

# Rate–Distortion Function via Minimum Mean Square Error Estimation

Neri Merhav

**Abstract**—We derive a simple general parametric representation of the rate–distortion function of a memoryless source, where both the rate and the distortion are given by integrals whose integrands include the minimum mean square error (MMSE) of the distortion  $\Delta = d(X, Y)$  based on the source symbol  $X$ , with respect to a certain joint distribution of these two random variables. At first glance, these relations may seem somewhat similar to the I-MMSE relations due to Guo, Shamai and Verdú, but they are, in fact, quite different. The new relations among rate, distortion, and MMSE are discussed from several aspects, and more importantly, it is demonstrated that they can sometimes be rather useful for obtaining non-trivial upper and lower bounds on the rate–distortion function, as well as for determining the exact asymptotic behavior for very low and for very large distortion. Analogous MMSE relations hold for channel capacity as well.

**Index Terms**—Rate–distortion function, Legendre transform, estimation, minimum mean square error.

## I. INTRODUCTION

IT has been well known for many years that the derivation of the rate–distortion function of a given source and distortion measure, does not lend itself to closed form expressions, even in the memoryless case, except for a few very simple examples [1],[2],[3],[5]. This has triggered the derivation of some upper and lower bounds, both for memoryless sources and for sources with memory.

One of the most important lower bounds on the rate–distortion function, which is applicable for difference distortion measures (i.e., distortion functions that depend on their two arguments only through the difference between them), is the Shannon lower bound in its different forms, e.g., the discrete Shannon lower bound, the continuous Shannon lower bound, and the vector Shannon lower bound. This family of bounds is especially useful for semi-norm–based distortion measures [5, Section 4.8]. The Wyner–Ziv lower bound [14] for a source with memory is a convenient bound, which is based on the rate–distortion function of the memoryless source formed from the product measure pertaining to the single-letter marginal distribution of the original source and it may be combined elegantly with the Shannon lower bound. The autoregressive lower bound asserts that the rate–distortion function of an autoregressive source is lower bounded by the rate–distortion function of its innovation process, which is again, a memoryless source.

Upper bounds are conceptually easier to derive, as they may result from the performance analysis of a concrete coding

scheme, or from random coding with respect to (w.r.t.) an arbitrary random coding distribution, etc. One well known example is the Gaussian upper bound, which upper bounds the rate–distortion function of an arbitrary memoryless (zero–mean) source w.r.t. the squared error distortion measure by the rate–distortion function of the Gaussian source with the same second moment. If the original source has memory, then the same principle generalizes with the corresponding Gaussian source having the same autocorrelation function as the original source [1, Section 4.6].

In this paper, we focus on a simple general parametric representation of the rate–distortion function which seems to set the stage for the derivation of a rather wide family of both upper bounds and lower bounds on the rate–distortion function. In this parametric representation, both the rate and the distortion are given by integrals whose integrands include the minimum mean square error (MMSE) of the distortion based on the source symbol, with respect to a certain joint distribution of these two random variables. More concretely, given a memoryless source designated by a random variable (RV)  $X$ , governed by a probability function<sup>1</sup>  $p(x)$ , a reproduction variable  $Y$ , governed by a probability function  $q(y)$ , and a distortion measure  $d(x, y)$ , the rate and the distortion can be represented parametrically via a real parameter  $s \in [0, \infty)$  as follows:

$$\begin{aligned} D_s &= D_0 - \int_0^s d\hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\ &= D_\infty + \int_s^\infty d\hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \end{aligned} \quad (1)$$

and

$$\begin{aligned} R_q(D_s) &= \int_0^s d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\ &= R_q(D_\infty) - \int_s^\infty d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X), \end{aligned} \quad (2)$$

where  $D_s$  is the distortion pertaining to parameter value  $s$ ,  $R_q(D_s)$  is the rate–distortion function w.r.t. reproduction distribution  $q$ , computed at  $D_s$ ,  $\Delta = d(X, Y)$ , and  $\text{mmse}_{\hat{s}}(\Delta|X)$  is the MMSE of estimating  $\Delta$  based on  $X$ , where the joint probability function of  $(X, \Delta)$  is induced by the following joint probability function of  $(X, Y)$ :

$$p_s(x, y) = p(x) \cdot w_s(y|x) = p(x) \cdot \frac{q(y)e^{-sd(x,y)}}{Z_x(s)} \quad (3)$$

N. Merhav is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, 32000, Israel. E-mail: merhav@ee.technion.ac.il.

<sup>1</sup>Here, and throughout the sequel, the term “probability function” refers to a probability mass function in the discrete case and to a probability density function in the continuous case.

where  $Z_x(s)$  is a normalization constant, given by  $\int dy q(y) e^{-sd(x,y)}$  in the continuous case, or  $\sum_y q(y) e^{-sd(x,y)}$  in the discrete case.

At first glance, eq. (2) looks somewhat similar to the I-MMSE relation of [6], which relates the mutual information between the input and the output of an additive white Gaussian noise (AWGN) channel and the MMSE of estimating the channel input based on the noisy channel output. As we discuss later on, however, eq. (2) is actually very different from the I-MMSE relation in many respects. In this context, it is important to emphasize that a relation analogous to (2) applies also to channel capacity, as will be discussed in the sequel.

The relations (1) and (2) have actually already been raised in a companion paper [9] (see also [10] for a conference version). Their derivation there was triggered and inspired by certain analogies between the rate-distortion problem and statistical mechanics, which were the main theme of that work. However, the significance and the usefulness of these rate-distortion-MMSE relations were not explored in [9] and [10].

It is the purpose of the present work to study these relations more closely and to demonstrate their utility, which is, as said before, in deriving upper and lower bounds. The underlying idea is that bounds on  $R_q(D)$  (and sometimes also on  $R(D) = \min_q R_q(D)$ ) may be obtained via relatively simple bounds on the MMSE of  $\Delta$  based on  $X$ . These bounds can either be simple technical bounds on the expression of the MMSE itself, or bounds that stem from pure estimation-theoretic considerations. For example, upper bounds may be derived by analyzing the MMSE of a certain sub-optimum estimator, e.g., a linear estimator, which is easy to analyze. Lower bounds can be taken from the available plethora of lower bounds offered by estimation theory, e.g., the Cramér–Rao lower bound.

Indeed, an important part of this work is a section of examples, where it is demonstrated how to use the proposed relations and derive explicit bounds from them. In one of these examples, we derive two sets of upper and lower bounds, one for a certain range of low distortions and the other, for high distortion values. At both edge-points of the interval of distortion values of interest, the corresponding upper and lower bound asymptotically approach the limiting value with the same leading term, and so, they sandwich the exact asymptotic behavior of the rate-distortion function, both in the low distortion limit and in the high distortion limit.

The outline of this paper is as follows. In Section II, we establish notation conventions. In Section III, we formally present the main result, prove it, and discuss its significance from the above-mentioned aspects. In Section IV, we provide a few examples that demonstrate the usefulness of the MMSE relations. Finally, in Section V, we summarize and conclude.

## II. NOTATION CONVENTIONS

Throughout this paper, RV's will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. For example,  $X$  is a random

variable,  $x$  is a specific realization of  $X$ , and  $\mathcal{X}$  is the alphabet in which  $X$  and  $x$  take on values. This alphabet may be finite, countably infinite, or a continuum, like the real line  $\mathbb{R}$  or an interval  $[a, b] \subset \mathbb{R}$ .

Sources and channels will be denoted generically by the letter  $p$ , or  $q$ , which will designate also their corresponding probability functions, i.e., a probability density function (pdf) in the continuous case, or a probability mass function (pmf) in the discrete case. Information-theoretic quantities, like entropies and mutual informations, will be denoted according to the usual conventions of the information theory literature, e.g.,  $H(X)$ ,  $I(X; Y)$ , and so on. If a RV is continuous-valued, then its differential entropy and conditional differential entropy will be denoted with  $h$  instead of  $H$ , i.e.,  $h(X)$  is the conditional differential entropy of  $X$ ,  $h(X|Y)$  is the conditional differential entropy of  $X$  given  $Y$ , and so on. The expectation operator will be denoted, as usual, by  $\mathbf{E}\{\cdot\}$ .

Given a source RV  $X$ , governed by a probability function  $p(x)$ ,  $x \in \mathcal{X}$ , a reproduction RV  $Y$ , governed by a probability function  $q(y)$ ,  $y \in \mathcal{Y}$ , and a distortion measure  $d : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ , we define the rate-distortion function of  $X$  w.r.t. distortion measure  $d$  and reproduction distribution  $q$  as

$$R_q(D) \triangleq \min I(X; Y), \quad (4)$$

where  $X \sim p$  and the minimum is across all channels  $\{w(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$  that satisfy  $\mathbf{E}\{d(X, Y)\} \leq D$  and  $\mathbf{E}\{w(y|X)\} = q(y)$  for all  $y \in \mathcal{Y}$ . Clearly, the rate-distortion function,  $R(D)$ , is given by  $R(D) = \inf_q R_q(D)$ . We will also use the notation  $\Delta \triangleq d(X, Y)$ . Obviously, since  $X$  and  $Y$  are RV's, then so is  $\Delta$ .

## III. MMSE RELATIONS: BASIC RESULT AND DISCUSSION

Throughout this section, our definitions will assume that both  $\mathcal{X}$  and  $\mathcal{Y}$  are finite alphabets. Extensions to continuous alphabets will be obtained by a limit of fine quantizations, with summations eventually being replaced by integrations.

Referring to the notation defined in Section II, for a given positive real  $s$ , define the conditional probability function

$$w_s(y|x) \triangleq \frac{q(y) e^{-sd(x,y)}}{Z_x(s)} \quad (5)$$

where

$$Z_x(s) \triangleq \sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)} \quad (6)$$

and the joint pmf

$$p_s(x, y) = p(x) w_s(y|x). \quad (7)$$

Further, let

$$\begin{aligned} \text{mmse}_s(\Delta|X) &= \mathbf{E}_s\{\Delta - \mathbf{E}\{\Delta|X\}\}^2 \\ &= \mathbf{E}_s\{[d(X, Y) - \mathbf{E}_s\{d(X, Y)|X\}]^2\} \end{aligned} \quad (8)$$

where  $\mathbf{E}_s\{\cdot\}$  is the expectation operator w.r.t.  $\{p_s(x, y)\}$ , and defining  $\psi(x)$  as the conditional expectation  $\mathbf{E}_s\{d(x, Y)|X = x\}$  w.r.t.  $\{w_s(y|x)\}$ ,  $\mathbf{E}_s\{d(X, Y)|X\}$  is defined as  $\psi(X)$ .

Our main result, in this section, is the following (the proof appears in the Appendix):

*Theorem 1:* The function  $R_q(D)$  can be represented parametrically via the parameter  $s \in [0, \infty)$  as follows:

(a) The distortion is obtained by

$$\begin{aligned} D_s &= D_0 - \int_0^s d\hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\ &= D_\infty + \int_s^\infty d\hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \end{aligned} \quad (9)$$

where

$$D_0 = \sum_{x,y} p(x)q(y)d(x,y) \quad (10)$$

and

$$D_\infty = \sum_x p(x) \min_y d(x,y). \quad (11)$$

(b) The rate is given by

$$\begin{aligned} R_q(D_s) &= \int_0^s d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\ &= R_q(D_\infty) - \int_s^\infty d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X). \end{aligned} \quad (12)$$

In the remaining part of this section, we discuss the significance and the implications of Theorem 1 from several aspects.

#### *Some General Technical Comments*

The parameter  $s$  has the geometric meaning of the negative local slope of the function  $R_q(D)$ . This is easily seen by taking the derivatives of (9) and (12), i.e.,  $dR_q(D_s)/ds = s \cdot \text{mmse}_s(\Delta|X)$  and  $dD_s/ds = -\text{mmse}_s(\Delta|X)$ , whose ratio is  $R'_q(D_s) = -s$ . This means also that the parameter  $s$  plays the same role as in the well known parametric representations of [1] and [5], which is to say that it can also be thought of as the Lagrange multiplier of the minimization of  $[I(X;Y) + sE\{d(X,Y)\}]$  subject to the reproduction distribution constraint.

On a related note, we point out that Theorem 1 is based on the following representation of  $R_q(D)$ :

$$R_q(D) = -\min_{s \geq 0} \left[ sD + \sum_{x \in \mathcal{X}} p(x) \ln Z_x(s) \right], \quad (13)$$

which we prove in the Appendix as the first step in the proof of Theorem 1.

It should be emphasized that the pmf  $q$ , that plays a role in the definition of  $w_{\hat{s}}(y|x)$  (and hence also the definition of  $\text{mmse}_{\hat{s}}(\Delta|X)$ ) should be kept *fixed* throughout the integration, independently of the integration variable  $\hat{s}$ , since it is the same pmf as in the definition of  $R_q(D)$ . Thus, even if  $q$  is known to be optimum for a given target distortion  $D$  (and then it yields  $R(D)$ ), the pmf  $q$  must be kept unaltered throughout the integration, in spite of the fact that for other values of  $\hat{s}$  (which correspond to other distortion levels), the optimum reproduction pmf might be different. In particular, note that the marginal of  $Y$ , that is induced from the joint pmf

$p_s(x,y)$ , may not necessarily agree with  $q$ . Thus,  $p_s(x,y)$  should only be considered as an auxiliary joint distribution that defines  $\text{mmse}_{\hat{s}}(\Delta|X)$ .

#### *Using Theorem 1 for Bounds on $R_q(D)$*

As was briefly explained in the Introduction (and will also be demonstrated in the next section), Theorem 1 may set the stage for the derivation of upper and lower bounds to  $R_q(D)$  for a general reproduction distribution  $q$ , and hence also for the rate-distortion function  $R(D)$  when the optimum  $q$  is happened to be known or is easily derivable (e.g., from symmetry and convexity considerations).

The basic underlying idea is that bounds on  $R_q(D)$  may be induced from bounds on  $\text{mmse}_{\hat{s}}(\Delta|X)$  across the integration interval. The bounds on the MMSE may either be derived from purely technical considerations, upon analyzing the expression of the MMSE directly, or by using estimation-theoretic tools. In the latter case, lower bounds may be obtained from fundamental lower bounds to the MMSE, like the Bayesian Cramér-Rao bound, or more advanced lower bounds available from the estimation theory literature, for example, the Weiss-Weinstein bound [12],[13], whenever applicable. Upper bounds may be obtained by analyzing the mean square error (MSE) of a specific (sub-optimum) estimator, which is relatively easy to analyze, or more generally by analyzing the performance of the best estimator within a certain limited class of estimators, like the class of linear estimators of the ‘observation’  $X$ , or a certain fixed function of  $X$ .

In Theorem 1 we have deliberately presented two integral forms for both the rate and the distortion. As  $D_s$  is monotonically decreasing and  $R_q(D_s)$  is monotonically increasing in  $s$ , the integrals at the first lines of both eqs. (9) and (12), which include relatively small values of  $\hat{s}$ , naturally lend themselves to derivation of bounds in the low-rate (high distortion) regime, whereas the second lines of these equations are more suitable in low-distortion (high resolution) region. For example, to derive an upper bound on  $R_q(D)$  in the high-distortion range, one would need a lower bound on  $\text{mmse}_{\hat{s}}(\Delta|X)$  to be used in the first line of (9) and an upper bound on  $\text{mmse}_{\hat{s}}(\Delta|X)$  to be substituted into the first line of (12). If one can then derive, from the former, an upper bound on  $s$  as a function of  $D$ , and substitute it into the upper bound on the rate in terms on  $s$ , then this will result in an upper bound to  $R_q(D)$ . A similar kind of reasoning is applicable to the derivation of other types of bounds. This point will be demonstrated mainly in Examples C and D in the next section.

#### *Comparison to the I-MMSE Relations*

In the more conceptual level, item (b) of Theorem 1 may remind the familiar reader about well-known results due to Guo, Shamai and Verdú [6], which are referred to as I-MMSE relations (as well as later works that generalize these relations). The similarity between eq. (12) and the I-MMSE relation (in its basic form) is that in both cases a mutual information is expressed as an integral whose integrand includes the MMSE of a certain random variable (or vector) given some

observation(s). However, to the best of our judgment, this is the only similarity.

In order to sharpen the comparison between the two relations, it is instructive to look at the special case where all random variables are Gaussian and the distortion measure is quadratic: In the context of Theorem 1, consider  $Y$  to be a zero-mean Gaussian RV with variance  $\sigma_y^2$ , and let  $d(x, y) = (x - y)^2$ . As will be seen in Example B of the next section, this then means that  $w_s(y|x)$  can be described by the additive Gaussian channel  $Y = aX + Z$ , where  $a = 2s\sigma_y^2/(1 + 2s\sigma_y^2)$  and  $Z$  is a zero-mean Gaussian RV, independent of  $X$ , and with variance  $\sigma_y^2/(1 + 2s\sigma_y^2)$ . Here, we have  $\Delta = (Y - X)^2 = [Z - (1 - a)X]^2$ . Thus, the integrand of (12) includes the MMSE in estimating  $[Z - (1 - a)X]^2$  based on the *channel input*  $X$ . It is therefore about estimating a certain function of  $Z$  and  $X$ , where  $X$  is the observation at hand and  $Z$  is independent of  $X$ .

This is very different from the paradigm of the I-MMSE relation: there the channel is  $Y = \sqrt{\text{snr}}X + Z$ , where  $Z$  is standard normal, the integration variable is  $\text{snr}$ , and the estimated RV is  $X$  (or equivalently,  $Z$ ) based on the *channel output*,  $Y$ . Also, by comparing the two channels, it is readily seen that the integration variable  $s$ , in our setting, can be related to the integration variable,  $\text{snr}$ , of the I-MMSE relation according to

$$\text{snr} = \frac{4s^2}{\sigma_y^2(1 + 2s\sigma_y^2)}, \quad (14)$$

and so, the relation between the two integration variables is highly non-linear. We therefore observe that the two MMSE results are fairly different.

#### Analogous MMSE Formula for Channel Capacity

Eq. (13) can be understood conveniently as an achievable rate using a simple random coding argument (see Appendix): The coding rate  $R$  should be (slightly larger than) the large deviations rate function of the probability of the event  $\{\sum_{i=1}^n d(x_i, Y_i) \leq nD\}$ , where  $(x_1, \dots, x_n)$  is a typical source sequence and  $(Y_1, \dots, Y_n)$  are drawn i.i.d. from  $q$ . As is well known, a similar random coding argument applies to channel coding (see also [8]): Channel capacity can be obtained as the large deviations rate function of the event  $\{\sum_{i=1}^n d(X_i, y_i) \leq nD\}$ , where now  $(y_1, \dots, y_n)$  is a channel output sequence typical to  $q$ ,  $(X_1, \dots, X_n)$  are drawn i.i.d. according to a given input pmf  $\{p(x)\}$ , the distortion measure is chosen to be  $d(x, y) = -\ln w(y|x)$  ( $\{w(y|x)\}$  being the channel transition probabilities) and  $D = H(Y|X)$ . Thus, the analogue of (13) is

$$C_p = -\min_{s \geq 0} \left[ sH(Y|X) + \sum_{y \in Y} q(y) \ln Z_y(s) \right] \quad (15)$$

where

$$Z_y(s) = \sum_{x \in \mathcal{X}} p(x) w^s(y|x) \quad (16)$$

and the minimizing  $s$  is always  $s^* = 1$ . Consequently, the

corresponding integrated MMSE formula would read

$$C_p = \int_0^1 ds \cdot s \cdot \text{mmse}_s[\ln p(Y|X)|Y], \quad (17)$$

where  $\text{mmse}_s[\ln p(Y|X)|Y]$  is defined w.r.t. the joint pmf

$$q_s(x, y) = q(y) v_s(x|y) = q(y) \cdot \frac{p(x) w^s(y|x)}{Z_y(s)}. \quad (18)$$

Eq. (17) seems to be less useful than the analogous rate-distortion formulas, for a very simple reason: Since the channel is given, then once the input pmf  $p$  is given too (which is required for the use of (17)), one can simply compute the mutual information, which is easier than applying (17). This is different from the situation in the rate-distortion problem, where even if both  $p$  and  $q$  are given, in order to compute  $R_q(D)$  in the direct way, one still needs to minimize the mutual information w.r.t. the channel between  $X$  and  $Y$ . Eq. (17) is therefore presented here merely for the purpose of drawing the duality.

#### Analogies With Statistical Mechanics

As was shown in [11] and further advocated in [8], the Legendre relation (13) has a natural statistical-mechanical interpretation, where  $Z_x(s)$  plays the role of a partition function of a system (indexed by  $x$ ),  $d(x, y)$  is an energy function (Hamiltonian) and  $s$  plays the role of inverse temperature (normally denoted by  $\beta$  in the Physics literature). The minimizing  $s$  is then the equilibrium inverse temperature when  $|\mathcal{X}|$  systems (each indexed by  $x$ , with  $n(x) = np(x)$  particles and Hamiltonian  $\mathcal{E}_x(y) = d(x, y)$ ) are brought into thermal contact and a total energy of  $nD$  is split among them. In this case,  $-R_q(D)$  is the thermodynamical entropy of the combined system and the MMSE, which is  $dD_s/ds$ , is intimately related to the heat capacity of the system.

An alternative, though similar, interpretation was given in [9],[10], where the parameter  $s$  was interpreted as being proportional to a generalized force acting on the system (e.g., pressure or magnetic field), and the distortion variable is the conjugate physical quantity influenced by this force (e.g., volume in the case of pressure, or magnetization in the case of a magnetic field). In this case, the minimizing  $s$  means the equal force that each one of the various subsystems is applying on the others when they are brought into contact and they equilibrate (e.g., equal pressures between two volumes of a gas separated by piston which is free to move). In this case,  $-R_q(D)$  is interpreted as the free energy of the system, and the MMSE formulas are intimately related to the fluctuation-dissipation theorem in statistical mechanics.

More concretely, it was shown in [9] that given a source distribution and a distortion measure, we can describe (at least conceptually) a concrete physical system that emulates the rate-distortion problem in the following manner: When no force is applied to the system, its total length is  $nD_0$ , where  $n$  is the number of particles in the system (and also the block length in the rate-distortion problem), and  $D_0$  is as defined above. If one applies to the system a contracting force, that increases from zero to some final value  $\lambda$ , such that the length

of the system shrinks to  $nD$ , where  $D < D_0$  is analogous to a prescribed distortion level, then the following two facts hold true: (i) An *achievable lower bound* on the total amount of mechanical work that must be carried out by the contracting force in order to shrink the system to length  $nD$ , is given by

$$W \geq nkTR_q(D), \quad (19)$$

where  $k$  is Boltzmann's constant and  $T$  is the temperature. (ii) The final force  $\lambda$  is related to  $D$  according to  $\lambda = kTR'_q(D)$ , where  $R'_q(\cdot)$  is the derivative of  $R_q(\cdot)$ . Thus, the rate-distortion function plays the role of a fundamental limit, not only in Information Theory, but in Physics as well.

#### IV. EXAMPLES

In this section, we provide a few examples for the use of Theorem 1. The first two examples are simple and well known, and their purpose is just to demonstrate how to use this theorem in order to calculate rate-distortion functions. The third example is aimed to demonstrate how Theorem 1 can be useful as a new method to evaluate the behavior of a certain rate-distortion function (which is apparently not straightforward to derive otherwise) at both the low distortion (a.k.a. high resolution) regime and the high distortion regime. Specifically, we first derive, for this example, upper and lower bounds on  $R(D)$ , which are applicable in certain ranges of high-distortion. These bounds have the same asymptotic behavior as  $D$  tends to its maximum possible value, and so, they sandwich the exact high-distortion asymptotic behavior of the true rate-distortion function. A similar analysis is then carried out in the low distortion range, and again, the two bounds have the same limiting behavior in the very low distortion limit. In the fourth and last example, we show how Theorem 1 can easily be used to evaluate the high-resolution behavior of the rate distortion function for a general power-law distortion measure of the form  $d(x, y) = |x - y|^r$ .

##### A. Binary Symmetric Source and Hamming Distortion

Perhaps the simplest example is that of the binary symmetric source (BSS) and the Hamming distortion measure. In this case, the optimum  $q$  is also symmetric. Here  $\Delta = d(X, Y)$  is a binary RV with

$$\Pr\{\Delta = 1|X = x\} = \frac{e^{-s}}{1 + e^{-s}} \quad (20)$$

independently of  $x$ . Thus, the MMSE estimator of  $d(X, Y)$  based on  $X$  is

$$\hat{\Delta} = \frac{e^{-s}}{1 + e^{-s}}, \quad (21)$$

regardless of  $X$ , and so the resulting MMSE (which is simply the variance in this case) is easily found to be

$$\text{mmse}_s(\Delta|X) = \frac{e^{-s}}{(1 + e^{-s})^2}. \quad (22)$$

Accordingly,

$$D = \frac{1}{2} - \int_0^s \frac{e^{-\hat{s}} d\hat{s}}{(1 + e^{-\hat{s}})^2} = \frac{e^{-s}}{1 + e^{-s}} \quad (23)$$

and

$$\begin{aligned} R(D) &= \int_0^s \frac{\hat{s}e^{-\hat{s}} d\hat{s}}{(1 + e^{-\hat{s}})^2} \\ &= \ln 2 + \frac{se^s}{1 + e^s} - \ln(1 + e^s) \\ &= \ln 2 - h_2\left(\frac{e^s}{1 + e^s}\right) \\ &= \ln 2 - h_2(D), \end{aligned} \quad (24)$$

where  $h_2(u) = -u \ln u - (1 - u) \ln(1 - u)$  is the binary entropy function.

##### B. Quadratic distortion and Gaussian Reproduction

Another classic example concerns a general source with  $\sigma_x^2 = E\{X^2\} < \infty$ , the quadratic distortion  $d(x, y) = (x - y)^2$ , and a Gaussian reproduction distribution, namely,  $q(y)$  is the pdf of a zero-mean Gaussian RV with variance  $\sigma_y^2 = \sigma_x^2 - D$ , for a given  $D < \sigma_x^2$ . In this case, it well known that  $R_q(D) = \frac{1}{2} \ln \frac{\sigma_x^2}{D}$  (even without assuming that the source  $X$  is Gaussian). We now demonstrate how this result is obtained from the MMSE formula of Theorem 1.<sup>2</sup>

First, observe that since  $q(y)$  is the pdf pertaining to  $\mathcal{N}(0, \sigma_x^2 - D)$ , then

$$w_s(y|x) = \frac{q(y)e^{-s(y-x)^2}}{\int_{-\infty}^{+\infty} dy' q(y')e^{-s(y'-x)^2}} \quad (25)$$

is easily found to correspond to the Gaussian additive channel

$$Y = \frac{2s(\sigma_x^2 - D)}{1 + 2s(\sigma_x^2 - D)} \cdot X + Z \quad (26)$$

where  $Z$  is a zero-mean Gaussian RV with variance  $\sigma_z^2 = (\sigma_x^2 - D)/[1 + 2s(\sigma_x^2 - D)]$ , and  $Z$  is uncorrelated with  $X$ . Now,

$$\begin{aligned} \Delta &= (Y - X)^2 \\ &= \left[ Y - \frac{2s(\sigma_x^2 - D)}{1 + 2s(\sigma_x^2 - D)} \cdot X - \frac{X}{1 + 2s(\sigma_x^2 - D)} \right]^2 \\ &= (Z - \alpha X)^2 \\ &= Z^2 - 2\alpha XZ + \alpha^2 X^2 \end{aligned} \quad (27)$$

where  $\alpha \triangleq 1/[1 + 2s(\sigma_x^2 - D)]$ . Thus, the MMSE estimator of  $\Delta$  given  $X$  is obtained by

$$\begin{aligned} \hat{\Delta} &= \mathbf{E}\{\Delta|X\} \\ &= \mathbf{E}\{Z^2|X\} - 2\alpha X \mathbf{E}\{Z|X\} + \alpha^2 X^2 \\ &= \mathbf{E}\{Z^2\} - 2\alpha X \mathbf{E}\{Z\} + \alpha^2 X^2 \\ &= \mathbf{E}\{Z^2\} + \alpha^2 X^2 \\ &= \sigma_z^2 + \alpha^2 X^2, \end{aligned} \quad (28)$$

<sup>2</sup>We are not arguing here that this is the simplest way to calculate  $R_q(D)$  in this example, the purpose is merely to demonstrate how Theorem 1 can be used.

which yields

$$\begin{aligned}
& \text{mmse}_s\{\Delta|X\} \\
&= \mathbf{E}\{(\hat{\Delta} - \Delta)^2\} \\
&= \mathbf{E}\{(\sigma_z^2 + \alpha^2 X^2 - Z^2 + 2\alpha XZ - \alpha^2 X^2)^2\} \\
&= 2\sigma_z^4 + 4\alpha^2 \sigma_x^2 \sigma_z^2 \\
&= \frac{2(\sigma_x^2 - D)^2}{[1 + 2s(\sigma_x^2 - D)]^2} + \frac{4\sigma_x^2(\sigma_x^2 - D)}{[1 + 2s(\sigma_x^2 - D)]^3}. \quad (29)
\end{aligned}$$

Now, in our case,  $D_0 = \sigma_x^2 + \sigma_y^2 = 2\sigma_x^2 - D$ , and so, for  $s = 1/(2D)$ , we get

$$\begin{aligned}
D_s &= D_0 - \int_0^s d\hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\
&= 2\sigma_x^2 - D - \\
&\quad 2(\sigma_x^2 - D)^2 \int_0^{1/2D} \frac{d\hat{s}}{[1 + 2\hat{s}(\sigma_x^2 - D)]^2} - \\
&\quad 4\sigma_x^2(\sigma_x^2 - D) \int_0^{1/2D} \frac{d\hat{s}}{[1 + 2\hat{s}(\sigma_x^2 - D)]^3} \\
&= 2\sigma_x^2 - D + \\
&\quad (\sigma_x^2 - D) \left[ \frac{1}{1 + 2\hat{s}(\sigma_x^2 - D)} \right]_0^{1/2D} + \\
&\quad \sigma_x^2 \left\{ \frac{1}{[1 + 2\hat{s}(\sigma_x^2 - D)]^2} \right\}_0^{1/2D} \quad (30)
\end{aligned}$$

which, after some straightforward algebra, gives  $D_s = D$ . I.e.,  $s$  and  $D$  are indeed related by  $s = 1/(2D)$ , or  $D = 1/(2s)$ . Finally,

$$\begin{aligned}
R_q(D) &= \int_0^s d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\
&= 2(\sigma_x^2 - D)^2 \int_0^{1/2D} \frac{\hat{s} d\hat{s}}{[1 + 2\hat{s}(\sigma_x^2 - D)]^2} + \\
&\quad 4\sigma_x^2(\sigma_x^2 - D) \int_0^{1/2D} \frac{\hat{s} d\hat{s}}{[1 + 2\hat{s}(\sigma_x^2 - D)]^3} \\
&= \frac{1}{2} \{ \ln[1 + 2s(\sigma_x^2 - D)] + \\
&\quad \frac{1}{1 + 2s(\sigma_x^2 - D)} \}^{1/2D} + \\
&\quad \frac{\sigma_x^2}{\sigma_x^2 - D} \left[ \frac{1}{2[1 + 2s(\sigma_x^2 - D)]^2} - \right. \\
&\quad \left. \frac{1}{1 + 2s(\sigma_x^2 - D)} \right]^{1/2D} \quad (31)
\end{aligned}$$

which yields, after a simple algebraic manipulation,  $R_q(D) = \frac{1}{2} \ln \frac{\sigma_x^2}{D}$ .

### C. Quadratic Distortion and Binary Reproduction

In this example, we again assume the quadratic distortion measure, but now, instead of Gaussian reproduction code-words, we impose binary reproduction,  $y \in \{-a, +a\}$ , where  $a$  is a given constant.<sup>3</sup> Clearly, if the pdf of the source  $X$  is symmetric about the origin, then the best output distribution

is also symmetric, i.e.,  $q(+a) = q(-a) = 1/2$ . Thus,  $R_q(D) = R(D)$  for every  $D$ , given this choice of  $q$ . The channel  $w_s(y|x)$  is now given by

$$w_s(y|x) = \frac{e^{-s(y-x)^2}}{e^{-s(x-a)^2} + e^{-s(x+a)^2}} = \frac{e^{2sxy}}{2 \cosh(2asx)}. \quad (32)$$

Note that in this case, the minimum possible distortion (obtained for  $s \rightarrow \infty$ ) is given by  $D_\infty = \mathbf{E}\{[X - a \text{sgn}(X)]^2\}$ . Thus, the rate-distortion function is actually defined only for  $D \geq D_\infty$ . The maximum distortion of interest is  $D_0 = \sigma_x^2 + a^2$ , pertaining to the choice  $s = 0$ , where  $X$  and  $Y$  are independent. To the best of our knowledge, there is no closed form expression for  $R(D)$  in this example. The parametric representation of  $D_s$  and  $R(D_s)$ , both as functions of  $s$ , does not seem to lend itself to an explicit formula of  $R(D)$ . The reason is that

$$\begin{aligned}
D_s &= \mathbf{E}\{(Y - X)^2\} \\
&= \sigma_x^2 + a^2 - 2\mathbf{E}\{XY\} \\
&= \sigma_x^2 + a^2 - 2\mathbf{E}\{X \cdot \mathbf{E}\{Y|X\}\} \\
&= \sigma_x^2 + a^2 - 2a\mathbf{E}\{X \tanh(2asX)\} \quad (33)
\end{aligned}$$

and there is no apparent closed-form expression of  $s$  a function of  $D$ , which can be substituted into the expression of  $R(D_s)$ .

Consider the MMSE estimator of  $\Delta = (Y - X)^2 = X^2 + a^2 - 2XY$ :

$$\begin{aligned}
\hat{\Delta} &= \mathbf{E}\{(Y - X)^2|X\} \\
&= X^2 + a^2 - 2X\mathbf{E}\{Y|X\} \\
&= X^2 + a^2 - 2aX \tanh(2asX). \quad (34)
\end{aligned}$$

The MMSE is then

$$\begin{aligned}
\text{mmse}_s(\Delta|X) &= \mathbf{E}\{[2X(Y - a \tanh(2asX))]^2\} \\
&= 4a^2[\sigma_x^2 - \mathbf{E}\{X^2 \tanh^2(2asX)\}]. \quad (35)
\end{aligned}$$

We first use this expression to obtain upper and lower bounds on  $R(D)$  which are asymptotically exact in the range of high distortion levels (small  $s$ ). Subsequently, we do the same for the range of low distortion (large  $s$ ).

*High Distortion.* Consider first the high distortion regime. For small  $s$ , we can safely upper bound  $\tanh^2(2asX)$  by  $(2asX)^2$  and get

$$\begin{aligned}
\text{mmse}_s(\Delta|X) &\geq 4a^2(\sigma_x^2 - 4a^2s^2\mathbf{E}\{X^4\}) \\
&= 4a^2\sigma_x^2 - 16a^4\rho_x^4s^2 \quad (36)
\end{aligned}$$

where  $\rho_x^4 \triangleq \mathbf{E}\{X^4\}$ . This results in the following lower bound to  $R(D_s)$ :

$$\begin{aligned}
R(D_s) &= \int_0^s d\hat{s} \cdot \hat{s} \cdot \text{mmse}_{\hat{s}}(\Delta|X) \\
&\geq \int_0^s d\hat{s} \cdot \hat{s} [4a^2\sigma_x^2 - 16a^4\rho_x^4\hat{s}^2] \\
&= 2a^2\sigma_x^2s^2 - 4a^4\rho_x^4s^4 \triangleq r(s). \quad (37)
\end{aligned}$$

To get a lower bound to  $D_s$ , we need an upper bound to the MMSE. An obvious upper bound (which is tight for small  $s$ )

<sup>3</sup>The derivation, in this example, can be extended to apply also to larger finite reproduction alphabets.

is given by  $4a^2\sigma_x^2$ , which yields:

$$\begin{aligned} D_s &= D_0 - \int_0^s d\hat{s} \cdot \text{mmse}_s(\Delta|X) \\ &\geq D_0 - \int_0^s d\hat{s} \cdot (4a^2\sigma_x^2) \\ &= D_0 - 4a^2\sigma_x^2 s \end{aligned} \quad (38)$$

or

$$s \geq \frac{D_0 - D_s}{4a^2\sigma_x^2}. \quad (39)$$

Consider now the range  $s \in [0, \sigma_x/(2a\rho_x^2)]$ , which is the range where  $r(s)$  is monotonically increasing as a function of  $s$ . In this range, a lower bound on  $s$  would yield a lower bound on  $r(s)$ , and hence a lower bound to  $R(D_s)$ . Specifically, for  $s \in [0, \sigma_x/(2a\rho_x^2)]$ , we get

$$\begin{aligned} R(D_s) &\geq r(s) \\ &\geq r\left(\frac{D_0 - D_s}{4a^2\sigma_x^2}\right) \\ &= \frac{(D_0 - D_s)^2}{8a^2\sigma_x^2} - \frac{\rho_x^4(D_0 - D_s)^4}{64a^4\sigma_x^8}. \end{aligned} \quad (40)$$

In other words, we obtain the lower bound

$$R(D) \geq \frac{(D_0 - D)^2}{8a^2\sigma_x^2} - \frac{\rho_x^4(D_0 - D)^4}{64a^4\sigma_x^8} \triangleq R_L(D). \quad (41)$$

for the range of distortions  $D \in [D_0 - 2a\sigma_x^3/\rho_x^2, D_0]$ . It is obvious that, at least in some range of high distortion levels, this bound is better than the Shannon lower bound,

$$R_S(D) = h(X) - \frac{1}{2} \ln(2\pi e D), \quad (42)$$

where  $h(X)$  is the differential entropy of  $X$ . This can be seen right away from the fact that  $R_S(D)$  vanishes at  $D = (2\pi e)^{-1}e^{2h(X)} \leq \sigma_x^2$ , whereas the bound  $R_L(D)$  of (41) vanishes at  $D_0 = \sigma_x^2 + a^2$ , which is strictly larger.

By applying the above-mentioned upper bound to the MMSE in the rate equation, and the lower bound to the MMSE – in the distortion equation, we can also get an upper bound to  $R(D)$  in the high-distortion range, in a similar manner. Specifically,

$$R(D_s) \leq \int_0^s d\hat{s} \cdot \hat{s}(4a^2\sigma_x^2) = 2a^2\sigma_x^2 s^2, \quad (43)$$

and

$$\begin{aligned} D_s &\leq D_0 - \int_0^s d\hat{s}(4a^2\sigma_x^2 - 16a^4\rho_x^4\hat{s}^2) \\ &= D_0 - 4a^2\sigma_x^2 s + \frac{16}{3}a^4\rho_x^4 s^3 \triangleq \delta(s). \end{aligned} \quad (44)$$

Considering again the range  $s \in [0, \sigma_x/(2a\rho_x^2)]$ , where  $\delta(s)$  is monotonically decreasing, the inverse function  $\delta^{-1}(D)$  is monotonically decreasing as well, and so an upper bound on  $R(D)$  will be obtained by substituting  $\delta^{-1}(D)$  instead of  $s$  in the bound on the rate, i.e.,  $R(D) \leq 2a^2\sigma_x^2[\delta^{-1}(D)]^2$ . To obtain an explicit expression for  $\delta^{-1}(D)$ , we need to solve a cubic equation in  $s$  and select the relevant solution among the three. Fortunately, since this cubic equation has no quadratic term, the expression of the solution can be found

trigonometrically and it is relatively simple (see, e.g., [7, p. 9]): Specifically, the cubic equation  $s^3 + As + B = 0$  has solutions of the form  $s = m \cos \theta$ , where  $m = 2\sqrt{-A/3}$  and  $\theta$  is any solution to the equation  $\cos(3\theta) = \frac{3B}{Am}$ . In other words, the three solutions to the above cubic equation are  $s_i = m \cos \theta_i$ , where

$$\theta_i = \frac{1}{3} \cos^{-1} \left( \frac{3B}{Am} \right) + \frac{2\pi(i-1)}{3}, \quad i = 1, 2, 3, \quad (45)$$

with  $\cos^{-1}(t)$  being defined as the unique solution to the equation  $\cos \alpha = t$  in the range  $\alpha \in [0, \pi]$ . In our case,

$$A = -\frac{3\sigma_x^2}{4a^2\rho_x^4}, \quad B = \frac{3(D_0 - D)}{16a^4\rho_x^4}, \quad (46)$$

and so, the relevant solution for  $s$  (i.e., the one that tends to zero as  $D \rightarrow D_0$ ), which is  $\delta^{-1}(D)$ , is given by

$$\begin{aligned} \delta^{-1}(D) &= \frac{\sigma_x}{a\rho_x^2} \cos \left[ \frac{1}{3} \cos^{-1} \left( \frac{3\rho_x^2(D - D_0)}{4a\sigma_x^3} \right) + \frac{4\pi}{3} \right] \\ &= \frac{\sigma_x}{a\rho_x^2} \cos \left[ \frac{1}{3} \left( \frac{\pi}{2} + \sin^{-1} \left( \frac{3\rho_x^2(D_0 - D)}{4a\sigma_x^3} \right) \right) + \frac{4\pi}{3} \right] \\ &= \frac{\sigma_x}{a\rho_x^2} \sin \left[ \frac{1}{3} \sin^{-1} \left( \frac{3\rho_x^2(D_0 - D)}{4a\sigma_x^3} \right) \right], \end{aligned} \quad (47)$$

where  $\sin^{-1}(t)$  is defined as the unique solution to the equation  $\sin \alpha = t$  in the range  $\alpha \in [-\pi/2, \pi/2]$ . This yields the upper bound

$$\begin{aligned} R(D) &\leq \frac{2\sigma_x^4}{\rho_x^4} \sin^2 \left[ \frac{1}{3} \sin^{-1} \left( \frac{3\rho_x^2(D_0 - D)}{4a\sigma_x^3} \right) \right] \\ &\triangleq R_U(D). \end{aligned} \quad (48)$$

for the range of distortions  $D \in [D_0 - 4a\sigma_x^3/(3\rho_x^2), D_0]$ .

For very small  $s$ , since the upper and the lower bound to the MMSE asymptotically coincide (namely,  $\text{mmse}_s(\Delta|X) \approx 4a^2\sigma_x^2$ ), then both  $R_U(D)$  and  $R_L(D)$  exhibit the same behavior near  $D = D_0$ , and hence so does the true rate-distortion function,  $R(D)$ , which is

$$R(D) \approx \frac{(D_0 - D)^2}{8a^2\sigma_x^2} \quad (49)$$

or, stated more rigorously,

$$\lim_{D \uparrow D_0} \frac{R(D)}{(D_0 - D)^2} = \frac{1}{8a^2\sigma_x^2}. \quad (50)$$

Note that the high-distortion behavior of  $R(D)$  depends on the pdf of  $X$  only via its second order moment  $\sigma_x^2$ . On the other hand, the upper and lower bounds,  $R_U(D)$  and  $R_L(D)$ , depend only on  $\sigma_x^2$  and the fourth order moment,  $\rho_x^4$ .

In Fig. 1, we display the upper bound  $R_U(D)$  (solid curve) and the lower bound  $R_L(D)$  (dashed curve) for the choice  $\sigma_x^2 = a^2 = 1$  (hence  $D_0 = \sigma_x^2 + a^2 = 2$ ) and  $\rho_x^4 = 3$ , which is suitable for the Gaussian source. The range of displayed distortions,  $[1.25, 2]$ , is part of the range where both bounds are valid in this numerical example. As can be seen, the functions  $R_L(D)$  and  $R_U(D)$  are very close throughout the interval  $[1.7, 2]$ , which is a fairly wide range of distortion levels. The corresponding Shannon lower bound, in this case,

which is  $R_S(D) = \max\{0, \frac{1}{2} \ln \frac{1}{D}\}$ , vanishes for all  $D \geq 1$  and hence also in the range displayed in the graph.

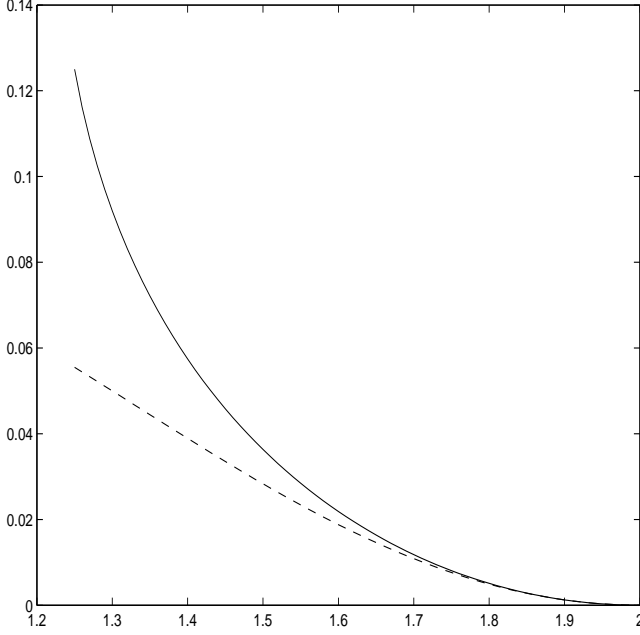


Fig. 1. The upper bound  $R_U(D)$  (solid curve) and the lower bound  $R_L(D)$  (dashed curve) in the high-distortion regime for  $\sigma_x^2 = a^2 = 1$  and  $\rho_x^4 = 3$ . The Shannon lower bound vanishes in this distortion range.

*Low Distortion.* We now consider the small distortion regime, where  $s$  is very large. Define the function

$$f(u) = \left( \frac{1-u}{1+u} \right)^2 \quad u \in [0, 1] \quad (51)$$

and consider the Taylor series expansion of  $f(u)$  around  $u = 0$ , which, for the sake of convenience, will be represented as

$$f(u) = 1 - \sum_{n=1}^{\infty} \phi_n u^n \quad (52)$$

The coefficients  $\{\phi_n\}$  will be determined explicitly in the sequel. Now, clearly,  $\tanh^2(2asx) \equiv f(e^{-4as|x|})$ , and so we have

$$\begin{aligned} \text{mmse}_s(\Delta|X) &= 4a^2 [\sigma_x^2 - \mathbf{E}\{X^2 f(\exp\{-4as|X|\})\}] \\ &= 4a^2 \left[ \sigma_x^2 - \mathbf{E} \left\{ X^2 \left( 1 - \sum_{n=1}^{\infty} \phi_n e^{-4ans|X|} \right) \right\} \right] \\ &= 4a^2 \sum_{n=1}^{\infty} \phi_n \mathbf{E} \left\{ X^2 e^{-4ans|X|} \right\}. \end{aligned} \quad (53)$$

To continue from this point, we will have to let  $X$  assume a certain pdf. For convenience, let us select  $X$  to have the Laplacian pdf with parameter  $\theta$ , i.e.,

$$p(x) = \frac{\theta}{2} e^{-\theta|x|}. \quad (54)$$

We then obtain

$$\begin{aligned} \text{mmse}_s(\Delta|X) &= 2a^2 \theta \sum_{n=1}^{\infty} \phi_n \int_{-\infty}^{+\infty} x^2 e^{-(\theta+4ans)|x|} dx \\ &= 8a^2 \theta \sum_{n=1}^{\infty} \frac{\phi_n}{(\theta + 4ans)^3}. \end{aligned} \quad (55)$$

Thus,

$$\begin{aligned} R(D_s) &= R(D_{\infty}) - \int_s^{\infty} d\hat{s} \cdot \hat{s} \cdot \text{mmse}_s(\Delta|X) \\ &= 1 - 8a^2 \theta \sum_{n=1}^{\infty} \phi_n \cdot \int_s^{\infty} \frac{d\hat{s} \cdot \hat{s}}{(\theta + 4an\hat{s})^3} \\ &= 1 - \frac{\theta}{2} \sum_{n=1}^{\infty} \frac{\phi_n}{n^2} \left[ \frac{1}{\theta + 4ans} - \frac{\theta}{2(\theta + 4ans)^2} \right]. \end{aligned} \quad (56)$$

Thus far, our derivation has been exact. We now make an approximation that applies for large  $s$  by neglecting the terms proportional to  $(\theta + 4ans)^{-2}$  and by neglecting  $\theta$  compared to  $4ans$  in the denominators of  $1/(\theta + 4ans)$ . This results in the approximation

$$R(D_s) \approx \tilde{R}(D_s) \triangleq 1 - \frac{\theta}{8as} \sum_{n=1}^{\infty} \frac{\phi_n}{n^3}. \quad (57)$$

Let us denote  $C \triangleq \frac{\theta}{8a} \sum_{n=1}^{\infty} \frac{\phi_n}{n^3}$ . Then,  $\tilde{R}(D_s) = 1 - C/s$ . Applying a similar calculation to  $D_s = D_{\infty} + \int_s^{\infty} d\hat{s} \cdot \text{mmse}_{hs}(\Delta|X)$ , yields, in a similar manner, the approximation

$$D_s \approx \tilde{D}_s \triangleq D_{\infty} + \frac{C}{2s^2}. \quad (58)$$

It is easy now to express  $s$  as a function of  $D$  and substitute into the rate equation to obtain

$$R(D) \approx 1 - \sqrt{2C(D - D_{\infty})}. \quad (59)$$

Finally, it remains to determine the coefficients  $\{\phi_n\}$  and then the constant  $C$ . The coefficients can easily be obtained by using the identity  $(1+u)^{-1} = \sum_{n=0}^{\infty} (-1)^n u^n$  ( $u \in [0, 1]$ ), which yields, after simple algebra,  $\phi_n = 4n(-1)^{n+1}$ . Thus,

$$C = \frac{\theta}{2a} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2 \theta}{24a}. \quad (60)$$

and we have obtained a precise characterization of  $R(D)$  in the high-resolution regime:

$$\lim_{D \downarrow D_{\infty}} \frac{1 - R(D)}{\sqrt{D - D_{\infty}}} = \sqrt{2C} = \frac{\pi}{2} \cdot \sqrt{\frac{\theta}{3a}}. \quad (61)$$

By applying a somewhat more refined analysis, one obtains (similarly as in the above derivation in the high distortion regime) upper and lower bounds to  $R(D_s)$  and  $D_s$ , this time, as polynomials in  $1/s$ . These again lend themselves to the derivation of upper and lower bounds on  $R(D)$ , which are applicable in certain intervals of low distortion. Specifically, the resulting upper bound is

$$R(D) \leq 1 - \sqrt{2C(D - D_{\infty})} + C_1(D - D_{\infty}), \quad (62)$$



where  $C_1 = \frac{9\theta}{\pi^2 a} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^3}$ , and it is valid in the range  $D \in [D_\infty, D_\infty + C/(2C_1^2)]$ . The obtained lower bound is

$$R(D) \geq 1 - \frac{\sqrt{6C(D - D_\infty)}}{2 \cos \left[ \frac{1}{3} \sin^{-1} \left( 2C_1 \sqrt{\frac{6(D - D_\infty)}{C}} \right) + \frac{\pi}{6} \right]}, \quad (63)$$

and it applies to the range  $D \in [D_\infty, D_\infty + C/(12C_1^2)]$ . Both bounds have the same leading term in asymptotic behavior, which supports eq. (61). The details of this derivation are omitted since they are very similar to those of the high-distortion analysis.

#### D. High Resolution for a General $L^r$ Distortion Measure

Consider the case where the distortion measure is given by the  $L^r$  metric,  $d(x, y) = |x - y|^r$  for some fixed  $r > 0$ . Let the reproduction symbols be selected independently at random according to the uniform pdf

$$q(y) = \begin{cases} \frac{1}{2A} & |y| \leq A \\ 0 & \text{elsewhere} \end{cases} \quad (64)$$

Then

$$w_s(y|x) = \frac{e^{-s|y-x|^r}}{\int_{-A}^{+A} dy' \cdot e^{-s|y'-x|^r}} \quad (65)$$

and so

$$\begin{aligned} D_s &= \int_{-\infty}^{+\infty} dx p(x) \cdot \frac{\int_{-A}^{+A} dy \cdot |x - y|^r e^{-s|y-x|^r}}{\int_{-A}^{+A} dy \cdot e^{-s|y-x|^r}} \\ &= - \int_{-\infty}^{+\infty} dx p(x) \cdot \frac{\partial}{\partial s} \ln \left[ \int_{-A}^{+A} dy \cdot e^{-s|y-x|^r} \right] \end{aligned} \quad (66)$$

Now, in the high-resolution limit, where  $s$  is very large, the integrand  $e^{-s|y-x|^r}$  decays very rapidly as  $y$  takes values away from  $x$ , and so, for every  $x \in (-A, +A)$  (which for large enough  $A$ , is the dominant interval for the outer integral over  $p(x)dx$ ), the boundaries,  $-A$  and  $+A$ , of the inner integral can be extended to  $-\infty$  and  $+\infty$  within a negligible error term (whose derivative w.r.t.  $s$  is negligible too). Having done this, the inner integral no longer depends on  $x$ , which also means that the outer integration over  $x$  becomes superfluous. This results in

$$\begin{aligned} D_s &= - \frac{\partial}{\partial s} \ln \left[ \int_{-\infty}^{+\infty} dy \cdot e^{-s|y|^r} \right] \\ &= - \frac{\partial}{\partial s} \ln \left[ s^{-1/r} \int_{-\infty}^{+\infty} d(s^{1/r} y) e^{-|s^{1/r} y|^r} \right] \\ &= - \frac{\partial}{\partial s} \ln \left[ s^{-1/r} \int_{-\infty}^{+\infty} dt \cdot e^{-|t|^r} \right] \\ &= - \frac{\partial}{\partial s} \ln(s^{-1/r}) \\ &= \frac{1}{rs}. \end{aligned} \quad (67)$$

Thus,

$$\text{mmse}_s(\Delta|X) = - \frac{dD_s}{ds} = \frac{1}{rs^2}, \quad (68)$$

which yields

$$\frac{dR_q(D_s)}{dD_s} = s \cdot \text{mmse}_s(\Delta|X) = \frac{1}{rs} \quad (69)$$

and so

$$\begin{aligned} R_q(D_s) &= K + \frac{1}{r} \ln s \\ &= K + \frac{1}{r} \ln \left( \frac{1}{rD_s} \right) \end{aligned} \quad (70)$$

where  $K$  is an integration constant. We have therefore obtained that in the high-resolution limit, the rate-distortion function w.r.t.  $q$  behaves according to

$$R_q(D) = K' - \frac{1}{r} \ln D. \quad (71)$$

with  $K' = K - (\ln r)/r$ . While this simple derivation does not determine yet the constant  $K'$ , it does provide the correct characteristics of the dependence of  $R_q(D)$  upon  $D$  for small  $D$ . For the case of quadratic distortion, where  $r = 2$ , one easily identifies the familiar factor of  $1/2$  in front of the log-distortion term.

The exact constant  $K$  (or  $K'$ ) can be determined by returning to the original expression of  $R_q(D)$  as the Legendre transform of the log-moment generating function of the distortion (eq. (13), and setting there  $s = 1/(rD)$  as the minimizing  $s$  for the given  $D$ . The resulting expression turns out to be

$$K' = \ln \left[ \frac{rA}{\Gamma(1/r)} \right] - \frac{1}{r} \ln(er). \quad (72)$$

#### V. CONCLUSION

In this paper, we derived relations between the rate-distortion function  $R_q(D)$  and the MMSE in estimating the distortion given the source symbol. These relations have been discussed from several aspects, and it was demonstrated how they can be used to obtain upper and lower bounds on  $R_q(D)$ , as well as the exact asymptotic behavior in very high and very low distortion.

The bounds derived in our examples were induced from purely mathematical bounds on the expression of the MMSE directly. We have not explored, however, examples of bounds on  $R_q(D)$  that stem from estimation-theoretic bounds on the MMSE, as was described in Section III. In future work, it would be interesting to explore the usefulness of such bounds as well. Another interesting direction for further work would be to make an attempt to extend our results to rate-distortion functions pertaining to more involved settings, such as successive refinement coding, and situations that include side information.

#### APPENDIX

##### Proof of Theorem 1.

Consider a random selection of a codebook of  $M = e^{nR}$  codewords, where the various codewords are drawn independently, and each codeword,  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , is drawn according to the product measure  $Q(\mathbf{y}) = \prod_{i=1}^n q(y_i)$ . Let  $\mathbf{x} = (x_1, \dots, x_n)$  be a typical source vector, i.e., the number of times each symbol  $x \in \mathcal{X}$  appears in  $\mathbf{x}$  is (very close to)  $np(x)$ . We now ask what is the probability of the event  $\{\sum_{i=1}^n d(x_i, Y_i) \leq nD\}$ ? As this is a large deviations event

whenever  $D < \sum_{x,y} p(x)q(y)d(x,y)$ , this probability must decay exponentially with some rate function  $I_q(D) > 0$ , i.e.,

$$I_q(D) = \lim_{n \rightarrow \infty} \left[ -\frac{1}{n} \ln \Pr \left\{ \sum_{i=1}^n d(x_i, Y_i) \leq nD \right\} \right]. \quad (73)$$

The function  $I_q(D)$  can be determined in two ways. The first is by the method of types [3], which easily yields

$$I_q(D) = \min[I(X; Y') + D(q' \| q)], \quad (74)$$

where the  $Y'$  is an auxiliary random variable governed by  $q'(y) = \sum_{x \in \mathcal{X}} p(x)w(y|x)$  and the minimum is over all conditional pmf's  $\{w(y|x)\}$  that satisfy the inequality  $\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} w(y|x)d(x,y) \leq D$ . The second method is based on large deviations theory [4] (see also [8]), which yields

$$I_q(D) = -\min_{s \geq 0} \left[ sD + \sum_{x \in \mathcal{X}} p(x) \ln Z_x(s) \right]. \quad (75)$$

We first argue that  $I_q(D) = R_q(D)$ . The inequality  $I_q(D) \leq R_q(D)$  is obvious, as  $R_q(D)$  is obtained by confining the minimization over the channels in (74) so as to comply with the additional constraint that  $\sum_{x \in \mathcal{X}} p(x)w(y|x) = q(y)$  for all  $y \in \mathcal{Y}$ . The reversed inequality,  $I_q(D) \geq R_q(D)$ , is obtained by the following coding argument: On the one hand, a trivial extension of the converse to the rate-distortion coding theorem [2, p. 317], shows that  $R_q(D)$  is a lower bound to the rate-distortion performance of any code that satisfies  $\frac{1}{n} \sum_{i=1}^n \Pr\{Y_i = y\} = q(y)$  for all  $y \in \mathcal{Y}$ .<sup>4</sup> On the other hand, we next show that  $I_q(D)$  is an achievable rate for codes in this class.

Consider the the random coding mechanism described in the first paragraph of this proof, with  $R = I_q(D) + \epsilon$ , with  $\epsilon > 0$  being arbitrarily small. Since the probability that for a single randomly drawn codeword,  $\Pr\{\sum_{i=1}^n d(x_i, Y_i) \leq nD\}$  is of the exponential order of  $e^{-nI_q(D)}$ , then the random selection of a codebook of size  $e^{n[I_q(D)+\epsilon]}$  constitutes  $e^{n[I_q(D)+\epsilon]}$  independent trials of an experiment whose probability of success is of the exponential order of  $e^{-nI_q(D)}$ . Using standard random coding arguments, the probability that at least one codeword, in that codebook, would fall within distance  $nD$  from the given typical  $x$  becomes overwhelmingly large as  $n \rightarrow \infty$ . Since this randomly selected codebook satisfies also  $\frac{1}{n} \sum_{i=1}^n \Pr\{Y_i = y\} \rightarrow q(y)$  in probability (as  $n \rightarrow \infty$ ) for all  $y \in \mathcal{Y}$  (by the weak law of large numbers), then  $I_q(D)$  is an achievable rate within the class of codes that satisfy  $\frac{1}{n} \sum_{i=1}^n \Pr\{Y_i = y\} \rightarrow q(y)$  for all  $y$ .

Thus,  $I_q(D) \geq R_q(D)$ , which together with the reversed inequality proved above, yields the equality  $I_q(D) = R_q(D)$ .

<sup>4</sup>To see why this is true, consider the functions  $\delta_k(y)$ ,  $y, k \in \mathcal{Y}$  (each of which is defined as equal one for  $y = k$  and zero otherwise) as  $|\mathcal{Y}|$  distortion measures, indexed by  $k \in \mathcal{Y}$ , and consider the rate-distortion function w.r.t. the usual distortion constraint and the  $|\mathcal{Y}|$  additional "distortion constraints"  $\mathbf{E}\{\delta_k(Y)\} \leq q(k)$  for all  $k \in \mathcal{Y}$ , which, when satisfied, they all must be achieved with equality (since they must sum to unity). The rate-distortion function w.r.t. these  $|\mathcal{Y}| + 1$  constraints, which is exactly  $R_q(D)$ , is easily shown (using the standard method) to be jointly convex in  $D$  and  $q$ .

Consequently, according to eq. (75), we have established the relation<sup>5</sup>

$$R_q(D) = -\min_{s \geq 0} \left[ sD + \sum_{x \in \mathcal{X}} p(x) \ln Z_x(s) \right]. \quad (76)$$

As this minimization problem is a convex problem ( $\ln Z_x(s)$  is convex in  $s$ ), the minimizing  $s$  for a given  $D$  is obtained by taking the derivative of the r.h.s., which leads to

$$\begin{aligned} D &= -\sum_{x \in \mathcal{X}} p(x) \cdot \frac{\partial \ln Z_x(s)}{\partial s} \\ &= \sum_{x \in \mathcal{X}} p(x) \cdot \frac{\sum_{y \in \mathcal{Y}} q(y) d(x,y) e^{-sd(x,y)}}{\sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)}}. \end{aligned} \quad (77)$$

This equation yields the distortion level  $D$  for a given value of the minimizing  $s$  in eq. (76). Let us then denote

$$D_s = \sum_{x \in \mathcal{X}} p(x) \cdot \frac{\sum_{y \in \mathcal{Y}} q(y) d(x,y) e^{-sd(x,y)}}{\sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)}}. \quad (78)$$

This notation obviously means that

$$R_q(D_s) = -sD_s - \sum_{x \in \mathcal{X}} p(x) \ln Z_x(s). \quad (79)$$

Taking the derivative of (78), we readily obtain

$$\begin{aligned} \frac{dD_s}{ds} &= \sum_{x \in \mathcal{X}} p(x) \frac{\partial}{\partial s} \left[ \frac{\sum_{y \in \mathcal{Y}} q(y) d(x,y) e^{-sd(x,y)}}{\sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)}} \right] \\ &= -\sum_{x \in \mathcal{X}} p(x) \left[ \frac{\sum_{y \in \mathcal{Y}} q(y) d^2(x,y) e^{-sd(x,y)}}{\sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)}} - \left( \frac{\sum_{y \in \mathcal{Y}} q(y) d(x,y) e^{-sd(x,y)}}{\sum_{y \in \mathcal{Y}} q(y) e^{-sd(x,y)}} \right)^2 \right] \\ &= -\sum_{x \in \mathcal{X}} p(x) \cdot \text{Var}_s\{d(x,Y)|X=x\} \\ &= -\text{mmse}_s(\Delta|X), \end{aligned} \quad (80)$$

where  $\text{Var}_s\{d(x,Y)|X=x\}$  is the variance of  $d(x,Y)$  w.r.t. the conditional pmf  $\{w_s(y|x)\}$ . The last line follows from the fact the expectation of  $\text{Var}_s\{d(X,Y)|X\}$  w.r.t.  $\{p(x)\}$  is exactly the MMSE of  $d(X,Y)$  based on  $X$ . The integral forms of this equation are then precisely as in part (a) of the theorem with the corresponding integration constants. Finally, differentiating both sides of eq. (79), we get

$$\begin{aligned} \frac{dR(D_s)}{ds} &= -s \cdot \frac{dD_s}{ds} - D_s - \sum_{x \in \mathcal{X}} p(x) \cdot \frac{\partial \ln Z_x(s)}{\partial s} \\ &= -s \cdot \frac{dD_s}{ds} - D_s + D_s \\ &= -s \cdot \frac{dD_s}{ds} \\ &= s \cdot \text{mmse}_s(\Delta|X), \end{aligned} \quad (81)$$

which when integrated back, yields part (b) of the theorem. This completes the proof of Theorem 1.

<sup>5</sup> Eq. (76) appears also in [5, p. 90, Corollary 4.2.3], with a completely different proof, for the special case where  $q$  minimizes both sides of the equation (and hence it refers to  $R(D)$ ). However, the extension of that proof to a generic  $q$  is not apparent to be straightforward because here the minimization over the channels is limited by the reproduction distribution constraint.

## REFERENCES

- [1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, New Jersey, U.S.A., 1971.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (second edition), John Wiley & Sons, Inc., New York, 2005.
- [3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, John and Bartlett Publishers, 1993.
- [5] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, 1990.
- [6] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.
- [7] M. Fogiel, *Handbook of Mathematical, Scientific, and Engineering Formulas, Tables, Functions, Graphs, Transforms*, Research and Education Association, Piscataway, New Jersey, U.S.A., 1997.
- [8] N. Merhav, "An identity of Chernoff bounds with an interpretation in statistical physics and applications in information theory," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3710–3721, August 2008.
- [9] N. Merhav, "Another look at the physics of large deviations with application to rate-distortion theory," [http://arxiv.org/PS\\_cache/arxiv/pdf/0908/0908.3562v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/0908/0908.3562v1.pdf).
- [10] N. Merhav, "On the physics of rate-distortion theory," to appear in *Proc. ISIT 2010*, Austin, Texas, U.S.A., June 2010.
- [11] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.
- [12] E. Weinstein and A. J. Weiss, "Lower bounds on the mean square estimation error," *Proc. of the IEEE*, vol. 73, no. 9, pp. 1433–1434, September 1985.
- [13] A. J. Weiss, *Fundamental Bounds in Parameter Estimation*, Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, June 1985.
- [14] A. D. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Trans. Inform. Theory*, vol. 17, no. 5, pp. 508–513, September 1971.