on values in the same alphabet, i.e., $\mathcal{V} = \mathcal{U} = \{0, 1, \ldots, K-1\}$ and the distortion function is

$$\rho(u,v) = \begin{cases} 0 & v = u \\ 1 & v = (u+1) \bmod K \\ \infty & \text{elsewhere} \end{cases} \qquad (57)$$

which means that errors other than $v = (u+1)\text{mod } K$ are strictly forbidden. Therefore the test channel pertaining to the source must be of the form

$$P_{V|U}(v|u) = \begin{cases} 1-\epsilon_u & v = u \\ \epsilon_u & v = (u+1) \bmod K \\ 0 & \text{elsewhere} \end{cases} \tag{58}$$

where $\{\epsilon_u\}$ are parameters taking values in $[0,1]$ and complying with the distortion constraint

$$\boldsymbol{E}\{\rho(U,V)\} = \frac{1}{K}\sum_{u=0}^{K-1}\epsilon_u \le d. \tag{59}$$

The communication channel $P_{Y|X}$ is a noise–free L–ary channel, i.e., its input and output alphabets are $\mathcal{X} = \mathcal{Y} = \{0,1,\ldots,L-1\}$ with $P_{Y|X}(y|x) = 1$ for $y = x$, and $P_{Y|X}(y|x) = 0$ otherwise.

Obviously, the case $K \le L$ is not interesting because the data can be conveyed error–free by trivially connecting the source to the channel. In the other extreme, where $K > 2L$, there must be some channel input symbol to which at least three source symbols are mapped. In such a case, it is impossible to avoid at least one of the forbidden errors in the reconstruction. Thus, the interesting cases are those for which $L < K \le 2L$, or equivalently, $\theta \in (1,2]$, where $\theta \triangleq K/L$.

In Appendix B, we derive a lower bound on the distortion, as a function of θ, by using the data processing inequality pertaining to the above defined choice of J_Q. Since J_Q depends also on s, then so does the corresponding lower bound on the distortion, which will be denoted by $d_s(\theta)$. In particular, for $s = 0$ we obtain the lower bound

$$d_0(\theta) = \frac{1}{2} - \frac{1}{2}\sqrt{2\theta - \theta^2} \tag{60}$$

whereas for $s \to \infty$, the resulting lower bound is

$$d_\infty(\theta) = \lim_{s\to\infty} d_s(\theta) = \frac{1}{2} - \frac{1}{2\theta}\sqrt{2\theta - \theta^2}, \tag{61}$$

which is larger, and hence tighter, for every $\theta \ge 1$. It is interesting to compare this also to the classical data processing theorem: Since the (ordinary) rate–distortion function of the source, in this example, is:

$$R(d) = \log K - h_2(d), \tag{62}$$

where $h_2(d)$ is the binary entropy function

$$h_2(d) = -d\log d - (1-d)\log(1-d) \tag{63}$$

and the channel capacity is

$$C = \log L, \tag{64}$$

then the ordinary data processing theorem yields

$$h_2(d) \geq \log \theta. \tag{65}$$

which means that the resulting lower bound is

$$d_{ordinary}(\theta) = h_2^{-1}(\log \theta), \tag{66}$$

where $h_2^{-1}(\cdot)$ is the inverse of $h_2(\cdot)$ in the range where $h_2(\cdot)$ is monotonically increasing. Since

$$h_2(d) \geq 4d(1-d) \tag{67}$$

and

$$2\left(1 - \frac{1}{\theta}\right) \geq \log_2 \theta \tag{68}$$

within the relevant range of θ, the bound $d_\infty(\theta)$ is also better than $d_{ordinary}(\theta)$ for this case. However, the bound $d_0(\theta)$ turns out to be inferior to $d_{ordinary}(\theta)$. Graphs of the three distortion bounds are depicted in Fig. 1. This completes the example. $\square$

Finally, we should comment that the monotonicity result concerning $V(t)$ contains as special cases, not only the H–theorem, as well as all other earlier mentioned monotonicity results, but also the 1975 Zakai–Ziv generalized data processing [25]. Consider a Markov chain $U \to V \to W$, where U, V and W are random variables that take on values in (finite) alphabets, $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$, respectively. Let us now map between the Markov chain (U, V, W) and the Markov process $\{X_t\}$ in the following manner: $(u, v) \in \mathcal{U} \times \mathcal{V}$ is assigned to the state x' of the process at time t, whereas $(u, w) \in \mathcal{U} \times \mathcal{W}$ corresponds[10] to x at time $t + 1$. Now, defining accordingly,

$$\begin{aligned} \mu_t^0(x') &= P_{UV}(u, v), & (69) \\ \mu_t^1(x') &= P_U(u)P_V(v), & (70) \\ \mu_{t+1}^0(x) &= P_{UW}(u, w), & (71) \end{aligned}$$

---

[10] While $\mathcal{V}$ and $\mathcal{W}$ may be different (finite) alphabets, x and x', of the original Markov process, must take on values in the same alphabet. Assuming, without loss of generality, that $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$ and $\mathcal{W} = \{1, 2, \ldots, |\mathcal{W}|\}$, then for the purpose of this mapping, we can unify these alphabets to be both $\{1, 2, \ldots, \max\{|\mathcal{V}|, |\mathcal{W}|\}\}$ and complete the missing elements of the extended transition matrix $P_{W|V}(w|v)$ in a consistent manner, according to the actual support of each distribution. We omit further technical details herein.

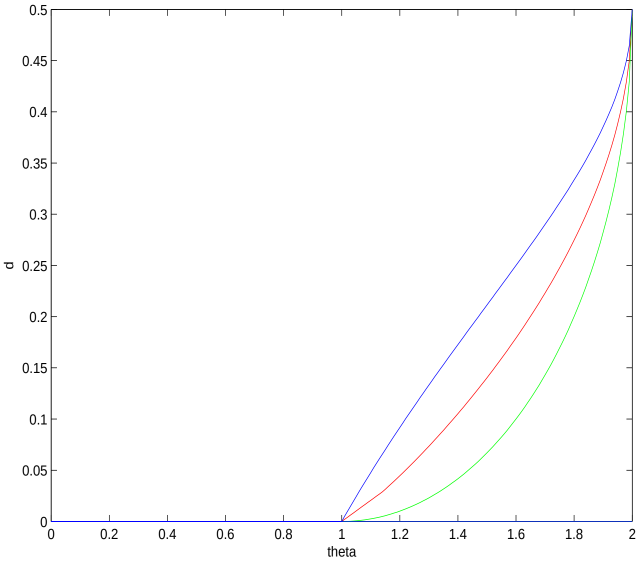

Figure 3: Various lower bounds on the distortion as a function of θ in the range $\theta \in [1,2]$: The upper–most curve (blue) is $d_\infty(\theta)$, the lower–most (green) is $d_0(\theta)$, and third curve (red) is $d_{ordinary}(\theta)$.

and

$$\mu_{t+1}^1(x) = P_U(u)P_W(w), \tag{72}$$

then due to the Markov property of (U,V,W), both measures satisfy the recursion with $P_{W|V}(w|v)$ playing the role[11] of $P(x|x')$. I.e.,

$$\begin{aligned} P_{UW}(u,w) &\triangleq \mu_{t+1}^0(x) \\ &= \sum_{x'} \mu_t^0(x')P(x|x') \\ &= \sum_{v} P_{UV}(u,v)P_{W|V}(w|v) \end{aligned} \tag{73}$$

and

$$\begin{aligned} P_U(u)P_W(w) &\triangleq \mu_{t+1}^1(x) \\ &= \sum_{x'} \mu_t^1(x')P(x|x') \\ &= \sum_{v} P_U(u)P_V(v)P_{W|V}(w|v) \end{aligned} \tag{74}$$

---

[11] Consider the component u of $x' = (u,v)$ and $x = (u,w)$ simply as an index.

Thus, for $Q(z) = -\ln z$, the monotonicity of $V(t)$ is nothing but the data processing of the classical mutual information. For a general function Q of one variable ($k = 1$), this gives the generalized data processing theorem of [24]. Furthermore, letting Q be a general convex function of k variables, and $\mu_t^0(x') = P_{U,V}(u,v)$ as before, we get the more general form of the data processing inequality of [25].

The above extension of the H–theorem gives rise to a seemingly more general data processing theorem than in [25], as it is not necessary to let $\mu_t^0(x)$ be the actual joint probability distribution. However, when looking at the entire class of convex functions with an arbitrary number of arguments, this is not really more general, as the corresponding generalized mutual information can readily be transformed back to the form of the 1975 Zakai–Ziv information functional using again the perspective operation. Indeed, as mentioned in the Introduction and shown in [25, Theorem 7.1], the class of generalized mutual information measures studied therein cannot be improved upon in the sense that there always exist choices of Q and $\{\mu^i\}$ that provide tight bounds on the distortion of the optimum system.

4 Summary and Conclusion

The main contributions of this work can be summarized as follows: First, we have established a unified framework and a relationship between (a generalized version of) the second law of thermodynamics and the generalized data processing theorems of Zakai and Ziv. This unified framework turns out to strengthen and expand both of these pieces of theory: Concerning the second law of thermodynamics, we have identified a significantly more general information measure, which is a monotonic function of time, when it operates on a Markov process. As for the generalized Ziv–Zakai data processing theorem, we have proposed a wider class of information measures obeying the data processing theorem, which includes free parameters that may be optimized so as to tighten the distortion bounds.

Acknowledgments

Interesting discussions with J. Ziv and M. Zakai are acknowledged with thanks. Useful comments made by the Associate Editor, Pascal Vontobel, and by the anonymous referees are acknowledged

with thanks.

Appendix A

In this appendix, we provide a direct proof of the data processing inequality for the information measure

$$J_Q(X;Y) = \sum_{x,y} \mu^0_{XY}(x,y) Q\left(\frac{\mu^1_{XY}(x,y)}{\mu^0_{XY}(x,y)}\right). \tag{A.1}$$

For a Markov chain $X \to Y \to Z$, we assume that eq. (44) holds. Denoting $\mu^i_{XYZ}(x,y,z) \triangleq \mu^i_{XY}(x,y)P_{Z|Y}(z|y)$, $i = 0,1$, we define

$$\mu^0_{Y|X,Z}(y|x,z) = \frac{\mu^0_{XYZ}(x,y,z)}{\mu^0_{XZ}(x,z)} \tag{A.2}$$

where

$$\mu^0_{XZ}(x,z) = \sum_y \mu^0_{XYZ}(x,y,z). \tag{A.3}$$

Obviously, $\mu^0_{Y|X,Z}(y|x,z) \geq 0$ and $\sum_y \mu^0_{Y|X,Z}(y|x,z) = 1$ for all x and z. Thus,

$$\begin{aligned}
J_Q(X;Y) &= \sum_{x,y} \mu^0_{XY}(x,y) Q\left(\frac{\mu^1_{XY}(x,y)}{\mu^0_{XY}(x,y)}\right) && \text{(A.4)}\\
&= \sum_{x,y,z} \mu^0_{XY}(x,y)P_{Z|Y}(z|y) Q\left(\frac{\mu^1_{XY}(x,y)P_{Z|Y}(z|y)}{\mu^0_{XY}(x,y)P_{Z|Y}(z|y)}\right) && \text{(A.5)}\\
&= \sum_{x,y,z} \mu^0_{XYZ}(x,y,z) Q\left(\frac{\mu^1_{XYZ}(x,y,z)}{\mu^0_{XYZ}(x,y,z)}\right) && \text{(A.6)}\\
&= \sum_{x,z} \mu^0_{XZ}(x,z) \sum_y \mu^0_{Y|XZ}(y|x,z) Q\left(\frac{\mu^1_{XYZ}(x,y,z)}{\mu^0_{XYZ}(x,y,z)}\right) && \text{(A.7)}\\
&\geq \sum_{x,z} \mu^0_{XZ}(x,z) Q\left(\sum_y \mu^0_{Y|XZ}(y|x,z) \cdot \frac{\mu^1_{XYZ}(x,y,z)}{\mu^0_{XYZ}(x,y,z)}\right) && \text{(A.8)}\\
&= \sum_{x,z} \mu^0_{XZ}(x,z) Q\left(\sum_y \frac{\mu^1_{XYZ}(x,y,z)}{\mu^0_{XZ}(x,z)}\right) && \text{(A.9)}\\
&= \sum_{x,z} \mu^0_{XZ}(x,z) Q\left(\frac{\mu^1_{XZ}(x,z)}{\mu^0_{XZ}(x,z)}\right) && \text{(A.10)}\\
&= J_Q(X;Z). && \text{(A.11)}
\end{aligned}$$

A similar proof applies to a modified version of Theorem 1, where $V(t)$ is redefined to be the summation over the entire alphabet, not just over symbols with positive state–state probability.

In the above proof, simply replace $\mu^0_{XY}(x,y)$, $\mu^1_{XY}(x,y)$, $\mu^0_{XZ}(x,z)$, $\mu^1_{XZ}(x,z)$, $P_{Z|Y}(z|y)$ and the summation over y, by $\mu^0_t(x)$, $\boldsymbol{\mu}_t(x)$, $\mu^0_{t+1}(x)$, $\boldsymbol{\mu}_{t+1}(x)$, $P(x'|x)$, and a summation over x, respectively.

Appendix B

In this appendix, we derive a distortion bound based on the generalized data processing theorem, in the spirit of [24] and [25], where we now have the parameter s as a degree of freedom.

As for the source, let us suppose that in addition to the distortion constraint, we impose the constraint that the distribution of the reproduction variable V, just like U, must be uniform over its alphabet, namely, $P_V(v) = 1/K$ for all $v \in \mathcal{V}$. In this case,

$$\begin{aligned}
-J_Q(U;V) &= \sum_{u,v} P_U(u)P_V(v)\sqrt{s + \frac{P_{V|U}(v|u)}{P_V(v)}} \\
&= \frac{1}{K^2}\sum_{u=0}^{K-1}\left[\sqrt{s+K\epsilon_u} + \sqrt{s+K(1-\epsilon_u)} + (K-2)\sqrt{s}\right] \\
&= \frac{1}{K^2}\sum_{u=0}^{K-1}\left[\sqrt{s+K\epsilon_u} + \sqrt{s+K(1-\epsilon_u)}\right] + \left(1-\frac{2}{K}\right)\sqrt{s} \\
&\le \frac{1}{K^2}\cdot K\left[\sqrt{s+Kd} + \sqrt{s+K(1-d)}\right] + \left(1-\frac{2}{K}\right)\sqrt{s} \\
&= \frac{1}{K}\left[\sqrt{s+Kd} + \sqrt{s+K(1-d)}\right] + \left(1-\frac{2}{K}\right)\sqrt{s}, \qquad \text{(B.1)}
\end{aligned}$$

where the inequality follows from the fact that the maximum of the function

$$\sum_u [\sqrt{s+K\epsilon_u} + \sqrt{s+K(1-\epsilon_u)}],$$

which is concave in $\{\epsilon_u\}$, subject to the distortion constraint (59), is achieved when $\epsilon_u = d$ for all $u \in \mathcal{U}$. Thus,

$$\begin{aligned}
R_Q(d) &\triangleq \min\{J_Q(U;V):\ \boldsymbol{E}\rho(U,V) \le D\} \qquad \text{(B.2)}\\
&= -\frac{1}{K}\left[\sqrt{s+Kd} - \sqrt{s+K(1-d)}\right] - \left(1-\frac{2}{K}\right)\sqrt{s}. \qquad \text{(B.3)}
\end{aligned}$$

As for the channel, we have:

$$-J_Q(X;Y) = \sum_{x,y} P_X(x)P_Y(y)\sqrt{s + \frac{P_{Y|X}(y|x)}{P_Y(y)}}$$

$$
\begin{aligned}
&= \sum_{x' \neq x} P_X(x) P_X(x') \sqrt{s} + \sum_x P_X^2(x) \sqrt{s + \frac{1}{P_X(x)}} \\
&= \sqrt{s} \left[1 - \sum_x P_X^2(x) \right] + \sum_x P_X^2(x) \sqrt{s + \frac{1}{P_X(x)}} \\
&= \sqrt{s} + \sum_x P_X^2(x) \left(\sqrt{s + \frac{1}{P_X(x)}} - \sqrt{s} \right) \\
&= \sqrt{s} + \sum_x P_X^2(x) \cdot \frac{1/P_X(x)}{\sqrt{s + 1/P_X(x)} + \sqrt{s}} \\
&= \sqrt{s} + \sum_x \frac{P_X(x)}{\sqrt{s + 1/P_X(x)} + \sqrt{s}}. \qquad \text{(B.4)}
\end{aligned}
$$

The function $f(t) = t/[\sqrt{s + 1/t} + \sqrt{s}]$ is convex in t (for fixed s) since $f''(t) \geq 0$ for all $t \geq 0$, as can readily be verified. Thus, $-J_Q(X;Y)$ is minimized by the uniform distribution $P_X(x) = 1/L$, $\forall x$, which leads to the 'capacity' expression:

$$
C_Q = -\sqrt{s} - \frac{1}{\sqrt{s} + \sqrt{s + L}}. \qquad \text{(B.5)}
$$

Applying now the data processing theorem,

$$
R_Q(d) \leq C_Q, \qquad \text{(B.6)}
$$

we obtain, after rearranging terms

$$
\sqrt{s + Kd} + \sqrt{s + K(1-d)} \geq \frac{K}{\sqrt{s} + \sqrt{s + L}} + 2\sqrt{s}. \qquad \text{(B.7)}
$$

Squaring both sides, we have:

$$
2s + K + 2\sqrt{(s + Kd)[s + K(1-d)]} \geq \left[\frac{K}{\sqrt{s} + \sqrt{s + L}} + 2\sqrt{s} \right]^2 \qquad \text{(B.8)}
$$

or

$$
2\sqrt{(s + Kd)[s + K(1-d)]} \geq \left[\frac{K}{\sqrt{s} + \sqrt{s + L}} + 2\sqrt{s} \right]^2 - 2s - K, \qquad \text{(B.9)}
$$

which after squaring again and applying some further straightforward algebraic manipulations, gives eventually the following inequality on the distortion d:

$$
4d(1-d) \geq \psi(s), \qquad \text{(B.10)}
$$

where

$$
\psi(s) \triangleq \frac{1}{K^2} \left[\left(\frac{K}{\sqrt{s} + \sqrt{s + L}} + 2\sqrt{s} \right)^2 - 2s - K \right]^2 - \frac{4s(s + K)}{K^2}. \qquad \text{(B.11)}
$$

The resulting lower bound on the distortion is the smaller of the two solutions of the equation $4d(1-d) = \psi(s)$, which is

$$d_s \triangleq \frac{1}{2} - \frac{1}{2}\sqrt{1-\psi(s)}. \tag{B.12}$$

Thus, the larger is $\psi(s)$, the better is the bound. The choice $s = 0$, which corresponds to the usual Ziv–Zakai bound for $Q(z) = -\sqrt{z}$, yields

$$\psi(0) = \frac{1}{K^2}\left[\left(\frac{K}{\sqrt{L}}\right)^2 - K\right]^2 = \left(\frac{K}{L} - 1\right)^2 = (\theta - 1)^2, \tag{B.13}$$

which depends on K and L only via $\theta = K/L$. When substituted into eq. (B.12), it yields

$$d_0 = d_0(\theta) = \frac{1}{2} - \frac{1}{2}\sqrt{2\theta - \theta^2}. \tag{B.14}$$

However, it turns out that $s = 0$ is not the best choice of s. We next examine the limit $s \to \infty$. To this end, we derive a lower bound to $\psi(s)$ which is more convenient to analyze in this limit. Note that for $s \geq L/8$, it is guaranteed that the expression in the square brackets of the expression defining $\psi(s)$, is positive, which means that an upper bound on $\sqrt{s+L}$ would yield a lower bound to $\psi(s)$. Thus, upper bounding $\sqrt{s+L}$ by

$$\sqrt{s+L} = \sqrt{s} \cdot \sqrt{1 + L/s} \leq \sqrt{s}\left(1 + \frac{L}{2s}\right),$$

we get

$$\begin{aligned}
K^2\psi(s) &= \left[\left(\frac{K}{\sqrt{s} + \sqrt{s+L}} + 2\sqrt{s}\right)^2 - 2s - K\right]^2 - 4s^2 - 4Ks \\
&\geq \left[\left(\frac{K}{\sqrt{s}(2 + L/2s)} + 2\sqrt{s}\right)^2 - 2s - K\right]^2 - 4s^2 - 4Ks \\
&= K^2\left(\frac{4s - L}{4s + L}\right)^2 + \frac{16K^4s^2}{(4s+L)^4} - \frac{8KLs}{4s+L} + \frac{16K^2s^2}{(4s+L)^2} + \frac{8K^3s(4s-L)}{(4s+L)^3} \\
&\triangleq K^2\psi_0(s),
\end{aligned} \tag{B.15}$$

where between the second and the third lines, we have skipped some standard algebraic operations. Taking now the limit $s \to \infty$, we obtain

$$\psi_\infty = \lim_{s\to\infty} \psi_0(s) = \frac{1}{K^2}(K^2 + 0 - 2KL + K^2 + 0) = 2\left(1 - \frac{L}{K}\right) = 2\left(1 - \frac{1}{\theta}\right), \tag{B.16}$$

which again depends on K and L only via θ, and yields a bound better than $d_0(\theta)$ for $\theta \geq 1$. In particular,

$$d_\infty(\theta) = \lim_{s\to\infty} d_s = \frac{1}{2} - \frac{1}{2\theta}\sqrt{2\theta - \theta^2}. \tag{B.17}$$

References

[1] M. S. Ali and D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc.*, Ser. B, no. 28, pp. 131–140, 1966.

[2] D. Andelman, "Bounds according to a generalized data processing theorem," M.Sc. dissertation, Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel, October 1974.

[3] G. B. Bağci, "The physical meaning of Rényi relative entropies," arXiv:cond-mat/0703008v1, March 1, 2007.

[4] A. H. W. Beck, *Statistical Mechanics, Fluctuations and Noise*, Edward Arnold Publishers, 1976.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] H. R. Brown, W. Myrvold and J. Uffink, "Boltzmann's H–theorem, its discontents, and the birth of statistical mechanics," *Studies in History and Philosophy of Modern Physics*, vol. 40, pp. 174–191, 2009.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, second edition, John Wiley & Sons, 2006.

[8] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 95–108, 1963.

[9] I. Csiszár, "Information–type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[10] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2 (1–4), pp. 191–213, 1972.

[11] I. Csiszár and P. Shields, "Information theory and statistics: a tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 417–528, 2004.

[12] B. Dacorogna and P. Maréchal, "The role of perspective functions in convexity, polyconvexity, rank–one convexity and separate convexity," http://caa.epfl.ch/publications/2008-The_role_of_perspective_functions_in_convexity.pdf

[13] E. T. Jaynes, "Violation of Boltzmann's H–theorem in real gases," *Phys. Rev. A*, vol. 4, no. 2, pp. 747–750, August 1971.

[14] M. Kardar, *Statistical Physics of Particles*, Cambridge University Press, 2007.

[15] F. P. Kelly, *Reversibility and Stochastic Networks*, J. Wiley & Sons, 1979.

[16] C. Kittel, *Elementary Statistical Physics*, John Wiley & Sons, 1958.

[17] F. Liese and I. Vajda, *Convex Statistical Distances*, Leipzig, Teubner, 1987.

[18] F. Liese and I. Vajda, "On divergence and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[19] T. Morimoto, "Markov processes and the H—theorem," *J. Phys. Soc. Jap.*, vol. 12, pp. 328–331, 1963.

[20] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inform. Theory*, vol. 54, no. 3, pp. 964–975, March 2008.

[21] A. Pérez–Madrid, "Molecular theory of irreversibility," *Journal of Chemical Physics*, vol. 123, 204148, 2005. q

[22] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, 1965.

[23] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw–Hill, 1979.

[24] J. Ziv and M. Zakai, "On functionals satisfying a data–processing theorem," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 3, pp. 275–283, May 1973.

[25] M. Zakai and J. Ziv, "A generalization of the rate–distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, pp. 87–123, 1975.