Communications in Information and Systems
Vol. 11, No. 4, pp. 343-368, 2011

003

ON OPTIMUM STRATEGIES FOR MINIMIZING THE EXPONENTIAL MOMENTS OF A LOSS FUNCTION

NERI MERHAV*

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, the basic observation that we make and use is about simple sufficient conditions for a strategy to be optimum in the exponential moment sense. This observation 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. 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, 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., [11, Chapters 5 and 10]). (ii) Gambling and portfolio theory [11, 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., [60],[61]). (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., [54]). (v) Prediction, sequential decision problems (see, for example, [42]) and stochastic control problems [7], 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,

*Department of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 32000, ISRAEL, E-mail: merhav@ee.technion.ac.il

$\boldsymbol{E}\exp\{\alpha\ell(X,s)\}$ ($\alpha > 0$), have received much less attention than they probably deserve in this context, at least in information theory and signal processing. In the realm of the theory of optimization and stochastic control, on the other hand, the problem of minimizing exponential moments has received much more attention, and it is well–known as the *risk–sensitive* or *risk–averse* cost function (see, e.g., [15], [18], [23], [28], [56], [57] and many references therein), where one of the main motivations for using the exponential function of $\ell(X,s)$ is to impose a penalty, or a risk, that is extremely sensitive to large values of $\ell(X,s)$, hence the qualifier "risk–sensitive" in the name of this criterion. Another motivation is associated with robustness properties of the resulting risk–sensitive optimum controllers [4], [20]. There are, in fact, a few additional motivations for examining strategies that minimize exponential moments, which are also relevant to many problem areas of information theory, communications and statistical signal processing. First and foremost, the exponential moment, $\boldsymbol{E}\exp\{\alpha\ell(X,s)\}$, 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 $\boldsymbol{E}\exp\{\alpha\ell(X,s)\}$ 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.[1] 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 [22], [29], [31], [37], [44], [53], [59], exponential moments related to guessing [1], [2], [3], [35], [40], [45], large deviations properties of parameter/signal estimators, [8], [32], [46], [51], [52], [62], and more.

It is natural to ask, in view of the foregoing discussion, how we can 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 basic observation, in this paper, provides a simple relationship between the two criteria. In particular, in Section 2, 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

[1] Uniform minimization of $\boldsymbol{E}\exp\{\alpha\ell(X,s)\}$ by a strategy s^* for all $\alpha \geq 0$ is stronger than minimization of the first moment $\boldsymbol{E}\ell(X,s)$ in the sense that the former implies the latter but the converse is not necessarily true: Let $f(\alpha) = \boldsymbol{E}\exp\{\alpha\ell(X,s^*)\}$ and $g(\alpha) = \boldsymbol{E}\exp\{\alpha\ell(X,s)\}$ for an arbitrary s. Since $f(0) = g(0) = 1$, the fact that $f(\alpha) \leq g(\alpha)$ for all $\alpha \geq 0$, implies $f'(0) \leq g'(0)$, which are the corresponding first moments (provided that the derivatives exist). Moreover if $f'(0) = g'(0)$, then $f''(0) \leq g''(0)$, which are the second moments, etc.

sense, for a possibly different probability distribution, which we characterize.

The main message of this expository paper is that the combination of these sufficient conditions sets the stage for a useful tool that can be used to solve concrete problems of minimizing exponential moments, and it may give a fresh look and a new insight into these problems. 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. In other applications, however, this may not be quite the case directly, yet the set of optimality conditions may still be useful: 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.

At the heart of this paper stands a section of application examples (Section 3). One example for the use of the proposed tool, that will be demonstrated in detail (and in more generality) later on, is the following: Given n independent and identically distributed (i.i.d.) Gaussian measurements, $X_1, \ldots, X_n$, with mean θ, the sample mean, $s(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, 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., $\boldsymbol{E}\exp\{\alpha[s(X_1, \ldots, X_n) - \theta]^2\}$ for all $\alpha \geq 0$ for which this expectation is finite. Another example belongs to the realm of universal lossless data compression and the famous minimum description length principle (MDL), due to Rissanen [50], who characterized the minimum achievable "price of universality" as a redundancy term of about $\frac{1}{2}\log n$ (without normalization) for each unknown parameter of the information source to be compressed. Here we show how this result (both the converse theorem and the direct theorem) extends from the expected code–length criterion to all exponential moments of the code–length. Yet another example is the memoryless Gaussian joint source–channel coding problem with the quadratic distortion measure and without bandwidth expansion: As is well known, minimum expected distortion can be achieved by optimum linear scalar encoders and decoders. When it comes to exponential moments of the distortion, the behavior is somewhat more complicated. As long as one forces a linear scalar encoder, the optimum decoder is also linear. However, once the constraint of linear encoding is relaxed, both the optimum encoder and the optimum decoder are no longer linear (see also [13] for a related work). Our above–mentioned basic principle sheds some insight on this problem too.

We next devote some attention to the asymptotic regime (Section 4). Consider the case where X is a random vector of dimension n, $X = (X_1, \ldots, X_n)$, 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., $\ell(X, s) = \sum_{i=1}^{n} l(X_i, s)$. 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"[2] to the empirical distribution of $(X_1, \ldots, X_n)$, 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 the above–mentioned extension of Rissanen's famous results in universal data compression [50].

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 (Section 5): Even in the ordinary setting, of seeking strategies that minimize $\boldsymbol{E}\{\ell(X,s)\}$, 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 on 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. The Basic Observation.

Let X be a random variable taking on values in a certain alphabet $\mathcal{X}$, and drawn according to a given probability distribution P. The alphabet $\mathcal{X}$ may either be finite, countable, or a continuous set. In the latter case, P (as well as other probability functions on $\mathcal{X}$) denotes a density with respect to a certain measure μ on $\mathcal{X}$, say the counting measure in the discrete case, or the Lebesgue measure in the continuous case. Let the variable s designate a *strategy*, or an *action*, chosen from some space $\mathcal{S}$ of allowed strategies. The term "strategy", in our context, means a mathematical object that, depending on the application, may be either a scalar variable, a vector, an infinite sequence, a function (of X), or a function of another random variable/vector that is statistically dependent on X. 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 $\boldsymbol{E}\{\cdot\}$ 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 $\boldsymbol{E}_Q\{\cdot\}$. 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 $\boldsymbol{E}_P\{\cdot\}$.

[2] The precise meaning of this will be clarified in the sequel.

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

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

$$Z(s) \stackrel{\Delta}{=} \boldsymbol{E}_P \exp\{\alpha\ell(X,s)\} < \infty. \tag{1}$$

A strategy $s \in \mathcal{S}$ minimizes $\boldsymbol{E}_P \exp\{\alpha\ell(X,s)\}$ 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 $\boldsymbol{E}_Q\{\ell(X,s)\}$ over $\mathcal{S}$.*
2. *The probability distribution Q is given by*

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

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

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

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 Observation 1. Consider the following chain of inequalities:

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

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 $D(Q\|P) \stackrel{\Delta}{=} \boldsymbol{E}_Q \ln[Q(X)/P(X)]$ being the relative entropy between Q and P), the third line is by condition 1, and the next equality results from condition 2: On substituting $Q^*(x) = P(x)e^{\alpha\ell(x,s^*)}/Z(s^*)$ into $D(Q^*\|P)$, one readily obtains $D(Q^*\|P) = \alpha\boldsymbol{E}_{Q^*}\ell(X,s^*) - \ln Z(s^*)$. This completes the proof of Observation 1. □

Discussion: Several comments are in order at this point.

Partially related results have appeared in the literature of optimization and control (cf. [23, Theorem 4.9]). However, in [23], a much more complicated and more involved paradigm (of controlling finite–state Markov processes) has been considered, and the results therein do not seem to be equivalent to Observation 1. Moreover, since the setting here is much simpler, then so is the proof, which is not only short, but also almost free of regularity conditions (as opposed to [23] and [15]). The only regularity condition needed here is that $Z(s) < \infty$ for some $s \in \mathcal{S}$. Obviously, without this condition, the problem under consideration is meaningless and empty in the first place.

Note that for a given s, Jensen's inequality

$$\boldsymbol{E}_Q \exp\left\{\alpha\ell(X,s) + \ln\frac{P(X)}{Q(X)}\right\} \geq \exp\left\{\alpha\boldsymbol{E}_Q\ell(X,s) - D(Q\|P)\right\} \tag{5}$$

of (4) (but with Q^* being replaced by a generic measure Q), 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 the expression $\alpha\boldsymbol{E}_Q\ell(X,s) - D(Q\|P)$ is maximized by this choice of $Q(x) = P(x)e^{\alpha\ell(x,s)}/Z(s)$, a fact which can also be seen from a direct maximization of this expression using standard methods. This leads directly to the well–known identity (see, e.g., [15, Proposition 2.3]):

$$\boldsymbol{E}_P \exp\{\alpha\ell(X,s)\} = \exp\{\max_Q[\alpha\boldsymbol{E}_Q\ell(X,s) - D(Q\|P)]\}, \tag{6}$$

which is also intimately related to the well–known Laplace principle [19] in large deviations theory, or more generally, to Varadhan's integral lemma [17, Section 4.3].

In view of eq. (6), another look at the problem of minimizing $\boldsymbol{E}_P\exp\{\alpha\ell(X,s)\}$ reveals that it is equivalent[3] to the minimax problem

$$\min_s \max_Q F(s,Q) \tag{7}$$

where

$$F(s,Q) \stackrel{\Delta}{=} \alpha\boldsymbol{E}_Q\ell(X,s) - D(Q\|P). \tag{8}$$

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

$$\min_{s\in\mathcal{S}} \max_Q F(s,Q) = \max_Q \min_{s\in\mathcal{S}} F(s,Q). \tag{9}$$

This equality between the minimax and the maximin means that there is a saddle point (s^*, Q^*), where s^* is a solution of the minimax problem on the left–hand side and Q^* is a solution to the maximin problem on the right–hand side. As mentioned above, 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 Observation 1. By the same token, the

[3] See also [15].

inner minimization over s on the right–hand side obviously minimizes $\boldsymbol{E}_{Q^*}\ell(X,s)$, which is condition 1. This means then that the two conditions of Observation 1 are actually equivalent to the conditions for the existence of a saddle point of $F(s,Q)$. This can be considered as an alternative proof of Observation 1. The original proof above, however, is much simpler: Not only is it shorter, but furthermore, it requires neither acquaintance with eq. (6) nor with the theory of minimax/maximin optimization and saddle points.

When does eq. (9) hold? In general, the well–known sufficient conditions for

$$\min_{u\in\mathcal{U}}\max_{v\in\mathcal{V}} f(u,v) = \max_{v\in\mathcal{V}}\min_{u\in\mathcal{U}} f(u,v) \tag{10}$$

are that $\mathcal{U}$ and $\mathcal{V}$ are convex sets (with u being allowed to take on values freely in $\mathcal{U}$, independently of v and vice versa), and that f is convex in u and concave in v. In our case, since the function $F(s,Q)$ 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. There are, of course, milder sufficient conditions that allow commutation of minimization and maximization (see, e.g,, [49, Chapters 36 and 37]).

Note that in retrospect, the minimax formulation of $F(s,Q)$ also suggests yet another motivation for the exponential moment minimization, which is *robustness against uncertainty* in P. Imagine a problem of first moment minimization of $\ell(X,s)$, where the *real* distribution of X, which is denoted Q, is uncertain, and it is only known to be at some neighborhood of a given nominal distribution P. The region of uncertainty is defined in terms under the "metric" $D(Q\|P)$, namely, it is the set of all distributions $\{Q\}$ such that $D(Q\|P)\le\epsilon$. In such an uncertainty situation, to be on the safe side, we opt to minimize the worst–case expected loss, which means

$$\min_{s}\max_{\{Q:\ D(Q\|P)\le\epsilon\}} \boldsymbol{E}_Q\ell(X,s). \tag{11}$$

The inner maximization problem of $F(s,Q)$ is nothing but the Lagrangian version of (and hence equivalent to) the inner maximization in the last expression, where α serves as a Lagrange multiplier. Thus, the choice of α corresponds to the radius of uncertainty ϵ around the nominal distribution P. This explains why risk–sensitive solutions are robust.

A similar, but somewhat different, criterion pertaining to exponential moments, which is reasonable to the same extent, is the dual problem of $\max_{s\in\mathcal{S}}\boldsymbol{E}\exp\{-\alpha\ell(X,s)\}$ (again, with $\alpha>0$), which is called, in the jargon of stochastic control, a *risk–seeking* cost criterion, as opposed to the risk-sensitive criterion discussed so far. 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 $\boldsymbol{E}\exp\{\alpha\ell(X,s)\}$ which, in many cases, is finite only for a limited range of α. For the same considerations as before, here we have:

$$\begin{aligned}\max_s \boldsymbol{E}\exp\{-\alpha\ell(X,s)\} &= \max_s \exp\{\max_Q[-\alpha\boldsymbol{E}_Q\ell(X,s) - D(Q\|P)]\}\\ &= \exp\{-\min_s \min_Q[\alpha\boldsymbol{E}_Q\ell(X,s) + D(Q\|P)]\},\end{aligned} \tag{12}$$

and so the optimality conditions relating s and Q are similar to those of Observation 1 (with α replaced by $-\alpha$), except that now we have a double minimization problem rather than a minimax problem. However, it should be noted that here the conditions of Observation 1 are only necessary conditions, as for the above equalities to hold, the pair (s, Q) should *globally* minimize the function $F(s, Q)$, unlike the earlier case, where only a saddle point was sought.[4] 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 $[\alpha\boldsymbol{E}_Q\ell(X,s) + D(Q\|P)]$ over Q), then one finds $s_1 = \arg\min_{s\in\mathcal{S}} \boldsymbol{E}_{Q_0}\{\ell(X,s)\}$, and so on. It is obvious that $\boldsymbol{E}\exp\{-\alpha\ell(X,s_i)\}$, $i = 0,1,2,\ldots$, increases (and hence improves) from iteration to iteration. This is different from the minimax situation we encountered earlier, where successive improvements are not guaranteed.

3. Applications.

Observation 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 proved optimum by checking the conditions. In yet some other cases, Observation 1 suggests an iterative algorithm for solving the problem. In this section, we will see a few application examples of all these types. Some of these examples could have been also solved directly (and/or the resulting conclusions may even already be known from the literature), without using Observation 1, but for others, this does not seem to be a trivial task. In any case, Observation 1 may suggest a fresh look and some new insight into the problem.

It should be emphasized once again that in all application examples below, the optimization of the exponential moments is motivated, as discussed in Sections 1 and 2, by several desirable features, which are summarized as follows: (i) Sensitivity to high risks. (ii) Robustness. (iii) Intimate relationship to large deviations performance. (iv) Stronger notion of optimality (compared to first moment only) whenever a single strategy s^* minimizes the exponential moment loss uniformly for all $\alpha \geq 0$. The example discussed in Subsection 3.2 is slightly exceptional since a risk–seeking

[4] In 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.

(rather than a risk–sensitive) cost function is used therein, and so, the corresponding motivations are somewhat different. The details will be given therein.

3.1. 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 $\ell(x, s) \stackrel{\triangle}{=} -\ln s(x)$. 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 $\boldsymbol{E}_Q\{-\ln s(X)\}$ is $s_Q = Q$. Thus, by eq. (3), we seek a distribution Q such that

$$Q(x) \propto P(x)\exp\{-\alpha \ln Q(x)\} = \frac{P(x)}{[Q(x)]^\alpha} \tag{13}$$

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

$$s_Q(x) = Q(x) = \frac{[P(x)]^{1/(1+\alpha)}}{\sum_{x'\in\mathcal{X}}[P(x')]^{1/(1+\alpha)}}. \tag{14}$$

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 Observation 1, our purpose in this example was to show how Observation 1 gives the desired solution very easily by solving a very simple equation. The resulting optimum exponential moment is well–known to be characterized in terms the Rényi entropy of order $1/(1+\alpha)$, namely,

$$\boldsymbol{E}e^{\alpha\ell(X,s_Q)} = e^{\alpha H_{1/(1+\alpha)}(P)} \tag{15}$$

where

$$H_{1/(1+\alpha)}(P) = \frac{1+\alpha}{\alpha}\ln\left(\sum_{x\in\mathcal{X}}[P(x)]^{1/(1+\alpha)}\right). \tag{16}$$

As expected (see footnote no. 1), the limit $\alpha \to 0$ recovers the ordinary Shannon entropy $H(P)$, which corresponds to the minimization of the first moment of $\ell(X, s)$. Note that the identity [1, eqs. (6)-(7)], [2, p. 1044, top]:

$$\alpha H_{1/(1+\alpha)}(P) = \sup_Q[\alpha H(Q) - D(Q\|P)] \tag{17}$$

suggests a possible interpretation that Observation 1 (in particular, $\min_s \max_Q F(s, Q)$) is actually always implicitly used when solving the above exponential moment problem.

The objective of minimizing exponential moments of lossless coding length functions has been studied in several papers in the literature, with the concrete motivation of its relation to buffer overflow problems, both in the lossless case and the lossy case (see, e.g., [22], [29], [31], [37], [43], [59] and references therein). Both in lossless and lossy source coding, exponential moments of length functions of long strings of symbols clearly serve as Chernoff bounds on probabilities of excess length in source coding. These Chernoff bounds become exponentially tight upon optimization of the parameter α, whose optimum value depends on the threshold above which the excess length event is defined. In the lossy case, the dual problem of characterizing the fastest achievable exponential rate of the probability of excess distortion was first addressed by Marton [34].

3.2. Quantization. Consider the problem of quantizing a real–valued random variable X, drawn by P, into M reproduction levels, $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{M-1}$, and let the distortion metric $d(x,\hat{x})$ be quadratic, i.e., $d(x,\hat{x}) = (x-\hat{x})^2$. The ordinary problem of optimum quantizer design is about the choice of a function $s : \mathcal{X} \to \{\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{M-1}\}$, that minimizes $\boldsymbol{E}_P[X - s(X)]^2$, i.e., in this case, $\ell(x,s) = [x - s(x)]^2$.

As is well known [25], [27], [33], in general, this problem lacks a closed–form solution, and the customary approach is to apply an iterative algorithm for quantizer design, a.k.a. the Lloyd–Max algorithm [36],[5] which alternates between two sets of necessary conditions for optimality: the nearest–neighbor condition, according to which $s(x)$ should be the reproduction level that is closest to x (i,e., the one that minimizes $(x - \hat{x}_i)^2$ over i), and the centroid condition, which means that $\hat{x}_i$ should be the conditional expectation of X given that X falls in the interval of values of x that are to be mapped to the i–th quantization level.

Consider now the criterion of maximizing the negative exponential moment $\boldsymbol{E}_P e^{-\alpha[X-s(X)]^2}$, i.e., the risk–seeking cost criterion, which was described and motivated in the last paragraph of Section 2. Here the centroid condition is no longer relevant since it is no longer a mean square error problem and the analogue to the centroid condition, when it comes to exponential moments, does not admit a closed–form expression in terms of the density of X and the given partition of the real line to quantization intervals. However, in light of the discussion in Section 2, one can use the fact that this problem is equivalent to the double minimization of

$$G(s,Q) \stackrel{\Delta}{=} \alpha \boldsymbol{E}_Q[X - s(X)]^2 + D(Q\|P) \tag{18}$$

over s and Q. This suggests an iterative algorithm, in the spirit of the Lloyd–Max algorithm, which consists of two nested loops: The outer loop alternates between minimizing s for a given Q, on the one hand, and minimizing Q for a given s, on the other hand. As explained in Section 2, these two minimizations are, in principle, nothing but the conditions of Observation 1, just with α being replaced by $-\alpha$. The

[5] The Lloyd–Max algorithm was first invented by Lloyd in 1957, but was not published at the time.

inner loop implements the former ingredient of minimizing $\boldsymbol{E}_Q[X-s(X)]^2$ over s for a given Q, which is again implementable by the standard iterative procedure that was described in the previous paragraph. Of course, it is not guaranteed that such an iterative algorithm would converge to the global minimum of $G(s,Q)$, just like in the case of ordinary iterative quantizer design.

While the minimax of $F(s,Q)$ was motivated in Section 2 as being equivalent to robust, worst–case minimization of the first moment of $\ell(X,s)$, the double minimization of $G(s,Q)$ can similarly be paralleled to *best–case* minimization, i.e.,

$$\min_s \min_{\{Q:\ D(Q\|P)\le\epsilon\}} \boldsymbol{E}_Q\ell(X,s). \tag{19}$$

This is relevant in situations where there is an available option of some supporting data pre–processing (before quantization) that provides a certain freedom to control and shape the distribution Q of the data X, so as to slightly deviate from the original distribution P within a limited radius ϵ in the divergence domain. This pre–processing may be of several possible forms, for example, a dithering scheme [30], a non–linearity (such as a hard–limiter or a compander, see e.g., [5], [9], [10], [24], [30, pp. 135–159]), or a data embedding algorithm (i.e., a watermarking or a data hiding scheme, see e.g., [12], [21]), just to name a few. While the minimax problem was motivated by situations where it is a hostile party that may control Q, the double minimization problem is motivated by situations where a friendly party shapes Q for us.

As a simple example for a combination of s and Q that are matched (in the above sense), consider the case where P is the uniform distribution over the interval $[-A,+A]$. The optimum MMSE quantizer in this case is uniform as well: The interval $[-A,A]$ is partitioned evenly into M sub-intervals, each of size $2A/M$ and the reproduction level $\hat{x}_i$, pertaining to each sub-interval, is its midpoint. What happens when the exponential moment is considered? Let us 'guess' that the same quantizer s remains optimum. Then,

$$Q(x)\propto 1\{|x|\le A\}\cdot\exp\{-\alpha(x-s(x))^2\} = 1\{|x|\le A\}\cdot\exp\{-\alpha\min_i(x-\hat{x}_i)^2\}, \tag{20}$$

which means that Q has "Gaussian peaks" at all reproduction points $\{\hat{x}_i\}$. It is plausible that the same uniform quantizer s continues to be optimum (or at least nearly so) for Q, and hence these s and Q match each other. Moreover, at least for large α, this quantizer nearly attains the rate–distortion function of Q at distortion level $D=1/(2\alpha)$. To see why this is true, observe that for large α, the factor $\exp\{-\alpha\min_i(x-\hat{x}_i)^2\} = \max_i\exp\{-\alpha(x-\hat{x}_i)^2\}$ is well approximated by $\sum_i\exp\{-\alpha(x-\hat{x}_i)^2\}$, which after normalization of Q, becomes essentially a mixture of M evenly weighted Gaussians, where the i–th mixture component is centered at $\hat{x}_i$, $i=0,1,\ldots,M-1$. This mixture approximation of Q can then be viewed as a convolution between the uniform discrete distribution on $\{\hat{x}_i\}$ and the Gaussian density $\mathcal{N}(0, 1/(2\alpha))$. Thus, for $D \le 1/(2\alpha)$, the rate–distortion function of Q agrees with the Shannon lower bound (see, e.g., [26, Chapter 4]), which is

$$\begin{aligned} R_L(D) &= h(Q) - \frac{1}{2}\log(2\pi e D) \\ &\approx \log M + \frac{1}{2}\log\left(\frac{\pi e}{\alpha}\right) - \frac{1}{2}\log(2\pi e D) \\ &= \log M + \frac{1}{2}\log\left(\frac{1}{2\alpha D}\right), \end{aligned} \tag{21}$$

which, for $D = 1/(2\alpha)$, indeed gives a coding rate of $\log M$, just like the uniform quantizer. Here, $h(Q)$ stands for the differential entropy pertaining to Q.

Finally, consider the case where the source vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is to be quantized by a sequential causal quantizer with memory, i.e., X_t is quantized into one of M quantization levels, which are now allowed to depend on past outputs of the quantizer $\hat{X}_1, \ldots, \hat{X}_{t-1}$. Here, the relevant exponential moment criterion would be $\boldsymbol{E}_P \exp\{-\alpha \sum_{t=1}^n [X_t - s(X_t|\hat{X}_1, \ldots, \hat{X}_{t-1})]^2\}$. As is shown in [43], however, whenever the source P is memoryless, the allowed dependence of the current quantization on the past does not improve the exponential moment performance, i.e., the optimum quantizer of this type makes use of the current symbol X_t only. This means that the causal vector quantization problem actually degenerates back to the scalar quantization problem considered in the previous paragraphs. This continues to be true even if variable–rate coding is allowed, except that then time–sharing between at most two quantizers must also be allowed. These results of [43], for the exponential moment criterion, are analogous to those of Neuhoff and Gilbert [48] for the ordinary criterion of expected code length for a given distortion (or expected distortion at fixed rate).

3.3. Non–Bayesian Parameter Estimation. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ be a Gaussian random vector with mean vector $\theta\boldsymbol{u}$, where $\theta \in \mathbb{R}$ and $\boldsymbol{u} = (u_1, \ldots, u_n)^T$ is a given deterministic vector. Let the non–singular $n \times n$ covariance matrix of $\boldsymbol{X}$ be given by Λ. The vector $\boldsymbol{X}$ can then be thought of as a set of measurements (contaminated by non–white Gaussian noise) of a signal, represented by a known vector $\boldsymbol{u}$, but with an unknown gain θ to be estimated.[6] This is a classical problem in non–Bayesian estimation theory. It is well known that for this kind of a model, among all unbiased estimators of θ, the one that minimizes the mean square error (or equivalently, the estimation error variance) is the maximum likelihood (ML) estimator, which in this case, is easily found to be given by

$$s(\boldsymbol{x}) = \frac{\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{x}}{\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u}}. \tag{22}$$

[6] In fact, the model and the analysis can be extended to the case where θ is a vector and $\boldsymbol{u}$ is a matrix with compatible dimensions, which then becomes a general linear regression problem with correlated Gaussian noise. For the sake of simplicity, however, the derivations are demonstrated for the one–dimensional case.

Does this estimator also minimize $\boldsymbol{E}_\theta \exp\{\alpha[s(\boldsymbol{X})-\theta]^2\}$ among all unbiased estimators and for all values of α in the allowed range? Here, $\boldsymbol{E}_\theta$ denotes expectation when the true parameter is θ.

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 this estimator indeed minimizes also $\boldsymbol{E}_\theta \exp\{\alpha[s(\boldsymbol{X})-\theta]^2\}$ and then check whether it satisfies the conditions of Observation 1. Denoting $\boldsymbol{v} = \Lambda^{-1}\boldsymbol{u}/(\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u})$, the corresponding probability measure Q, which will be denoted here by Q_θ, is given by

$$\begin{aligned} Q_\theta(\boldsymbol{x}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\theta\boldsymbol{u})^T\Lambda^{-1}(\boldsymbol{x}-\theta\boldsymbol{u}) + \alpha\left(\boldsymbol{v}^T\boldsymbol{x}-\theta\right)^2\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\theta\boldsymbol{u})^T\Lambda^{-1}(\boldsymbol{x}-\theta\boldsymbol{u}) + \alpha\left[\boldsymbol{v}^T(\boldsymbol{x}-\theta\boldsymbol{u})\right]^2\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\theta\boldsymbol{u})^T\Lambda^{-1}(\boldsymbol{x}-\theta\boldsymbol{u}) + \alpha(\boldsymbol{x}-\theta\boldsymbol{u})^T\boldsymbol{v}\boldsymbol{v}^T(\boldsymbol{x}-\theta\boldsymbol{u})\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\theta\boldsymbol{u})^T(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{x}-\theta\boldsymbol{u})\right\}, \qquad (23) \end{aligned}$$

where α is chosen small enough such that the matrix $(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)$ is still positive definite. Now, the ML estimator of θ under Q_θ is given by

$$\begin{aligned} s_Q(\boldsymbol{x}) &= \frac{\boldsymbol{u}^T(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{x}}{\boldsymbol{u}^T(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{u}} \\ &= \frac{\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{x}-2\alpha\boldsymbol{u}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{x}}{\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u}-2\alpha\boldsymbol{u}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{u}} \\ &= \frac{[1-2\alpha/(\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u})]\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{x}}{[1-2\alpha/(\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u})]\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u}} \\ &= s(\boldsymbol{x}), \qquad (24) \end{aligned}$$

where in the third line we have used the fact that $\boldsymbol{u}^T\boldsymbol{v} = 1$. In other words, we are back to the original estimator we started from, which means that our 'guess' was successful. Indeed, the MSE of $s(\boldsymbol{x})$ under Q_θ, which is $\boldsymbol{v}^T(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)^{-1}\boldsymbol{v}$, can easily be shown to be identical to the Cramér–Rao lower bound under Q_θ, which is given by $1/[\boldsymbol{u}^T(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{u}]$. We can therefore summarize our conclusion in the following proposition:

PROPOSITION 1. *Let $\boldsymbol{X}$ be a Gaussian random vector with a mean vector $\theta\boldsymbol{u}$ and a non–singular covariance matrix Λ. Let a be the supremum of all values of α such that the matrix $(\Lambda^{-1}-2\alpha\boldsymbol{v}\boldsymbol{v}^T)$ is positive definite, where $\boldsymbol{v} = \Lambda^{-1}\boldsymbol{u}/(\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u})$. Then, among all unbiased estimators of θ, the estimator $s(\boldsymbol{x}) = \boldsymbol{v}^T\boldsymbol{x}$ uniformly minimizes the exponential moment $\boldsymbol{E}\exp\{\alpha[s(\boldsymbol{X})-\theta]^2\}$ for all $\alpha \in (0, a)$. This minimum of the exponential moment is given by*

$$\boldsymbol{E}\exp\left\{\alpha\left(\boldsymbol{v}^T\boldsymbol{X}-\theta\right)^2\right\} = \frac{1}{\sqrt{det(I-2\alpha\boldsymbol{v}\boldsymbol{v}^T\Lambda)}}. \qquad (25)$$

It is easy to see that for $\alpha \to 0$, the limit of $\frac{1}{\alpha}\ln \boldsymbol{E}\exp\{\alpha(\boldsymbol{v}^T\boldsymbol{X}-\theta)^2\}$ recovers the mean–square error $\mathrm{tr}\{\boldsymbol{v}\boldsymbol{v}^T\Lambda\} = \boldsymbol{v}^T\Lambda\boldsymbol{v}$, as expected.

An alternative way to understand Proposition 1 could be directly in the spirit of the proof of Observation 1:

$$\begin{aligned}\boldsymbol{E}_{P_\theta} e^{\alpha[s(\boldsymbol{X})-\theta]^2} &= \boldsymbol{E}_{Q_\theta}\exp\left\{\alpha[s(\boldsymbol{X})-\theta]^2 + \ln\frac{P_\theta(\boldsymbol{X})}{Q_\theta(\boldsymbol{X})}\right\}\\ &\geq \exp\left\{\alpha\boldsymbol{E}_{Q_\theta}[s(\boldsymbol{X})-\theta]^2 - D(Q_\theta\|P_\theta)\right\} \qquad (26)\end{aligned}$$

where we have now denoted expectations w.r.t. P_θ and Q_θ by $\boldsymbol{E}_{P_\theta}$ and $\boldsymbol{E}_{Q_\theta}$, respectively. Now, $\boldsymbol{E}_{Q_\theta}[s(\boldsymbol{X})-\theta]^2$ is lower bounded by the Cramér–Rao lower bound under Q_θ, which is as said, given by $\boldsymbol{v}^T(\Lambda^{-1} - 2\alpha\boldsymbol{v}\boldsymbol{v}^T)^{-1}\boldsymbol{v}$, and $D(Q_\theta\|P_\theta)$ (defined for n–vectors) is easy to compute. The result coincides, of course, with the r.h.s. of eq. (25).

Finally, we point out that here, since the same estimator minimizes the exponential moment for all $\alpha \geq 0$, we have the full best achievable characteristic function. Consequently, in the limit $n \to \infty$, we also have the best achievable large deviations performance in the sense of asymptotically minimizing of probabilities of the form $\Pr\{|s(\boldsymbol{X})-\theta| \geq R\}$ for all $R > 0$. Related results by Kester and Kallenberg [32] (and some subsequent works) concern the asymptotic large deviations optimality of the ML estimator, among all consistent estimators, for the case of exponential families of i.i.d. measurements. Here, on the other hand, the model is not i.i.d., the result holds (for exponential moments) for all n, and not only asymptotically, and the class of competing estimators is the class of unbiased estimators.

3.4. Gaussian–Quadratic Joint Source–Channel Coding. Consider the Gaussian memoryless source

$$P_U(\boldsymbol{u}) = (2\pi\sigma_u^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_u^2}\sum_{i=1}^n u_i^2\right\} \qquad (27)$$

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

$$P_Z(\boldsymbol{z}) = (2\pi\sigma_z^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_z^2}\sum_{i=1}^n z_i^2\right\}. \qquad (28)$$

In the ordinary joint source–channel coding problem, one seeks an encoder and decoder that would minimize $D = \frac{1}{n}\sum_{i=1}^n \boldsymbol{E}\{(U_i - V_i)^2\}$, where $\boldsymbol{V} = (V_1, \ldots, V_n)$ is the reconstruction at the decoder. It is very well known that the best achievable distortion, in this case, is given by

$$D = \frac{\sigma_u^2}{1+\Gamma/\sigma_z^2}, \qquad (29)$$

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, $\boldsymbol{E}\exp\{\alpha\sum_i(U_i-V_i)^2\}$? It is natural to wonder whether simple linear transmitters and receivers, of the kind defined in the previous paragraph, are still optimum. This question is intimately related to the problem of optimizing joint source–channel excess distortion exponents [13] (which is still open) and hence also to the related problem of jointly optimum modulation and estimation, when examined from the large deviations perspective [39], as opposed to the more customary mean square error regime [58, Chap. 8].

The random object X, in this example, is the pair of vectors $(\boldsymbol{U},\boldsymbol{Z})$, where $\boldsymbol{U}$ is the source vector and $\boldsymbol{Z}$ 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 $\boldsymbol{x}=f(\boldsymbol{u})$ and a decoding function $\boldsymbol{v}=g(\boldsymbol{y})$. The class $\mathcal{S}$ is then the set of all pairs of functions $\{f,g\}$, where f satisfies the power constraint $\boldsymbol{E}_P\{\|f(\boldsymbol{U})\|^2\}\le n\Gamma$. Condition 2 of Observation 1 tells us that the modified probability distribution of $\boldsymbol{u}$ and $\boldsymbol{z}$ should be of the form

$$Q(\boldsymbol{u},\boldsymbol{z})\propto P_U(\boldsymbol{u})P_Z(\boldsymbol{z})\exp\left\{\alpha\sum_{i=1}^{n}[u_i-g_i(f(\boldsymbol{u})+\boldsymbol{z})]^2\right\} \tag{30}$$

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

Clearly, if we continue to restrict the encoder to be a scalar linear amplifier, $f(\boldsymbol{u})=\sqrt{\Gamma/\sigma_u^2}\cdot\boldsymbol{u}$, which simply exploits the allowed power Γ, and the only remaining room for optimization concerns the decoder g, then we are basically dealing with a problem of pure Bayesian estimation in the Gaussian regime, and then the optimum choice of the decoder (estimator) continues to be the same linear decoder as before (see [47]).[7] However, if we further 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 Observation 1 are no longer met for any linear functions f and g. The key reason is that while $Q(\boldsymbol{u},\boldsymbol{z})$ of eq. (30) continues to be Gaussian (though now U_i and Z_i are correlated) when f and g are linear, the power constraint, $\boldsymbol{E}_P\{\|\boldsymbol{X}\|^2\}\le n\Gamma$, when expressed as an expectation w.r.t. Q, becomes $\boldsymbol{E}_Q\{\|f(\boldsymbol{U})\|^2P(\boldsymbol{U})/Q(\boldsymbol{U})\}\le n\Gamma$, but "power" function $\|f(\boldsymbol{u})\|^2P(\boldsymbol{u})/Q(\boldsymbol{u})$, with P and Q being Gaussian densities, is no longer the

[7] This can also be obtained by applying Observation 1 to the problem of Bayesian estimation in the Gaussian regime under the exponential moment criterion. See Section 3.2 in the original version of this paper [38].

usual quadratic function of $f(\boldsymbol{u})$ for which there is a linear encoder and decoder that is optimum.

While Observation 1 merely furnishes sufficient conditions for optimality, it is not difficult to see that linear encoders and decoders cannot be optimal in the exponential moment sense, using a simple information–theoretic 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

$$d(\boldsymbol{u},\boldsymbol{v}) = \exp\left\{\alpha\sum_{i=1}^{n}(u_i - v_i)^2\right\}. \tag{31}$$

For such a joint source–channel coding system to be optimal, the induced channel $P(\boldsymbol{v}|\boldsymbol{u})$ must [6, p. 31, eq. (2.5.13)] be proportional to

$$P(\boldsymbol{v})\exp\{-\beta d(\boldsymbol{u},\boldsymbol{v})\} = P(\boldsymbol{v})\exp\left[-\beta\exp\left\{\alpha\sum_{i}(u_i - v_i)^2)\right\}\right] \tag{32}$$

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 $P(\boldsymbol{v})$ would be Gaussian, and then $P(\boldsymbol{v}|\boldsymbol{u})$ 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 $\boldsymbol{u}$ to $\boldsymbol{v}$, which is very different, and therefore cannot be optimum.

Of course, the minimum of $\boldsymbol{E}\exp\{\alpha\sum_i(U_i - V_i)^2\}$ 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 $\boldsymbol{E}_P\exp\{\alpha\ell(X,s)\}$ 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 $\boldsymbol{X} = (X_1,\ldots,X_n)$, governed by a product distribution function

$$P(\boldsymbol{x}) = \prod_{i=1}^{n}P(x_i), \tag{33}$$

where each component x_i of the vector $\boldsymbol{x} = (x_1,\ldots,x_n)$ takes on values in a finite set $\mathcal{X}$. If $\ell(\boldsymbol{x},s)$ grows linearly[8] with n for a given empirical distribution of $\boldsymbol{x}$ and a

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

given $s \in \mathcal{S}$, then it is expected that the exponential moment $\boldsymbol{E} \exp\{\alpha \ell(\boldsymbol{x}, s)\}$ would behave, at least asymptotically, as an exponential function of n. In particular, for a given s, let us assume that the limit

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

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

$$E(s^*, \alpha, P) \leq E(s, \alpha, P) \tag{34}$$

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. (34) 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 $\mathcal{X}$,[9] We denote by T_Q the type class pertaining to an empirical distribution Q, namely, the set of vectors $\boldsymbol{x} \in \mathcal{X}^n$ whose empirical distribution is Q.

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

(a) For every type class T_Q and every $\boldsymbol{x} \in T_Q$, $\ell(\boldsymbol{x}, s^*) \leq n[\lambda(Q) + o(1)]$, where $o(1)$ designates a (positive) sequence that tends to zero as $n \to \infty$.

(b) For every type class T_Q and every $s \in \mathcal{S}$,

$$\left| T_Q \cap \{\boldsymbol{x} : \ \ell(\boldsymbol{x}, s) \geq n[\lambda(Q) - o(1)]\} \right| \geq e^{-no(1)} |T_Q|. \tag{35}$$

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

$$E(s^*, \alpha, P) = \max_Q [\alpha \lambda(Q) - D(Q \| P)], \tag{36}$$

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 $\boldsymbol{x}$ or on its type class,[10] yet the performance of s^*, for every $\boldsymbol{x} \in T_Q$, "adapts" to the empirical distribution Q of $\boldsymbol{x}$ in a way, that according to condition (b),

[9] It should be understood that extensions of the following discussion to more general classes of sources, like the class of Markov sources, is essentially straightforward in many cases, although there may be elements that might require some more caution.

[10] As before, s^* is chosen without observing the data first.

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 Observation 1. First, observe that in order to guarantee asymptotic optimality of s^*, condition 2 of Observation 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 $\alpha\ell(\boldsymbol{X},s)+\ln[P(\boldsymbol{X})/Q(\boldsymbol{X})]$ completely degenerate (i.e., a constant for every realization $\boldsymbol{x}$, as in condition 2 of Observation 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 $Q^*(\boldsymbol{x})$ to be the memoryless source induced by the dominant Q^*, this is indeed precisely what happens under conditions (a) and (b), which imply that

$$\begin{aligned}\alpha\ell(\boldsymbol{x},s^*)+\ln\frac{P(\boldsymbol{x})}{Q^*(\boldsymbol{x})} &\approx n\alpha\lambda(Q)+\sum_{i=1}^{n}\ln\frac{P(x_i)}{Q^*(x_i)}\\ &= n\alpha\lambda(Q)+n\sum_{x\in\mathcal{X}}Q^*(x)\ln\frac{P(x)}{Q^*(x)}\\ &= n[\alpha\lambda(Q^*)-D(Q^*\|P)], \qquad (37)\end{aligned}$$

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 $\boldsymbol{E}\exp\{\alpha\ell(\boldsymbol{x},s^*)\}$. 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 Observation 1. Moreover, since s^* "adapts" to every T_Q, in the sense explained above, then this has the flavor of the maximin problem discussed in Section 2, where s is allowed to be optimized for each and every Q. Since the minimizing s, in the maximin 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 $\boldsymbol{X}$ 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 $d:\mathcal{X}\times\hat{\mathcal{X}}\to\mathrm{I\!R}$, $\hat{\mathcal{X}}$ 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 $\ell(\boldsymbol{x},s)$ designate the distortion between the source vector $\boldsymbol{x}$ 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 [14, 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 $\boldsymbol{x}$ in two parts, the first of which is a header that describes the index of the type class T_Q of

$\boldsymbol{x}$ (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 $\boldsymbol{x}$ is encoded by $\ell(\boldsymbol{x}, s)$ 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 Subsection 3.1), 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 [50] in universal data compression: According to [50], given a length function of a lossless data compression $\ell(\boldsymbol{x}, s)$ (s being the data compression scheme), and given a parametric class of sources of n–vectors, $\{P_\theta^n\}$, indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$, a lower bound on $\boldsymbol{E}_{P_\theta^n}\ell(\boldsymbol{X}, s)$, that applies to most[11] values of θ, is given by

$$\boldsymbol{E}_{P_\theta^n}\ell(\boldsymbol{X}, s) \geq H(P_\theta^n) + (1-\epsilon)\frac{k}{2}\log n, \tag{38}$$

where $\epsilon > 0$ is arbitrarily small (for large n), $H(P_\theta^n)$ is the entropy of $\boldsymbol{X}$ associated with P_θ^n, and $\boldsymbol{E}_{P_\theta^n}\{\cdot\}$ is the expectation under P_θ^n. 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 $(1 + \epsilon)$.

Now, for a given source P_θ^n, let us define Q_θ^n as being the source probability function that is proportional to $(P_\theta^n)^{1/(1+\alpha)}$. Then, as a lower bound, we have

$$\begin{aligned}
\ln \boldsymbol{E}_{P_\theta^n} \exp\{\alpha\ell(\boldsymbol{X}, s)\} &= \max_Q\, [\alpha \boldsymbol{E}_Q \ell(\boldsymbol{X}, s) - nD(Q\|P_\theta^n)] \\
&\geq \alpha \boldsymbol{E}_{Q_\theta^n}\ell(\boldsymbol{X}, s) - D(Q_\theta^n\|P_\theta^n) \\
&\geq \alpha\left[H(Q_\theta^n) + (1-\epsilon)\frac{k}{2}\log n\right] - D(Q_\theta^n\|P_\theta^n) \\
&= \alpha H_{1/(1+\alpha)}(P_\theta^n) + \alpha(1-\epsilon)\frac{k}{2}\log n, \qquad (39)
\end{aligned}$$

where the third line follows from Rissanen's lower bound (for most sources), and where $H_u(P_\theta^n)$ is Rényi's entropy of order u, namely,

$$H_u(P_\theta^n) = \frac{1}{1-u}\ln\left\{\sum_{\boldsymbol{x}\in\mathcal{X}^n}[P_\theta^n(\boldsymbol{x})]^u\right\}. \tag{40}$$

[11] "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.

Consider now the case where $\{P_\theta^n,\ \theta \in \Theta\}$ is the class of all memoryless sources over $\mathcal{X}$, where the parameter vector θ designates $k = |\mathcal{X}| - 1$ letter probabilities. In this case, since the source is completely defined by the single–letter probabilities, we can omit the superscript n of P_θ^n and denote the source by P_θ. Define a two–part code s^*, which first encodes the index of the type class Q and then the index of $\boldsymbol{x}$ within the type class. The corresponding length function is given by

$$\ell(\boldsymbol{x}, s^*) = \ln |T_Q| + k \log n \approx n\hat{H}(\boldsymbol{x}) + \frac{k}{2} \log n, \tag{41}$$

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

$$\begin{aligned}\ln \boldsymbol{E}_{P_\theta} \exp\{\alpha\ell(\boldsymbol{X}, s)\} &= \ln \boldsymbol{E}_{P_\theta} \exp\{\alpha n\hat{H}(\boldsymbol{X})\} + \alpha\frac{k}{2} \log n \\ &= \ln \boldsymbol{E}_{P_\theta} \exp\{\alpha \min_Q[-\ln Q(\boldsymbol{X})]\} + \alpha\frac{k}{2} \log n \\ &\le \min_Q \ln \boldsymbol{E}_{P_\theta} \exp\{-\alpha \ln Q(\boldsymbol{X})\} + \alpha\frac{k}{2} \log n \\ &= n\alpha H_{1/(1+\alpha)}(P_\theta) + \alpha\frac{k}{2} \log n, \end{aligned} \tag{42}$$

and then it essentially achieves the lower bound. Rissanen's result is now obtained a special case of this, by dividing both sides of the inequality by α and then taking the limit $\alpha \to 0$.

We next summarize these findings in the form of a theorem, which is an exponential–moment counterpart of [50, Theorem 1]. The converse part (part (a)) can actually be extended to even more general classes of sources, which are not even necessarily parametric, using the results of [41], where the expression $(k \log n)/(2n)$ is replaced, more generally, by the capacity of the "channel" from θ to $\boldsymbol{X}$, as defined by the class of sources $\{P_\theta\}$ when viewed as a set of conditional distributions of $\boldsymbol{X}$ given θ. For the sake of simplicity, the direct part (part (b)) of this theorem is formalized for the class of all memoryless sources with a given finite alphabet, parametrized by the letter probabilities, but it can also be extended to wider classes of sources, like Markov sources of a given order.

Theorem 1.

(a) Converse part: Let $\mathcal{P} = \{P_\theta^n,\ \theta \in \Theta\}$ be a parametric class of finite–alphabet memoryless sources, indexed by a parameter θ that takes on values in a compact subset Θ of $\mathbb{R}^k$. Let the central limit theorem hold, under Q_θ^n, for the ML estimator of each θ in the interior of Θ. If $\ell(\boldsymbol{x}, s)$ is a length function of a code s satisfying the Kraft inequality, then for every $\alpha > 0$ and $\epsilon > 0$,

$$\frac{1}{n\alpha} \ln \boldsymbol{E}_{P_\theta^n} \exp\{\alpha\ell(\boldsymbol{X}, s)\} \ge H_{1/(1+\alpha)}(P_\theta^n) + (1 - \epsilon) \cdot \frac{k \log n}{2n}, \tag{43}$$

for all points $\theta \in \Theta$ except for a set $\mathcal{A}_\epsilon(n) \subset \Theta$ whose Lebesgue measure vanishes as $n \to \infty$ for every fixed $\epsilon > 0$.

(b) *Direct part: For the case where $\mathcal{P}$ is the class of all memoryless sources with a given alphabet of size $k+1$ and θ designates the vector of the k first letter probabilities, there exists a universal lossless data compression code s^*, whose length function $\ell(\boldsymbol{x}, s^*)$ satisfies the reversed inequality where the factor $(1-\epsilon)$ is replaced by $(1+\epsilon)$.*

It is interesting to note that Rissanen's universal coding redundancy theorem and its above extension, given by Theorem 1 above, can be harnessed to address universal prediction problems. Rissanen [50, Section VI] has already carried out this program for the minimum mean square error criterion. In particular, let P_θ by a family of Gaussian ARMA processes $\{X_n\}$, where θ is the vector $(\sigma^2, a_1, \ldots, a_p, b_1, \ldots, b_q)$ of the innovation variance σ^2, the autoregressive parameters $(a_1, \ldots, a_p)$ and the moving average parameters $(b_1, \ldots, b_q)$. In [50, Theorem 2], the following result was stated and proved (with a few minor modifications in the formulation): Let $\hat{X}_i = f(X_1, \ldots, X_{i-1})$ be an arbitrary predictor of X_i, which is independent of the unknown parameter vector θ. Then,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}_\theta(\hat{X}_i - X_i)^2 \geq \sigma^2\left[1 + (1-\epsilon)(p+q)\frac{\ln n}{n}\right] \tag{44}$$

for all points θ except in a set $A_\epsilon(n)$ whose Lebesgue measure tends to zero as $n \to \infty$. It is also pointed out in [50], that this lower bound is achievable (with the sign of ϵ switched) at least for AR processes by on-line maximum likelihood estimation of the AR parameters at each time instant i [16].

Considering now the exponential moment criterion, we readily obtain by the same technique

$$\begin{aligned}
\boldsymbol{E}_\theta \exp\left\{\alpha\sum_{i=1}^{n}(\hat{X}_i - X_i)^2\right\} &\geq \exp\left\{\alpha\sum_{i=1}^{n}\boldsymbol{E}_{\tilde{\theta}}(\hat{X}_i - X_i)^2 - D(P_{\tilde{\theta}}^n \| P_\theta^n)\right\} \\
&\geq \exp\left\{\alpha \cdot n\tilde{\sigma}^2\left[1 + (1-\epsilon)(p+q)\frac{\ln n}{n}\right] - D(P_{\tilde{\theta}}^n \| P_\theta^n)\right\} \\
&\stackrel{\Delta}{=} \exp\left\{n\alpha_n \cdot \tilde{\sigma}^2 - D(P_{\tilde{\theta}}^n \| P_\theta^n)\right\}
\end{aligned}$$

where $P_{\tilde{\theta}}$, for an arbitrary $\tilde{\theta}$, plays the earlier role of Q, and where we have defined $\alpha_n = \alpha[1 + (1-\epsilon)(\ln n)/n]$. The lower bound can now be maximized over $\tilde{\theta}$. For the sake of simplicity, let us particularize this result to the special case of a first–order autoregressive process, where $\theta = (\sigma^2, a)$ and $\tilde{\theta}$ denotes an arbitrary alternative parameter value $(\tilde{\sigma}^2, \tilde{a})$. In this case, the divergence is easily derived to be

$$D(P_{\tilde{\theta}}^n \| P_\theta^n) = \frac{n}{2}\left[\frac{\tilde{\sigma}^2}{\sigma^2}\left(1 + \frac{(\tilde{a}-a)^2}{1-\tilde{a}^2}\right) - \ln\frac{\tilde{\sigma}^2}{\sigma^2} - 1\right], \tag{45}$$

and so the exponent of the lower bound, $n\alpha_n\tilde{\sigma}^2 - D(P_{\tilde{\theta}}^n \| P_\theta^n)$, is maximized by $\tilde{\theta}$ with $\tilde{\sigma}^2 = \sigma^2/(1 - 2\alpha_n\sigma^2)$ and $\tilde{a} = a$. Of course, in these derivations, α (and α_n) must be limited to be strictly less than $1/(2\sigma^2)$. On substituting this back into the lower bound, we finally obtain

$$\begin{aligned} & \frac{1}{n}\ln\left[\boldsymbol{E}_\theta \exp\left\{\alpha\sum_{i=1}^{n}(\hat{X}_i - X_i)^2\right\}\right] \\ & \geq \frac{\alpha_n\sigma^2}{1-2\alpha_n\sigma^2} - f(1-2\alpha_n\sigma^2) \\ & = \frac{\alpha\sigma^2}{1-2\alpha\sigma^2} - f(1-2\alpha\sigma^2) + \frac{(1-\epsilon)\alpha\sigma^2}{1-2\alpha\sigma^2}\cdot\frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right) \end{aligned} \tag{46}$$

where $f(x) \stackrel{\Delta}{=} \frac{1}{2}(\frac{1}{x} - \ln\frac{1}{x} - 1)$, and where the last passage is a first order approximation that follows from a Taylor series expansion around α. While the first two terms in the last expression form a lower bound that applies even when a is known (which is achievable using the ordinary the linear predictor $\hat{X}_t = aX_{t-1}$), we have not addressed the achievability of the full universal prediction lower bound (which includes the extra redundancy term) when a is unknown, in the sense of achieving the compatible coefficient in front of the $(\ln n)/n$ term.

Our last example corresponds to a secrecy system. A sequence $\boldsymbol{x}$ is to be communicated to a legitimate decoder which shares with the transmitter a random key $\boldsymbol{z}$ of nR purely random bits. The encoder transmits an encrypted message $\boldsymbol{y} = \phi(\boldsymbol{x}, \boldsymbol{z})$, which is an invertible function of $\boldsymbol{x}$ given $\boldsymbol{z}$, and hence decipherable by the legitimate decoder. An eavesdropper, which has no access to the key $\boldsymbol{z}$, submits a sequence of guesses concerning $\boldsymbol{x}$ 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 $G(\boldsymbol{X}, s)$, i.e., $\boldsymbol{E}\{G^\alpha(\boldsymbol{X}, s)\}$? In this case, $\ell(\boldsymbol{x}, s) = \ln G(\boldsymbol{x}, s)$. As is shown in [40], there exists a guessing strategy s^*, which for every $\boldsymbol{x} \in T_Q$, gives $\ell(\boldsymbol{x}, s^*) \approx n\min\{H(Q), R\}$, 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 $\lambda(Q) = \min\{H(Q), R\}$.

It should be pointed out that in [40], as well as in other related works on various settings of the guessing problem [1], [2], [3], [45], the technique proposed by Observation 1 was actually already used (at least implicitly) to address all these problems.

5. Lower Bounds on Exponential Moments. As explained in the Introduction, even in the ordinary setting, of the quest for minimizing $\boldsymbol{E}\{\ell(X, s)\}$, 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 on the expectation of $\ell(X, s)$, possibly using a change of measure, in the spirit of the proof of Observation

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 [55]), 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 $\boldsymbol{E}_\theta \exp\{\alpha(\hat{\theta} - \theta)^2\}$, for any unbiased estimator of θ, where as before, $\boldsymbol{E}_\theta$ denotes expectation w.r.t. P_θ. Consider the following chain of inequalities, which holds for any $\theta' \in \Theta$:

$$\begin{aligned}
\boldsymbol{E}_\theta \exp\{\alpha(\hat{\theta}-\theta)^2\} &= \boldsymbol{E}_{\theta'} \exp\left\{\alpha(\hat{\theta}-\theta)^2 + \ln\frac{P_\theta(X)}{P_{\theta'}(X)}\right\} \\
&\ge \exp\left\{\alpha \boldsymbol{E}_{\theta'}(\hat{\theta}-\theta)^2 - D(P_{\theta'}\|P_\theta)\right\} \\
&= \exp\left\{\alpha \boldsymbol{E}_{\theta'}(\hat{\theta}-\theta')^2 + \alpha(\theta-\theta')^2 - D(P_{\theta'}\|P_\theta)\right\} \\
&\ge \exp\left\{\alpha \mathrm{CRB}(\theta') + \alpha(\theta-\theta')^2 - D(P_{\theta'}\|P_\theta)\right\}, \qquad (47)
\end{aligned}$$

where $\mathrm{CRB}(\theta)$ is the Cramér–Rao bound for unbiased estimators, computed at θ (i.e., $\mathrm{CRB}(\theta) = 1/I(\theta)$, 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

$$\ln \boldsymbol{E}_\theta \exp\{\alpha(\hat{\theta}-\theta)^2\} \ge \sup_{\theta'\in\Theta}\left[\alpha\mathrm{CRB}(\theta') + \alpha(\theta'-\theta)^2 - D(P_{\theta'}\|P_\theta)\right]. \qquad (48)$$

More generally, if $\theta = (\theta_1, \ldots, \theta_k)^T$ 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

$$\ln \boldsymbol{E}_\theta \exp\{\alpha^T(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\alpha\} \ge \sup_{\theta'\in\Theta}\left[\alpha^T I^{-1}(\theta')\alpha + [\alpha^T(\theta'-\theta)]^2 - D(P_{\theta'}\|P_\theta)\right], \qquad (49)$$

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, IT–42: 1(1996), pp. 99–105.

[2] E. Arikan and N. Merhav, *Guessing subject to distortion,* IEEE Trans. Inform. Theory, 44:3(1998), pp. 1041–1056.

[3] E. Arikan and N. Merhav, *Joint source–channel coding and guessing with application to sequential decoding,* IEEE Trans. Inform. Theory, 44:5(1998), pp. 1756–1769.

[4] R. Atar, P. Dupuis, and A. Shwartz, *An escape–time criterion for queueing networks: asymptotic risk–sensitive control via differential games,* Mathemtics of Operation Research, 28:4(2003), pp. 801–835.

[5] M. Bakshi and D. R. Fuhrmann, *Improving the visual quality of JPEG-encoded images via companding, J. Electronics Imaging*, 6:2(1997), pp. 189–197.

[6] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice Hall, Englewood Cliffs, New Jersey, U.S.A., 1971.

[7] D. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I and II, 3rd ed. Nashua, NH: Athena Scientific, 2007.

[8] J. P. N. Bishwal, *Large deviations and Berry–Esseen inequalities for estimators in nonlinear nonhomogeneous diffusions,* REVSTAT Statistical Journal, 5:3(2007), pp. 249–267.

[9] J. A. Buclew, *Companding and random quantization in several dimensions, IEEE Trans. Inform. Theory*, IT–27:2(1981), pp. 207–211.

[10] J. A. Buclew, *A note on optimal multidimensional companders,* IEEE Trans. Inform. Theory, IT–29:2(1983), p. 279.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second Edition, John Wiley & Sons, Hoboken, New Jersey, U.S.A., 2006.

[12] B. Chen and G. W. Wornell, *Quantization index modulation: a class of provably good methods for digital watermarking and information embedding,* IEEE Trans. Inform. Theory, 47:4(2001), pp. 1423–1443.

[13] I. Csiszár, *On the error exponent of source-channel transmission with a distortion threshold,* IEEE Trans. Inform. Theory, IT–28:6(1982), pp. 823–828.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, U.S.A., 1981.

[15] P. Dai Pra, L. Meheghini, and W. J. Runggaldier, *Connections between stochastic control and dynamic games,* Mathematics of Control, Signals and Systems, 9(1996), pp. 303–326.

[16] L. D. Davisson, *The prediction error of stationary Gaussian time series of unknown covariance,* IEEE Trans. Inform. Theory, IT–11:4(1965), pp. 527–532.

[17] A. Dembo and O. Zeitouni, *Large DEviations and Applications*, Jones and Bartlett Publishers, London 1993.

[18] G. B. Di Masi and L. Stettner, *Risk–sensitive control of discrete–time Markov processes with infinite horizon,* SIAM Journal on Control and Optimization, 38:1(1999), pp. 61–78.

[19] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, 1997.

[20] P. Dupuis, M. R. James, and I. Petersen, *Robust properties of risk–sensitive control,* Mathematics of Control, Signals and Systems, 13(2000), pp. 318–322.

[21] J. J. Eggers and B. Girod, *Quantization effects on digital watermarks,* Signal Processing, 81(2001), pp. 239–263.

[22] 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, IT–32:6(1986), pp. 839–845.

[23] W. H. Fleming and D. Hernández–Hernández, *Risk–sensitive control of finite state ma-*

chines on an infinite horizon I, SIAM Journal on Control and Optimization, 35:5(1997), pp. 1790–1810.

[24] A. Gersho, *Principles of quantization,* IEEE Transactions on Circuits and Systems, CAS–25:7(1978), pp. 427–436.

[25] R. M. Gray, *Vector quantization,* IEEE Transactions on Acoustics, Speech, and Signal Processing, Mag: 4–29, April 1984.

[26] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, Norwell, MA, U.S.A., 1990.

[27] R. M. Gray and D. L. Neuhoff, *Quantization,* IEEE Trans. Inform. Theory, IT–44:6(1998), pp. 2325–2383.

[28] R. Howard and J. Matheson, *Risk–sensitive Markov decision processes,* Management Science, 18(1972), pp. 356–369.

[29] P. A. Humblet, *Generalization of Huffman coding to minimize the probability of buffer overflow,* IEEE Transactions on Information Theory, IT–27:2(1981), pp. 230–232.

[30] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice–Hall, Englewood Cliffs, 1984.

[31] F. Jelinek, *Buffer overflow in variable length coding of fixed rate sources,* IEEE Transactions on Information Theory, IT–14:3(1968), pp. 490–501.

[32] A. D. M. Kester and W. C. M. Kallenberg, *Large deviations of estimators,* Annals of Mathmatical Statistics, 14:2(1986), pp. 648–664.

[33] Y. Linde, A. Buzo, and R. M. Gray, *An algorithm for vector quantizer design,* IEEE Transactions on Communications, COM–28:1(1980), pp. 84–95.

[34] K. Marton, *Error exponent for source coding with a fidelity criterion,* IEEE Trans. Inform. Theory, IT–20:2(1974), pp. 197–199.

[35] J. L. Massey, *Guessing and entropy,* Proc. 1994 IEEE Int. Symp. Inform. Theory, (ISIT '94), p. 204, Trondheim, Norway, 1994.

[36] J. Max, *Quantizing fro minimum distortion,* IRE Trans. on Inform. Theory, pp. 7–12, March 1960.

[37] N. Merhav, *Universal coding with minimum probability of code word length overflow,* IEEE Trans. Inform. Theory, 37:3(1991), pp. 556–563.

[38] N. Merhav, *On optimum strategies for minimizing the exponential moments of a given cost function,* http://arxiv.org/PS_cache/arxiv/pdf/1103/1103.2882v1.pdf

[39] N. Merhav, *On optimum parameter modulation–estimation from a large deviations perspective,* submitted to IEEE Trans. Inform. Theory, March 2012.
http://arxiv.org/pdf/1203.4358.pdf

[40] N. Merhav and E. Arikan, *The Shannon cipher system with a guessing wiretapper,* IEEE Trans. Inform. Theory, 45:6(1999), pp. 1860–1866.

[41] N. Merhav and M. Feder, *A strong version of the redundancy–capacity theorem of universal coding,* IEEE Trans. Inform. Theory, 41:3(1995), pp. 714-722.

[42] N. Merhav and M. Feder, *Universal prediction,* IEEE Trans. Inform. Theory, 44:6(1998), pp. 2124–2147.

[43] N. Merhav and I. Kontoyiannis, *Source coding exponents for zero–delay coding with finite memory,* IEEE Trans. Inform. Theory, 49:3(2003), pp. 609–625.

[44] 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*, 38:1(1992), pp. 135–140.

[45] N. Merhav, R. M. Roth, and E. Arikan, *Hierarchical guessing with a fidelity criterion,* IEEE Trans. Inform. Theory, 45:1(1999), pp. 330–337.

[46] M. N. Mishra and B. L. S. Prakasa Rao, *Large deviation probabilities for maximum likelihood estimator and Bayes estimator of a parameter for fractional Ornstein–Uhlenbeck*

type process, Bulletin of Information and Cybernetics, 38(2006), pp. 71–83.

[47] J. B. Moore, R. J. Elliott, and S. Dey, *Risk–sensitive generalizations of minimum variance estimation and control,* Journal of Mathematical Systems and Control, 7:1(1997), pp. 1–15.

[48] D. L. Neuhoff and R. K. Gilbert, *Causal source codes,* IEEE Trans. Inform. Theory, IT–28:5(1982), pp. 701–713.

[49] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, U.S.A., 1972.

[50] J. Rissanen, *Universal coding, information, prediction, and estimation,* IEEE Transactions on Information Theory, IT–30:4(1984), pp. 629–636.

[51] S. Sherman, *Non–mean square error criteria,* IRE Transactions on Information Theory, IT-4(1958), pp. 125–126.

[52] A. Sieders and K. Dzhaparidze, *A large deviations result for parameter estimators and its application to non–linear regression analysis,* Annals of Statistics, 15:3(1987), pp, 1031–1049.

[53] O. Uchida and T. S. Han, *The optimal overflow and underflow probabilities with variable–length coding for the general source,* preprint 1999.

[54] H. van Trees, *Detection, Estimation, and Modulation Theory*, Part I, John Wiley & Sons, New York, 1968.

[55] A. J. Weiss and E. Weinstein, *A lower bound on the mean square error in random parameter estimation,* IEEE Trans. Inform. Theory, IT–31:5(1985), pp. 680–682.

[56] P. Whittle, *Risk–sensitive linear/quadratic/Gaussian control,* Advances in Applied Probability, 13(1981), pp. 764–777.

[57] P. Whittle, *A risk–sensitive maximum principle,* Systems and Control Letters, 15(1990), pp. 183–192.

[58] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, 1965. Reissued by Waveland Press, 1990.

[59] A. D. Wyner, *On the probability of buffer overflow under an arbitrary bounded input-output distribution,* SIAM Journal on Applied Mathematics, 27:4(1974), pp. 544–570.

[60] A. D. Wyner, *Another look at the coding theorem of information theory- a tutorial,* Proc. IEEE, 58:6(1980), pp. 894–913.

[61] A. D. Wyner, *Fundamental limits in information theory,* Proc. IEEE, 69:2(1981), pp. 239–251.

[62] A. D. Wyner and J. Ziv, *On communication of analog data from a bounded source space,* Bell Systems Technical Journal, 48(1969), pp. 3139–3172.