



IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

On Optimum Strategies for Minimizing the Exponential Moments of a Given Cost

Neri Merhav

CCIT Report #785
March 2011

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

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



On Optimum Strategies for Minimizing the Exponential Moments of a Given Cost Function

Neri Merhav

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

Abstract

We consider a general problem of finding a strategy that minimizes the exponential moment of a given cost function, with an emphasis on its relation to the more common criterion of minimization the expectation of the first moment of the same cost function. In particular, our main result is a theorem that gives simple sufficient conditions for a strategy to be optimum in the exponential moment sense. This theorem may be useful in various situations, and application examples are given. We also examine the asymptotic regime and investigate universal asymptotically optimum strategies in light of the aforementioned sufficient conditions, as well as phenomena of irregularities, or phase transitions, in the behavior of the asymptotic performance, which can be viewed and understood from a statistical–mechanical perspective. Finally, we propose a new route for deriving lower bounds on exponential moments of certain cost functions (like the square error in estimation problems) on the basis of well known lower bounds on their expectations.

Index Terms: loss function, exponential moment, large deviations, phase transitions, universal schemes.

1 Introduction

Many problems in information theory, communications, statistical signal processing, and related disciplines can be formalized as being about the quest for a strategy s that minimizes (or maximizes) the expectation of a certain cost function, $\ell(X, s)$, where X is a random variable (or a random vector). Just a few examples of this generic paradigm are the following: (i) Lossless and lossy data compression, where X symbolizes the data to be compressed, s is the data compression scheme, and $\ell(X, s)$ is the length of the compressed binary representation, or the distortion (in the lossy

case) or a linear combination of both (see, e.g., [6, Chapters 5 and 10]). (ii) Gambling and portfolio theory [6, Chapters 6 and 16], where cost function is logarithm of the wealth relative. (iii) Lossy joint source–channel coding, where X collectively symbolizes the randomness of source and the channel, s is the encoding–decoding scheme and $\ell(X, s)$ is the distortion in the reconstruction (see, e.g., [27],[28]). (iv) Bayesian estimation of a random variable based on measurements, where X designates jointly the desired random variable and the measurements, s is the estimation function and $\ell(X, s)$ is the error function, for the example, the squared error. Non–Bayesian estimation problems can be considered similarly (see, e.g., [24]). (v) Prediction, sequential decision problems (see, for example, [18]) and stochastic control problems [5], such as the linear quadratic Gaussian (LQG) problem, as well as general Markov decision processes, are also formalized in terms of selecting strategies in order to minimize the expectation of a certain loss function.

While the criterion of minimizing the expected value of $\ell(X, s)$ has been predominantly the most common one, the exponential moments of $\ell(X, s)$, namely, $\mathbf{E} \exp\{\alpha \ell(X, s)\}$ ($\alpha > 0$), have received much less attention than they probably deserve in this context. There are a few motivations for examining strategies that minimize exponential moments. First, $\mathbf{E} \exp\{\alpha \ell(X, s)\}$, as a function of α , is obviously the moment–generating function of $\ell(X, s)$, and as such, it provides the full information about the entire distribution of this random variable, not just its first order moment. Thus, in particular, if we are fortunate enough to find a strategy that uniformly minimizes $\mathbf{E} \exp\{\alpha \ell(X, s)\}$ for all $\alpha \geq 0$ (and there are examples that this may be the case), then this is much stronger than just minimizing the first moment. Secondly, exponential moments are intimately related to large–deviations rate functions, and so, the minimization of exponential moments may give us an edge on minimizing probabilities of (undesired) large deviations events of the form $\Pr\{\ell(X, s) \geq L_0\}$ (for some threshold L_0), or more precisely, on maximizing the exponential rate of decay of these probabilities. There are several works along this line, especially in contexts related to buffer overflow in data compression [11],[12], [13],[15],[19],[23],[26], and exponential moments related to guessing [1],[2], [3],[14],[17],[20].

It is natural to ask, in view of the foregoing discussion, how can we harness the existing body of knowledge concerning optimization of strategies for minimizing the first moment of $\ell(X, s)$, which is quite mature in many applications, in our quest for optimum strategies that minimize exponential moments. Our main basic result, in this paper, is a simple theorem that relates the two criteria. In

particular, we furnish sufficient conditions that the optimum strategy in the exponential moment sense can be found in terms of the optimum strategy in the first moment sense, for a possibly different probability distribution, which our theorem characterizes.

In some applications, these sufficient conditions for optimality in the exponential moment sense, yield an equation in s , whose solution is the desired optimum strategy. It is clear then that in these applications, the optimality conditions provide a concrete tool for deriving the optimum solution. In other applications, however, this may not be quite the case directly, yet the set of optimality conditions may still serve as a useful tool: More often than not, in a given instance of the problem under discussion, one may have a natural intuitive guess concerning the optimum strategy, and then the optimality conditions can be used to prove that this is the case. One example for this, that will be demonstrated in detail later on, is the following: Given n independent and identically distributed (i.i.d.) Gaussian observations, $X_1, \dots, X_n$, with mean θ , the sample mean, $s(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$, is the optimum unbiased estimator of θ , not merely in the mean squared error sense (as is well known), but also in the sense of minimizing all exponential moments of the squared error, i.e., $\mathbf{E} \exp\{\alpha[s(X_1, \dots, X_n) - \theta]^2\}$ for all $\alpha \geq 0$ for which this expectation is finite.

We next devote some attention to the asymptotic regime. Consider the case where X is a random vector of dimension n , $X = (X_1, \dots, X_n)$, governed by a product-form probability distribution, and $\ell(X, s)$ grows linearly for a given empirical distribution of X , for example, when $\ell(X, s)$ is additive, i.e., $\ell(X, s) = \sum_{i=1}^n \ell(X_i, s)$. In this case, the exponential moments of $\ell(X, s)$ typically behave (at least asymptotically) like exponential functions of n . If we can then select a strategy s that somehow “adapts”¹ to the empirical distribution of $(X_1, \dots, X_n)$, then such strategies may be universally optimum (or asymptotically optimum in the sense of achieving the minimum exponential rate of the exponential moment) in that they depend on neither the underlying probability distribution, nor on the parameter α . This is demonstrated in several examples, one of which is an extension of a well known result by Rissanen in universal data compression [22].

An interesting byproduct of the use of the exponential moment criterion in the asymptotic regime is the possible existence of phase transitions: It turns out that the asymptotic exponential rate of $\mathbf{E} \exp\{\alpha \ell(X_1, \dots, X_n, s)\}$ as a function of n , may not be a smooth function of α and/or the

¹The precise meaning of this will be clarified in the sequel.

parameters of the underlying probability distribution even when the model under discussion seems rather simple and ‘innocent.’ This is best understood from a statistical–mechanical perspective, because in some cases, the calculation of the exponential moment is clearly analogous to that of the partition function of a certain physical system of interacting particles, which is known to exhibit phase transitions. It is demonstrated that at least in certain cases, these phase transitions are not merely an artifact of a badly chosen strategy, but they appear even when the optimum strategy is used, and hence these phase transitions are inherent in the model.

We end this paper by touching upon yet another aspect of the exponential moment criterion, which we do not investigate very thoroughly here, but we believe it is interesting and therefore certainly deserves a further study in the future: Even in the ordinary setting, of seeking strategies that minimize $\mathbf{E}\{\ell(X, s)\}$, optimum strategies may not always be known, and then lower bounds are of considerable importance as a reference performance figure. This is *a-fortiori* the case when exponential moments are considered. One way to obtain non-trivial bounds on exponential moments is via lower bounds the expectation of $\ell(X, s)$, using the techniques developed in this paper. We demonstrate this idea in the context of a lower bound on the expected exponentiated squared error of an unbiased parameter estimator, on the basis of the Cramér–Rao bound (CRB), but it should be understood that, more generally, the same idea can be applied on the basis of other well-known bounds of the mean-square error (Bayesian and non-Bayesian) in parameter estimation, and in signal estimation, as well as in other problem areas.

2 Basic Optimality Conditions

Let X be a random variable taking on values in a certain alphabet $\mathcal{X}$, and drawn according to a given probability distribution P . Let the variable s designate a *strategy* chosen from some space $\mathcal{S}$ of allowed strategies. The term “strategy” in our context is fairly generic: it may be a scalar variable, a vector, an infinite sequence, a function (of X), a partition of $\mathcal{X}$, a coding scheme for X , and so on. Associated with each $x \in \mathcal{X}$ and $s \in \mathcal{S}$, is a loss $\ell(x, s)$. The function $\ell(x, s)$ is called the *loss function*, or the *cost function*. The operator $\mathbf{E}\{\cdot\}$ will be understood as the expectation operator with respect to (w.r.t.) the underlying distribution P , and whenever we refer to the expectation w.r.t. another probability distribution, say, Q , we use the notation $\mathbf{E}_Q\{\cdot\}$.

Nonetheless, occasionally, when there is more than one probability distribution playing a role at the same time and we wish to emphasize that the expectation is taken w.r.t. P , then to avoid confusion, we may denote this expectation by $\mathbf{E}_P\{\cdot\}$.

For a given $\alpha > 0$, consider the problem of minimizing $\mathbf{E} \exp\{\alpha \ell(X, s)\}$ across $s \in \mathcal{S}$. The following theorem relates the optimum s for this problem to the optimum s for the problem of minimizing $\mathbf{E}_Q\{\ell(X, s)\}$ w.r.t. another probability distribution Q .

Theorem 1 *Assume that there exists a strategy $s \in \mathcal{S}$ for which*

$$Z(s) \triangleq \mathbf{E}_P \exp\{\alpha \ell(X, s)\} < \infty. \quad (1)$$

A strategy $s \in \mathcal{S}$ minimizes $\mathbf{E}_P \exp\{\alpha \ell(X, s)\}$ if there exists a probability distribution Q on $\mathcal{X}$ that satisfies the following two conditions at the same time:

1. *The strategy s minimizes $\mathbf{E}_Q\{\ell(X, s)\}$ over $\mathcal{S}$.*
2. *The probability distribution Q is given by*

$$Q(x) = \frac{P(x)e^{\alpha \ell(x, s)}}{Z(s)}. \quad (2)$$

An equivalent formulation of Theorem 1 is the following: denoting by s_Q a strategy that minimizes $\mathbf{E}_Q\{\ell(X, s)\}$ over $\mathcal{S}$, then the theorem asserts that s_Q minimizes $\mathbf{E}_P \exp\{\alpha \ell(X, s)\}$ over $\mathcal{S}$ if

$$Q(x) \propto P(x)e^{\alpha \ell(x, s_Q)}, \quad (3)$$

where by $A(x) \propto B(x)$, we mean that $A(x)/B(x)$ is a constant, independent of x .

Proof. Let $s \in \mathcal{S}$ be arbitrary and let (s^*, Q^*) satisfy conditions 1 and 2 of Theorem 1. Consider the following chain of inequalities:

$$\begin{aligned} \mathbf{E}_P \exp\{\alpha \ell(X, s)\} &= \mathbf{E}_{Q^*} \exp\left\{\alpha \ell(X, s) + \ln \frac{P(X)}{Q^*(X)}\right\} \\ &\geq \exp\{\alpha \mathbf{E}_{Q^*} \ell(X, s) - D(Q^* \| P)\} \\ &\geq \exp\{\alpha \mathbf{E}_{Q^*} \ell(X, s^*) - D(Q^* \| P)\} \\ &= \exp\left\{\alpha \mathbf{E}_{Q^*} \ell(X, s^*) - \mathbf{E}_{Q^*} \ln \frac{e^{\alpha \ell(X, s^*)}}{Z(s^*)}\right\} \\ &= Z(s^*) = \mathbf{E}_P \exp\{\alpha \ell(X, s^*)\}, \end{aligned} \quad (4)$$

where the first equality results from a change of measure (multiplying and dividing $e^{\alpha\ell(X,s)}$ by $Q^*(X)$), the second line is by Jensen's inequality and the convexity of the exponential function (with $D(Q\|P) \triangleq \mathbf{E}_Q \ln[Q(X)/P(X)]$ being the relative entropy between Q and P), the third line is by condition 1 of Theorem 1, and the remaining equalities result from condition 2: On substituting $Q^*(x) = P(x)e^{\alpha\ell(x,s^*)}/Z(s^*)$ into $D(Q^*\|P)$, one readily obtains $D(Q^*\|P) = \alpha\mathbf{E}_{Q^*}\ell(X, s^*) - \ln Z(s^*)$. This completes the proof of Theorem 1. $\square$

Observe that for a given s , Jensen's inequality in the second line of (4), becomes an equality for $Q(x) = P(x)e^{\alpha\ell(x,s)}/Z(s)$, since for this choice of Q , the random variable that appears in the exponent, $\alpha\ell(X, s) + \ln \frac{P(X)}{Q(X)}$, becomes degenerate (constant with probability one). Since the original expression is independent of Q , such an equality in Jensen's inequality means that $\alpha\mathbf{E}_Q\ell(X, s) - D(Q\|P)$ is maximized by this choice of Q , a fact which can also be seen from a direct maximization of this expression using standard methods. Thus, we have a simple identity for every s :

$$\mathbf{E}_P \exp\{\alpha\ell(X, s)\} = \exp\{\alpha \max_Q [\mathbf{E}_Q\ell(X, s) - D(Q\|P)]\}. \quad (5)$$

This identity will prove useful in several places throughout the sequel.

Suppose next that the set $\mathcal{S}$ and the loss function $\ell(x, s)$ are such that:

$$\min_{s \in \mathcal{S}} \max_Q [\alpha\mathbf{E}_Q\ell(X, s) - D(Q\|P)] = \max_Q \min_{s \in \mathcal{S}} [\alpha\mathbf{E}_Q\ell(X, s) - D(Q\|P)]. \quad (6)$$

This equality between the min-max and the max-min means that there is a saddle point (s^*, Q^*) , where s^* is a solution of the min-max problem on the left-hand side and Q^* is a solution to the max-min problem on the right-hand side. It is easy to check that the maximizing Q in the inner maximization on the left-hand side is $Q^*(x) = P(x)e^{\alpha\ell(x,s^*)}/Z(s^*)$, which is condition 2 of Theorem 1. By the same token, the inner minimization over s on the right-hand side obviously minimizes $\mathbf{E}_{Q^*}\ell(X, s)$, which is condition 1. It follows then that if the min-max and the max-min are equal, then the saddle point satisfies the conditions of Theorem 1, and hence the corresponding s^* is optimum. Note also that when eq. (6) holds, the conditions of Theorem 1 become also necessary conditions for optimality: Suppose that s^* is optimum. Then, by eq. (5), it must solve the minimax problem on the left-hand side of eq. (6). But if eq. (6) holds then there is a saddle point, and s^* if the first coordinate of this saddle point, (s^*, Q^*) . But then s^* and Q^* must be related according to the conditions of Theorem 1, as explained above.

When does eq. (6) hold? In general, the well-known sufficient conditions for

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} f(u, v) = \max_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} f(u, v) \quad (7)$$

are that $\mathcal{U}$ and $\mathcal{V}$ are convex sets (with $\mathcal{U}$ being independent of v and $\mathcal{V}$ being independent of u), and that f is convex in u and concave in v . In our case, since the function $f(s, Q) = \alpha \mathbf{E}_Q \ell(X, s) - D(Q \| P)$ is always concave in Q , this sufficient condition would automatically hold whenever $\ell(x, s)$ is convex in s (for every fixed x), provided that $\mathcal{S}$ is a space in which convex combinations can be well defined, and that $\mathcal{S}$ is a convex set.

Maximizing Negative Exponential Moments. A similar, but somewhat different, criterion pertaining to exponential moments, which is reasonable to the same extent, is the dual problem of $\max_{s \in \mathcal{S}} \mathbf{E} \exp\{-\alpha \ell(X, s)\}$ (again, with $\alpha > 0$). If $\ell(x, s)$ is non-negative for all x and s , this has the advantage that the exponential moment is finite for all $\alpha > 0$, as opposed to $\mathbf{E} \exp\{\alpha \ell(X, s)\}$ which, in many cases, is finite only for a limited range of α . For the same considerations as before, here we have:

$$\begin{aligned} \max_s \mathbf{E} \exp\{-\alpha \ell(X, s)\} &= \max_s \exp\{\max_Q [-\alpha \mathbf{E}_Q \ell(X, s) - D(Q \| P)]\} \\ &= \exp\{-\min_s \min_Q [\alpha \mathbf{E}_Q \ell(X, s) + D(Q \| P)]\}, \end{aligned} \quad (8)$$

and so the optimality conditions relating s and Q are similar to those of Theorem 1 (with α replaced by $-\alpha$), except that now we have a double minimization problem rather than a min-max problem. However, it should be noted that here the conditions of Theorem 1 are only necessary conditions, as for the above equalities to hold, the pair (s, Q) should *globally* minimize the function $[\alpha \mathbf{E}_Q \ell(X, s) + D(Q \| P)]$, unlike the earlier case, where only a saddle point was sought.² On the other hand, another advantage of this criterion, is that even if one cannot solve explicitly the equation for the optimum s , then the double minimization naturally suggests an iterative algorithm: starting from an initial guess $s_0 \in \mathcal{S}$, one computes $Q_0(x) \propto P(x) \exp\{-\alpha \ell(x, s_0)\}$ (which minimizes $[\alpha \mathbf{E}_Q \ell(X, s) + D(Q \| P)]$ over $\{Q\}$), then one finds $s_1 = \arg \min_{s \in \mathcal{S}} \mathbf{E}_{Q_0} \{\ell(X, s)\}$, and so on. It is obvious that $\mathbf{E} \exp\{-\alpha \ell(X, s_i)\}$, $i = 0, 1, 2, \dots$, increases (and hence improves) from iteration to iteration. This is different from the min-max situation we encountered earlier, where successive improvements are not guaranteed.

²In other words, it is not enough now that s and Q are in ‘equilibrium’ in the sense that s is a minimizer for a given Q and vice versa.

3 A Few Examples

Theorem 1 tells us that if we are fortunate enough to find a strategy $s \in \mathcal{S}$ and a probability distribution Q , which are ‘matched’ to one another (in the sense defined by the above conditions), then we have solved the problem of minimizing the exponential moment. Sometimes it is fairly easy to find such a pair (s, Q) by solving an equation. In other cases, there might be a natural guess for the optimum s , which can be proven optimum by checking the conditions. In this section, we will see examples of both types. Some of these examples could have been also solved directly, without using Theorem 1, but for others, this does not seem to be a trivial task. In some of the examples, it turns out that the same optimum strategy that minimizes expected loss, is also optimum in the sense of minimizing all exponential moments, but this is, of course, not always the case.

3.1 Example 1: Lossless Data Compression

We begin with a very simple example. Let X be a random variable taking on values in a finite alphabet $\mathcal{X}$, let s be a probability distribution on $\mathcal{X}$, i.e., a vector $\{s(x), x \in \mathcal{X}\}$ with $\sum_{x \in \mathcal{X}} s(x) = 1$ and $s(x) \geq 0$ for all $x \in \mathcal{X}$, and let $\ell(x, s) \triangleq -\ln s(x)$. This example is clearly motivated by lossless data compression, as $-\ln s(x)$ is the length function (in nats) pertaining to a uniquely decodable code that is induced by a distribution s , ignoring integer length constraints. In this problem, one readily observes that the optimum s for minimizing $\mathbf{E}_Q\{-\ln s(X)\}$ is $s_Q = Q$. Thus, by eq. (3), we seek a distribution Q such that

$$Q(x) \propto P(x) \exp\{-\alpha \ln Q(x)\} = \frac{P(x)}{[Q(x)]^\alpha} \quad (9)$$

which means $[Q(x)]^{1+\alpha} \propto P(x)$, or equivalently, $Q(x) \propto [P(x)]^{1/(1+\alpha)}$. More precisely,

$$s_Q(x) = Q(x) = \frac{[P(x)]^{1/(1+\alpha)}}{\sum_{x' \in \mathcal{X}} [P(x')]^{1/(1+\alpha)}}, \quad (10)$$

and the expectation of $-\ln s_Q(X)$ yields the Rényi entropy. Note that here $\ell(x, s)$ is convex in s and so, the minimax condition holds. While this result is well known and it could have been obtained even without using Theorem 1, our purpose in this example was to show how Theorem 1 gives the desired solution even more easily than with the direct method, by solving a very simple equation.

3.2 Example 2: Bayesian Estimation

Let (X, Y) be random variables, where Y is distributed according to a given density $P(y)$ and the conditional density of X given Y is given by

$$P(x|y) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x - \phi(y))^2 \right\}, \quad (11)$$

where $\phi(y)$ is a given function. We are seeking the optimum linear estimator of X based on the observation Y in the sense of minimizing the exponential moment of squared error. In other words, we seek a real number s that minimizes $\mathbf{E} \exp\{\alpha(X - sY)^2\}$, where $\alpha \in (0, 1/2)$. Once again, the loss function is convex in s . According to the second condition of Theorem 1, Q should be of the form

$$\begin{aligned} Q(x, y) &\propto P(y) \cdot \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x - \phi(y))^2 + \alpha(x - sy)^2 \right\} \\ &= P(y) \exp \left\{ \frac{\alpha}{1 - 2\alpha} [\phi(y) - sy]^2 \right\} \cdot Q(x|y) \\ &\propto \tilde{P}(y) \cdot Q(x|y) \end{aligned} \quad (12)$$

where $Q(x|y)$ is a Gaussian distribution with mean $[\phi(y) - 2\alpha sy]/(1 - 2\alpha)$ and variance $1/(1 - 2\alpha)$ and

$$\tilde{P}(y) \propto P(y) \exp \left\{ \frac{\alpha}{1 - 2\alpha} [\phi(y) - sy]^2 \right\}. \quad (13)$$

On the other hand, by the first condition of Theorem 1, s should be the coefficient pertaining to the optimum linear estimator of X based on Y under Q , which is

$$s = \frac{\mathbf{E}_Q(XY)}{\mathbf{E}_Q(Y^2)} = \frac{\mathbf{E}_Q\{Y \cdot \mathbf{E}_Q(X|Y)\}}{\mathbf{E}_Q(Y^2)}. \quad (14)$$

But since $Q(x|y)$ is Gaussian with mean $[\phi(y) - 2\alpha sy]/(1 - 2\alpha)$ as said, then this is exactly the inner expectation at the numerator, and so, we obtain

$$s = \frac{1}{1 - 2\alpha} \left[\frac{\mathbf{E}_Q\{Y \cdot \phi(Y)\}}{\mathbf{E}_Q(Y^2)} - 2\alpha s \right], \quad (15)$$

or equivalently,

$$s = \frac{\mathbf{E}_Q\{Y \phi(Y)\}}{\mathbf{E}_Q(Y^2)}. \quad (16)$$

But since these expectations involve only the random variable Y whose marginal under Q is $\tilde{P}$, then the expectations are actually taken under $\tilde{P}$, i.e.,

$$s = \frac{\mathbf{E}_{\tilde{P}}\{Y\phi(Y)\}}{\mathbf{E}_{\tilde{P}}(Y^2)}. \quad (17)$$

Note that this is different from the solution to the ordinary MMSE problem, where the solution is given by the same expression, but with $\tilde{P}$ being replaced by P . It should be kept in mind that $\tilde{P}$ depends, in general, on s , then so does the right hand side of the last equation. We have therefore obtained an equation whose solution $s = s_Q$ is the optimum coefficient in the sense of minimum $\mathbf{E}_P \exp\{\alpha(X - sY)^2\}$. Let us now examine a few simple special cases.

Consider first the case $\phi(y) = s_0 y$, for some real constant s_0 . In this case, the right hand-side of eq. (17) is trivially equal to s_0 , which means that $s_Q = s_0$. This means that whenever (X, Y) is a Gaussian vector, the linear MMSE estimator minimizes also all exponential moments of the squared error (among all linear estimators).

Consider next the case where

$$P(y) = \frac{1}{2}\delta(y-1) + \frac{1}{2}\delta(y+1), \quad (18)$$

and denote $\phi_+ \triangleq \phi(+1)$ and $\phi_- \triangleq \phi(-1)$. Then

$$\begin{aligned} \tilde{P}(y) &= \frac{\exp\left\{\frac{\alpha}{1-2\alpha}(\phi_+ - s)^2\right\}}{\exp\left\{\frac{\alpha}{1-2\alpha}(\phi_+ - s)^2\right\} + \exp\left\{\frac{\alpha}{1-2\alpha}(\phi_- + s)^2\right\}} \cdot \delta(y-1) + \\ &\quad \frac{\exp\left\{\frac{\alpha}{1-2\alpha}(\phi_- + s)^2\right\}}{\exp\left\{\frac{\alpha}{1-2\alpha}(\phi_+ - s)^2\right\} + \exp\left\{\frac{\alpha}{1-2\alpha}(\phi_- + s)^2\right\}} \cdot \delta(y+1). \end{aligned} \quad (19)$$

Since $\mathbf{E}_{\tilde{P}}(Y^2) = 1$, the equation in s reads

$$s = \frac{\phi_+ \exp\left\{\frac{\alpha}{1-2\alpha}(\phi_+ - s)^2\right\} - \phi_- \exp\left\{\frac{\alpha}{1-2\alpha}(\phi_- + s)^2\right\}}{\exp\left\{\frac{\alpha}{1-2\alpha}(\phi_+ - s)^2\right\} + \exp\left\{\frac{\alpha}{1-2\alpha}(\phi_- + s)^2\right\}} \quad (20)$$

For $\phi_+ = -\phi_-$, we get $s = \phi_+$, which is expected since this is actually the linear case discussed above, with $s_0 = \phi_+$. For $\phi_+ = \phi_- \triangleq \phi$, the equation reads

$$s = -\phi \tanh\left(\frac{2\alpha\phi s}{1-2\alpha}\right) \quad (21)$$

and the only solution is $s = 0$, which makes sense since X and Y are independent in this case. For $\alpha \rightarrow 0$, the ordinary MMSE linear estimator is recovered, whose coefficient is $s = (\phi_+ - \phi_-)/2$, as expected.

Returning to the general setting of this example, let us examine what would happen if we expand the scope and allow a general, non-linear estimator. In this case, we seek a general function $s(y)$ such that

$$Q(x, y) \propto P(y) \exp \left\{ -\frac{1}{2}(x - \phi(y))^2 \right\} \cdot \exp \{ \alpha(x - s(y))^2 \}. \quad (22)$$

If $\alpha \in (0, 1/2)$ and we guess $s(y) = \phi(y)$, we obtain

$$Q(x, y) \propto P(y) \exp \left\{ -\left(\frac{1}{2} - \alpha \right) (x - \phi(y))^2 \right\} \quad (23)$$

for which the conditional mean, $s_Q(y) = \mathbf{E}_Q(X|Y = y)$, is indeed $\phi(y)$, and so, the conditions of Theorem 1 are satisfied. It follows then that for our example, $s(y) = \mathbf{E}(X|Y = y) = \phi(y)$, minimizes not only the MSE, but also all exponential moments of the squared error, $\mathbf{E} \exp \{ \alpha(X - s(Y))^2 \}$ for $0 < \alpha < 1/2$.

The same idea applies to somewhat more general situations. Let $\rho(t)$ be an even function, which is monotonically non-decreasing for $t \geq 0$, and steeply enough so that $\int_{-\infty}^{+\infty} dt e^{-\beta \rho(t)} < \infty$ for all $\beta > \beta_0$, where $\beta_0 > 0$ is a certain constant. Suppose that $P(x, y) \propto P(y) \exp[-\beta \rho(x - \phi(y))]$ for some $\beta > \beta_0$, and we are interested in minimizing the exponential moment $\mathbf{E} \exp \{ \alpha \rho(X - s(Y)) \}$. Then, for every $\alpha \in (0, \beta - \beta_0)$, the choice $s(y) = \phi(y)$ leaves $Q(x|y)$ symmetric about $x = \phi(y)$. If $\tau = 0$ minimizes $\int_{-\infty}^{+\infty} dt \rho(t - \tau) e^{-\beta \rho(t)}$ for every $\beta > \beta_0$ (which is true in many cases), then the estimator $s(y) = \phi(y)$ minimizes all exponential moments of $\rho(X - s(Y))$. This can be even further generalized to cases where $P(x|y) \propto \exp[-\beta \rho_1(x - \phi(y))]$ for a given symmetric function ρ_1 that may be different from the function ρ for which we wish to minimize the exponential moment.

The above considerations extend also to signal estimation (prediction, filtering, etc.): Consider two jointly wide-sense stationary Gaussian processes, $\{(X_n, Y_n)\}$. Given $(\dots, Y_{-1}, Y_0, Y_1, \dots)$, each X_t is Gaussian, with conditional mean given by $\mathbf{E}\{X_t | \dots, Y_{-1}, Y_0, Y_1, \dots\} = \sum_{i=-\infty}^{\infty} h_i Y_{t-i}$, $\{h_i\}$ being the impulse response of the non-causal Wiener filter. From the same reasons as before, the exponential moments of the square error are also minimized by the non-causal Wiener filter. It is not clear, however, whether the causal Wiener filter minimizes the exponentiated square error among all causal filters, unless the non-causal Wiener filter happens to coincide with the causal

one. Optimum linear prediction of Gaussian processes in the ordinary mean square error sense are also optimum in the mean exponentiated squared error sense.

3.3 Example 3: Non-Bayesian Estimation

Let $X_1, X_2, \dots, X_n$ be i.i.d. Gaussian random variables with mean θ and variance σ^2 . It is very well known that among all unbiased estimators of θ , the one that minimizes the mean square error (or equivalently, the estimation error variance) is the sample mean $s(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$. Does the sample mean estimator also minimize $\mathbf{E} \exp\{\alpha[s(X_1, \dots, X_n) - \theta]^2\}$ among all unbiased estimators and for all values of α in the allowed range?

Once again, the class $\mathcal{S}$ of all unbiased estimators is clearly a convex set and $(s - \theta)^2$ is convex in s . Let us ‘guess’ that the sample mean indeed minimizes also $\mathbf{E} \exp\{\alpha[s(X_1, \dots, X_n) - \theta]^2\}$ and then check whether it satisfies the conditions of Theorem 1. The corresponding probability measure Q , which will be denoted here by Q_θ , is given by

$$\begin{aligned} Q_\theta(x_1, \dots, x_n) &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 + \alpha \left(\frac{1}{n} \sum_{i=1}^n x_i - \theta \right)^2 \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 + \alpha \left[\frac{1}{n} \sum_{i=1}^n (x_i - \theta) \right]^2 \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{x} - \theta \mathbf{u})^T W (\mathbf{x} - \theta \mathbf{u}) \right\}, \end{aligned} \quad (24)$$

where $\mathbf{x} \triangleq (x_1, \dots, x_n)^T$, $\mathbf{u} = (1, 1, \dots, 1)^T \in \mathbb{R}^n$ and

$$W = I - \frac{2\alpha\sigma^2}{n^2} \mathbf{u} \mathbf{u}^T, \quad (25)$$

I being the $n \times n$ identity matrix. The maximum likelihood estimator of θ under Q_θ is given by

$$s(\mathbf{x}) = \frac{\mathbf{u}^T W \mathbf{x}}{\mathbf{u}^T W \mathbf{u}} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (26)$$

namely, the sample mean. It can easily be shown to achieve the Cramér–Rao lower bound under Q_θ , which is $\sigma^2/(\mathbf{u}^T W \mathbf{u})$. Thus, the sample mean estimator is an optimum unbiased estimator for Q_θ and hence it satisfies the conditions of Theorem 1. The answer to the question of the previous paragraph is then affirmative. The best achievable performance, in the exponential moment sense,

is given by

$$\mathbf{E} \exp \left\{ \alpha \left(\frac{1}{n} \sum_{i=1}^n X_i - \theta \right)^2 \right\} = \frac{1}{\sqrt{\det(W)}}. \quad (27)$$

3.4 Example 4: The Gaussian Joint Source–Channel Coding Problem

Consider the Gaussian memoryless source

$$P_U(\mathbf{u}) = (2\pi\sigma_u^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_u^2} \sum_{i=1}^n u_i^2 \right\} \quad (28)$$

and the Gaussian memoryless channel $\mathbf{y} = \mathbf{x} + \mathbf{z}$, where the noise is distributed according to

$$P_Z(\mathbf{z}) = (2\pi\sigma_z^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_z^2} \sum_{i=1}^n z_i^2 \right\}. \quad (29)$$

In the ordinary joint source–channel coding problem, one seeks an encoder and decoder that would minimize $D = \frac{1}{n} \sum_{i=1}^n \mathbf{E}\{(U_i - V_i)^2\}$, where $\mathbf{V} = (V_1, \dots, V_n)$ is the reconstruction at the decoder.

It is very well known that the best achievable distortion, in this case, is given by

$$D = \frac{\sigma_u^2}{1 + \Gamma/\sigma_z^2}, \quad (30)$$

where Γ is the maximum power allowed at the transmitter, and it may be achieved by a transmitter that simply amplifies the source by a gain factor of $\sqrt{\Gamma/\sigma_u^2}$ and a receiver that implements linear MMSE estimation of U_i given Y_i , on a symbol-by-symbol basis.

What happens if we replace the criterion of expected distortion by the criterion of the exponential moment on the distortion, $\mathbf{E} \exp\{\alpha \sum_i (U_i - V_i)^2\}$? It is natural to wonder whether simple linear transmitters and receivers, of the kind defined in the previous paragraph, are still optimum.

The random object X , in this example, is the pair of vectors $(\mathbf{U}, \mathbf{Z})$, where $\mathbf{U}$ is the source vector and $\mathbf{Z}$ is the channel noise vector, which under $P = P_U \times P_Z$, are independent Gaussian i.i.d. random vectors with zero mean and variances σ_u^2 and σ_z^2 , respectively, as said. Our strategy s consists of the choice of an encoding function $\mathbf{x} = f(\mathbf{u})$ and a decoding function $\mathbf{v} = g(\mathbf{y})$. The class $\mathcal{S}$ is then the set of all pairs of functions $\{f, g\}$, where f satisfies the power constraint $\mathbf{E}_P\{\|f(\mathbf{U})\|^2\} \leq n\Gamma$. Condition 2 of Theorem 1 tells us that the modified probability distribution of $\mathbf{u}$ and $\mathbf{z}$ should be of the form

$$Q(\mathbf{u}, \mathbf{z}) \propto P_U(\mathbf{u})P_Z(\mathbf{z}) \exp \left\{ \alpha \sum_{i=1}^n [u_i - g_i(f(\mathbf{u}) + \mathbf{z})]^2 \right\} \quad (31)$$

where g_i is restriction of g to the i -th component of $\mathbf{v}$.

Clearly, if we continue to restrict the encoder f to be linear, with a gain of $\sqrt{\Gamma/\sigma_u^2}$, which simply exploits the allowed power Γ , and the only remaining room for optimization concerns the decoder g , then we are basically back to the previous example of Bayesian estimation in the Gaussian regime, and the optimum choice of the decoder is a linear one, exactly like in the traditional mean square error case (from the same consideration as in the Bayesian estimation example). However, once we extend the scope and allow f to be a non-linear encoder, then the optimum choice of f and g would no longer remain linear like in the expected distortion case. It is not difficult to see that the conditions of Theorem 1 are no longer met for any linear functions f and g . The key reason is that while $Q(\mathbf{u}, \mathbf{z})$ of eq. (31) continues to be Gaussian (though now U_i and Z_i are correlated) when f and g are linear, the power constraint, $\mathbf{E}_P\{\|\mathbf{X}\|^2\} \leq n\Gamma$, when expressed as an expectation w.r.t. Q , becomes $\mathbf{E}_Q\{\|f(\mathbf{U})\|^2 P(\mathbf{U})/Q(\mathbf{U})\} \leq n\Gamma$, but “power” function $\|f(\mathbf{u})\|^2 P(\mathbf{u})/Q(\mathbf{u})$, with P and Q being Gaussian densities, is no longer the usual quadratic function of $f(\mathbf{u})$ for which there is a linear encoder and decoder that is optimum.

Another way to see that linear encoders and decoders are suboptimal, is to consider the following argument: For a given n , the expected exponentiated squared error is minimized by a joint source-channel coding system, defined over a super-alphabet of n -tuples, with respect to a distortion measure, defined in terms of a single super-letter, as

$$d(\mathbf{u}, \mathbf{v}) = \exp \left\{ \alpha \sum_{i=1}^n (u_i - v_i)^2 \right\}. \quad (32)$$

For such a joint source-channel coding system to be optimal, the induced channel $P(\mathbf{v}|\mathbf{u})$ must [4, p. 31, eq. (2.5.13)] be proportional to

$$P(\mathbf{v}) \exp\{-\beta d(\mathbf{u}, \mathbf{v})\} = P(\mathbf{v}) \exp \left[-\beta \exp \left\{ \alpha \sum_i (u_i - v_i)^2 \right\} \right] \quad (33)$$

for some $\beta > 0$, which is the well-known structure of the optimum test channel that attains the rate-distortion function for the Gaussian source and the above defined distortion measure. Had the aforementioned linear system been optimum, the optimum output distribution $P(\mathbf{v})$ would be Gaussian, and then $P(\mathbf{v}|\mathbf{u})$ would remain proportional to a double exponential function of $\sum_i (u_i - v_i)^2$. However, the linear system induces instead a Gaussian channel from $\mathbf{u}$ to $\mathbf{v}$, which is very different, and therefore cannot be optimum.

Of course, the minimum of $\mathbf{E} \exp\{\alpha \sum_i (U_i - V_i)^2\}$ can be approached by separate source- and channel coding, defined on blocks of super-letters formed by n -tuples. The source encoder is an optimum rate-distortion code for the above defined ‘single-letter’ distortion measure, operating at a rate close to the channel capacity, and the channel code is constructed accordingly to support the same rate.

4 Universal Asymptotically Optimum Strategies

The optimum strategy for minimizing $\mathbf{E}_P \exp\{\alpha \ell(X, s)\}$ depends, in general, on both P and α . It turns out, however, that this dependence on P and α can sometimes be relaxed if one gives up the ambition of deriving a strictly optimum strategy, and resorts to asymptotically optimum strategies.

Consider the case where, instead of one random variable X , we have a random vector $\mathbf{X} = (X_1, \dots, X_n)$, governed by an product form probability function

$$P(\mathbf{x}) = \prod_{i=1}^n P(x_i), \quad (34)$$

where each component x_i of the vector $\mathbf{x} = (x_1, \dots, x_n)$ takes on values in a finite set $\mathcal{X}$. If the $\ell(\mathbf{x}, s)$ grows linearly³ with n for a given empirical distribution of $\mathbf{x}$ and a given $s \in \mathcal{S}$, then it is expected that the exponential moment $\mathbf{E} \exp\{\alpha \ell(\mathbf{x}, s)\}$ would behave, at least asymptotically, as an exponential function of n . In particular, for a given s , the limit

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbf{E}_P \exp\{\alpha \ell(\mathbf{X}, s)\}$$

exists. Let us denote this limit by $E(s, \alpha, P)$. An *asymptotically optimum* strategy is then a strategy s^* for which

$$E(s^*, \alpha, P) \leq E(s, \alpha, P) \quad (35)$$

for every $s \in \mathcal{S}$. An asymptotically optimum strategy s^* is called *universal asymptotically optimum* w.r.t. a class $\mathcal{P}$ of probability distributions, if s^* is independent of α and P , yet it satisfies eq. (35) for all α in the allowed range, every $s \in \mathcal{S}$, and every $P \in \mathcal{P}$. In this section, we take $\mathcal{P}$ to be the class of all memoryless sources with a given finite alphabet $\mathcal{X}$. We denote by T_Q the type class pertaining to an empirical distribution Q , namely, the set of vectors $\mathbf{x} \in \mathcal{X}^n$ whose empirical distribution is Q .

³This happens, for example, when ℓ is additive, i.e., $\ell(\mathbf{x}, s) = \sum_{i=1}^n \ell(x_i, s)$.

Suppose there exists a strategy s^* and a function $\lambda : \mathcal{P} \rightarrow \mathbb{R}$ such that following two conditions hold:

- (a) For every type class T_Q and every $\mathbf{x} \in T_Q$, $\ell(\mathbf{x}, s^*) \leq n[\lambda(Q) + o(n)]$, where $o(n)$ designates a (positive) sequence that tends to zero as $n \rightarrow \infty$.
- (b) For every type class T_Q and every $s \in \mathcal{S}$,

$$\left| T_Q \cap \{\mathbf{x} : \ell(\mathbf{x}, s) \geq n[\lambda(Q) - o(n)] \} \right| \geq e^{-no(n)} |T_Q|. \quad (36)$$

It is then a straightforward exercise to show, using the method of types, that s^* is a universal asymptotically optimum strategy w.r.t. $\mathcal{P}$, with

$$E(s^*, \alpha, P) = \max_Q [\alpha \lambda(Q) - D(Q \| P)], \quad (37)$$

where condition (a) supports the direct part and condition (b) supports the converse part. The interesting point here then is not quite in the last statement, but in the fact that there are quite a few application examples where these two conditions hold at the same time.

Before we provide such examples, however, a few words are in order concerning conditions (a) and (b). Condition (a) means that there is a choice of s^* , that does *not* depend on $\mathbf{x}$ or on its type class,⁴ yet the performance of s^* , for every $\mathbf{x} \in T_Q$, “adapts” to the empirical distribution Q of $\mathbf{x}$ in a way, that according to condition (b), is “essentially optimum” (i.e., cannot be improved significantly), at least for a considerable (non-exponential) fraction of the members of T_Q . It is instructive to relate conditions (a) and (b) above to conditions 1 and 2 of Theorem 1. First, observe that in order to guarantee asymptotic optimality of s^* , condition 2 of Theorem 1 can be somewhat relaxed: For Jensen’s inequality in (4) to remain exponentially tight, it is no longer necessary to make the random variable $\alpha \ell(\mathbf{X}, s) + \ln[P(\mathbf{X})/Q(\mathbf{X})]$ completely degenerate (i.e., a constant for every realization $\mathbf{x}$, as in condition 2 of Theorem 1), but it is enough to keep it essentially fixed across a considerably large subset of the dominant type class, T_{Q^*} , i.e., the one whose empirical distribution Q^* essentially achieves the maximum of $[\alpha \lambda(Q) - D(Q \| P)]$. Taking $Q^*(\mathbf{x})$ to be the memoryless source induced by the dominant Q^* , this is indeed precisely what happens under

⁴As before, s^* is chosen without observing the data first.

conditions (a) and (b), which imply that

$$\begin{aligned}
\alpha \ell(\mathbf{x}, s^*) + \ln \frac{P(\mathbf{x})}{Q^*(\mathbf{x})} &\approx n\alpha\lambda(Q) + \sum_{i=1}^n \ln \frac{P(x_i)}{Q^*(x_i)} \\
&= n\alpha\lambda(Q) + n \sum_{x \in \mathcal{X}} Q^*(x) \ln \frac{P(x)}{Q^*(x)} \\
&= n[\alpha\lambda(Q^*) - D(Q^* \| P)],
\end{aligned} \tag{38}$$

for (at least) a non-exponential fraction of the members of T_{Q^*} , namely, a subset of T_{Q^*} that is large enough to maintain the exponential order of the (dominant) contribution of T_{Q^*} to $\mathbf{E} \exp\{\alpha \ell(\mathbf{x}, s^*)\}$. Loosely speaking, the combination of conditions (a) and (b) also means then that s^* is essentially optimum for (this subset of) T_{Q^*} , which is a reminiscence of condition 1 of Theorem 1. Moreover, since s^* “adapts” to every T_Q , in the sense explained above, then this has the flavor of the max-min problem discussed in Section 2, where s is allowed to be optimized for each and every Q . Since the minimizing s , in the max-min problem, is independent of P and α , this also explains the universality property of such a strategy.

Let us now discuss a few examples. The first example is that of fixed-rate rate-distortion coding. A vector $\mathbf{X}$ that emerges from a memoryless source P is to be encoded by a coding scheme s with respect to a given additive distortion measure, based on a single-letter distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}$, $\hat{\mathcal{X}}$ being the reconstruction alphabet. Let $D_Q(R)$ denote the distortion-rate function of a memoryless source Q (with a finite alphabet $\mathcal{X}$) relative to the single-letter distortion measure d and let $\ell(\mathbf{x}, s)$ designate the distortion between the source vector $\mathbf{x}$ and its reproduction, using a rate-distortion code s . It is not difficult to see that this example meets conditions (a) and (b) with $\lambda(Q) = D_Q(R)$: Condition (a) is based on the type covering lemma [7, Section 2.4], according to which each type class T_Q can be completely covered by essentially e^{nR} ‘spheres’ of radius $nD_Q(R)$ (in the sense of d), centered at the reproduction vectors. Thus s^* can be chosen to be a scheme that encodes $\mathbf{x}$ in two parts, the first of which is a header that describes the index of the type class T_Q of $\mathbf{x}$ (whose description length is proportional to $\log n$) and the second part encodes the index of the codeword within T_Q , using nR nats. Condition (b) is met since there is no way to cover T_Q with exponentially less than e^{nR} spheres within distortion less than $D_Q(R)$.

By the same token, consider the dual problem of variable-rate coding within a maximum allowed distortion D . In this case, every source vector $\mathbf{x}$ is encoded by $\ell(\mathbf{x}, s)$ nats, and this time, conditions

(a) and (b) apply with the choice $\lambda(Q) = R_Q(D)$, which is the rate–distortion function of Q (the inverse function of $D_Q(R)$). The considerations are similar to those of the first example. It is interesting to particularize this example, of variable–rate coding, to the lossless case, $D = 0$ (thus revisiting Example 1), where $R_Q(0) = H_Q$, the empirical entropy associated with Q . In this case, a more refined result can be obtained, which extends a well known result due to Rissanen [22] in universal data compression: According to [22], given a length function of a lossless data compression $\ell(\mathbf{x}, s)$ (s being the data compression scheme), and given a parametric class of sources $\{P_\theta\}$, indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$, a lower bound on $\mathbf{E}_\theta \ell(\mathbf{X}, s)$, that applies to most⁵ values of θ , is given by

$$\mathbf{E}_\theta \ell(\mathbf{X}, s) \geq nH_\theta + (1 - \epsilon) \frac{k}{2} \log n, \quad (39)$$

where $\epsilon > 0$ is arbitrarily small (for large n), H_θ is the entropy associated with P_θ , and $\mathbf{E}_\theta\{\cdot\}$ is the expectation under P_θ . On the other hand, the same expression is achievable, by a number of universal coding schemes, provided that the factor $(1 - \epsilon)$ in the above expression is replaced by $(1 + \epsilon)$. Consider now the case where $\{P_\theta, \theta \in \Theta\}$ is the class of all memoryless sources over $\mathcal{X}$, where the parameter vector θ designates $k = |\mathcal{X}| - 1$ letter probabilities. As for a lower bound, we have

$$\begin{aligned} \ln \mathbf{E}_P \exp\{\alpha \ell(\mathbf{X}, s)\} &\geq \max_Q [\alpha \mathbf{E}_Q \ell(\mathbf{X}, s) - nD(Q\|P)] \\ &\geq \max_Q \left\{ \alpha \left[nH_Q + (1 - \epsilon) \frac{k}{2} \ln n \right] - nD(Q\|P) \right\} \\ &= n \max_Q [\alpha H_Q - D(Q\|P)] + \alpha(1 - \epsilon) \frac{k}{2} \ln n \\ &= n\alpha H_{1/(1+\alpha)}(P) + \alpha(1 - \epsilon) \frac{k}{2} \ln n, \end{aligned} \quad (40)$$

where the second line follows from Rissanen’s lower bound (for most sources), and where $H_u(P)$ is Rényi’s entropy of order u , namely,

$$H_u(P) = \frac{1}{1 - u} \ln \left[\sum_{x \in \mathcal{X}} P(x)^u \right]. \quad (41)$$

Consider now a two–part code s^* , which first encodes the index of the type class Q and then the index of $\mathbf{x}$ within the type class. The corresponding length function is given by

$$\ell(\mathbf{x}, s^*) = \ln |T_Q| + k \ln n \approx n\hat{H}(\mathbf{x}) + \frac{k}{2} \ln n, \quad (42)$$

⁵“Most values of θ ” means all values of θ with the possible exception of a subset of Θ whose Lebesgue measure tends to zero as n tends to infinity.

where $\hat{H}(\mathbf{x})$ is the empirical entropy pertaining to $\mathbf{x}$, and where the approximate inequality is easily obtained by the Sterling approximation. Then,

$$\begin{aligned}
\ln \mathbf{E}_P \exp\{\alpha \ell(\mathbf{X}, s)\} &= \ln \mathbf{E} \exp\{\alpha n \hat{H}(\mathbf{X})\} + \alpha \frac{k}{2} \ln n \\
&= \ln \mathbf{E}_P \exp\{\alpha \min_Q [-\ln Q(\mathbf{X})]\} + \alpha \frac{k}{2} \ln n \\
&\leq \min_Q \ln \mathbf{E}_P \exp\{-\alpha \ln Q(\mathbf{X})\} + \alpha \frac{k}{2} \ln n \\
&= n\alpha H_{1/(1+\alpha)}(P) + \alpha \frac{k}{2} \ln n,
\end{aligned} \tag{43}$$

and then it essentially achieves the lower bound. Rissanen's result is now a special case of this, corresponding to $\alpha \rightarrow 0$.

Our last example corresponds to a secrecy system. A sequence $\mathbf{x}$ is to be communicated to a legitimate decoder which shares with the transmitter a random key $\mathbf{z}$ of nR purely random bits. The encoder transmits an encrypted message $\mathbf{y} = \phi(\mathbf{x}, \mathbf{z})$, which is an invertible function of $\mathbf{x}$ given $\mathbf{z}$, and hence decipherable by the legitimate decoder. An eavesdropper, which has no access to the key $\mathbf{z}$, submits a sequence of guesses concerning $\mathbf{x}$ until it receives an indication that the last guess was correct (e.g., a correct guess of a password admits the eavesdropper into a secret system). For the best possible encryption function ϕ , what would be the optimum guessing strategy s^* that the eavesdropper may apply in order to minimize the α -th moment of the number of guesses $G(\mathbf{X}, s)$, i.e., $\mathbf{E}\{G^\alpha(\mathbf{X}, s)\}$? In this case, $\ell(\mathbf{x}, s) = \ln G(\mathbf{x}, s)$. As is shown in [17], there exists a guessing strategy s^* , which for every $\mathbf{x} \in T_Q$, gives $\ell(\mathbf{x}, s^*) \approx n \min\{H_Q, R\}$, a quantity that essentially cannot be improved upon by any other guessing strategy, for most members of T_Q . In other words, conditions (a) and (b) apply with $\lambda(Q) = \min\{H_Q, R\}$.

5 Phase Transitions

Another interesting aspect of the asymptotic behavior of the exponential moment is the possible appearance of phase transitions, i.e., irregularities in the exponent function $E(s, \alpha, P)$ even in some very simple and 'innocent' models. By irregularities, we mean a non-smooth behavior, namely, discontinuities in the derivatives of $E(s, \alpha, P)$ with respect to α and/or the parameters of the source P .

One example that exhibits phase transitions is that of the secrecy system, mentioned in the last paragraph of the previous section. As is shown in [17], the optimum exponent $E(s^*, \alpha, p)$ for this case consists of two phase transitions as a function of R (namely, three different phases). In particular,

$$E(s^*, \alpha, P) = \begin{cases} \alpha R & R < H(P) \\ (\alpha - \theta_R)R + \theta_R H_{1/(1+\theta_R)}(P) & H(P) \leq R \leq H(P_\alpha) \\ \alpha H_{1/(1+\alpha)}(P) & R > H(P_\alpha) \end{cases} \quad (44)$$

where P_α is the distribution defined by

$$P_\alpha(x) \triangleq \frac{P^{1/(1+\alpha)}(x)}{\sum_{x' \in \mathcal{X}} P^{1/(1+\alpha)}(x')}, \quad (45)$$

$H(Q)$ is the Shannon entropy associated with a distribution Q , $H_u(Q)$ is the Rényi entropy of order u as defined before, and θ_R is the unique solution of the equation $R = H(P_\theta)$ for R in the range $H(P) \leq R \leq H(P_\alpha)$. But this example may not really be extremely surprising due to the non-smoothness of the function $\lambda(Q) = \min\{H_Q, R\}$.

It may be somewhat less expected, however, to witness phase transitions also in some very simple and ‘innocent’ looking models. One way to understand the phase transitions in these cases, comes from the statistical-mechanical perspective. It turns out that in some cases, the expression of the exponential moment is analogous to that of a partition function of a certain many-particle physical system with interactions, which may exhibit phase transitions and these phase transitions correspond the above-mentioned irregularities.

We now demonstrate a very simple model, which has phase transitions. Consider the case where $\mathbf{X}$ is a binary vector whose components take on values in $\mathcal{X} = \{-1, +1\}$, and which is governed by a binary memoryless source P_μ with probabilities $\Pr\{X_i = +1\} = 1 - \Pr\{X_i = -1\} = (1 + \mu)/2$ (μ designating the expected ‘magnetization’ of each binary spin X_i , to make the physical analogy apparent). The probability of $\mathbf{x}$ under P_μ is thus easily shown to be given by

$$P_\mu(\mathbf{x}) = \left(\frac{1+\mu}{2}\right)^{(n+\sum_i x_i)/2} \cdot \left(\frac{1-\mu}{2}\right)^{(n-\sum_i x_i)/2} = \left(\frac{1-\mu^2}{4}\right)^{n/2} \cdot \left(\frac{1+\mu}{1-\mu}\right)^{\sum_i x_i/2}. \quad (46)$$

Consider the estimation of the parameter μ by the ML estimator

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (47)$$

How does the exponential moment of $\mathbf{E}_\mu \exp\{\alpha n(\hat{\mu} - \mu)^2\}$ behave like? A straightforward derivation yields

$$\begin{aligned} \mathbf{E}_\mu \exp\{\alpha n(\hat{\mu} - \mu)^2\} &= \left(\frac{1 - \mu^2}{4}\right)^{n/2} e^{n\alpha\mu^2} \sum_{\mathbf{x}} \left(\frac{1 + \mu}{1 - \mu}\right)^{\sum_i x_i/2} \exp\left\{\frac{\alpha}{n} \left(\sum_i x_i\right)^2 - 2\alpha\mu \sum_i x_i\right\} \\ &= \left(\frac{1 - \mu^2}{4}\right)^{n/2} e^{n\alpha\mu^2} \sum_{\mathbf{x}} \exp\left\{\left(\frac{1}{2} \ln \frac{1 + \mu}{1 - \mu} - 2\alpha\mu\right) \sum_i x_i + \frac{\alpha}{n} \left(\sum_i x_i\right)^2\right\}. \end{aligned}$$

The last summation over $\{\mathbf{x}\}$ is exactly the partition function pertaining to the *Curie–Weiss model* of spin arrays in statistical mechanics (see, e.g., [21, Subsection 2.5.2]), where the magnetic field is given by

$$B = \frac{1}{2} \ln \frac{1 + \mu}{1 - \mu} - 2\alpha\mu \quad (48)$$

and the coupling coefficient for every pair of spins is $J = 2\alpha$. It is well known that this model exhibits phase transitions pertaining to spontaneous magnetization below a certain critical temperature. In particular, using the method of types [7], this partition function can be asymptotically evaluated as being of the exponential order of

$$\exp\left\{n \cdot \max_{|m| \leq 1} \left[h_2\left(\frac{1 + m}{2}\right) + Bm + \frac{J}{2} \cdot m^2\right]\right\},$$

where $h_2(\cdot)$ is the binary entropy function, which stands for the exponential order of the number of configurations $\{\mathbf{x}\}$ with a given value of $m = \frac{1}{n} \sum_i x_i$. This expression is clearly dominated by a value of m (the dominant magnetization m^*) which maximizes the expression in the square brackets, i.e., it solves the equation

$$m = \tanh(Jm + B), \quad (49)$$

or in our variables,

$$m = \tanh\left(2\alpha m + \frac{1}{2} \ln \frac{1 + \mu}{1 - \mu} - 2\alpha\mu\right). \quad (50)$$

For $\alpha < 1/2$, there is only one solution and there is no spontaneous magnetization (paramagnetic phase). For $\alpha > 1/2$, however, there are three solutions, and only one of them dominates the partition function, depending on the sign of B , or equivalently, on whether $\alpha > \alpha_0(\mu) \triangleq \frac{1}{4\mu} \ln \frac{1 + \mu}{1 - \mu}$ or $\alpha < \alpha_0(\mu)$ and according to the sign of μ . Accordingly, there are five different phases in the plane spanned by α and μ . The paramagnetic phase $\alpha < 1/2$, the phases $\{\mu > 0, 1/2 < \alpha < \alpha_0(\mu)\}$ and

$\{\mu < 0, \alpha > \alpha_0(\mu)\}$, where the dominant magnetization m is positive, and the two complementary phases, $\{\mu < 0, 1/2 < \alpha < \alpha_0(\mu)\}$ and $\{\mu > 0, \alpha > \alpha_0(\mu)\}$, where the dominant magnetization is negative. Thus, there is a multi-critical point where the boundaries of all five phases meet, which is the point $(\mu, \alpha) = (0, 1/2)$. The phase diagram is depicted in Fig. 1.

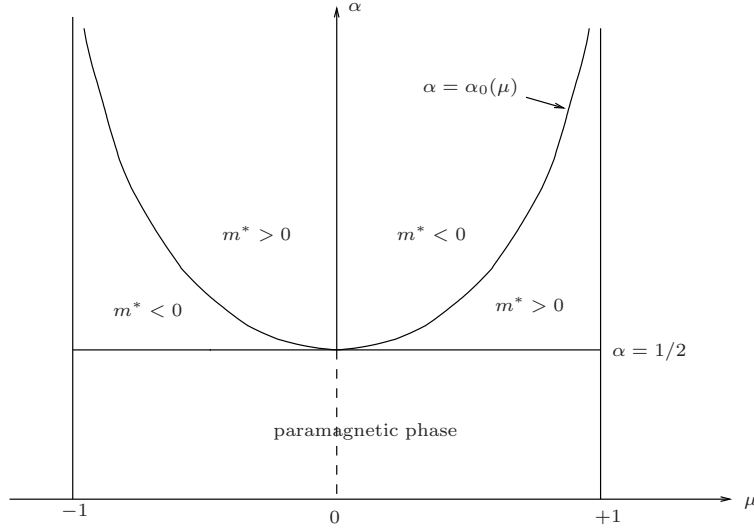


Figure 1: Phase diagram in the plane of (μ, α) .

Yet another example of phase transitions is that of fixed-rate lossy data compression, discussed in the previous section. To demonstrate this explicitly, consider the binary symmetric source (BSS) and the Hamming distortion measure d , and consider a random selection of a rate- R code by ne^{nR} independent fair coin tosses, one for each of the n components of every one of the e^{nR} codewords. It was shown in [16] that the asymptotic exponent of the negative exponential moment, $\mathbf{E} \exp\{-\alpha \sum_i d(U_i, V_i)\}$ (where the expectation is w.r.t. both the source and the random code selection), is given by the following expression, which obviously exhibits a (second order) phase transition:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbf{E} \exp \left\{ -\alpha \sum_i d(U_i, V_i) \right\} = \begin{cases} -\alpha \delta(R) & \alpha \leq \alpha(R) \\ -\alpha + \ln(1 + e^\alpha) + R - \ln 2 & \alpha > \alpha(R) \end{cases} \quad (51)$$

where $\delta(R)$ is the distortion-rate function of the BSS w.r.t. the Hamming distortion measure and

$$\alpha(R) = \ln \frac{1 - \delta(R)}{\delta(R)}. \quad (52)$$

The analysis in [16] is based on the random energy model (REM), [8],[9],[10], a well-known statistical-mechanical model of spin glasses with strong disorder, which is known to exhibit phase transitions.

Moreover, it is shown in [16] that ensembles of codes that have an hierarchical structure may have more than one phase transition.

6 Lower Bounds on Exponential Moments

As explained in the Introduction, even in the ordinary setting, of the quest for minimizing $\mathbf{E}\{\ell(X, s)\}$, optimum strategies may not always be known, and then useful lower bounds are very important. This is definitely the case when exponential moments are considered, because the exponential moment criterion is even harder to handle. To obtain non-trivial bounds on exponential moments, we propose to harness lower bounds the expectation of $\ell(X, s)$, possibly using a change of measure, in the spirit of the proof of Theorem 1 and the previous example of a lower bound on universal lossless data compression. We next demonstrate this idea in the context of a lower bound on the expected exponentiated squared error of an unbiased estimator, on the basis of the Cramér–Rao bound (CRB). The basic idea, however, is applicable more generally, e.g., by relying on other well-known Bayesian/non-Bayesian bounds on the mean-square error (e.g., the Weiss–Weinstein bound for Bayesian estimation [25]), as well as in bounds on signal estimation (filtering, prediction, etc.), and in other problem areas as well. Further investigation in the line may be of considerable interest.

Consider a parametric family of probability distributions $\{P_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$ being the parameter set, and suppose that we are interested in a lower bound on $\mathbf{E}_\theta \exp\{\alpha(\hat{\theta} - \theta)^2\}$, for any unbiased estimator of θ , where as before, $\mathbf{E}_\theta$ denotes expectation w.r.t. P_θ . Consider the following chain of inequalities, which holds for any $\theta' \in \Theta$:

$$\begin{aligned} \mathbf{E}_\theta \exp\{\alpha(\hat{\theta} - \theta)^2\} &= \mathbf{E}_{\theta'} \exp\left\{\alpha(\hat{\theta} - \theta)^2 + \ln \frac{P_\theta(X)}{P_{\theta'}(X)}\right\} \\ &\geq \exp\left\{\alpha \mathbf{E}_{\theta'}(\hat{\theta} - \theta)^2 - D(P_{\theta'} \| P_\theta)\right\} \\ &= \exp\left\{\alpha \mathbf{E}_{\theta'}(\hat{\theta} - \theta')^2 + \alpha(\theta - \theta')^2 - D(P_{\theta'} \| P_\theta)\right\} \\ &\geq \exp\left\{\alpha \text{CRB}(\theta') + \alpha(\theta - \theta')^2 - D(P_{\theta'} \| P_\theta)\right\}, \end{aligned} \quad (53)$$

where $\text{CRB}(\theta)$ is the Cramér–Rao bound for unbiased estimators, computed at θ (i.e., $\text{CRB}(\theta) = 1/I(\theta)$, where $I(\theta)$ is the Fisher information). Since this lower bound applies for every $\theta' \in \Theta$, one can take its supremum over $\theta' \in \Theta$ and obtain

$$\ln \mathbf{E}_\theta \exp\{\alpha(\hat{\theta} - \theta)^2\} \geq \sup_{\theta' \in \Theta} [\alpha \text{CRB}(\theta') + \alpha(\theta' - \theta)^2 - D(P_{\theta'} \| P_\theta)]. \quad (54)$$

More generally if $\theta = (\theta_1, \dots, \theta_k)^T$ is a parameter vector (thus $\theta \in \Theta \subseteq \mathbb{R}^k$) and $\alpha \in \mathbb{R}^k$ is an arbitrary deterministic (column) vector, then

$$\ln \mathbf{E}_\theta \exp\{\alpha^T(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \alpha\} \geq \sup_{\theta' \in \Theta} [\alpha^T I^{-1}(\theta') \alpha + [\alpha^T(\theta' - \theta)]^2 - D(P_{\theta'} \| P_\theta)], \quad (55)$$

where here $I(\theta)$ is the Fisher information matrix and $I^{-1}(\theta)$ is its inverse.

It would be interesting to further investigate bounds of this type, in parameter estimation in particular, and in other problem areas in general, and to examine when these bounds may be tight and useful.

Acknowledgment

Interesting discussions with Rami Atar are acknowledged with thanks.

References

- [1] E. Arikan, “An inequality on guessing and its application to sequential decoding,” *IEEE Trans. Inform. Theory*, vol. IT-42, no. 1, pp. 99–105, January 1996.
- [2] E. Arikan and N. Merhav, “Guessing subject to distortion,” *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.
- [3] E. Arikan and N. Merhav, “Joint source–channel coding and guessing with application to sequential decoding,” *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1756–1769, September 1998.
- [4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice Hall, Englewood Cliffs, New Jersey, U.S.A., 1971.
- [5] D. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I and II, 3rd ed. Nashua, NH: Athena Scientific, 2007.
- [6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second Edition, John Wiley & Sons, Hoboken, New Jersey, U.S.A., 2006.

- [7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, U.S.A., 1981.
- [8] B. Derrida, “Random–energy model: limit of a family of disordered models,” *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, July 1980.
- [9] B. Derrida, “The random energy model,” *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.
- [10] B. Derrida, “Random–energy model: an exactly solvable model for disordered systems,” *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.
- [11] N. Farvardin and J. W. Modestino, “On overflow and underflow problems in buffer instrumented variable-length coding of fixed-rate memoryless sources,” *IEEE Transactions on Information Theory*, vol. IT–32, no. 6, pp. 839–845, November 1986.
- [12] P. A. Humblet, “Generalization of Huffman coding to minimize the probability of buffer overflow,” *IEEE Transactions on Information Theory*, vol. IT–27, no. 2, pp. 230–232, March 1981.
- [13] F. Jelinek, “Buffer overflow in variable length coding of fixed rate sources,” *IEEE Transactions on Information Theory*, vol. IT–14, no. 3, pp. 490–501, May 1968.
- [14] J. L. Massey, “Guessing and entropy,” *Proc. 1994 IEEE Int. Symp. Inform. Theory*, (ISIT ‘94), p. 204, Trondheim, Norway, 1994.
- [15] N. Merhav, “Universal coding with minimum probability of code word length overflow,” *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 556–563, May 1991.
- [16] N. Merhav, “The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes,” *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1250–1268, March 2009.
- [17] N. Merhav and E. Arikan, “The Shannon cipher system with a guessing wiretapper,” *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1860–1866, September 1999.

- [18] N. Merhav and M. Feder, “Universal prediction,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.
- [19] N. Merhav and D. L. Neuhoff, “Variable-to-fixed length codes provide better large deviations performance than fixed-to-variable length codes,” *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 135–140, January 1992.
- [20] N. Merhav, R. M. Roth, and E. Arikan, “Hierarchical guessing with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 330–337, January 1999.
- [21] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, 2009.
- [22] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Transactions on Information Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
- [23] O. Uchida and T. S. Han, “The optimal overflow and underflow probabilities with variable-length coding for the general source,” preprint 1999.
- [24] H. van Trees, *Detection, Estimation, and Modulation Theory*, Part I, John Wiley & Sons, New York, 1968.
- [25] A. J. Weiss and E. Weinstein, “A lower bound on the mean square error in random parameter estimation,” *IEEE Trans. Inform. Theory*, vol. IT-31, no. 5, pp. 680–682, September 1985.
- [26] A. D. Wyner, “On the probability of buffer overflow under an arbitrary bounded input-output distribution,” *SIAM Journal on Applied Mathematics*, vol. 27, no. 4, pp. 544–570, December 1974.
- [27] A. D. Wyner, “Another look at the coding theorem of information theory- a tutorial,” *Proc. IEEE*, vol. 58, no. 6, pp. 894–913, June 1980.
- [28] A. D. Wyner, “Fundamental limits in information theory,” *Proc. IEEE*, vol. 69, no. 2, pp. 239–251, February 1981.