



**IRWIN AND JOAN JACOBS**  
**CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES**

# **Physics of the Shannon Limits**

**Neri Merhav**

**CCIT Report # 725**  
**March 2009**

■ ■ ■ ■ Electronics  
■ ■ ■ ■ Computers  
■ ■ ■ ■ Communications

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



# Physics of the Shannon Limits

Neri Merhav

Department of Electrical Engineering  
Technion – Israel Institute of Technology  
Technion City, Haifa 32000, Israel  
Email: merhav@ee.technion.ac.il

**Abstract**— We provide a simple physical interpretation, in the context of the second law of thermodynamics, to the information inequality (a.k.a. the Gibbs' inequality, which is also equivalent to the log-sum inequality), asserting that the relative entropy between two probability distributions cannot be negative. Since this inequality stands at the basis of the data processing theorem (DPT), and the DPT in turn is at the heart of most, if not all, proofs of converse theorems in Shannon theory, it is observed that conceptually, the roots of fundamental limits of Information Theory can actually be attributed to the laws of physics, in particular, the second law of thermodynamics, and indirectly, also the law of energy conservation. By the same token, in the other direction: one can view the second law as stemming from information-theoretic principles.

**Index Terms**— Gibbs' inequality, data processing theorem, entropy, second law of thermodynamics, divergence, relative entropy, mutual information.

## I. INTRODUCTION

While the laws of physics draw the boundaries between the possible and the impossible in Nature, the coding theorems of Information Theory, or more precisely, their converse parts, draw the boundaries between the possible and the impossible in the design and performance of coded communication systems and in data processing. A natural question that may arise, in view of these two facts, is whether there is any relationship between them. It is the purpose of this work to touch upon this question and to make an attempt to provide at least a partial answer.

Perhaps the most fundamental inequality in Information Theory is the so called *information inequality* (cf. e.g., [1, Theorem 2.6.3, p. 28]), which asserts that the relative entropy (a.k.a. the Kullback–Leibler divergence) between two probability distributions over the same alphabet  $P = \{P(x), x \in \mathcal{X}\}$  and  $Q = \{Q(x), x \in \mathcal{X}\}$ ,

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

can never be negative, and a similar fact applies to probability density functions with the summation across  $\mathcal{X}$  being replaced by integration.

The *log-sum inequality* (LSI) [1, Theorem 2.7.1, p. 31], which asserts that for two sets of non-negative numbers,  $(a_1, a_2, \dots, a_n)$  and  $(b_1, b_2, \dots, b_n)$ :

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \left( \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right),$$

is completely equivalent<sup>1</sup> to the information inequality, although proved in [1] in a rather different manner.

Yet another name for the same inequality, which is more frequently encountered in the jargon of physicists, is the *Gibbs' inequality*: When the information inequality is applied to two probability distributions of the Boltzmann form (cf. Section IV below), it yields an interesting inequality concerning their corresponding free energies (cf. e.g., [2, Section 5.6, pp. 143–146]), which serves as a useful tool for obtaining good bounds on the free energy of a complex system, when its exact value is difficult to calculate.

In this work, we provide a simple physical interpretation to this inequality of the free energies, and thereby also to the information inequality, or the log-sum inequality. This physical interpretation is directly related to the second law of thermodynamics, which asserts that the entropy of an isolated physical system cannot decrease: According to this interpretation, the divergence between two probability distributions is proportional to the energy dissipated in the system when it undergoes an irreversible process, and hence converts this energy loss into entropy production, or heat. Thus, the non-negativity of the relative entropy is related to the non-negativity of this entropy change, which is, as said, the second law of thermodynamics.

Since the mutual information can be thought of as an instance of the relative entropy, and so can the difference between two mutual informations defined along a Markov chain, then the data processing theorem (DPT) can, of course, also be given the very same physical interpretation. Considering the fact that the DPT is pivotal to most, if not all, converse theorems in Information Theory, this means that, in fact, the fundamental limits of Information Theory can, at least conceptually, be attributed to the laws of physics, in particular, to the second law of thermodynamics.<sup>2</sup> The rate loss in any suboptimal coded communication system, is given the meaning of irreversibility and entropy production in a corresponding physical system. Optimum (or nearly optimum) communication systems are corresponding to reversible processes (or lack of any process at all) with no entropy

<sup>1</sup>The information inequality is obtained from the LSI when  $(a_1, a_2, \dots, a_n)$  and  $(b_1, b_2, \dots, b_n)$  both sum to unity, and conversely, the LSI is obtained from the information inequality, by applying the latter to the probability distributions  $P_i = a_i / \sum_j a_j$  and  $Q_i = b_i / \sum_j b_j$ .

<sup>2</sup>Another law of physics that plays a role here, at least indirectly, is the law of energy conservation, because our derivations are all based on the Boltzmann–Gibbs distribution of equilibrium statistical mechanics, and this distribution, in turn, is derived on the basis of the energy conservation law.

production. Stated in somewhat different words, had there been a communication system that violated a fundamental limit (e.g., beating the entropy, or channel capacity), then in principle, one could have constructed a physical system that violates the second law, and vice versa.

The outline of the remaining part of the paper is as follows. In Section II, we give some basic background in statistical physics. Section III reviews the role of the DPT in many of the converse theorems in the Shannon theory. In Section III, we offer a physical interpretation to the Gibbs' inequality and show how it applies to the DPT in two different scenarios. Finally, in Section IV, we discuss relationships between reversible processes in physics and error exponents of classical Neyman–Pearson hypothesis testing.

## II. PHYSICS BACKGROUND

Consider a physical system with  $n$  particles, which at any time instant, can be found in any one out of a variety of microscopic states (or *microstates*, for short). The microstate is defined by the full physical information about all  $n$  particles, e.g., the positions, momenta, angular momenta, spins, etc., depending on the type of the physical system. In particular, a microstate is designated by  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , where each  $x_i$  may itself be a vector, consisting of all the relevant physical state variables (such as the above) for particle number  $i$  at a given time instant. Associated with every microstate  $\mathbf{x}$ , there is an energy function, a.k.a. the *Hamiltonian*,  $\mathcal{E}(\mathbf{x})$ . For example, in the case of the ideal gas,  $x_i = (\mathbf{p}_i, \mathbf{r}_i)$ , where  $\mathbf{p}_i$  and  $\mathbf{r}_i$ , both three dimensional vectors, are the momentum and the position of particle number  $i$ , respectively, and

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \left[ \frac{\|\mathbf{p}_i\|^2}{2m} + mgz_i \right], \quad (1)$$

where  $m$  is the mass of each particle,  $g$  is the gravitation constant, and  $z_i$  is the height – one of the components of  $\mathbf{r}_i$ .

One of the most fundamental results in statistical physics (based on the law of energy conservation and the postulate that all microstates of the same energy are equiprobable) asserts that, when a system lies in thermal equilibrium with the environment (heat bath), the probability of finding the system at state  $\mathbf{x}$  is given by the *Boltzmann–Gibbs* distribution

$$P(\mathbf{x}) = \frac{e^{-\beta\mathcal{E}(\mathbf{x})}}{Z(\beta)} \quad (2)$$

where  $\beta = 1/(kT)$ ,  $k$  being Boltzmann's constant and  $T$  being temperature, and  $Z(\beta)$  is the normalization constant, called the *partition function*, which is given by

$$Z(\beta) = \sum_{\mathbf{x}} e^{-\beta\mathcal{E}(\mathbf{x})}, \text{ or } Z(\beta) = \int d\mathbf{x} e^{-\beta\mathcal{E}(\mathbf{x})}, \quad (3)$$

depending on whether  $\mathbf{x}$  is discrete or continuous. The partition function is a key quantity from which many important macroscopic physical quantities can be derived. For example, the average internal energy w.r.t. (2) is

$$E = \mathbf{E}\{\mathcal{E}(\mathbf{X})\} = -\frac{d \ln Z(\beta)}{d\beta}, \quad (4)$$

the entropy (in units of  $k$ ) pertaining to (2) is

$$\Sigma(\beta) \triangleq \frac{S(\beta)}{k} = -\mathbf{E}\{\ln P(\mathbf{X})\} = \ln Z(\beta) + \beta \cdot E, \quad (5)$$

and the *free energy* is given by

$$F(\beta) = -\frac{\ln Z(\beta)}{\beta}. \quad (6)$$

From eq. (5), one readily obtains the well known relationship

$$F = E - ST.$$

Thus, any change in the internal energy, along a fixed temperature (isothermal) process, is given by

$$\Delta E = \Delta F + T\Delta S,$$

in other words, it consists of two components: the first is the change in the free energy,  $\Delta F$ , and the second pertains to entropy production,  $T\Delta S$ . By the first law of thermodynamics, which is actually, the law of energy conservation,

$$\Delta E = \Delta Q + \Delta W,$$

namely, the origins of any change in the internal energy may be a combination of the heat  $\Delta Q$  transferred into the system and the work  $\Delta W$  applied to it. According to the thermodynamical definition, the entropy difference,  $\Delta S$ , between two macroscopic states  $A$  and  $B$ , is defined as  $\int_A^B dQ/T$ , where the integration is along a *quasi-static* or *reversible* process, i.e., a process that is slow enough such that, along the way, the system is kept always very close to equilibrium. By the *Clausius theorem* (cf. e.g., [2, p. 13]), in the above described isothermal process,  $\Delta S$  is never smaller than  $\Delta Q/T$ , with equality when the process is reversible. Thus, by comparing the two expressions of  $\Delta E$ , we immediately observe that  $\Delta W \geq \Delta F$ .

The free energy is then given a meaning of crucial importance in thermodynamics and statistical physics: The difference,  $\Delta F$ , between the free energies associated with two equilibrium points pertaining to the same temperature (but with two different values of some other control parameter, such as pressure or magnetic field) has the physical meaning of the *minimum* amount of work that should be applied to the system in order to transfer it between these two equilibria along an isothermal process, and this minimum is attained when the process is reversible.<sup>3</sup> Equivalently, the negative free-energy difference,  $-\Delta F$ , is the *maximum* amount of work that can be exploited from the system in an isothermal process, and this maximum is achieved, again, if the process is reversible. The second law of thermodynamics, as mentioned earlier, asserts that the entropy of an isolated system cannot decrease.

<sup>3</sup>This fact is also known as the *minimum work principle*.

### III. THE DATA PROCESSING THEOREM AND FUNDAMENTAL LIMITS

As mentioned earlier, our observations apply to any fundamental limit, or converse theorem, that makes use of the information inequality, in one way or another. However, even if we confine our attention only to those that use it explicitly in the form of the DPT, it is not difficult to appreciate the fact that we already cover many of the fundamental limits, if not all of them. Here are just a few examples.

*Lossy/lossless source coding:* Consider a source vector  $U^N = (U_1, \dots, U_N)$  compressed into a bitstream  $X^n = (X_1, \dots, X_n)$  from which the decoder generates a reproduction  $V^N = (V_1, \dots, V_N)$  with distortion  $\sum_{i=1}^N E\{d(U_i, V_i)\} \leq ND$ . Then, by the DPT,  $I(U^N; V^N) \leq I(X^n; V^N) \leq H(X^n)$ , where  $I(U^N; V^N)$  is further lower bounded by  $NR(D)$  and  $H(X^n) \leq n$ , which together lead to the converse to the lossy data compression theorem, asserting that the compression ratio  $n/N$  cannot be less than  $R(D)$ . Lossless compression is obtained, of course, as a special case where  $D = 0$ .

*Channel coding under bit error probability:* Let  $U^N = (U_1, \dots, U_N)$  be drawn from the binary symmetric course (BSS), designating  $M = 2^N$  equiprobable messages of length  $N$ . The encoder maps  $U^N$  into a channel input vector  $X^n$ , which in turn, is sent across the channel. The receiver observes  $Y^n$ , a noisy version of  $X^n$ , and decodes the message as  $V^N$ . Let  $P_b = \frac{1}{N} \sum_{i=1}^N \Pr\{V_i \neq U_i\}$  designate the bit error probability. Then, by the DPT,  $I(U^N; V^N) \leq I(X^n; Y^n)$ , where  $I(X^n; Y^n)$  is further upper bounded by  $nC$ ,  $C$  being the channel capacity, and  $I(U^N; V^N) = H(U^N) - H(U^N|V^N) \geq N - \sum_{i=1}^N H(U_i|V_i) \geq N - \sum_i h_2(\Pr\{V_i \neq U_i\}) \geq N[1 - h_2(P_b)]$ . Thus, for  $P_b$  to vanish, the coding rate,  $N/n$  should not exceed  $C$ .

*Channel coding under block error probability – Fano’s inequality:* This is the same as in the previous item, except that the error performance is the block error probability  $P_B = \Pr\{V^N \neq U^N\}$ . This time,  $H(U^N|V^N)$ , which is identical to  $H(U^N, E|V^N)$ , with  $E \triangleq \mathcal{I}\{V^N \neq U^N\}$  ( $\mathcal{I}$  being the indicator function), is decomposed as  $H(E|V^N) + H(U^N|V^N, E)$ , where the first term is upper bounded by 1 and the second term is upper bounded by  $P_B \log(2^N - 1) < NP_B$ , owing to the fact that the maximum of  $H(U^N|V^N, E = 1)$  is obtained when  $U^N$  is distributed uniformly over all  $V^N \neq U^N$ . Putting these facts all together, we obtain Fano’s inequality  $P_B \geq 1 - 1/n - C/R$ , where  $R = N/n$  is the coding rate. Thus, the DPT directly supports Fano’s inequality, which in turn is the main tool for proving converses to channel coding theorems in a large variety of communication situations, including network configurations.

*Joint source–channel coding and the separation principle:*

In a joint source–channel situation, where the source vector  $U^N$  is mapped into a channel input vector  $X^n$  and the channel output vector  $Y^n$  is decoded into a reconstruction  $V^N$ , the DPT gives rise to the chain of inequalities  $NR(D) \leq I(U^N; V^N) \leq I(X^n; Y^n) \leq nC$ , which is the converse to the joint source–channel coding theorem, whose direct part can be achieved by separate source- and channel coding. The first two examples above are special cases of this.

*Conditioning reduces entropy:* Perhaps even more often than the term “data processing theorem” can be found as part of a proof of a converse theorem, one encounters an equivalent of this theorem under the slogan “conditioning reduces entropy”. This in turn is part of virtually every converse proof in the literature. Indeed, if  $(X, U, V)$  is a triple of RV’s, then this statement means that  $H(X|V) \geq H(X|U, V)$ . If, in addition,  $X \rightarrow U \rightarrow V$  is a Markov chain, then  $H(X|U, V) = H(X|U)$ , and so,  $H(X|V) \geq H(X|U)$ , which in turn is equivalent to the more customary form of the DPT,  $I(X; U) \geq I(X; V)$ , obtained by subtracting  $H(X)$  from both sides of the entropy inequality. In fact, as we shall see shortly, it is this entropy inequality that lends itself more naturally to a physical interpretation. Moreover, we can think of the conditioning–reduces–entropy inequality as another form of the DPT even in the absence of the aforementioned Markov condition, because  $X \rightarrow (U, V) \rightarrow V$  is always a Markov chain.

### IV. PHYSICS OF THE INFORMATION INEQUALITY & DPT

We consider two forms of the information inequality an the DPT, one corresponding to an isothermal process and one – to an adiabatic process (fixed amount of heat).

#### A. Isothermal Version

Consider a system, with a microstate  $\mathbf{x}$ , which may have two possible Hamiltonians –  $\mathcal{E}_0(\mathbf{x})$  and  $\mathcal{E}_1(\mathbf{x})$ . Let  $Z_i(\beta)$ , denote the partition function pertaining to  $\mathcal{E}_i(\cdot)$ , that is,  $Z_i(\beta) = \sum_{\mathbf{x}} e^{-\beta \mathcal{E}_i(\mathbf{x})}$ ,  $i = 0, 1$ , where  $\beta = 1/(kT)$  is the inverse temperature. Since  $\beta$  is fixed throughout this section, we will also use the shorthand notation  $Z_i$  for the partition function. Let  $P_i(\mathbf{x})$  denote the Boltzmann–Gibbs distribution (cf. eq. (2)) pertaining to  $Z_i$ ,  $i = 0, 1$  (both for the same given value of  $\beta$ ). Applying the information inequality to  $P_0$  and  $P_1$ , we get:

$$\begin{aligned} 0 &\leq D(P_0||P_1) = \sum_{\mathbf{x}} P_0(\mathbf{x}) \ln \left[ \frac{e^{-\beta \mathcal{E}_0(\mathbf{x})}/Z_0}{e^{-\beta \mathcal{E}_1(\mathbf{x})}/Z_1} \right] \\ &= \ln Z_1 - \ln Z_0 + \beta \mathbf{E}_0\{\mathcal{E}_1(\mathbf{X}) - \mathcal{E}_0(\mathbf{X})\} \end{aligned} \quad (7)$$

where  $\mathbf{E}_0\{\cdot\}$  denotes the expectation operator w.r.t.  $P_0$ . After a minor algebraic rearrangement, this becomes:

$$\begin{aligned} \mathbf{E}_0\{\mathcal{E}_1(\mathbf{X}) - \mathcal{E}_0(\mathbf{X})\} &\geq kT \ln Z_0 - kT \ln Z_1 \\ &\equiv F_1 - F_0, \end{aligned} \quad (8)$$

where  $F_i$  is the free energy pertaining to  $P_i$ ,  $i = 0, 1$  (cf. eq. 6)).

We now offer the following physical interpretation to this inequality: Imagine that a system with Hamiltonian  $\mathcal{E}_0(\mathbf{x})$  is

in equilibrium for all  $t < 0$ ,<sup>4</sup> but then, at time  $t = 0$ , the Hamiltonian changes *abruptly* from the  $\mathcal{E}_0(\mathbf{x})$  to  $\mathcal{E}_1(\mathbf{x})$  (e.g., by suddenly applying a force, like pressure or a magnetic field, to the system), which means that if the system is found at state  $\mathbf{x}$  at time  $t = 0$ , additional energy of  $W(\mathbf{x}) = \mathcal{E}_1(\mathbf{x}) - \mathcal{E}_0(\mathbf{x})$  is suddenly ‘injected’ into it. This additional energy can be thought of as work performed on the system, or as supplementary potential energy. Of course,  $W(\mathbf{x})$  is a random variable due to the randomness of  $\mathbf{x}$ . Since this passage between  $\mathcal{E}_0$  and  $\mathcal{E}_1$  is abrupt, and the microstate  $\mathbf{x}$  does not change instantaneously, the expectation of  $W(\mathbf{X})$  should be taken w.r.t.  $P_0$ , and this average is exactly what we have at the left-hand side eq. (8). The Gibbs’ inequality tells us then that this average work is at least as large as  $\Delta F = F_1 - F_0$ , the increase in free energy, in compliance to the explanation in Section II. The difference

$$E_0\{W(\mathbf{X})\} - \Delta F = kT \cdot D(P_0 \| P_1) \geq 0$$

is due to the irreversible nature of this abrupt energy injection, and this irreversibility means an increase of the total entropy of the system and its environment.<sup>5</sup> Thus, the Gibbs’ inequality is, in fact, a version of the second law of thermodynamics, and the relative entropy is given a very simple physical significance. We next consider two examples.

*Example 1 – Fixed-to-variable compression and the Ising model.* A natural information-theoretic example for this can be easily motivated by the interpretation of the relative entropy as the rate loss (or, the redundancy) due to mismatch in fixed-to-variable lossless data compression: Suppose that  $\mathbf{X} \in \{-1, +1\}^n$  emerges from a first-order Markov source  $P_0(\mathbf{x}) = \prod_{i=1}^n P_0(x_i | x_{i-1})$ , where

$$P_0(x | x') = \frac{\exp\{Jx \cdot x'\}}{Z_0}, \quad x, x' \in \{-1, +1\},$$

and where  $J$  is a given constant and

$$Z_0 = 2 \cosh(J).$$

However, the code designer designs a Shannon code according to  $P_1(\mathbf{x}) = \prod_{i=1}^n P_1(x_i | x_{i-1})$ , where

$$P_1(x | x') = \frac{\exp\{Jx \cdot x' + Kx\}}{\zeta(x')}, \quad x, x' \in \{-1, +1\}$$

where  $K$  is another given constant and  $\zeta(x)$  is the appropriate normalization factor given by

$$\zeta(x) = \begin{cases} 2 \cosh(J + K) & x = +1 \\ 2 \cosh(J - K) & x = -1 \end{cases}$$

<sup>4</sup>Since the information inequality applies to any pair of distributions, it is conceivable that the interpretation we offer may remain relevant even beyond the realm of systems in equilibrium. Indeed, even if the system is away from equilibrium, when it is nevertheless in steady state (in the sense that macroscopic physical quantities are time-invariant), the negative logarithm of the density function can be given the meaning of an *effective Hamiltonian* [3]. This, however, is beyond the scope of this work.

<sup>5</sup>See also [4], [5], [6] and references therein, where the same conclusions are reached from a more general perspective of irreversible processes, but under certain limiting assumptions on the physical system.

Considering the fact that  $x \in \{-1, +1\}$ ,  $\zeta(x)$  can also be written in a unified way as

$$\zeta(x) = Z_1 \cdot \left[ \frac{\cosh(J + K)}{\cosh(J - K)} \right]^{x/2}.$$

where

$$Z_1 = 2\sqrt{\cosh(J + K) \cosh(J - K)}.$$

From the physics point of view, both  $P_0$  and  $P_1$  can be thought of as Boltzmann–Gibbs distributions with inverse temperature  $\beta = 1$ : For the former, we define the Hamiltonian as

$$\begin{aligned} \mathcal{E}_0(\mathbf{x}) &\triangleq -n \ln Z_0 - \sum_{i=1}^n \ln P_0(x_i | x_{i-1}) \\ &= -J \cdot \sum_i x_{i-1} x_i \end{aligned} \quad (9)$$

which can be thought of as the energy pertaining to nearest-neighbor interactions between spins in a one-dimensional array, that is, the one-dimensional *Ising model* (see, e.g., [7, Sect. 1.8]) with a coupling coefficient  $J$ , in the absence of a magnetic field. On the other hand, for  $P_1$  we define:

$$\begin{aligned} \mathcal{E}_1(\mathbf{x}) &\triangleq -n \ln Z_1 - \sum_{i=1}^n \ln P_1(x_i | x_{i-1}) \\ &= -J \sum_i x_{i-1} x_i - K \sum_i x_i - \\ &\quad \frac{1}{2} \left[ \ln \frac{\cosh(J - K)}{\cosh(J + K)} \right] \cdot \sum_i x_{i-1} \\ &\approx -J \sum_i x_{i-1} x_i - \\ &\quad \left( K + \frac{1}{2} \ln \frac{\cosh(J - K)}{\cosh(J + K)} \right) \cdot \sum_i x_i \\ &\triangleq -J \sum_i x_{i-1} x_i - B \sum_i x_i \end{aligned} \quad (10)$$

where in the approximate equality we neglected ‘‘edge effects’’ that make the (relatively) small difference between  $\sum_i x_i$  and  $\sum_i x_{i-1}$  (for large  $n$ ). This is the same Ising model as before, but now also with a magnetic field  $B$ . Thus,

$$\mathcal{E}_1(\mathbf{x}) - \mathcal{E}_0(\mathbf{x}) = -B \sum_i x_i$$

is the energy injected by an abrupt application of the magnetic field  $B$ . We have therefore demonstrated that the entropy production due to the irreversibility of this abrupt magnetic field is (within the additive constant,  $\Delta F = 1 \cdot (\ln Z_0 - \ln Z_1)$ ) proportional to the redundancy of the mismatched code.

*Example 2 – Run-length coding and the grand-canonical ensemble.* The Boltzmann–Gibbs distribution of eq. (2), a.k.a. the *canonical distribution*, is the equilibrium distribution of a system that is allowed to exchange heat energy with its environment at a fixed temperature  $T$ . It also assumes that the system has a fixed number of particles  $n$ , and a fixed volume  $V$ , whenever the volume is a relevant factor.

When the system is allowed to exchange with the environment, not only energy, but also matter, namely, particles, then eq. (2) is extended to the *grand-canonical distribution* [2, Sect. 4.9], whose microstate is defined as  $(\mathbf{x}, n)$ , where  $n$  is now a random variable, and  $\mathbf{x}$  is defined as before for the given  $n$ . According to this distribution,

$$P(\mathbf{x}, n) = \frac{e^{\beta(\mu n - \mathcal{E}(\mathbf{x}))}}{\Xi(\beta, \mu)}$$

where

$$\Xi(\beta, \mu) = \sum_{n \geq 0} e^{\beta \mu n} \sum_{\mathbf{x}} e^{-\beta \mathcal{E}(\mathbf{x})} \triangleq \sum_{n \geq 0} e^{\beta \mu n} Z(\beta, n)$$

is the *grand partition function*. The parameter  $\mu$ , which is called the *chemical potential*, controls the average number of particles in the system. Note that  $P(\mathbf{x}, n)$  can be thought of as  $P(n) \cdot P(\mathbf{x}|n)$  where  $P(\mathbf{x}|n)$  obeys the canonical distribution for the given  $n$  and  $P(n)$  is proportional to  $e^{\beta \mu n} Z(\beta, n)$ . It is well known (see, e.g., [2]) that  $kT \ln \Xi(\beta, \mu)$  gives the equilibrium pressure–volume product of the system,  $PV$ . Now let  $P_0(\mathbf{x}, n)$  and  $P_1(\mathbf{x}, n)$  be two grand-canonical distributions that differ only in the chemical potentials,  $\mu_i$ ,  $i = 0, 1$ , respectively. Applying the information inequality, we get

$$\begin{aligned} 0 &\leq D(P_0 \| P_1) \\ &= \ln \Xi(\beta, \mu_1) - \ln \Xi(\beta, \mu_0) + \\ &\quad \beta(\mu_0 - \mu_1) \mathbf{E}_0\{N\} \end{aligned} \quad (11)$$

where  $N$  designates the random number of particles. Dividing by  $\beta$  and rearranging terms, this becomes:

$$P_1 V \geq P_0 V + (\mu_1 - \mu_0) \mathbf{E}_0\{N\},$$

and after dividing by  $V$  (which is assumed fixed), we get:

$$P_1 \geq P_0 + (\mu_1 - \mu_0) \mathbf{E}_0\{\rho\},$$

where  $\rho = N/V$  is the density of particles.

A natural information–theoretic analogue of this is run–length coding: Given a 0–1 binary memoryless source with a very high probability of ‘0’, which we shall designate by  $e^\mu$  ( $\mu < 0$ ,  $\beta = 1$ ), the idea is to encode the number  $N$  of successive zeroes between every two consecutive ones. Clearly, the distribution of  $N$  is exponential

$$\Pr\{N = n\} = \frac{e^{\mu n}}{\Xi(\mu)}$$

where, with a slight abuse of notation, we define

$$\Xi(\mu) = \frac{1}{1 - e^\mu},$$

and where we have assumed  $\mathcal{E}(\mathbf{x}) = -\ln P(\mathbf{x}|n)$ , and so,  $Z(1, n) = 1$  for all  $n$ . Thus, when applying run–length coding, the price of mismatch in  $\mu$  is parallel to the difference between the two sides of the above pressure inequality, where the ‘pressure’ in run–length coding is proportional to  $-\ln(1 - e^\mu)$ . As  $\mu \uparrow 0$ , the pressure increases, and more ‘particles’ (i.e., runs of zeroes) enter into the system, which means that the runlengths becomes larger. Thus, we have

demonstrated an analogy between run–length coding and the physics of the grand–canonical ensemble: the log–probability of ‘0’ plays the role the chemical potential whereas the log–probability of ‘1’ is associated with pressure. This concludes Example 2.

Returning to the general framework, let us now see how the Gibbs’ inequality is related to the DPT. Consider a triple of random variables  $(X, U, V)$  which form a Markov chain  $X \rightarrow U \rightarrow V$ . The DPT asserts that  $I(X; U) \geq I(X; V)$ . We can obtain the DPT as a special case of the Gibbs’ inequality because

$$\begin{aligned} I(X; U) - I(X; V) &= H(X|V) - H(X|U) \\ &= \mathbf{E}\{D(P_{X|U,V}(\cdot|U, V) \| P_{X|V}(\cdot|V))\} \end{aligned}$$

where the expectation is w.r.t. the randomness of  $(U, V)$ . Thus, For a given realization  $(u, v)$  of  $(U, V)$ , consider the Hamiltonians  $\mathcal{E}_0(x) = -\ln P(x|u) = -\ln P(x|u, v)$  and  $\mathcal{E}_1(x) = -\ln P(x|v)$ , pertaining to a single ‘particle’ whose state is  $x$ . Let us also set  $\beta = 1$ . Thus, for a given  $(u, v)$ :

$$\begin{aligned} \mathbf{E}_0\{W(X)\} &= \sum_x P(x|u, v) [\ln P(x|u) - \ln P(x|v)] \\ &= H(X|V = v) - H(X|U = u) \end{aligned} \quad (12)$$

and after further averaging w.r.t.  $(U, V)$ , the average work becomes  $H(X|V) - H(X|U) = I(X; U) - I(X; V)$ . Concerning the free energies, we have

$$\begin{aligned} Z_0(1) &= \sum_x \exp\{-1 \cdot [-\ln P(x|u, v)]\} \\ &= \sum_x P(x|u, v) = 1 \end{aligned} \quad (13)$$

and similarly,

$$Z_1(1) = \sum_x P(x|v) = 1$$

which means that  $F_0(1) = F_1(1) = 0$ , and so  $\Delta F = 0$  as well. So by the Gibbs’ inequality, the average work,  $I(X; U) - I(X; V)$ , cannot be smaller than the free–energy difference, which in this case vanishes, namely,  $I(X; U) - I(X; V) \geq 0$ , which is the DPT. Note that in this case, there is a maximum degree of irreversibility: The identity  $I(X; U) - I(X; V) = H(X|V) - H(X|U)$  means that whole average work,  $W = I(X; U) - I(X; V)$ , goes for entropy increase  $T\Delta\Sigma = 1 \cdot [H(X|V) - H(X|U)]$ , whereas the free energy remains unchanged, as mentioned earlier. Moreover, the entire entropy increase goes to the system under discussion, and none of it goes to the environment.

At this point a comment is in order: The rate loss of a suboptimal communication system, when viewed from the DPT perspective, may be attributed to two possible factors: one factor comes from a possible mismatch between actual distributions and optimum distributions in the information–theoretic sense, for example, the encoder may not induce the capacity–achieving channel input distribution or the test channel of the rate–distortion function. The other factor is a

possible gap between mutual informations along the Markov chain ( $I(X;U)$  may be strictly larger than  $I(X;V)$ ), which actually means *information loss*, and which is *irreversible* ( $U$  cannot be retrieved from  $V$ ). It is the latter kind of loss that is parallel to the irreversible free energy loss and dissipation.

From a more general physical perspective, we can think of the Hamiltonian

$$\mathcal{E}_\lambda(x) = \mathcal{E}_0(x) + \lambda[\mathcal{E}_1(x) - \mathcal{E}_0(x)]$$

as a linear interpolation between the two extremes,  $\lambda = 0$  and  $\lambda = 1$ , pertaining to  $\mathcal{E}_0$  and  $\mathcal{E}_1$ , and then  $\lambda$  can be thought of as a control parameter or a ‘force’ that influences the system. The Jarzynsky equality (cf. e.g., [4] and references therein) tells that under certain conditions on the system and the environment, and given any protocol for a temporal change in  $\lambda$ , designated by  $\{\lambda_t\}$ , for which  $\lambda_t = 0$  for all  $t < 0$ , and  $\lambda_t = 1$  for all  $t \geq \tau$  ( $\tau \geq 0$ ), the work  $W$  applied to the system is a RV that satisfies

$$\mathbf{E}\{e^{-\beta W}\} = e^{-\beta \Delta F}.$$

By Jensen’s inequality,

$$\mathbf{E}\{e^{-\beta W}\} \geq \exp(-\beta \mathbf{E}\{W\}),$$

which then gives  $\mathbf{E}\{W\} \geq \Delta F$ , for an arbitrary protocol  $\{\lambda_t\}$ . The Gibbs’ inequality is then a special case, where  $\lambda_t$  is given by the unit step function, but it applies regardless of the assumptions of [4]. At the other extreme, when  $\lambda_t$  changes very slowly, corresponding to a reversible process,  $W$  approaches determinism, and then Jensen’s inequality becomes tight. In the limit of an arbitrarily slow process, this yields  $W = \Delta F$ , with no increase in entropy.

### B. Adiabatic Version

Thus far, we discussed an isothermal process, where the change was attributed to the Hamiltonian – a transition from  $\mathcal{E}_0$  to  $\mathcal{E}_1$ . In the special case where the two Hamiltonians are proportional to one another, namely, when  $\mathcal{E}_1(x)/\mathcal{E}_0(x) = \text{const.}$ , independent of  $x$ , one can, of course, still consider it as an isothermal process and refer the change in the Hamiltonian to that of a multiplicative control parameter  $\lambda$ , as before (e.g., the harmonic potential  $\frac{\lambda}{2}x^2$ ). But perhaps even more natural, in this case, is to refer the change to temperature. In this case, there is no external mechanical work, and the change in the internal energy of the system comes solely from heat: We replace a heat bath (large environment) with temperature  $T_0 = 1/(k\beta_0)$  by a heat bath with a higher temperature  $T_1 = 1/(k\beta_1)$ . If we apply the Gibbs’ inequality to this special case, this amounts to

$$\ln Z(\beta_1) \geq \ln Z(\beta_0) + (\beta_0 - \beta_1)\mathbf{E}_0\{\mathcal{E}_0(X)\}$$

which is easily shown (cf. eq. (5)) to be equivalent to

$$\begin{aligned} \Delta \Sigma &\equiv \Sigma(\beta_1) - \Sigma(\beta_0) \\ &\geq \beta_1[\mathbf{E}_1\{\mathcal{E}_0(X)\} - \mathbf{E}_0\{\mathcal{E}_0(X)\}] \equiv \frac{\Delta Q}{kT_1}, \end{aligned} \quad (14)$$

where  $\Sigma(\beta_0)$  and  $\Sigma(\beta_1)$  are the equilibrium entropies (in units of  $k$ ) pertaining to  $\beta_0$  and  $\beta_1$ , respectively, and  $\Delta Q$  is the amount of heat injected into the system, assuming there is no mechanical work. This inequality is a special case of the Clausius theorem (mentioned earlier), which in its general form, asserts that  $\Delta S = k\Delta \Sigma$  is never smaller than  $\int dQ/T$  for any process, with equality in the case of a reversible process. The expression  $\Delta Q/T_1$  is the result of this integral when the heat bath of temperature  $T_0$  is abruptly replaced by one with temperature  $T_1$ . An alternative interpretation of this inequality is, again, as an instance of the second law: The entropy of our system increases by  $\Delta S$  and the entropy of the (new) heat bath decreases by  $\Delta Q/T_1$ , thus the net entropy change of the combined system (which is assumed isolated),  $\Delta S - \Delta Q/T_1$ , must be non-negative.

In the information-theoretic context, the relevant situation is one where  $P(x|u, v) = P(x|u)$  and  $P(x|v) = \int du P(x|u, v)P(u|v)$  can be represented as Boltzmann distributions with the same Hamiltonian, but which may differ in temperature and possibly in shifts (by  $u$  or  $v$ ). I.e.,

$$P(x|u, v) = P(x|u) = \frac{e^{-\beta_0 \mathcal{E}(x-u)}}{Z(\beta_0)};$$

$$P(x|v) = \frac{e^{-\beta_1 \mathcal{E}(x-v)}}{Z(\beta_1)} \quad \beta_1 < \beta_0$$

This turns out to be the case when  $X$ ,  $U$  and  $V$  are related by a cascade of two additive channels of the same family (e.g., a degraded broadcast channel), one from  $V$  to  $U$  and the other from  $U$  to  $X$  (or in the other direction). Two classical examples are those when both channels are binary and symmetric (with possibly two different crossover parameters), and when they are both Gaussian (with possibly different noise variances). Other examples of these properties could pertain to any choice of an infinitely divisible random variable as a noise model in both channels, like the Poisson RV, the binomial RV, and so on.

Using again the Gibbs’ inequality as before, we now get, for given  $u$  and  $v$ :

$$\begin{aligned} \ln Z(\beta_1) &\geq \ln Z(\beta_0) + \beta_0 \mathbf{E}_{\beta_0, u, v}\{\mathcal{E}(X - u)\} - \\ &\quad \beta_1 \mathbf{E}_{\beta_0, u, v}\mathcal{E}(X - v), \end{aligned} \quad (15)$$

where  $\mathbf{E}_{\beta_0, u, v}$  denotes expectation w.r.t.  $P(x|u, v)$  as defined above. Now, assuming shift-invariance of integrals over  $x$  (as is the case in the BSC and Gaussian examples mentioned above),  $\mathbf{E}_{\beta_0, u, v}\{\mathcal{E}(X - u)\} = \mathbf{E}_{\beta_0, 0, 0}\{\mathcal{E}(X)\} \triangleq \mathbf{E}_{\beta_0}\{\mathcal{E}(X)\}$ , independently of  $u$  and  $v$ . As for the third term, from the above relation between  $P(x|u, v)$  and  $P(x|v)$ , it is apparent that after averaging  $\mathbf{E}_{\beta_0, u, v}\{\mathcal{E}(X - v)\}$  (which is independent of  $u$ ) w.r.t.  $P(u|v)$ , it becomes  $\mathbf{E}_{\beta_1, v}\{\mathcal{E}(X - v)\} = \mathbf{E}_{\beta_1, 0}\{\mathcal{E}(X)\} \triangleq \mathbf{E}_{\beta_1}\{\mathcal{E}(X)\}$ . Thus, we get

$$\begin{aligned} \Sigma(\beta_1) &\equiv \ln Z(\beta_1) + \beta_1 \mathbf{E}_{\beta_1}\mathcal{E}(X) \\ &\geq \ln Z(\beta_0) + \beta_0 \mathbf{E}_{\beta_0}\{\mathcal{E}(X)\} \equiv \Sigma(\beta_0) \end{aligned} \quad (16)$$

This is then a special case of the inequality  $\Delta \Sigma \geq \Delta Q/(kT_1)$ , where  $\Delta Q = 0$ , namely, an *adiabatic process*, and then  $\Delta \Sigma \geq$

0, or  $\Delta S \geq 0$ . The information loss due to the DPT again has the physical interpretation of entropy increase, but this time it is purely due to temperature increase, rather than the dissipated work that we have seen before.

We end this section with two simple examples, namely, the Gaussian broadcast channel and the binary symmetric broadcast channel. In both examples, we view the mutual information difference, which is the entropy increase, as an integral of temperature, and thereby identify the corresponding heat capacity from the integrand.

*Example 3 – Gaussian degraded broadcast channel:* Consider a Gaussian degraded broadcast channel, i.e., a cascade of two independent additive white Gaussian noise (AWGN) channels, given by:

$$X = U + \frac{N_1}{\sqrt{\beta_0}}$$

and

$$U = V + N_2 \sqrt{\frac{1}{\beta_1} - \frac{1}{\beta_0}}, \quad \beta_1 < \beta_0,$$

where  $N_1$  and  $N_2$  are both zero-mean, unit-variance Gaussian RV's, independent of each other as well as of  $V$ , which in turn has an arbitrary density with  $E\{V^2\} < \infty$ . In this case,

$$\begin{aligned} \Delta S &= I(X; U) - I(X; V) \\ &= h(X|V) - h(X|U) \\ &= \frac{1}{2} \ln \left( \frac{2\pi e}{\beta_1} \right) - \frac{1}{2} \ln \left( \frac{2\pi e}{\beta_0} \right) \\ &= \frac{1}{2} \ln \frac{\beta_0}{\beta_1} \\ &= \frac{1}{2} \int_{\beta_1}^{\beta_0} \frac{d\beta}{\beta} \\ &= \int_{T_0}^{T_1} \frac{dT}{2T}, \end{aligned}$$

where in the last step, we changed the integration variable from  $\beta$  to  $T = 1/(\beta k)$ . As mentioned in Section II, in the thermodynamical definition, an entropy change is given by

$$\Delta S = k \Delta \Sigma = \int \frac{dQ}{T}$$

along a reversible process, but  $dQ = C(T)dT$ , where  $C(T)$  is the heat capacity (at constant volume), and so,

$$\Delta S = \int_{T_0}^{T_1} \frac{dT C(T)}{T}.$$

Thus, we identify the heat capacity pertaining the Gaussian broadcast channel as  $C(T) = k/2$ , independently of  $T$ , which is exactly the same as the heat capacity (per degree of freedom) of an ideal gas without gravitation (cf. e.g., [2, Sect. 4.4, p. 106]).<sup>6</sup> This is because the Gaussian channel, considered in

<sup>6</sup>The classical heat capacity per particle of an ideal gas at constant volume is actually  $C = 3k/2$ . The extra factor of 3 accounts for three degrees of freedom per particle, owing to the three dimensions of space.

this example, induces a quadratic Hamiltonian, just like that of the ideal gas (cf. the first term of eq. (1)).

It is instructive to examine also the case where the directions of the additive channels are reversed, or equivalently, to examine the difference  $I(U; V) - I(X; V)$  for the original channels defined above. Adopting the latter definition, and using the main results of [8], concerning the relation between  $I(U; V)$  and the minimum mean square error (MMSE),  $\text{mmse}(V|U)$ , in estimating  $V$  from  $U$  (and of course, similar relations for  $X$  and  $V$ ), we find that the increase in entropy is:

$$\begin{aligned} I(U; V) - I(X; V) &= \frac{1}{2} \int_0^{\beta_0} \text{mmse} \left( V \middle| V + \frac{N}{\sqrt{\beta}} \right) d\beta - \\ &\quad \frac{1}{2} \int_0^{\beta_1} \text{mmse} \left( V \middle| V + \frac{N}{\sqrt{\beta}} \right) d\beta \\ &= \frac{1}{2} \int_{\beta_1}^{\beta_0} \text{mmse} \left( V \middle| V + \frac{N}{\sqrt{\beta}} \right) d\beta \\ &= \int_{T_0}^{T_1} \frac{\text{mmse}(V|V + N\sqrt{kT})}{2kT^2} dT \quad (17) \end{aligned}$$

where  $N \sim \mathcal{N}(0, 1)$ . Thus, now we identify the heat capacity as

$$C(T) = \frac{\text{mmse}(V|V + N\sqrt{kT})}{2T}.$$

If, in addition,  $V$  is zero-mean, Gaussian, with variance  $\sigma_V^2$ , then

$$C(T) = \frac{k\sigma_V^2}{2(\sigma_V^2 + kT)}.$$

In the high-SNR regime ( $\sigma_V^2 \gg kT$ ), this gives  $C(T) \approx k/2$ , which is the same as before.

*Example 4 – binary symmetric degraded broadcast channel:* In a similar manner, consider the binary symmetric degraded broadcast channel, that is, a cascade of two binary symmetric channels,

$$X = U \oplus N_1; \quad U = V \oplus N_2,$$

where all RV's are binary  $\{0, 1\}$ ,  $\oplus$  designates addition modulo 2, and  $(X, N_1, N_2)$  are independent. In this case, the Hamiltonian is  $\mathcal{E}(x) = E_0 x$ ,  $x \in \{0, 1\}$ , where  $E_0$  is a constant (having the units of energy), and we have

$$\Pr\{N_1 = x\} = \frac{e^{-\beta_0 E_0 x}}{1 + e^{-\beta_0 E_0 x}} \quad x \in \{0, 1\}$$

and similarly,

$$\Pr\{N_1 \oplus N_2 = x\} = \frac{e^{-\beta_1 E_0 x}}{1 + e^{-\beta_1 E_0 x}}.$$

Here the heat capacity can be shown to be given by:

$$C(T) = \frac{E_0^2}{kT^2} \cdot \frac{e^{-E_0/(kT)}}{[1 + e^{-E_0/(kT)}]^2},$$

which agrees with the heat capacity of a system of two-level non-interacting particles (see, e.g. [2, Sect. 4.3, eq. (4.22)]).



## V. ERROR EXPONENTS AND REVERSIBLE PROCESSES

We mentioned the notion of a reversible process, and the question that might naturally arise, at this point, concerns the information-theoretic analogue of this term. This seems to have a direct relationship to the behavior of error exponents of hypothesis testing and the Neyman–Pearson lemma: Let  $P_0(x)$  and  $P_1(x)$  be two probability distributions (or densities, in the continuous case) of a random variable  $X$ , taking values in an alphabet  $\mathcal{X}$ . Given an observation  $x \in \mathcal{X}$ , one would like to decide whether it emerged from  $P_0$  or  $P_1$ . A decision rule is a partition of  $\mathcal{X}$  into two complementary regions  $\mathcal{X}_0$  and  $\mathcal{X}_1$ , such that whenever  $X \in \mathcal{X}_i$  one decides in favor of the hypothesis that  $X$  has emerged from  $P_i$ ,  $i = 0, 1$ . Associated with any decision rule, there are two kinds of error probabilities:  $P_0(\mathcal{X}_1)$  is the probability of deciding in favor of  $P_1$  while  $x$  has actually generated by  $P_0$ , and  $P_1(\mathcal{X}_0)$  is the opposite kind of error. The Neyman–Pearson problem is about the quest for the optimum decision rule in the sense of minimizing  $P_1(\mathcal{X}_0)$  subject to the constraint that  $P_0(\mathcal{X}_1) \leq \alpha$  for a prescribed constant  $\alpha \in [0, 1]$ . The Neyman–Pearson lemma asserts that the optimum decision rule, in this sense, is given by the likelihood ratio test (LRT)  $\mathcal{X}_0^* = (\mathcal{X}_1^*)^c = \{x : P_0(x)/P_1(x) \geq \mu\}$ , where the threshold  $\mu = \mu(\alpha)$  is tuned so as to meet the constraint  $P_0(\mathcal{X}_1) \leq \alpha$  with equality (assuming that this is possible).

Assume now that instead of one observation  $x$ , we have a vector  $\mathbf{x}$  of  $n$  i.i.d. observations  $(x_1, \dots, x_n)$ , emerging either all from  $P_0$ , or all from  $P_1$ . In this case, the error probabilities of the two kinds, pertaining to the LRT,  $P_0(\mathbf{x})/P_1(\mathbf{x}) \geq \alpha$ , can decay asymptotically exponentially, provided that  $\alpha = \alpha_n$  is chosen to decay exponentially with  $n$  (though not too fast), and the asymptotic exponents,  $e_0 = \lim_{n \rightarrow \infty} [-\frac{1}{n} \ln P_0(\mathcal{X}_1^*)]$  and  $e_1 = \lim_{n \rightarrow \infty} [-\frac{1}{n} \ln P_1(\mathcal{X}_0^*)]$  can be easily found (e.g., by using the method of types) to be

$$e_i(\lambda) = D(P_\lambda \| P_i) = \sum_{x \in \mathcal{X}} P_\lambda(x) \ln \frac{P_\lambda(x)}{P_i(x)}; \quad i = 0, 1$$

where

$$P_\lambda(x) = \frac{P_0^{1-\lambda}(x) P_1^\lambda(x)}{Z(\lambda)}$$

with

$$Z(\lambda) = \sum_{x \in \mathcal{X}} P_0^{1-\lambda}(x) P_1^\lambda(x)$$

and  $\lambda \in [0, 1]$  being a parameter (depending on  $\mu$ ) that controls the tradeoff between the error exponents of the two kinds: For  $\lambda = 0$ ,  $e_0(0) = 0$  and  $e_1(0) = D(P_0 \| P_1)$ . As  $\lambda$  grows from 0 to 1,  $e_0(\lambda)$  increases and  $e_1(\lambda)$  decreases. Finally, for  $\lambda = 1$ ,  $e_0(1) = D(P_1 \| P_0)$  and  $e_1(1) = 0$ .

From the physics point of view, given  $P_0$  and  $P_1$ , let us define the Hamiltonians,  $\mathcal{E}_0(x) = -\ln P_0(x)$  and  $\mathcal{E}_1(x) = -\ln P_1(x)$ , and let the inverse temperature be set to  $\beta = 1$ . Let  $P_\lambda(x)$  be defined as above, which can be referred to as the Boltzmann distribution with Hamiltonian  $\mathcal{E}_\lambda(x) = (1 - \lambda)\mathcal{E}_0(x) + \lambda\mathcal{E}_1(x)$  and  $\beta = 1$ . Let  $\lambda_t$ ,  $t \in [0, \tau]$ , be a function that starts from  $\lambda_0 = 0$  and ends at  $\lambda_\tau = 1$ . Now, assuming

that the conditions for the Jarzynsky equality hold in this case, the average work along the process, which is

$$\mathbf{E}\{W\} = \int_0^\tau d\lambda_t \cdot \mathbf{E}_{\lambda_t}\{\mathcal{E}_1(X) - \mathcal{E}_0(X)\},$$

cannot be smaller than  $\Delta F$ , which in this case vanishes. As said, equality  $\mathbf{E}\{W\} = \Delta F \equiv 0$  is attained for a reversible process.

Indeed, these relations can easily be seen to hold here and also be related to the error exponents of Neyman–Pearson testing, and even from a direct derivation, without recourse to physical considerations: Considering the Hamiltonians  $\mathcal{E}_i(x) = -\ln P_i(x)$ ,  $i = 0, 1$ , as mentioned above, we have:

$$\begin{aligned} \mathbf{E}\{W\} &= \int_0^\tau d\lambda_t \mathbf{E}_{\lambda_t} \ln \frac{P_0(X)}{P_1(X)} \\ &= \int_0^\tau d\lambda_t \sum_{x \in \mathcal{X}} P_{\lambda_t}(x) \ln \frac{P_0(x)}{P_1(x)} \\ &= \int_0^\tau d\lambda_t [D(P_{\lambda_t} \| P_1) - D(P_{\lambda_t} \| P_0)] \\ &= \int_0^\tau d\lambda_t [e_1(\lambda_t) - e_0(\lambda_t)] \end{aligned} \quad (18)$$

On the other hand, we can also rewrite the second line of the last chain of equalities as:

$$\mathbf{E}\{W\} = - \int_0^\tau d\lambda_t \cdot \left[ \frac{\partial \ln Z(\lambda)}{\partial \lambda} \right]_{\lambda=\lambda_t}. \quad (19)$$

Now, if  $\{\lambda_t\}$  is everywhere differentiable (which is analogue to a reversible process), this amounts to

$$\begin{aligned} \mathbf{E}\{W\} &= - \int_0^\tau dt \dot{\lambda}_t \cdot \left[ \frac{\partial \ln Z(\lambda)}{\partial \lambda} \right]_{\lambda=\lambda_t} \\ &= - \int_0^\tau dt \cdot \frac{d \ln Z(\lambda_t)}{dt} \\ &= \ln Z(\lambda_0) - \ln Z(\lambda_\tau) \\ &= \ln Z(0) - \ln Z(1) \\ &= \ln 1 - \ln 1 = 0. \end{aligned} \quad (20)$$

If, on the other hand,  $\{\lambda_t\}$  contains jump–discontinuities, then every such jump, say, from  $\lambda_1$  to  $\lambda_2$ , contributes to the integral a term of the form

$$d\lambda_t \cdot \left[ \frac{\partial \ln Z(\lambda)}{\partial \lambda} \right]_{\lambda=\lambda_t} = (\lambda_2 - \lambda_1) \cdot \left[ \frac{\partial \ln Z(\lambda)}{\partial \lambda} \right]_{\lambda=\lambda_1},$$

which is smaller than  $\ln Z(\lambda_2) - \ln Z(\lambda_1)$ , due to the convexity of the function  $\ln Z(\lambda)$ . Consequently, because of the minus sign, each such discontinuity increases  $\mathbf{E}\{W\}$  above zero. Thus, we indeed see that,

$$\int_0^\tau d\lambda_t e_1(\lambda_t) \geq \int_0^\tau d\lambda_t e_0(\lambda_t)$$

with equality in the differentiable (reversible) case. This in turn means that in this case,

$$\int_0^\tau dt \dot{\lambda}_t e_0(\lambda_t) = \int_0^\tau dt \dot{\lambda}_t e_1(\lambda_t).$$

The left- (resp. right-) hand side is simply  $\int_0^1 d\lambda e_0(\lambda)$  (resp.  $\int_0^1 d\lambda e_1(\lambda)$ ) which means that the areas under the graphs of the functions  $e_0$  and  $e_1$  are always the same.

While these integral relations between the error exponent functions have actually been derived without recourse to any physical considerations, it is the physical point of view that gives the trigger to point out these relations.

#### ACKNOWLEDGEMENT

The author thanks Shlomo Shamai for the suggesting the problem, as well as Yariv Kafri and Dov Levine for useful discussions and for bringing ref. [4] to his attention.

#### REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2006.
- [2] M. Kardar, *Statistical Physics of Particles*, Cambridge University Press, 2007.
- [3] T. S. Komatsu, N. Nakagawa, S.-I. Sasa, and H. Tasaki, “Representation of nonequilibrium steady states in large mechanical systems,” *J. Stat. Phys.*, vol. 134, pp. 401–423, 2009.
- [4] P. Pradhan, Y. Kafri, and D. Levine, “Non-equilibrium fluctuation theorems in the presence of local heating,” arXiv:0712.0339v2 [cond-mat.stat-mech] 3 Apr 2008.
- [5] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, “Dissipation: the phase-space perspective,” *Phys. Rev. Lett.*, vol. 98, 080602, 2007.
- [6] J. Horowitz and C. Jarzynsky, “An illustrative example of the relationship between dissipation and relative entropy,” arXiv:0901.0576v1 [cond-mat.stat-mech] 5 Jan 2009.
- [7] R. J. Baxter, *Exactly Solvable Models in Statistical Mechanics*, Academic Press, 1982.
- [8] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.