On the Physics of Rate–Distortion Theory

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

***Abstract*—We revisit and extend the physical interpretation recently given to a certain identity between large–deviations rate–functions, as well as applications of this identity to rate–distortion theory, as an instance of thermal equilibrium between several physical systems that are brought into contact. Our new interpretation, of mechanical equilibrium between these systems, is shown to have several advantages. This physical point of view also provides a trigger to the development of certain new alternative representations of the rate–distortion function and channel capacity.**

I. Introduction

In [5], an identity between two forms of a certain large deviations rate function (RF) was established, with applications in information theory. Inspired by a few earlier works (cf. e.g., [4], [8]), this identity was interpreted as *thermal equilibrium* between several physical systems in contact. In particular, the parameter that undergoes optimization of the Chernoff bound, i.e., the *Chernoff parameter*, plays the role of the *inverse temperature*. The corresponding RF is then identified with the entropy of the system.

While this physical interpretation is reasonable, it turns out that it leaves some room for improvement, and we mention here just two points. The first is that this interpretation is not generalizable to RF's of combinations of more than one rare event, where the number of Chernoff parameters is as the number of events. The reason is that in physics, there is one temperature parameter only. The other point is the following (more details will follow in Section 2): while the log–moment generating function, pertaining to the RF, naturally includes weighting, its physical analogue, which is the *partition function*, does not. If these weights are subjected to optimization, they may depend on the Chernoff parameter, i.e., on the temperature and the resulting expression can no longer really be viewed as a partition function.

We propose to interpret the above–mentioned identity of RF's as *mechanical equilibrium* (i.e., balance between mechanical forces), rather than thermal equilibrium, and then the Chernoff parameter plays the physical role of an external *force* applied to the system. In this paradigm, the RF has a natural interpretation as the *Helmholtz free energy* (HFE) of the system, rather than as entropy. Accordingly, since the rate–distortion function (RDF), and similarly, also channel capacity, can be thought of as RF's, they can also be interpreted as HFE's.

This interpretation has several advantages. First, it is consistent with the analogy between the HFE in physics and the Kullback–Leibler divergence in information theory (see, e.g., [1]), which plays a role as a RF when the large deviations analysis is approached by the method of types. Second, it is free of the limitations mentioned in the previous paragraph, as we will later. Third, it serves as a trigger to develop certain new representations of the RDF, and analogously, the channel capacity.

Since the RDF can be thought of as HFE, one of the representations of the RDF expresses it as (the minimum achievable) mechanical work carried out by the aforementioned force, along a 'distance' that is measured in terms of the distortion. Another representation is as an integral that involves the minimum mean square error (MMSE) in estimating the distortion given the source symbol, according to a certain joint distribution. The latter representation suggests a new route to upper and lower bounds on the RDF and channel capacity, using the plethora of bounds on MMSE, available from estimation theory. We have not explored these directions, however, in the framework of this work.

An additional byproduct of the this perspective is the following: Given a source distribution and a distortion measure, we can describe a physical system that emulates the rate–distortion problem in the following manner: When no force is applied to the system, its total length is $n\Delta_0$, where n is the number of particles in the system (and also the block length in the rate–distortion problem), and Δ_0 is the distortion corresponding to zero coding rate. If one applies to the system a contracting force, that increases from zero to some final value λ, such that the length shrinks to $n\Delta$, where $\Delta < \Delta_0$ is analogous to a prescribed distortion level, then the following two facts hold: (i) An *achievable lower bound* on the total amount of mechanical work carried out by the force in order to shrink the system to length $n\Delta$, is given by $W \geq nkTR_Q(\Delta)$, where k is Boltzmann's constant, T is the temperature, and $R_Q(\Delta)$ is the RDF w.r.t. an input distribution Q. (ii) The final force λ is related to Δ according to $\lambda = kTR'_Q(\Delta)$, where $R'_Q(\cdot)$ is the derivative of $R_Q(\cdot)$. Thus, $R(\Delta)$ is a fundamental limit, not only in information theory, but also in physics.

II. Physics Background and Preliminaries

Consider a physical system with n particles, which can be in a variety of microscopic states ('microstates'), defined by combinations of, e.g., positions, momenta, angular momenta, spins, etc., of all n particles. For each such microstate of the system, which we shall designate by a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, there is an associated energy, given by an Hamiltonian (energy function), $\mathcal{E}(\boldsymbol{x})$. One of the most fundamental results in statistical physics is that in thermal equilibrium, the probability of a microstate $\boldsymbol{x}$ is given by the *Boltzmann–Gibbs* (BG) distribution

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

where $\beta = 1/(kT)$, T being temperature, k being Boltzmann's constant, and $Z_n(\beta)$ is the *partition function*, given by $Z_n(\beta) = \sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})}$. The partition function is a key quantity from which many macroscopic physical quantities can be derived, for example, the HFE is $-\frac{1}{\beta}\ln Z_n(\beta)$, the average internal energy (i.e., the expectation of $\mathcal{E}(\boldsymbol{x})$ where $\boldsymbol{x}$ drawn is according (1)) is given by the negative derivative of $\ln Z_n(\beta)$, etc. One way to obtain eq. (1), is as the maximum entropy distribution under an energy constraint, where β plays the role of a Lagrange multiplier that controls this energy.

Under certain assumptions on the Hamiltonian, the following relations are well–known to hold (see, e.g., [3],[7]): Defining the per–particle entropy, $S(E)$, associated with per–particle energy $E = \mathcal{E}(\boldsymbol{x})/n$, as $\lim_{n\to\infty}[\ln\Omega(E)]/n$, (provided that the limit exists), where $\Omega(E)$ is the number of microstates $\{\boldsymbol{x}\}$ with energy level $\mathcal{E}(\boldsymbol{x}) = nE$, then as in the method of types, one can evaluate $Z_n(\beta)$ defined above, as $Z_n(\beta) = \sum_E \Omega(E)e^{-\beta E}$ which is of the exponential order of $\exp\{n\max_E[S(E)-\beta E]\}$. Defining $\phi(\beta) = \lim_{n\to\infty}\frac{\ln Z_n(\beta)}{n}$, and the HFE per–particle as $F(\beta) = -\frac{\phi(\beta)}{\beta}$, we obtain the Legendre transform (LT) relation $\phi(\beta) = \max_E[S(E)-\beta E]$, where here $E = E(\beta)$ is the maximizer of $[S(E)-\beta E]$. For a given β, the BG distribution has a sharp peak (for large n) at $E(\beta)$. Assuming that $S(\cdot)$ is concave (as is normally the case), the above LT relation can be inverted to $S(E) = \min_{\beta\ge 0}[\beta E + \phi(\beta)]$, and both relations can be identified with the thermodynamical definition of the HFE as $F = E - TS$. In the latter relation, the minimizing $\beta = \beta(E)$ (the inverse function of $E(\beta)$) is the equilibrium inverse temperature associated with E. The second law of thermodynamics asserts that in an isolated system (which does not exchange energy with its environment), the entropy cannot decrease, and hence in equilibrium, it reaches its maximum. When the system is allowed to exchange heat with the environment, this maximum entropy principle is replaced by the *minimum free energy* principle: The HFE cannot increase, and it reaches its minimum in equilibrium.

When the Hamiltonian is additive, i.e., $\mathcal{E}(\boldsymbol{x}) = \sum_i \mathcal{E}(x_i)$, then $P(\boldsymbol{x})$ has a product form, and then the above mentioned physical quantities per particle can be extracted from $n = 1$. In this case, the LT from $\phi(\beta)$ to $S(E)$, is similar to the LT that defines the RF pertaining to the probability of the event $\sum_{i=1}^n \mathcal{E}(x_i) \le nE$, thus the parameter of the Chernoff bound plays the role of inverse temperature in the corresponding physical system.

Another look at this correspondence between RF's and thermal equilibrium is this: If P is the above mentioned BG distribution and Q is an arbitrary distribution on $\boldsymbol{x}$, then (see [1]), then $D(Q\|P) = \beta(F_Q - F_P)$, where F_P and F_Q are, respectively, the HFE's pertaining to P and Q. The RF pertaining to an event is given by the minimum divergence under the constraints corresponding to this event, and so, it is equivalent to minimum free energy, i.e., thermal equilibrium by the second law.

Consider next a system as before, except that now the Hamiltonian is shifted by a quantity proportional a parameter λ, i.e., $\mathcal{E}(\boldsymbol{x},\boldsymbol{y}) = \mathcal{E}_0(\boldsymbol{x}) - \lambda\cdot\sum_{i=1}^n y_i$, where we have changed the notation of the (original) Hamiltonian to $\mathcal{E}_0(\boldsymbol{x})$, and where $\{y_i\}$ are some additional variables of the microstate. These new variables may either be dependent or independent of the original microstate variables $\{x_i\}$ The parameter λ is an external control parameter, i.e., a *driving force* that acts on the system via $\{y_i\}$.

Consider the partition function $\tilde{Z}_n(\beta,\lambda) = \sum_{\boldsymbol{x},\boldsymbol{y}} e^{-\beta[\mathcal{E}_0(\boldsymbol{x})-\lambda\sum_i y_i]}$. The *Gibbs free energy* (GFE) per particle is defined as $G_n(\beta,\lambda) = -\frac{1}{n}kT\ln\tilde{Z}_n(\beta,\lambda)$ and the asymptotic GFE per particle is $G(\beta,\lambda) = \lim_{n\to\infty} G_n(\beta,\lambda)$. To find the relation between between the HFE and the GFE, let $\Omega(E,Y) \sim e^{nS(E,Y)}$ denote the number of microstates $\{(\boldsymbol{x},\boldsymbol{y})\}$ for which $\sum_i \mathcal{E}_0(x_i) = nE$ and $\sum_i y_i = nY$. Then, defining $Z_n(\beta,Y) = \sum_{\{(\boldsymbol{x},\boldsymbol{y}):\ \sum_i y_i = nY\}} e^{-\beta\mathcal{E}_0(\boldsymbol{x})}$, the normalized HFE, $F_n(\beta,Y) = -\frac{1}{n}kT\ln Z_n(\beta,Y)$, and the corresponding asymptotic normalized HFE, $F(\beta,Y) = \lim_{n\to\infty} F_n(\beta,Y)$, we have (see [6] for the detailed derivation): $e^{-\beta nG_n(\beta,\lambda)} \doteq \exp\{n\beta\cdot\max_Y[\lambda Y - F(\beta,Y)\}$, where $\doteq$ denotes asymptotic equivalence in the exponential scale. This results in the LT relation $G(\beta,\lambda) = \min_Y[F(\beta,Y) - \lambda Y]$. Assuming that $F(\beta,Y)$ is convex in Y, the inverse LT is $F(\beta,Y) = \max_\lambda[G(\beta,\lambda) + \lambda Y]$, which yields (cf. [6]):

$$\begin{aligned} F(\beta,Y) &= kT\cdot\max_s\Bigg[sY - \\ &\qquad \lim_{n\to\infty}\frac{1}{n}\ln\left(\sum_{\boldsymbol{x},\boldsymbol{y}} e^{-\beta\mathcal{E}_0(\boldsymbol{x})}\cdot e^{s\sum_i y_i}\right)\Bigg] \end{aligned} \tag{2}$$

where we changed the optimization variable λ to $s = \beta\lambda$ for fixed β. Since s is proportional to λ, and λ designates force, we will refer to s also as 'force'. We will get back to eq. (2) soon.

We now proceed to provide a brief summary of [5]. As mentioned, the LT relation $S(E) = \min_{\beta\ge 0}[\beta E + \phi(\beta)]$ is similar to the RF of the event $\{\sum_i \mathcal{E}(x_i) \le nE\}$ for i.i.d. RV's $\{x_i\}$, governed by a given distribution P. The difference is that in the latter, $\ln\sum_x P(x)e^{-\beta\mathcal{E}(x)}$, that undergoes the LT, contains weighting by the probabilities $\{P(x)\}$, unlike the log–partition $\ln\sum_x e^{-\beta\mathcal{E}(x)}$, which does not. In [5] it was proposed to interpret $\{P(x)\}$ as being proportional to a factor of the multiplicity of states $\{x\}$ having the same $\mathcal{E}(x)$, i.e., as the *degeneracy*.

When considering applications of large deviations theory to information theory, one can view the RDF (and channel capacity) as the RF of the event $\{\sum_{i=1}^n d(x_i,\hat{x}_i) \le n\Delta\}$, where $\boldsymbol{x} = (x_1,\ldots,x_n)$ is a given typical source sequence and $\{\hat{x}_i\}$ are i.i.d. RV's drawn by a certain random coding distribution Q. As was observed in [5], there are two ways to express the large deviations RF of this event, which is also the RDF, $R_Q(\Delta)$, for the given Q: The first is by considering all distortion variables $\{d(x_i, \hat{x}_i)\}$ together, on the same footing, resulting in the expression

$$I(\Delta) = -\min_{\beta\geq 0}\left[\beta\Delta + \sum_x P(x)\ln\sum_{\hat{x}} Q(\hat{x})e^{-\beta d(x,\hat{x})}\right].$$

The second way is to separate the distortion contributions, $\{\Delta_x\}$, allocated to the various source letters $\{x\}$, which results in

$$I(\Delta) = -\max\sum_x P(x)\min_{\beta_x\geq 0}\left[\beta\Delta_x + \ln\sum_{\hat{x}} Q(\hat{x})e^{-\beta_x d(x,\hat{x})}\right]$$

where the maximum is subject to the constraint $\sum_x P(x)\Delta_x \leq \Delta$. The identity between these two expressions means that the outer maximum in the second expression is achieved when $\{\Delta_x\}$ are such that the minimizing $\{\beta_x\}$ are all the same, namely, thermal equilibrium between all subsystems indexed by x. Once again, $\{Q(\hat{x})\}$ can be interpreted as degeneracy, which is fine as long as Q is fixed. However, the real RDF is $R(\Delta) = \min_Q R_Q(\Delta)$, and the optimum Q may depend on β. Thus, Q can no longer be given the meaning of degeneracy, which in physics, has nothing to do with temperature.

Another limitation of interpreting β as temperature, is that it does not extend to two or more rare events. For instance, the RDF $R_Q(\Delta_1, \Delta_2)$, w.r.t. two simultaneous distortion constraints, with distortion measures d_1 and d_2, is given by

$$\begin{aligned} R_Q(\Delta_1,\Delta_2) \quad = \quad & -\min_{\beta_1\geq 0}\min_{\beta_2\geq 0}\left[\beta_1\Delta_1 + \beta_2\Delta_2 + \sum_{x\in\mathcal{X}} P(x)\times\right. \\ & \left.\ln\left(\sum_{\hat{x}} Q(\hat{x})e^{-\beta_1 d_1(x,\hat{x})-\beta_2 d_2(x,\hat{x})}\right)\right]. \quad (3)\end{aligned}$$

But this does not have any apparent physical interpretation because there is only one temperature in physics.

III. Large Deviations and Free Energy

In order to give a physical interpretation to the RF as the LT of the log–moment generating function, we use the LT that relates the HFE to the GFE, $G(\beta, \lambda)$ (cf. eq. (2)), rather than the one that relates the HFE to the entropy, $S(E)$. Thus, the Chernoff variable would be λ (or s) rather than β. Also, considering the temperature as being fixed throughout, we can view $\{Q(\hat{x})\}$ as part of the Hamiltonian $\mathcal{E}_0$, which now may depend on λ. This also allows combinations of two or more large deviations events since one may consider a system that is subjected to more than one force. Specifically, let us first compare the HFE expression (2) to the RF [2] of the simple large deviations event $\{\sum_i y_i \geq nY\}$ w.r.t. some probability distribution P:

$$I(Y) = \max_s\left[sY - \lim_{n\to\infty}\frac{1}{n}\ln\left(\sum_{\boldsymbol{y}} P(\boldsymbol{y})e^{s\sum_i y_i}\right)\right]$$

which in the case where $\{y_i\}$ are i.i.d. ($P(\boldsymbol{y}) = \prod_i P(y_i)$), boils down to $\max_s\left[sY - \ln\sum_y P(y)e^{sy}\right]$. Fixing the temperature T to some $T_0 = 1/(k\beta_0)$, taking $\boldsymbol{y} \equiv \boldsymbol{x}$ and $\mathcal{E}_0(\boldsymbol{x}) \equiv \mathcal{E}_0(\boldsymbol{y}) = -kT_0\ln P(\boldsymbol{y})$, we readily see that $I(Y)$ coincides with $F(\beta_0, Y)$ up to the factor of kT_0, which is immaterial. We observe then that the RF has a natural interpretation as the HFE (in units of kT_0) of a system with Hamiltonian $\mathcal{E}_0(\boldsymbol{y}) = -kT_0\ln P(\boldsymbol{y})$ and temperature T_0.

As said, the Chernoff parameter s has the meaning of a driving force that acts on the displacement variables $\{y_i\}$. For example, in the i.i.d. case, the force s required to shift the expectation of each y_i (and hence also of $\frac{1}{n}\sum_i y_i$) towards Y, which is the solution to the equation $Y = \frac{\partial}{\partial s}\ln\sum_y P(y)e^{sy}$ or equivalently, $Y = \sum_y P(y)\cdot ye^{sy}/[\sum_y P(y)\cdot e^{sy}]$. The LT relation between the log–partition function and $I(Y)$ induces a one–to–one mapping between Y and s which is defined by the above equation.

To emphasize this dependency, we henceforth denote the value of Y, corresponding to a given s, by $\langle y\rangle_s$, which symbolizes the fact that it is the expectation[1] of each y_i, denoted generically by y, w.r.t. the probability distribution $P_s = \{P_s(y)\}$, where $P_s(y) = P(y)e^{sy}/[\sum_{y'} P(y')e^{sy'}]$, i.e., $\langle y\rangle_s = \sum_y P(y)\cdot ye^{sy}/[\sum_y P(y)\cdot e^{sy}] = \frac{\partial}{\partial s}\ln\sum_y P(y)e^{sy}$. On substituting $\langle y\rangle_s$ instead of Y in the expression of $I(Y)$, we the RF as a function of s, i.e., $\hat{I}(s) = s\langle y\rangle_s - \ln\sum_y P(y)e^{sy}$. As shown in [6], $\hat{I}(s)$ can be represented as $\hat{I}(s) = \int_{\langle y\rangle_0}^{\langle y\rangle_s}\hat{s}\cdot\mathrm{d}\langle y\rangle_{\hat{s}}$. Now observe that the integrand is a product of the force, $\hat{s}$, and an infinitesimal displacement that it works upon, $\mathrm{d}\langle y\rangle_{\hat{s}} = \langle y\rangle_{\hat{s}} - \langle y\rangle_{\hat{s}-d\hat{s}}$ In physical terms, $\hat{s}\cdot\mathrm{d}\langle y\rangle_{\hat{s}}$ is therefore an infinitesimal contribution of the average *work* (in units of kT_0) done by the force $\hat{s}$ on the variables $\{y_i\}$. Thus, $\hat{I}(s) = \int\hat{s}\cdot\mathrm{d}\langle y\rangle_{\hat{s}}$ is the total amount of work done by the force $\hat{s}$, as it increases from zero to s during a slow process that allows the system to equilibrate after every infinitesimally small change in $\hat{s}$. In the language of physics, this is a *reversible process*, or a *quasi-static process*. Using the concavity of F as a function of s, it is easy to show that any protocol of changing $\hat{s}$ from 0 to s, in a way that includes abrupt changes in $\hat{s}$, would always yield an amount of work larger than or equal to $\hat{I}(s)$.

For an alternative integral expression, one observes that $\mathrm{d}\langle y\rangle_s/\mathrm{d}s = \langle y^2\rangle_s - \langle y\rangle_s^2 \stackrel{\triangle}{=} \mathrm{Var}_s\{y\}$, namely, the variance of y w.r.t. P_s. Thus, $\hat{I}(s) = \int_0^s\hat{s}\cdot\mathrm{Var}_{\hat{s}}\{y\}\mathrm{d}\hat{s}$ and $\langle y\rangle_s = \langle y\rangle_0 + \int_0^s\mathrm{Var}_{\hat{s}}\{y\}\mathrm{d}\hat{s}$. In the more general context considered here, this is a special case of the fluctuation–dissipation theorem in statistical physics [7, p. 32, eq. (2.44)]. We next discuss a physical example which is directly relevant for the rate–distortion problem.

Example [3, p. 134, Problem 13]: Consider a physical system, modeled as a one–dimensional array of n elements (depicted as small springs in Fig. 1), that are arranged along a straight line. Each element may independently be in one of two states, A or B The state of the i–th element, $i = 1, 2, \ldots, n$, is

[1] In the sequel, we use $\langle\cdot\rangle_s$ to denote other moments of y w.r.t. P_s as well.

labeled $\hat{x}_i \in \{A, B\}$. When an element is at state $\hat{x}$, its length is $y_{\hat{x}}$ and its internal energy is $\epsilon_{\hat{x}}$. A stretching force $\lambda > 0$ (or a contracting force, if $\lambda < 0$) is applied to one edge of the array, whereas the other edge is fixed to a wall. What is the expected (and most probable) total length $L = nY$ of the array at temperature T_0?

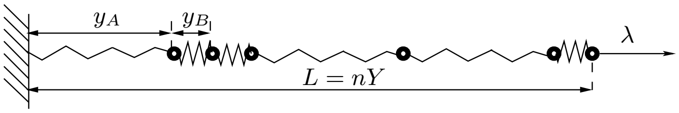

Fig. 1. One–dimensional array of two–state elements.

Since the elements are independent,

$$\begin{aligned}
&\tilde{Z}_n(\beta_0, \lambda) \\
&= \sum_{\hat{x}_1=0}^{1} \cdots \sum_{\hat{x}_n=0}^{1} \exp\left\{-\beta_0\left[\sum_i \epsilon_{\hat{x}_i} - \lambda \sum_i y_{\hat{x}_i}\right]\right\} \\
&= [e^{-\beta_0(\epsilon_A - \lambda y_A)} + e^{-\beta_0(\epsilon_B - \lambda y_B)}]^n, \qquad (4)
\end{aligned}$$

and so, $G_n(\beta_0, \lambda) = -nkT_0 \ln[e^{-\beta_0(\epsilon_A - \lambda y_A)} + e^{-\beta_0(\epsilon_B - \lambda y_B)}]$. The expected length is

$$\begin{aligned}
nY &= -n \cdot \frac{\partial G_n(\beta_0, \lambda)}{\partial \lambda} \\
&= \frac{n[y_A e^{-\beta_0(\epsilon_A - \lambda y_A)} + y_B e^{-\beta_0(\epsilon_B - \lambda y_B)}]}{e^{-\beta_0(\epsilon_A - \lambda y_A)} + e^{-\beta_0(\epsilon_B - \lambda y_B)}}. \qquad (5)
\end{aligned}$$

In terms of the foregoing discussion, $s = \beta_0 \lambda$ controls the expected length per element which is

$$Y = \langle y \rangle_s = \frac{y_A e^{-\beta_0 \epsilon_A + s y_A} + y_B e^{-\beta_0 \epsilon_B + s y_B}}{e^{-\beta_0 \epsilon_A + s y_A} + e^{-\beta_0 \epsilon_B + s y_B}}.$$

The HFE per element is then

$$F(\beta_0, Y) = -kT_0 \ln\left[e^{-\beta_0 \epsilon_A + s y_A} + e^{-\beta_0 \epsilon_B + s y_B}\right] + kT_0 s Y$$

where s is related to Y according to second to the last equation.

Consider now two arrays as above, labeled by $x \in \{a, b\}$, which consist of two different types of elements. Array x has $n(x)$ elements, and as before, each element of this array may be in one of two states, A or B. When an element of array x is at state $\hat{x}$, its length is $y_{\hat{x}|x}$ and its internal energy is $\epsilon_{\hat{x}|x}$. The two arrays are connected together to form a larger system with a total of $n = n(a) + n(b)$ elements, and this larger system is stretched (or shrinked) so that its edges are fixed at two points which are at distance nY_0 far apart. What is the contribution of each individual array to the total length, nY, and what is the force 'felt' by each one of them? Denoting $p_a = n(a)/n$ and $p_b = n(b)/n$, the total HFE per element is given by $p_a F_a(\beta_0, Y_a) + p_b F_b(\beta_0, Y_b)$, where the second term is equal to $p_b F_b(\beta_0, (Y_0 - p_a Y_a)/p_b)$, where F_a and F_b are the HFE's per element pertaining to the two arrays, respectively, and Y_a and Y_b are their normalized lengths. At equilibrium, Y_a minimizes this expression, and the minimizing Y_a solves the equation:

$$\left.\frac{\partial F_a(\beta_0, Y)}{\partial Y}\right|_{Y=Y_a} = \left.\frac{\partial F_b(\beta_0, Y)}{\partial Y}\right|_{Y=(Y_0 - p_a Y_a)/p_b}.$$

But the left–hand side is $\lambda_a = kT_0 s_a$, the force felt by array (a), and the right–hand side is $\lambda_b = kT_0 s_b$, the force felt by array (b). The last equation tells us that in mechanical equilibrium they are equal. In other words, the equilibrium values of Y_a and Y_b are adjusted such that $F_a(\beta_0, Y_a) = \max_\lambda[G_a(\beta_0, \lambda) + \lambda Y_a]$ and $F_b(\beta_0, Y_b) = \max_\lambda[G_b(\beta_0, \lambda) + \lambda Y_b]$ would be both maximized by the *same* value of λ (or, equivalently, s). In the next section, we will see how this example is directly applicable to the rate–distortion setting.

IV. Rate–Distortion

Consider now the rate–distortion problem. We are given a source sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ to be compressed, whose letters $\{x_i\}$ take on values in a finite alphabet $\mathcal{X}$ of size K. We assume that the source has a given empirical distribution $P = \{P(x),\ x \in \mathcal{X}\}$, i.e., each letter $x \in \mathcal{X}$ appears $n(x) = nP(x)$ times in $\boldsymbol{x}$. Next consider a random selection of a reproduction codeword $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$, where each reproduction symbol $\hat{x}_i$ is drawn i.i.d. from a distribution $Q = \{Q(\hat{x}),\ \hat{x} \in \hat{\mathcal{X}}\}$, where $\hat{\mathcal{X}}$ is a finite reproduction alphabet of size J. For the most part of our discussion, it is assumed that even if the desired distortion level varies, the random coding distribution Q is kept fixed, for the sake of simplicity.[2] It is well known that the RDF of the source P, w.r.t. a given distortion measure $d(x, \hat{x})$, is given by the RF of the large deviations event $\{\sum_{i=1}^n d(x_i, \hat{x}_i) \le n\Delta\}$.

Occasionally, we will work directly with the distortions $\{d(x_i, \hat{x}_i)\}$ incurred, which will be denoted by $\{\delta_i\}$ (playing the same of $\{y_i\}$). Accordingly, we define $Q(\delta|x) = \sum_{\{\hat{x}:\ d(x,\hat{x})=\delta\}} Q(\hat{x})$. The large deviations event under consideration is $\{\sum_{i=1}^n \delta_i \le n\Delta\}$, where $\{\delta_i\}$ are independent. For each $x \in \mathcal{X}$, $n(x) = nP(x)$ of these RV's are drawn from $Q(\delta|x)$. The RF, obtained when all $\{\delta_i\}$ are handled as a whole, is given by

$$I(\Delta) = \max_s \left[s\Delta - \sum_{x \in \mathcal{X}} P(x) \ln\left(\sum_\delta Q(\delta|x) e^{s\delta}\right)\right].$$

In analogy to the results of [5], another look is the following: Consider $\sum_{i:\ x_i = x} \delta_i$, which is the total distortion contributed by x. Clearly, the large deviations event occurs iff there exists a distortion allocation $\mathcal{D} = \{\Delta_x,\ x \in \mathcal{X}\}$ with $\sum_{x \in \mathcal{X}} P(x)\Delta_x \le \Delta$ such that $\sum_{i:\ x_i = x} \delta_i \le n(x)\Delta_x$ for all $x \in \mathcal{X}$. Thus, it can be thought of as the union (over all distortion allocations) of the intersections (over $\mathcal{X}$) of the independent events $\{\sum_{i:\ x_i = x} \delta_i \le n(x)\Delta_x\}$. As shown in [5], since the effective number of distortion allocations is polynomial in n, the probability is dominated by the worst

[2] A word of clarification is in order. While the optimum Q depends on s, or equivalently on Δ, later, we describe certain processes along which the distortion level varies, starting from high distortion Δ_0, and ending at a given distortion Δ. To make a statement concerning $R(\Delta)$, we can always pick the optimum Q for the target value Δ and keep it fixed, even when considering the higher distortion levels. Thus, in these processes, we will 'move' along the curve $R_Q(\cdot)$, which is the RDF with an output distribution Q, rather than $R(\cdot)$.

allocation, which yields

$$\tilde{I}(\Delta)=\min\sum_{x\in\mathcal{X}}P(x)\max_{s_x}\left[s_x\Delta_x-\ln\left(\sum_{\delta}Q(\delta|x)e^{s_x\delta}\right)\right],$$

where the minimum is under the constraint $\sum_{x\in\mathcal{X}}P(x)\Delta_x\leq\Delta$. We argue that $\tilde{I}(\Delta)=I(\Delta)$ (see proof in [6]) and hence both coincide with the RDF $R_Q(\Delta)$ w.r.t. Q.

The intuition behind this argument comes from interpreting the expressions of the RF's in the framework of the example of stretching/contracting concatenated arrays. Here, we have $|\mathcal{X}|=K$ arrays at temperature T_0, concatenated to form one larger system with a total of n elements. Each array is labeled by $x\in\mathcal{X}$ and contains $n(x)=nP(x)$ elements. Each such element may be in one of J states, labeled by $\hat{x}\in\hat{\mathcal{X}}$. The 'length' and the internal energy of an element of array x at state $\hat{x}$ are $\delta_{\hat{x}|x}=d(x,\hat{x})$ and $\epsilon_{\hat{x}|x}=-kT_0\ln Q(\hat{x})$, respectively. Upon identifying this mapping between the rate–distortion problem and the physical example, we immediately see that their mathematical formalisms, and hence also their properties, are the same. Indeed, the expression of $I(\Delta)$ is the HFE per element when the total length is shrinked to $n\Delta$. On the other hand, the expression of $\tilde{I}(\Delta)$ describes the *minimum* HFE across all partial length allocations $\{n(x)\Delta_x\}_{x\in\mathcal{X}}$ that comply with a total length not exceeding $n\Delta$. But this minimum HFE is achieved when all individual arrays 'feel' the same force s_x. Hence, the two expressions should coincide.

Comment: As noted in [5], our discussion in this section, as well as in the next section, applies to channel capacity too, provided that $P=\{P(x)\}$ is understood as the channel output distribution, $Q=\{Q(\hat{x})\}$ is the random (channel) coding distribution, the distortion measure is taken to be $d(x,\hat{x})=-\ln W(x|\hat{x})$, where W is the transition probability matrix associated with the memoryless channel, and the "distortion level" is set to $\Delta=-\sum_{x,\hat{x}}Q(\hat{x})W(x|\hat{x})\ln W(x|\hat{x})$. In this case, the maximizing s is always $s^*=-1$.

V. Integral Representations

In view of the above observations, it is interesting to represent the RDF as mechanical work carried out on the distortion variable along a reversible process, as well as in terms of the integrated variance of the distortion:

$$R_Q(\Delta)==\sum_{x\in\mathcal{X}}P(x)\cdot\int_0^s\mathrm{d}\hat{s}\cdot\hat{s}\cdot\mathrm{Var}_{\hat{s}|x}\{\delta\},\qquad(6)$$

where s is related to Δ via $\sum_{x\in\mathcal{X}}P(x)\left\langle\delta\right\rangle_{s|x}=\Delta$ and where $\left\langle\delta\right\rangle_{s|x}=\sum_\delta\delta Q(\delta|x)e^{s\delta}/[\sum_\delta Q(\delta|x)e^{s\delta}]$ and $\mathrm{Var}_{s|x}\{\delta\}=\left\langle\delta^2\right\rangle_{s|x}-\left\langle\delta\right\rangle_{s|x}^2$. The integrated variance formula above can also be represented as $R_Q(\Delta)=\int_0^s\mathrm{d}\hat{s}\cdot\hat{s}\cdot\sum_{x\in\mathcal{X}}P(x)\cdot\mathrm{Var}_{\hat{s}|x}\{\delta\}$, which is equivalent to $R_Q(\Delta)=\int_0^s\mathrm{d}\hat{s}\cdot\hat{s}\cdot\mathrm{mmse}(\hat{s})$, where $\mathrm{mmse}(s)$ is the minimum mean squared error (MMSE) in estimating δ based on x, when they are jointly distributed according to $P_s(x,\delta)=P(x)P_s(\delta|x)$, with $P_s(\delta|x)$ being defined as $P_s(\delta|x)=Q(\delta|x)e^{s\delta}/[\sum_{\delta'}Q(\delta'|x)e^{s\delta'}]$. The distortion, $\left\langle\delta\right\rangle_s$, which we also denote by Δ, can be represented by:

$$\begin{aligned}\Delta&\equiv\left\langle\delta\right\rangle_s=\sum_{x\in\mathcal{X}}P(x)\cdot\left[\left\langle\delta\right\rangle_{0|x}+\int_0^s\mathrm{d}\hat{s}\cdot\mathrm{Var}_{\hat{s}|x}\{\delta\}\right]\\&=\Delta_0+\int_0^s\mathrm{d}\hat{s}\cdot\mathrm{mmse}(\hat{s}).\qquad(7)\end{aligned}$$

As an example, consider the binary symmetric source (BSS) and the Hamming distortion measure. The optimum Q is also symmetric. Here δ is a binary RV with $\Pr\{\delta=1|x\}=e^s/(1+e^s)$ independent of x. Thus, the MMSE estimator of δ is $\hat{\delta}=e^s/(1+e^s)$, regardless of x, and so the resulting MMSE is easily found to be $\mathrm{mmse}(s)=e^s/(1+e^s)^2$. Accordingly, $\Delta=\frac{1}{2}+\int_0^s e^{\hat{s}}\mathrm{d}\hat{s}/(1+e^{\hat{s}})^2=e^s/(1+e^s)$ and

$$\begin{aligned}R(\Delta)&=\int_0^s\frac{\hat{s}e^{\hat{s}}\mathrm{d}\hat{s}}{(1+e^{\hat{s}})^2}=\ln 2+\frac{se^s}{1+e^s}-\ln(1+e^s)\\&=\ln 2-h_2(\Delta).\qquad(8)\end{aligned}$$

These relations can be generalized: Let $\theta=t(x,\hat{x})$ be a given function and let $\left\langle\theta\right\rangle_s$ denote the expectation of $t(x,\hat{x})$ w.r.t. the distribution $P_s(x,\hat{x})=P(x)Q(\hat{x})e^{sd(x,\hat{x})}/[\sum_{\hat{x}'}Q(\hat{x}')e^{sd(x,\hat{x}')}$. This characterizes the expected (and typical) value of $\frac{1}{n}\sum_{i=1}^n t(x_i,\hat{x}_i)$, where $\hat{x}=(\hat{x}_1,\ldots,\hat{x}_n)$ continues to be the codeword that encodes $\boldsymbol{x}$ from a rate–distortion code designed and operated with the metric d.[3] Then,

$$\left\langle\theta\right\rangle_s=\left\langle\theta\right\rangle_0+\int_0^s\mathrm{d}\hat{s}\cdot\sum_{x\in\mathcal{X}}P(x)\cdot\mathrm{Cov}_{s|x}\{\theta,\delta\},$$

where $\mathrm{Cov}_{s|x}\{\theta,\delta\}$ is the covariance between $\theta=t(x,\hat{x})$ and $\delta=d(x,\hat{x})$ w.r.t. $Q_s(\hat{x}|x)=Q(\hat{x})e^{sd(x,\hat{x})}/[\sum_{\hat{x}'}Q(\hat{x}')e^{sd(x,\hat{x}')}]$, for fixed x. This is an integral form of a more general version of the fluctuation–dissipation theorem, mentioned above.

References

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

[2] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, John and Bartlett Publishers, 1993.

[3] R. Kubo, *Statistical Mechanics*, North Holland Publishing Company, Amsterdam, 1961.

[4] D. McAllester, "A statistical mechanics approach to large deviations theorems," preprint, 2006. Available on-line at: [http://citeseer.ist.psu.edu/443261.html].

[5] N. Merhav, "An identity of Chernoff bounds with an interpretation in statistical physics and applications in information theory," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3710–3721, August 2008.

[6] N. Merhav, "Another look at the physics of large deviations with application to rate–distortion theory," submitted to *IEEE Trans. Inform. Theory*, August 2009. Also, in http://arxiv.org/PS_cache/arxiv/pdf/0908/0908.3562v1.pdf

[7] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, 2009.

[8] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

[3] As examples, consider the case where t is another distortion measure – although the codebook is designed and operated relative to d, its performance can also be judged relative to t. If $t(x,\hat{x})$ depends on $\hat{x}$ only, it may serve as a transmission power function $\Pi(\hat{x})$ (in joint source–channel coding) or it can be the length function $\ell(\hat{x})$ (in bits) of lossless compression for the individual reproduction symbols.