Physics of the Shannon Limits

Neri Merhav

Department of Electrical Engineering

Technion – Israel Institute of Technonolgy

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, to the second law of thermodynamics, and at least indirectly, also to 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.

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 $Q = \{Q(x), \ x \in \mathcal{X}\}$, $D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log[P(x)/Q(x)]$ cannot 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, $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$:

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

is completely equivalent[1] 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:[2] 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 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.

---

[1] The information inequality is obtained from the LSI when $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ 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 $Q_i = b_i / \sum_j b_j$.

[2] Another relevant law, at least indirectly, is energy conservation, as we rely on the Boltzmann–Gibbs distribution, which is derived from the energy conservation law.

II. STATISTICAL 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 *micorstates*, 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, 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 $\boldsymbol{x}$, there is an energy function, a.k.a. the *Hamiltonian*, $\mathcal{E}(\boldsymbol{x})$. For example, in the case of the ideal gas, $x_i = (\boldsymbol{p}_i, \boldsymbol{r}_i)$, where $\boldsymbol{p}_i$ and $\boldsymbol{r}_i$, both three dimensional vectors, are the momentum and the position of particle number i, respectively, and $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^n \frac{\|p_i\|^2}{2m} + mgz_i$, where m is the mass of each particle, g is the gravitation constant, and z_i is the height – one of the components of $\boldsymbol{r}_i$. One of the most fundamental results in statistical physics is that, when a system lies in thermal equilibrium with the environment (heat bath), the probability of finding the system at state $\boldsymbol{x}$ is given by the *Boltzmann–Gibbs* distribution

$$P(\boldsymbol{x}) = \frac{e^{-\beta \mathcal{E}(\boldsymbol{x})}}{Z(\beta)} \tag{1}$$

where $\beta = 1/(kT)$, k being Boltzmann's constant and T being temperature, and $Z(\beta)$ is the normalization constant, called the *partition function*, and given by $Z(\beta) = \sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}(\boldsymbol{x})}$ or $Z(\beta) = \int \mathrm{d}\boldsymbol{x}\, e^{-\beta \mathcal{E}(\boldsymbol{x})}$, depending on whether $\boldsymbol{x}$ is discrete or continuous. The partition function is is the key to calculate many important physical quantities. For example, the average energy w.r.t. (1) is

$$\bar{E} = \boldsymbol{E}\{\mathcal{E}(\boldsymbol{X})\} = -\frac{\mathrm{d}\ln Z(\beta)}{\mathrm{d}\beta}, \tag{2}$$

the entropy (in units of k) pertaining to (1) is

$$S(\beta) = \boldsymbol{E}\{-\ln P(\boldsymbol{X})\} = \ln Z(\beta) + \beta \cdot \bar{E}, \tag{3}$$

and the *free energy* is given by

$$F(\beta) = -\frac{\ln Z(\beta)}{\beta}. \tag{4}$$

The free energy is of crucial importance in thermodynamics and statistical physics: The difference, ΔF, between the free energies associated with two equilibirium 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 a fixed temperature (isothermal) process, and this minimum is attained when the process is *quasi–static* or *reversible*, i.e., slow enough such that, along the way, the system is kept always very close to equilibrium.[3] Equivalently, the negative free–energy difference, $-\Delta F$, is

[3]This fact is also known as the *minimum work principle*.

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 asserts that the entropy of an isolated system cannot decrease.

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 $U^N = (U_1, \ldots U_N)$ compressed into a bitstream $X^n = (X_1, \ldots, X_n)$ from which the decoder generates a reproduction $V^N = (V_1, \ldots, V_N)$ with distortion $\sum_{i=1}^N \boldsymbol{E}\{d(U_i, V_i)\} \leq ND$. 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 $R(D)$. The case of lossless compression is obtained as a special case where $D = 0$.

Channel coding under bit error probability: Let $U^N = (U_1, \ldots U_N)$ 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 $E \stackrel{\triangle}{=} \mathcal{I}\{V^N \neq U^N\}$ ($\mathcal{I}$ 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 reconsdtruction 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 DATA PROCESSING THEOREM

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

A. Isothermal Version of the Data Processing Theorem

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

$$0 \leq D(P_0 \| P_1) = \sum_x P_0(x) \ln \left[\frac{e^{-\beta \mathcal{E}_0(x)}/Z_0}{e^{-\beta \mathcal{E}_1(x)}/Z_1} \right]$$
$$= \ln Z_1 - \ln Z_0 + \beta E_0 \{\mathcal{E}_1(X) - \mathcal{E}_0(X)\} \quad (5)$$

where $E_0\{\cdot\}$ denotes the expectation operator w.r.t. P_0. After a minor algebraic rearrangement, this becomes:

$$E_0\{\mathcal{E}_1(X) - \mathcal{E}_0(X)\} \geq kT \ln Z_0 - kT \ln Z_1$$
$$\equiv F_1 - F_0, \quad (6)$$

where F_i is the free energy pertaining to P_i, $i = 0, 1$ (cf. eq. 4)).

We now offer the following physical interpretation to this inequality: Imagine that a system with Hamiltoinan $\mathcal{E}_0(x)$ is in equilibrium for all $t < 0$,[4] but then at time $t = 0$, the Hamitonian changes *abruptly* from the $\mathcal{E}_0(x)$ to $\mathcal{E}_1(x)$ (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 x at time $t = 0$, additional energy of $W(x) = \mathcal{E}_1(x) - \mathcal{E}_0(x)$ 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, $W(x)$ is a random variable due to the randomness of x. Since this passage between $\mathcal{E}_0$ and $\mathcal{E}_1$ is abrupt, and the microstate x does not change instantaneously, the expectation of $W(X)$ should be taken w.r.t. P_0, and this average is exactly what we have at the left–hand side eq. (6). 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 $E_0\{W(X)\} - \Delta F = kT \cdot D(P_0 \| P_1) \geq 0$ 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.[5] 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.

Now, let us 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

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

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):

$$E_0\{W(X)\} = \sum_x P(x|u, v)[\ln P(x|u) - \ln P(x|v)]$$
$$= H(X|V = v) - H(X|U = u) \quad (7)$$

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

---

[4]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 equilibirium, but this is beyond the scope of this work.

[5]See also [3], [4], [5] and references therein, where the same conclusions are reached from a more general perspective of irrreversible processes, but under certain limiting assumptions on the physical system.

the free energies, we have

$$Z_0 = \sum_x \exp\{-1 \cdot [-\ln P(x|u,v)]\}$$
$$= \sum_x P(x|u,v) = 1 \tag{8}$$

and similarly, $Z_1 = \sum_x P(x|v) = 1$, which means that $F_0 = F_1 = 0$, 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 $T\Delta S = 1 \cdot [H(X|V) - H(X|U)]$, 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 $\mathcal{E}_\lambda(\boldsymbol{x}) = \mathcal{E}_0(\boldsymbol{x}) + \lambda[\mathcal{E}_1(\boldsymbol{x}) - \mathcal{E}_0(\boldsymbol{x})]$ as a linear interpolation between the two extremes $\lambda = 0$ and $\lambda = 1$, pertaining to $\mathcal{E}_0$ and $\mathcal{E}_1$, and then λ can be thought of as a control parameter or a 'force' that influences the system. The Jarzynsky equality (cf. e.g., [3] and references therein) tells that under certain conditions on the system and the environment, and given any protocol for a temporal change in λ, 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 $\boldsymbol{E}\{e^{-\beta W}\} = e^{-\beta \Delta F}$. By Jensen's inequality, $\boldsymbol{E}\{e^{-\beta W}\} \geq \exp(-\beta \boldsymbol{E}\{W\})$, which then gives $\boldsymbol{E}\{W\} \geq \Delta F$, for an arbitrary protocol $\{\lambda_t\}$. The Gibbs' inequality is then a special case, where λ_t is given by the unit step function, but it applies regardless of the assumptions of [3]. At the other extreme, when λ_t changes very slowly, corresponding to a reversible process, W approaches determinism, and then Jensen's inequality becomes tight, which then gives (in the limit) $W = \Delta F$, with no increase in entropy.

B. Adiabatic Version of the Data Processing Theorem

So far, we discussed a process with fixed temperature (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, it more natural to refer the change to temperature. In this case, there is no external mechanical work necessarily, and the change in the internal energy of the system comes solely from heat: We replace a heat bath (large environement) 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

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

which is easily shown (cf. eq. (3)) to be equivalent to

$$\Delta S \equiv S(\beta_1) - S(\beta_0) \geq \beta_1[\boldsymbol{E}_1\{\mathcal{E}_0(X)\} - \boldsymbol{E}_0\{\mathcal{E}_0(X)\}] \equiv \frac{\Delta Q}{T_1},$$

where $S(\beta_0)$ and $S(\beta_1)$ are the equilibrium entropies pertaining to β_0 and β_1, respectively, and ΔQ 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*, which asserts that ΔS 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 ΔS 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. 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:

$$\ln Z(\beta_1) \geq \ln Z(\beta_0) + \beta_0 \boldsymbol{E}_{\beta_0,u,v}\{\mathcal{E}(X-u)\} - \beta_1 \boldsymbol{E}_{\beta_0,u,v}\mathcal{E}(X-v), \tag{9}$$

where $\boldsymbol{E}_{\beta_0,u,v}$ 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), $\boldsymbol{E}_{\beta_0,u,v}\{\mathcal{E}(X-u)\} = \boldsymbol{E}_{\beta_0,0,0}\{\mathcal{E}(X)\} \triangleq \boldsymbol{E}_{\beta_0}\{\mathcal{E}(X)\}$, 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 $\boldsymbol{E}_{\beta_0,u,v}\{\mathcal{E}(\boldsymbol{X}-v)\}$ (which is independent of u) w.r.t. $P(u|v)$, it becomes $\boldsymbol{E}_{\beta_1,v}\{\mathcal{E}(X-v)\} = \boldsymbol{E}_{\beta_1,0}\{\mathcal{E}(X)\} \triangleq \boldsymbol{E}_{\beta_1}\{\mathcal{E}(X)\}$. Thus, we get

$$
\begin{aligned}
S(\beta_1) &\equiv \ln Z(\beta_1) + \beta_1 \boldsymbol{E}_{\beta_1} \mathcal{E}(X)\} \\
&\geq \ln Z(\beta_0) + \beta_0 \boldsymbol{E}_{\beta_0}\{\mathcal{E}(X)\} \equiv S(\beta_0) \quad (10)
\end{aligned}
$$

where $S(\beta_0)$ and $S(\beta_1)$ are again the equilibrium entropies before and after the process. This is then a special case of the inequality $\Delta S \geq \Delta Q/T_1$, where $\Delta Q = 0$, namely, an *adiabatic process*, and then $\Delta S \geq 0$. 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.

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 $\boldsymbol{x}$ of n i.i.d. observations $(x_1, \ldots, x_n)$, 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, $P_0(\boldsymbol{x})/P_1(\boldsymbol{x}) \geq \alpha$, can decay asymptotically exponentially, provided that $\alpha = \alpha_n$ is chosen to decay exponentially with n (though not too fast), and the asymptotic exponents, $e_0 = \lim_{n\to\infty}[-\frac{1}{n}\ln P_0(\mathcal{X}_1^*)]$ and $e_1 = \lim_{n\to\infty}[-\frac{1}{n}\ln P_1(\mathcal{X}_0^*)]$ 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 μ) 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 λ 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 λ_t, $t \in [0,\tau]$, be a function that starts from $\lambda_0 = 0$ and ends at $\lambda_\tau = 1$. Then, assuming that the conditions for the Jarzynski equality hold in this case, the average work along the process, which is

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

cannot be smaller than ΔF, which in this case vanishes. As said, equality $\boldsymbol{E}\{W\} = \Delta F \equiv 0$ is attained for a reversible process. However, this average work can also easily seen to be related to the error exponents of Neyman–Pearson testing:

$$
\boldsymbol{E}\{W\} = \int_0^\tau \mathrm{d}\lambda_t[e_0(\lambda_t) - e_1(\lambda_t)].
$$

Thus, we see that $\int_0^\tau \mathrm{d}\lambda_t e_0(\lambda_t) \geq \int_0^\tau \mathrm{d}\lambda_t e_1(\lambda_t)$ with equality in the reversible case. Indeed, the last inequality can be also shown to hold using a direct derivation, and the equality is easily shown to hold whenever λ_t is differentiable for every $t \in [0,\tau]$, in which case, it becomes: $\int_0^\tau \mathrm{d}t\dot\lambda_t e_0(\lambda_t) = \int_0^\tau \mathrm{d}t\dot\lambda_t e_1(\lambda_t)$. The left– (resp. right–) hand side is simply $\int_0^1 \mathrm{d}\lambda e_0(\lambda)$ (resp. $\int_0^1 \mathrm{d}\lambda e_1(\lambda)$) 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 could have been derived without recourse to any physical considerations, it is the physical point of view that gives the trigger to point out these relations.

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] 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.
[4] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, "Dissipation: the phase–space perspective," *Phys. Rev. Lett.*, vol. 98, 080602, 2007.
[5] J. Horowitz and C. Jarzynski, "An illustrative example of the relationship between dissipation and relative entropy," arXiv:0901.0576v1 [cond-mat.stat-mech] 5 Jan 2009.