IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Exponential Error Bounds on Parameter Modulation-Estimation for Discrete Memoryless Channels

Neri Merhav

CCIT Report #822
December 2012

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

CCIT Report #822 December 2012

Exponential Error Bounds on Parameter Modulation–Estimation for Discrete Memoryless Channels

Neri Merhav

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

Abstract

We consider the problem of modulation and estimation of a random parameter U to be conveyed across a discrete memoryless channel. Upper and lower bounds are derived for the best achievable exponential decay rate of a general moment of the estimation error, $\boldsymbol{E}|\hat{U}-U|^{\rho}$, $\rho \geq 0$, when both the modulator and the estimator are subjected to optimization. These exponential error bounds turn out to be intimately related to error exponents of channel coding and to channel capacity. While in general, there is some gap between the upper and the lower bound, they asymptotically coincide both for very small and for very large values of the moment power ρ. This means that our achievability scheme, which is based on simple quantization of U followed by channel coding, is nearly optimum in both limits. Some additional properties of the bounds are discussed and demonstrated, and finally, an extension to the case of a multidimensional parameter vector is outlined, with the principal conclusion that our upper and lower bound asymptotically coincide also for a high dimensionality.

Index Terms: Parameter estimation, modulation, discrete memoryless channels, error exponents, random coding, data processing theorem.

1 Introduction

Consider the problem of conveying the value of a parameter u across a given discrete memoryless channel

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} p(y_t|x_t), \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ are the channel input and output vectors, respectively. Our main interest, in this work, is in the following questions: How well can one estimate u based on $\boldsymbol{y}$ when one is allowed to optimize, not only the estimator, but also the modulator, that is, the function $\boldsymbol{x}(u) = (x_1(u), \ldots, x_n(u))$ that maps u into a channel input vector? How fast does the estimation error decay as a function of n when the best modulator and estimator are used?

In principle, this problem, which is the discrete–time analogue of the classical problem of "waveform communication" (in the terminology of [15, Chap. 8]), can be viewed both from the information–theoretic and the estimation–theoretic perspectives. Classical results in neither of these disciplines, however, seem to suggest satisfactory answers.

From the information–theoretic point of view, if the parameter is random, call it U, this is actually a problem of joint source–channel coding, where the source emits a single variable U (or a fixed number of them when U is a vector), whereas the channel is allowed to be used many times (n is large). The separation theorem of classical information theory asserts that asymptotic optimality of separate source– and channel coding is guaranteed in the limit of long blocks. However, it refers to a regime of long blocks both in source coding and channel coding, whereas here the source block length is 1, and so, there is no hope to compress the source with performance that comes close to the rate–distortion function.

In the realm of estimation theory, on the other hand, there is a rich literature on Bayesian and non–Bayesian bounds, mostly concerning the mean square error (MSE) in estimating parameters from signals corrupted by an additive white Gaussian noise (AWGN) channel, as well as other channels (see, e.g., [12] and the introductions of [1], [2], and [14] for overviews on these bounds). Most of these bounds lend themselves to calculation for a *given* modulator $\boldsymbol{x}(u)$ and therefore they may give insights concerning optimum estimation for this specific modulator. They may not, however, be easy to use for the derivation of *universal* lower bounds, namely, lower bounds that

depend neither on the modulator nor on the estimator, which are relevant when both optimum modulators and optimum estimators are sought. Two exceptions to this rule (although usually, not presented as such) are families of bounds that stem from generalized data processing theorems (DPT's) [5], [6], [11], [16], [18], henceforth referred to as "DPT bounds", and bounds based on hypothesis testing and channel coding considerations [1], [3], [17], henceforth called "channel–coding bounds."

In this paper, we use both the channel–coding techniques and DPT techniques in order to derive lower bounds on general moments of the estimation error, $\boldsymbol{E}|\hat{U}-U|^\rho$, where U is a random parameter, $\hat{U}$ is its estimate, and the power ρ is an arbitrary positive real (not necessarily an integer). It turns out that when $\boldsymbol{x}(u)$ is subjected to optimization, $\boldsymbol{E}|\hat{U}-U|^\rho$ can decay exponentially rapidly as a function of n, and so, our focus is on the best achievable exponential rate of decay as a function of ρ, which we shall denote by $\mathcal{E}(\rho)$, that is,

$$\inf \boldsymbol{E}|\hat{U}-U|^\rho \approx e^{-n\mathcal{E}(\rho)}, \tag{2}$$

where the infimum is over all modulators and estimators.[1] Interestingly, both the upper and lower bounds on $\mathcal{E}(\rho)$ are intimately related to well–known exponential error bounds associated with channel coding, such as Gallager's random coding exponent (for small values of ρ) and the expurgated exponent function (for large values of ρ). In other words, we establish an estimation–theoretic meaning to these error exponent functions. In particular, under certain conditions, our channel–coding upper bound on $\mathcal{E}(\rho)$ (corresponding to a lower bound on $\boldsymbol{E}|\hat{U}-U|^\rho$) can be presented as

$$\overline{E}(\rho) = \begin{cases} E_0(\rho) & \rho < \rho_0 \\ E_{ex}(0) & \rho \geq \rho_0 \end{cases} \tag{3}$$

where $E_0(\rho) = \max_q E_0(\rho, q)$, $E_0(\rho, q)$ being Gallager's function, $E_{ex}(0)$ is the expurgated exponent at zero rate, and ρ_0 is value of ρ for which $E_0(\rho) = E_{ex}(0)$ (so that $\overline{E}(\rho)$ is continuous). In addition, we derive a DPT bound and discuss its advantages and disadvantages compared to the above bound.

We also suggest a lower bound, $\underline{E}(\rho)$, on $\mathcal{E}(\rho)$ (associated with upper bounds on $\inf \boldsymbol{E}|\hat{U}-U|^\rho$), which is achieved by a simple, separation–based modulation and estimation scheme. While there is a certain gap between $\overline{E}(\rho)$ and $\underline{E}(\rho)$ for every finite ρ, it turns out that this gap disappears (in the sense that the ratio $\underline{E}(\rho)/\overline{E}(\rho)$ tends to unity) both for large ρ and for small ρ, and so, we have

[1] This is still an informal and non–rigorous description. More precise definitions will be given in the sequel.

exact asymptotics of $\mathcal{E}(\rho)$ in these two extremes: For large ρ, $\mathcal{E}(\rho)$ tends to $E_{ex}(0)$ and for small ρ, $\mathcal{E}(\rho) \sim \rho C$, where C is the channel capacity. Our simple achievability scheme is then nearly optimum at both extremes, which means that a separation theorem essentially holds for very small and for very large values of ρ, in spite of the earlier discussion (see also [7, Section III.D]). The results are demonstrated for the example of a "very noisy channel," [4, Example 3, pp. 147–149], [13, pp. 155–158], which is convenient to analyze, as it admits closed–form expressions.

Finally, we suggest an extension of our results to the case of a multidimensional parameter vector $\boldsymbol{U} = (U_1, \ldots, U_d)$. It turns out that the effect of the dimension d is in reducing the effective value of ρ by a factor of d. In other words, $\overline{E}(\rho)$ is replaced by $\overline{E}(\rho/d)$ and the extension of the achievability result is straightforward. This means that for fixed ρ, the limit of large d (where the effective value ρ/d is very small) also admits exact asymptotics, where $\mathcal{E}(\rho) \sim \rho C/d$.

The outline of the paper is as follows. In Section 2, we define the problem formally and we establish notation conventions. In Section 3, we derive our main upper and lower bounds based on channel coding considerations. In Section 4, we derive our DPT bound and discuss it. Section 5 is devoted to the example of the very noisy channel, and finally, in Section 6 the multidimensional case is considered.

2 Notation Conventions and Problem Formulation

Throughout this paper, scalar random variables (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. A similar convention will apply to random vectors and their sample values which will be denoted with same symbols in a bold face font. For example, $y \in \mathcal{Y}$ is a realization of a random variable Y, whereas $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ (n being a positive integer and $\mathcal{Y}^n$ being the n–th Cartesian power of $\mathcal{Y}$) is a realization of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.

Let U be a uniformly distributed[2] random variable over the interval $[-1/2, +1/2]$, which we will also denote by $\mathcal{U}$. We refer to U as the parameter to be conveyed from the source to the destination, via a given noisy channel. A given realization of U will be denoted by u.

[2]This specific assumption concerning the density of U and its support is made for convenience only. Our results extend to more general densities.

A discrete memoryless channel (DMC) is characterized by a matrix of conditional probabilities $p = \{p(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, where the channel input and output alphabets, $\mathcal{X}$ and $\mathcal{Y}$, are assumed finite.[3] When a DMC $p = \{p(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ is fed by an input vector $\boldsymbol{x} \in \mathcal{X}^n$, it produces an output vector $\boldsymbol{y} \in \mathcal{Y}^n$ according to

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} p(y_t|x_t). \tag{4}$$

A modulator is a measurable mapping $\boldsymbol{x} = f_n(u)$ from $\mathcal{U} = [-1/2, +1/2]$ to $\mathcal{X}^n$ and an estimator is a mapping $\hat{u} = g_n(\boldsymbol{y})$ from $\mathcal{Y}^n$ back to $\mathcal{U}$. The random vector $f_n(U)$ will also be denoted by $\boldsymbol{X}$. Similarly, the random variable $g_n(\boldsymbol{Y})$ will also be denoted by $\hat{U}$. Our basic figure of merit for communication systems is the expectation of ρ–th power of the estimation error, i.e., $\boldsymbol{E}\{|\hat{U} - U|^\rho\}$, where ρ is a positive real (not necessarily an integer) and $\boldsymbol{E}\{\cdot\}$ is the expectation operator with respect to (w.r.t.) the randomness of U and $\boldsymbol{Y}$. The capability of attaining an exponential decay in $\boldsymbol{E}\{|\hat{U} - U|^\rho\}$ by certain choices of a modulator f_n and an estimator g_n, motivates the definition of the following exponential rates

$$\overline{\mathcal{E}}(\rho) = \limsup_{n\to\infty} \left[-\frac{1}{n} \ln \left(\inf_{f_n, g_n} \boldsymbol{E}\{|\hat{U} - U|^\rho\} \right) \right] \tag{5}$$

and

$$\underline{\mathcal{E}}(\rho) = \liminf_{n\to\infty} \left[-\frac{1}{n} \ln \left(\inf_{f_n, g_n} \boldsymbol{E}\{|\hat{U} - U|^\rho\} \right) \right]. \tag{6}$$

This paper is basically about the derivation of upper bounds on $\overline{\mathcal{E}}(\rho)$ and lower bounds on $\underline{\mathcal{E}}(\rho)$, with special interest in situations where these upper and lower bounds come close to each other.

3 Upper and Lower Bounds Based on Channel Coding

Let $q = \{q(x),\ x \in \mathcal{X}\}$ be a given probability vector of a random variable X taking on values in $\mathcal{X}$, and let $p = \{p(y|x),\ \mathcal{X},\ y \in \mathcal{Y}\}$ define the given DMC. Let $E_0(\rho, q)$ be the *Gallager function* [4, p. 138, eq. (5.6.14)], [13, p. 133, eq. (3.1.18)], defined as

$$E_0(\rho, q) = -\ln \left(\sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} q(x) p(y|x)^{1/(1+\rho)} \right]^{1+\rho} \right), \quad \rho \geq 0. \tag{7}$$

[3] The finite alphabet assumption is used mainly for reasons of simplicity. The extension to continuous alphabets is possible, though some caution should be exercised at several places.

Next, we define

$$E_0(\rho) = \max_q E_0(\rho, q), \tag{8}$$

where the maximum is over the entire simplex of probability vectors, and let $\overline{E}_0(\rho)$ be the upper concave envelope[4] (UCE) of $E_0(\rho)$. Next define

$$E_x(\varrho) = -\varrho \ln \left(\sum_{x,x' \in \mathcal{X}} q(x)q(x') \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]^{\varrho} \right) \tag{9}$$

where the parameter ϱ should be distinguished from the power ρ of the estimation error in discussion. The *expurgated exponent function* [4, p. 153, eq. (5.7.11)], [13, p. 146, eq. (3.3.13)] is defined as

$$E_{ex}(R) = \sup_{\varrho \geq 1} [E_x(\varrho) - \varrho R]. \tag{10}$$

It is well known (and a straightforward exercise to show) that

$$E_{ex}(0) = \sup_{\varrho \geq 1} E_x(\varrho) = \lim_{\varrho \to \infty} E_x(\varrho) = -\sum_{x,x' \in \mathcal{X}} q(x)q(x') \ln \left[\sum_{y \in \mathcal{Y}} \sqrt{p(y|x)p(y|x')} \right]. \tag{11}$$

Finally, define

$$\bar{E}(\rho) = \begin{cases} \overline{E}_0(\rho) & \rho \leq \rho_0 \\ E_{ex}(0) & \rho > \rho_0 \end{cases} \tag{12}$$

where ρ_0 is the (unique) solution to the equation $\overline{E}_0(\rho) = E_{ex}(0)$.

Our first theorem (see Appendix A for the proof) asserts that $\overline{E}(\rho)$ is an upper bound on the best achievable exponential decay rate of ρ–th moment of the estimation error.

Theorem 1 *Let U be uniformly distributed over $\mathcal{U} = [-1/2, +1/2]$ and let $p = \{p(y|x)\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ be a given DMC. Then, for every $\rho \geq 0$*

$$\overline{\mathcal{E}}(\rho) \leq \overline{E}(\rho). \tag{13}$$

We now proceed to present a lower bound $\underline{E}(\rho)$ to $\underline{\mathcal{E}}(\rho)$. Let R_- be the smallest R such that $E_{ex}(R)$ is attained with $\varrho = 1$ and let R_+ denote the largest R such that

$$E_r(R) = \max_{0 \leq \rho \leq 1} [E_0(\varrho, q) - \varrho R] \tag{14}$$

[4]While the Gallager function $E_0(\rho, q)$ is known to be concave in ρ for every fixed q [13, p. 134, eq. (3.2.5a)], we are not aware of an argument asserting that $E_0(\rho)$ is concave in general. On the other hand, there are many situations where $E_0(\rho)$ is, in fact, concave and then $\overline{E}_0(\rho) = E_0(\rho)$, for example, when the achiever q^* of $\max_q E_0(\rho, q)$ is independent of ρ, like the case of the binary input output–symmetric (BIOS) channel [13, p. 153].

is attained for $\varrho = 1$.[5] Next, define

$$\rho_+ = \frac{E_0(1) - R_+}{R_+} \tag{15}$$

$$\rho_- = \frac{E_0(1) - R_-}{R_-} \tag{16}$$

and finally,

$$\underline{E}(\rho) = \begin{cases} \sup_{0\le\varrho\le 1} \rho E_0(\varrho)/(\varrho+\rho) & \rho \le \rho_+ \\ \rho E_0(1)/(1+\rho) = \rho E_x(1)/(1+\rho) & \rho_+ < \rho \le \rho_- \\ \sup_{\varrho\ge 1} \rho E_x(\varrho)/(\varrho+\rho) & \rho > \rho_- \end{cases} \tag{17}$$

Our next theorem (see Appendix B for the proof) tells us that $\underline{E}(\rho)$ is a lower bound on the best attainable exponential decay rate of $\boldsymbol{E}\{|\hat{U} - U|^\rho\}$.

Theorem 2 *Let U be uniformly distributed over $\mathcal{U} = [-1/2, +1/2]$ and let $p = \{p(y|x)\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ be a given DMC. Then, for every $\rho \ge 0$*

$$\underline{\mathcal{E}}(\rho) \ge \underline{E}(\rho). \tag{18}$$

The derivations of both $\overline{E}(\rho)$ and $\underline{E}(\rho)$ rely on channel coding considerations. In particular, the derivation of $\overline{E}(\rho)$ builds strongly on the method of [7], which extends the derivation of the Ziv–Zakai bound [17] and the Chazan–Zakai–Ziv bound [3]. While the two latter bounds are based on considerations associated with binary hypotheses testing, here and in [7], the general idea is extended to exponentially many hypotheses pertaining to channel decoding.

We see that both bounds exhibit different types of behavior in different ranges of ρ (i.e., "phase transitions"), but in a different manner. For both $\overline{E}(\rho)$ and $\underline{E}(\rho)$ the behavior is related to the ordinary Gallager function in some range of small ρ, and to the expurgated exponent in a certain range of large ρ.

As can be seen in the proof of Theorem 2 (Appendix B), the communication system that achieves $\underline{E}(\rho)$ works as follows (see also [7], [8]): Define

$$R(\rho) = \frac{\underline{E}(\rho)}{\rho} = \begin{cases} \sup_{0\le\varrho\le 1} E_0(\varrho)/(\varrho+\rho) & \rho \le \rho_+ \\ E_0(1)/(1+\rho) = E_x(1)/(1+\rho) & \rho_+ < \rho \le \rho_- \\ \sup_{\varrho\ge 1} E_x(\varrho)/(\varrho+\rho) & \rho > \rho_- \end{cases} \tag{19}$$

[5] For example, in the case of the BSC with a crossover parameter p, $R_- = \ln 2 - h_2(Z/(1+Z))$, with $Z = \sqrt{4p(1-p)}$, and $R_+ = \ln 2 - h_2(\sqrt{p}/(\sqrt{p}+\sqrt{1-p}))$, where $h_2(x) = -x\ln x - (1-x)\ln(1-x)$ [13, pp. 151–152].

Construct a uniform grid of $M = e^{nR(\rho)}/2$ evenly spaced points along $\mathcal{U}$, denoted $\{u_1, u_2, \ldots, u_M\}$. If $\rho > \rho_-$ assign to each grid point u_i a codeword of a code of rate $R(\rho)$ that achieves the expurgated exponent $E_{ex}[R(\rho)]$ (see [4, Theorem 5.7.1] or [13, Theorem 3.3.1]). If $\rho \leq \rho_-$, do the same with a code that achieves $E_r[R(\rho)]$ (see [4, p. 139, Corollary 1] or [13, Theorem 3.2.1]). Given u, let $f_n(u)$ be the codeword $\boldsymbol{x}_i$ that is assigned to the grid point u_i, which is closest to u. Given $\boldsymbol{y}$, let $g_n(\boldsymbol{y})$ be the grid point u_j that corresponds to the codeword $\boldsymbol{x}_j$ that has been decoded based on $\boldsymbol{y}$ using the ML decoder for the given DMC.

Let us examine the behavior of these bounds as $\rho \to 0$ and as $\rho \to \infty$. For very large values of ρ), where the upper bound $\overline{E}(\rho)$ is obviously given by $E_{ex}(0)$, the lower bound is given by

$$\begin{aligned}
\lim_{\rho\to\infty} \underline{E}(\rho) &= \lim_{\rho\to\infty} \sup_{\varrho\geq 1} \frac{\rho E_x(\varrho)}{\varrho+\rho} && (20)\\
&\geq \lim_{\rho\to\infty} \frac{\rho E_x(\sqrt{\rho})}{\sqrt{\rho}+\rho} && (21)\\
&= \lim_{\rho\to\infty} E_x(\sqrt{\rho}) = E_{ex}(0), && (22)
\end{aligned}$$

which means that for large ρ all the exponents asymptotically coincide:

$$\lim_{\rho\to\infty} \underline{E}(\rho) = \lim_{\rho\to\infty} \underline{\mathcal{E}}(\rho) = \lim_{\rho\to\infty} \overline{\mathcal{E}}(\rho) = \lim_{\rho\to\infty} \overline{E}(\rho) = E_{ex}(0). \quad (23)$$

In the achievability scheme described above, $R(\rho)$ is a very low coding rate. On the other hand, for very small values of ρ, where $\overline{E}(\rho) = \overline{E}_0(\rho) = \rho C + o(\rho)$, C being the channel capacity, we have

$$\begin{aligned}
\lim_{\rho\to 0} \frac{\underline{E}(\rho)}{\rho} &= \lim_{\rho\to 0} \sup_{0\leq\varrho\leq 1} \frac{E_0(\varrho)}{\varrho+\rho} && (24)\\
&\geq \lim_{\rho\to 0} \frac{E_0(\sqrt{\rho})}{\sqrt{\rho}+\rho} && (25)\\
&= \lim_{\rho\to 0} \frac{E_0(\sqrt{\rho})}{\sqrt{\rho}} \cdot \frac{1}{1+\sqrt{\rho}} && (26)\\
&= \lim_{\rho\to 0} \frac{E_0(\sqrt{\rho})}{\sqrt{\rho}} = C, && (27)
\end{aligned}$$

which means that for small ρ all the exponents behave like ρC, i.e.,

$$\lim_{\rho\to 0} \frac{\underline{E}(\rho)}{\rho} = \lim_{\rho\to 0} \frac{\underline{\mathcal{E}}(\rho)}{\rho} = \lim_{\rho\to 0} \frac{\overline{\mathcal{E}}(\rho)}{\rho} = \lim_{\rho\to 0} \frac{\overline{E}(\rho)}{\rho} = C. \quad (28)$$

It is then interesting to observe that not only channel–coding error exponents, but also channel capacity plays a role in the characterization of the best achievable modulation–estimation performance. In the achievability scheme described above, $R(\rho)$ is a very high coding rate, very close to the capacity C.

4 Upper Bound Based on Data Processing Inequalities

We next derive an alternative upper bound on $\overline{\mathcal{E}}(\rho)$ that is based on generalized data processing inequalities, following Ziv and Zakai [18] and Zakai and Ziv [16]. The idea behind these works is that it is possible to define generalized mutual information functionals satisfying a DPT, by replacing the negative logarithm function of the ordinary mutual information, by a general convex function. This enables to obtain tighter distortion bounds for communication systems with short block length.

In [6] it was shown that the following generalized mutual information functional, between two generic random variables, A and B, admits a DPT for every positive integer k and for every vector $(\alpha_1, \ldots, \alpha_k)$ whose components are non–negative and sum to unity:

$$\tilde{I}(A;B) = -\boldsymbol{E}\left\{\sum_{b\in\mathcal{B}}\prod_{i=1}^{k} p(b|A_i)^{\alpha_i}\right\} = -\sum_{b\in\mathcal{B}}\prod_{i=1}^{k}\sum_{a_i\in\mathcal{A}} q(a_i)p(b|a_i)^{\alpha_i}. \tag{29}$$

In particular, since $U \to \boldsymbol{Y} \to \hat{U}$ is a Markov chain, then by the generalized DPT,

$$\tilde{I}(U;\hat{U}) \le \tilde{I}(U;\boldsymbol{Y}). \tag{30}$$

The idea is to further upper bound $I(U;\boldsymbol{Y})$ and to further lower bound $\tilde{I}(U;\hat{U})$ subject to the constraint $\boldsymbol{E}|\hat{U}-U|^{\rho} = D$, which leads to a generalized rate–distortion function, and thereby to obtain an inequality on $\boldsymbol{E}|\hat{U}-U|^{\rho}$. Specifically, $I(U;\boldsymbol{Y})$ is upper bounded as follows:

$$\begin{aligned}
\tilde{I}(U;\boldsymbol{Y}) &= -\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\prod_{i=1}^{k}\int_{-1/2}^{+1/2}\mathrm{d}u_i\, p(\boldsymbol{y}|f_n(u_i))^{\alpha_i} && (31)\\
&= -\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\prod_{i=1}^{k}\int_{-1/2}^{+1/2}\mathrm{d}u_i\prod_{t=1}^{n} p(y_t|[f_n(u_i)]_t)^{\alpha_i} && (32)\\
&= -\prod_{t=1}^{n}\sum_{y\in\mathcal{Y}}\prod_{i=1}^{k}\int_{-1/2}^{+1/2}\mathrm{d}u_i\, p(y_t|[f_n(u_i)]_t)^{\alpha_i} && (33)\\
&\le -\min_{q}\prod_{t=1}^{n}\sum_{y\in\mathcal{Y}}\prod_{i=1}^{k}\sum_{x_i\in\mathcal{X}} q(x_i)p(y_t|x_i)^{\alpha_i} && (34)\\
&= -\min_{q}\left[\sum_{y\in\mathcal{Y}}\prod_{i=1}^{k}\sum_{x_i\in\mathcal{X}} q(x_i)p(y|x_i)^{\alpha_i}\right]^n && (35)\\
&= -\exp\{-n\max_{q} E(\alpha_1,\ldots,\alpha_k,q)\}, && (36)
\end{aligned}$$

where $[f_n(u_i)]_t$ denotes the t–th component of the vector $\boldsymbol{x} = f_n(u_i)$ and where

$$E(\alpha_1, \ldots, \alpha_k, q) = -\ln \left[\sum_{y \in \mathcal{Y}} \prod_{i=1}^{k} \left(\sum_{x_i \in \mathcal{X}} q(x_i) p(y|x_i)^{\alpha_i} \right) \right]. \tag{37}$$

Note that for $k = 1 + \varrho$ (ϱ – integer),

$$\hat{E}\left(\frac{1}{1+\varrho}, \ldots, \frac{1}{1+\varrho}, q \right) = E_0(\varrho, q). \tag{38}$$

In Appendix C we show that

$$\min\{\tilde{I}(U, \hat{U}) : \ \boldsymbol{E}|\hat{U} - U|^\rho = D\} \stackrel{\Delta}{=} \tilde{R}(D) \geq -c \cdot D^{\sum_{i=1}^{k} \zeta_\rho(\alpha_i)} \tag{39}$$

where c is a constant that depends solely on ρ, k and $\alpha_1, \ldots, \alpha_k$, and where

$$\zeta_\rho(\alpha) = \left\{ \begin{array}{ll} \alpha & 0 \leq \alpha \leq \frac{1}{1+\rho} \\ \frac{1-\alpha}{\rho} & \frac{1}{1+\rho} \leq \alpha \leq 1 \end{array} \right. = \min\left\{ \alpha, \frac{1-\alpha}{\rho} \right\}. \tag{40}$$

The function $\tilde{R}(D)$ in eq. (39) is referred to as a "generalized rate–distortion function" in the terminology of [18] and [16]. Thus, from the generalized DPT,

$$\boldsymbol{E}|\hat{U} - U|^\rho \equiv D \geq c' \cdot e^{-n\overline{E}_{DPT}(\rho)} \tag{41}$$

where c' is another constant and

$$\overline{E}_{DPT}(\rho) \stackrel{\Delta}{=} \inf_{k>1} \inf_{\alpha_1, \ldots, \alpha_k} \sup_{q} \frac{E(\alpha_1, \ldots, \alpha_k, q)}{\sum_{i=1}^{k} \zeta_\rho(\alpha_i)}. \tag{42}$$

As an example, assume that the channel is such that the function $E_0(\varrho)$ is concave, so that $\overline{E}_0(\varrho) = E_0(\varrho)$. In this case, $\rho_0 \geq 1$ since $E_0(1) \leq E_{ex}(0)$ and $E_0(\varrho)$ is monotonically increasing. Now, let $\rho \leq \rho_0$ be an integer (for example, $\rho = 1$ is always a legitimate choice). Then,

$$\begin{aligned} \overline{E}(\rho) &= E_0(\rho) && (43) \\ &= \sup_q \frac{E(1/(1+\rho), \ldots, 1/(1+\rho), q)}{(1+\rho)\zeta_\rho(1/(1+\rho))} && (44) \\ &\geq \inf_{k>1} \inf_{\alpha_1, \ldots, \alpha_k} \sup_q \frac{\hat{E}(\alpha_1, \ldots, \alpha_k, q)}{\sum_{i=1}^{k} \zeta_\rho(\alpha_i)} && (45) \\ &= \overline{E}_{DPT}(\rho). && (46) \end{aligned}$$

Thus, at least in this case, the DPT bound is guaranteed to be no worse than the channel–coding bound $\overline{E}(\rho)$. Nonetheless, in our numerical studies, we have not found an example where the

DPT bound strictly improves on the channel–coding bound, i.e., $\overline{E}_{DPT}(\rho) < \overline{E}(\rho)$, and it remains an open question whether the DPT bound can offer improvement in any situation, thanks to its additional degrees of freedom. It should be pointed out that the vector $(\alpha_1, \ldots, \alpha_k)$ that achieves $E_{DPT}(\rho)$ is not always given by $(1/(k+1), \ldots, 1/(k+1))$ because the function $E(\alpha_1 \ldots, \alpha_k, q)$ is not convex in $(\alpha_1, \ldots, \alpha_k)$. At any rate, in all cases where the two bounds are equivalent, namely, $\overline{E}_{DPT}(\rho) = \overline{E}(\rho)$, this is interesting on its own right since the two bounds are obtained by two different techniques that are based on completely different considerations. One advantage of the DPT approach is that it seems to lend itself more comfortably to extensions that account for moments of more general functions of the estimation error, i.e., $\boldsymbol{E}\{g(|\hat{U} - U|)\}$, for a large class of monotonically increasing functions g. On the other hand, the optimization associated with calculation of the DPT bound is not trivial.

5 Example: Very Noisy Channel

As an example, we consider the so called *very noisy channel*, which is characterized by

$$p(y|x) = p(y)[1 + \epsilon(x, y)], \qquad |\epsilon(x, y)| \ll 1, \quad \forall x \in \mathcal{X},\ y \in \mathcal{Y}. \tag{47}$$

As is shown in [13, Sect. pp. 155–158], to the first order, we have the following relations

$$C = \frac{1}{2} \max_q \sum_{x,y} q(x)p(y)\epsilon^2(x, y) \tag{48}$$

$$E_0(\varrho) = \frac{\varrho}{1 + \varrho} \cdot C, \tag{49}$$

and therefore

$$E_r(R) = \max_{0 \le \varrho \le 1} \left(\frac{\varrho}{1 + \varrho} \cdot C - \varrho R \right) = \begin{cases} \frac{C}{2} - R & R < \frac{C}{4} \\ (\sqrt{C} - \sqrt{R})^2 & \frac{C}{4} \le R \le C \\ 0 & R > C \end{cases} \tag{50}$$

As for the expurgated exponent, we have

$$E_x(\varrho) = E_0(1) = \frac{C}{2} \tag{51}$$

and so,

$$E_{ex}(R) = \sup_{\varrho \ge 1} [E_x(\varrho) - \varrho R] = \frac{C}{2} - R \tag{52}$$

which means that expurgation does not help for very noisy channels. This implies that $\rho_0 = 1$ and so

$$\overline{E}(\rho) = \begin{cases} \frac{\rho}{1+\rho} \cdot C & \rho \leq 1 \\ \frac{C}{2} & \rho > 1 \end{cases} \tag{53}$$

As for the lower bound, we have the following: For $\rho < 1$,

$$\underline{E}(\rho) = \sup_{0 \leq \varrho \leq 1} \frac{\rho}{\rho + \varrho} \cdot \frac{\varrho}{1+\varrho} \cdot C = \frac{\rho}{(1+\sqrt{\rho})^2} \cdot C. \tag{54}$$

The same result is obtained, of course, from the solution to the equation $\rho R = (\sqrt{C} - \sqrt{R})^2$. For $\rho \geq 1$,

$$\underline{E}(\rho) = \sup_{\varrho \geq 1} \frac{\rho E_x(\varrho)}{\varrho + \rho} = \sup_{\varrho \geq 1} \frac{\rho}{\varrho + \rho} \cdot \frac{C}{2} = \frac{\rho}{1+\rho} \cdot \frac{C}{2}. \tag{55}$$

Thus, in summary

$$\underline{E}(\rho) = \begin{cases} \frac{\rho}{(1+\sqrt{\rho})^2} \cdot C & \rho < 1 \\ \frac{\rho}{1+\rho} \cdot \frac{C}{2} & \rho \geq 1 \end{cases} \tag{56}$$

We see how the bounds asymptotically coincide (in the sense that $\overline{E}(\rho)/\underline{E}(\rho) \approx 1$) both for very large values of ρ and for very small values of ρ (see Fig. 1).

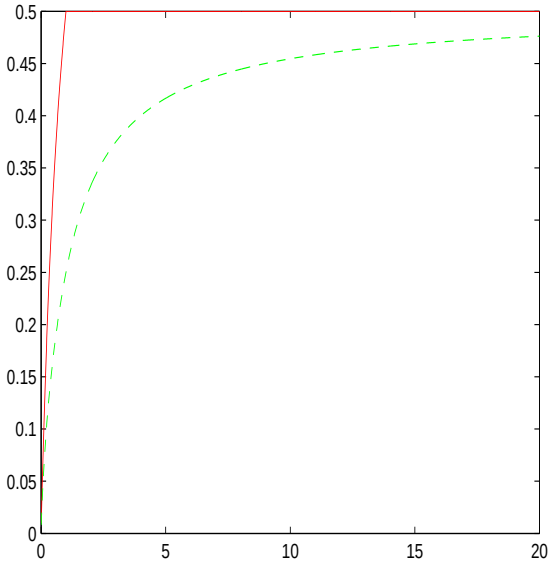

Figure 1: The upper bound $\overline{E}(\rho)/C$ (solid curve) and the lower bound $\underline{E}(\rho)/C$ (dashed curve) for the example of the very noisy channel.

As for the DPT bound, we have the following approximate analysis:

$$
\begin{aligned}
e^{-\sup_q E(\alpha_1,\ldots,\alpha_k,q)} &= \inf_q \sum_{y\in\mathcal{Y}} \prod_{i=1}^{k} \left[\sum_{x_i\in\mathcal{X}} q(x_i) p(y|x_i)^{\alpha_i} \right] && (57)\\
&= \inf_q \sum_{y\in\mathcal{Y}} \prod_{i=1}^{k} \left\{ p(y)^{\alpha_i} \left[\sum_{x_i\in\mathcal{X}} q(x_i)[1+\epsilon(x_i,y)]^{\alpha_i} \right] \right\} && (58)\\
&= \inf_q \sum_{y\in\mathcal{Y}} p(y) \prod_{i=1}^{k} \left[\sum_{x_i\in\mathcal{X}} q(x_i)[1+\epsilon(x_i,y)]^{\alpha_i} \right] && (59)\\
&\approx \inf_q \sum_{y\in\mathcal{Y}} p(y) \prod_{i=1}^{k} \left(\sum_{x_i\in\mathcal{X}} q(x_i) \left[1+\alpha_i\epsilon(x_i,y) - \frac{1}{2}\alpha_i(1-\alpha_i)\epsilon^2(x_i,y) \right] \right) && (60)\\
&= \inf_q \sum_{y\in\mathcal{Y}} p(y) \prod_{i=1}^{k} \left[1 - \frac{1}{2}\alpha_i(1-\alpha_i) \sum_{x_i\in\mathcal{X}} q(x_i)\epsilon^2(x_i,y) \right] && (61)\\
&\approx \inf_q \sum_{y\in\mathcal{Y}} p(y) \left[1 - \frac{1}{2}\sum_{i=1}^{k} \alpha_i(1-\alpha_i) \sum_{x_i\in\mathcal{X}} q(x_i)\epsilon^2(x_i,y) \right] && (62)\\
&= 1 - \frac{1}{2}\sum_{i=1}^{k} \alpha_i(1-\alpha_i) \sup_q \sum_{x_i\in\mathcal{X}} \sum_{y\in\mathcal{Y}} q(x_i)p(y)\epsilon^2(x_i,y) && (63)\\
&\approx 1 - C\sum_{i=1}^{k} \alpha_i(1-\alpha_i) && (64)\\
&= 1 - C\left(1 - \sum_{i=1}^{k} \alpha_i^2\right). && (65)
\end{aligned}
$$

where in the fifth line, we have used the identity $\sum_x q(x)\epsilon(x,y) = 0$ for all y with $p(y) > 0$ [13, p. 156, eq. (3.4.28)]. Thus,

$$
\sup_q E(\alpha_1,\ldots,\alpha_k,q) = -\ln\left[1 - C\left(1 - \sum_{i=1}^{k}\alpha_i^2\right)\right] \approx C\left(1 - \sum_{i=1}^{k}\alpha_i^2\right), \tag{66}
$$

and then

$$
\overline{E}_{DPT}(\rho) \approx C \cdot \inf_{k>1} \inf_{\alpha_1,\ldots,\alpha_k} \frac{1 - \sum_{i=1}^{k}\alpha_i^2}{\sum_{i=1}^{k}\zeta_\rho(\alpha_i)}. \tag{67}
$$

The very same expressions are obtained for the continuous–time AWGN channel with unlimited bandwidth, where $C = P/N_0$, P being the signal power and N_0 being the one–sided noise spectral

density. For $\rho = 1$ and $k = 2$, we have $\zeta_1(\alpha) = \min\{\alpha, 1-\alpha\}$:

$$\overline{E}_{DPT}(1) \leq C \cdot \inf_{0\leq\alpha\leq 1} \frac{1-\alpha^2-(1-\alpha)^2}{2\min\{\alpha,1-\alpha\}} \tag{68}$$

$$= C \cdot \inf_{0\leq\alpha\leq 1/2} \frac{2\alpha(1-\alpha)}{2\alpha} = \frac{C}{2}, \tag{69}$$

which agrees with $\overline{E}(\rho)$. For $\rho = 2$ and $k = 2$, the minimum is attained for $\alpha = 1/3$, and the result is $E_{DPT}(2) \leq 8C/9$. However for $k = 3$, the bound improves to $C/3$.

6 Extension to the Multidimensional Case

Consider now the case of a parameter vector $\boldsymbol{U} = (U_1, \ldots, U_d)$, uniformly distributed across the unit hypercube $[-1, 2, +1/2]^d$. A reasonable figure of merit in this case would be a linear combination of $\boldsymbol{E}\{|\hat{U}_i - U_i|^\rho\}$, $i = 1, 2, \ldots, d$. Since each one of these terms is exponential in n, it makes sense to let the coefficients of this linear combination also be exponential functions of n, as otherwise, the results will be exponentially insensitive to the choice of the coefficients. This means that we consider the criterion

$$\sum_{i=1}^{d} e^{nr_i} \cdot \boldsymbol{E}\{|\hat{U}_i - U_i|^\rho\}, \tag{70}$$

where, without loss of generality, we take $r_i \geq 0$, $\min_i r_i = 0$.

The derivation below is an extension of the derivation of the channel coding bound, given in Appendix A for the case $d = 1$. Therefore, a reader who is interested in the details is advised to read Appendix A first, or otherwise to skip directly to the final result in eq. (81) and the discussion that follows.

Let us define $R_i = (r_i + \gamma)/\rho$ for some constant $\gamma \geq 0$. Consider the following chain of

inequalities:

$$\begin{aligned}
\sum_{i=1}^{d} e^{nr_i} \cdot \boldsymbol{E}\{|\hat{U}_i - U_i|^{\rho}\} &\geq \sum_{i=1}^{d} e^{nr_i} \cdot e^{-n\rho R_i} \Pr\{|\hat{U}_i - U_i| \geq e^{-nR_i}\} && (71)\\
&= \sum_{i=1}^{d} e^{-n(\rho R_i - r_i)} \Pr\{|\hat{U}_i - U_i| \geq e^{-nR_i}\} && (72)\\
&= e^{-\gamma n} \sum_{i=1}^{d} \Pr\{|\hat{U}_i - U_i| \geq e^{-n(r_i+\gamma)/\rho}\} && (73)\\
&\geq e^{-\gamma n} \cdot \Pr \bigcup_{i=1}^{d} \left\{|\hat{U}_i - U_i| \geq e^{-n(r_i+\gamma)/\rho}\right\} && (74)\\
&\geq e^{-\gamma n} \cdot \exp\left\{-nE_{sl}\left(\frac{1}{\rho}\left[\sum_{i=1}^{d} r_i + \gamma d\right]\right)\right\}, && (75)
\end{aligned}$$

where the second line follows from Chebychev's inequality, the fifth line follows from the union bound, and the last line follows from the same arguments as in [7, Sect. IV.A]. Maximizing over γ, we get

$$\sum_{i=1}^{d} e^{nr_i} \cdot \boldsymbol{E}\{|\hat{U}_i - U_i|^{\rho}\} \geq \exp\left\{-n \min_{\gamma \geq 0}\left[\gamma + E_{sl}\left(\frac{1}{\rho}\left[\sum_{i=1}^{d} r_i + \gamma d\right]\right)\right]\right\}. \quad (76)$$

Defining $R = (\sum_{i=1}^{d} r_i + \gamma d)/\rho$, $R_{\min} = \sum_{i=1}^{d} r_i/\rho$ and $\bar{r} = R_{\min}/d$, the above minimization at the exponent becomes equivalent to

$$\begin{aligned}
&\min_{R \geq R_{\min}} \left[\frac{\rho R - \sum_i r_i}{d} + E_{sl}(R)\right] && (77)\\
= &\min_{R \geq R_{\min}} \left[\frac{\rho}{d} \cdot R + E_{sl}(R)\right] - \rho \bar{r} && (78)\\
= &\begin{cases} E_{sp}(R_{\rho/d}) + \frac{\rho}{d}(R_{\rho/d} - R_{\min}) & \rho/d \leq \rho_0 \\ E_{ex}(0) - \rho_0 R_{\min} & \rho/d > \rho_0 \end{cases} && (79)
\end{aligned}$$

where R_θ is defined as the achiever of $\min_{R \geq R_{\min}}[\theta R + E_{sp}(R)]$. Thus, the extension of the channel–coding bound to the d–dimensional case reads

$$\begin{aligned}
\overline{E}(\rho, d, r_1, \ldots, r_d) &= \begin{cases} E_{sp}(R_{\rho/d}) + \frac{\rho}{d} R_{\rho/d} - \frac{1}{d}\sum_{i=1}^{d} r_i & \rho \leq \rho_0 d \\ E_{ex}(0) - \frac{\rho_0}{\rho}\sum_{i=1}^{d} r_i & \rho > \rho_0 d \end{cases} && (80)\\
&= \begin{cases} E_0\left(\frac{\rho}{d}\right) - \frac{1}{d}\sum_{i=1}^{d} r_i & \rho/d \leq \rho_0 \\ E_{ex}(0) - \frac{\rho_0}{\rho}\sum_{i=1}^{d} r_i & \rho/d > \rho_0 \end{cases} && (81)
\end{aligned}$$

We see that when $r_i = 0$ for all i (i.e., all weights are 1), it is the same channel–coding bound as before, except that ρ is replaced by ρ/d, that is, $\overline{E}(\rho/d)$. For $\rho \to \infty$, the bound tends to $E_{ex}(0)$,

which can be approached again by a low–rate code for a Cartesian grid in the parameter space. At the other extreme, when d is very large compared to ρ, so ρ/d is small, construct a grid of $e^{n(C-\epsilon)/d} \times e^{n(C-\epsilon)/d} \times \ldots \times e^{n(C-\epsilon)/d}$, quantize $\boldsymbol{U}$ and assign to each grid point a codeword of a typical random code at rate $C-\epsilon$. Then the performance will be about $e^{-n\rho C/d}$. Therefore, as a corollary of the above result, we have

$$\sum_{i=1}^{d} \boldsymbol{E}\{|\hat{U}_i - U_i|^\rho\} \geq e^{-n[\overline{E}(\rho/d)+o(n)]}. \tag{82}$$

Appendix A

Proof of Theorem 1. We begin by using the Markov/Chebychev inequality:

$$\boldsymbol{E}|\hat{U}-U|^\rho \geq \Delta^\rho \Pr\{|\hat{U}-U| \geq \Delta\}. \tag{A.1}$$

Next we need to further lower bound $\Pr\{|\hat{U}-U| \geq \Delta\}$ and then maximize the r.h.s. over Δ. Equivalently, similarly as in [7], we may set $\Delta = e^{-nR}$ in the r.h.s. and maximize the bound w.r.t. R. Let $E(R)$ be the reliability function of the channel. Then, similarly[6] as in [7, Theorem 1], we have:

$$\Pr\{|\hat{U}-U| \geq e^{-nR}\} \geq e^{-n[E(R)+o(n)]} \tag{A.2}$$

and so,

$$\boldsymbol{E}|\hat{U}-U|^\rho \geq e^{-n\rho R} \cdot e^{-n[E(R)+o(n)]} = e^{-n[\rho R+E(R)+o(n)]}. \tag{A.3}$$

The best[7] lower bound is obtained by maximizing the r.h.s. over R, yielding

$$\begin{aligned} \boldsymbol{E}|\hat{U}-U|^\rho &\geq e^{-n\min_{R\geq 0}[\rho R+E(R)+o(n)]} \\ &\geq e^{-n\min_{R\geq 0}[\rho R+E_{sl}(R)+o(n)]} \end{aligned} \tag{A.4}$$

[6]While ref. [7] is primarily about the continuous time additive white Gaussian noise (AWGN) channel, the arguments in the proof of Theorem 1 therein are insensitive to this assumption. They hold verbatim here, provided that the observation time T in [7] is replaced by the block length n and the reliability function of the AWGN channel is replaced by that of the DMC considered here.

[7]The reader might suspect that the use of Chebychev's inequality yields a loose bound. Note, however, that even the exact relation $\boldsymbol{E}|\hat{U}-U|^\rho = \rho n \int_0^\infty \mathrm{d}R \cdot e^{-n\rho R} \cdot \Pr\{|\hat{U}-U| > e^{-nR}\}$, with $\Pr\{|\hat{U}-U| > e^{-nR}\} \geq e^{-n[E(R)+o(n)]}$, would yield, after saddle–point integration, exactly the same exponential order as presented above. The weak link here is, therefore, not the Chebychev inequality but the fact that there is no apparent single estimator, independent of R, that minimizes $\Pr\{|\hat{U}-U| > e^{-nR}\}$ uniformly for all R.

where $E_{sl}(R)$ is the exponent associated with the *straight line bound*, which is well known to be an upper bound on the reliability function $E(R)$ [9], [10], [13, Sect. 3.8], and which is given by

$$E_{sl}(R) = \begin{cases} E_{ex}(0) - \rho_0 R & 0 \le R \le R_0 \\ E_{sp}(R) & R_0 < R \le C \\ 0 & R > C \end{cases} \tag{A.5}$$

where

$$E_{sp}(R) = \sup_{\varrho \ge 0}[E_0(\varrho) - \varrho R] \tag{A.6}$$

is the *sphere–packing exponent*, ρ_0 is as defined in Theorem 1 and R_0 is the rate R at which $\mathrm{d}E_{sp}(R)/\mathrm{d}R = -\rho_0$, or equivalently, the solution to the equation $E_{sp}(R) = E_{ex}(0) - \rho_0 R$. Thus, according to the second line of eq. (A.4),

$$\overline{\mathcal{E}}(\rho) \le \min_{R \ge 0}[\rho R + E_{sl}(R)]. \tag{A.7}$$

For $\rho \ge \rho_0$, the minimum is obviously attained at $R = 0$, and so,

$$\overline{\mathcal{E}}(\rho) \le \rho \cdot 0 + E_{sl}(0) = E_{ex}(0). \tag{A.8}$$

For $\rho < \rho_0$, we use

$$\overline{\mathcal{E}}(\rho) \le \min_{R \ge 0}[\rho R + E_{sl}(R)] \le \min_{R \ge 0}[\rho R + E_{sp}(R)]. \tag{A.9}$$

The right–most side of eq. (A.9) is the Legendre–Fenchel transform (LFT) of $E_{sp}(R)$, which in turn (according to (A.6)), is the LFT of $E_0(\rho)$. Thus, the right–most side of (A.9) is given by the UCE of $E_0(\rho)$, which is $\overline{E}_0(\rho)$. Thus,

$$\overline{\mathcal{E}}(\rho) \le \begin{cases} \overline{E}_0(\rho) & \rho < \rho_0 \\ E_{ex}(0) & \rho \ge \rho_0 \end{cases} = \overline{E}(\rho). \tag{A.10}$$

This completes the proof of Theorem 1.

Appendix B

Proof of Theorem 2. Define

$$R(\rho) = \frac{\underline{E}(\rho)}{\rho} = \begin{cases} \sup_{0 \le \varrho \le 1} E_0(\varrho)/(\varrho + \rho) & \rho \le \rho_+ \\ E_0(1)/(1+\rho) = E_x(1)/(1+\rho) & \rho_+ < \rho \le \rho_- \\ \sup_{\varrho \ge 1} E_x(\varrho)/(\varrho + \rho) & \rho > \rho_- \end{cases} \tag{B.1}$$

Consider a grid of $M = e^{nR(\rho)}/2$ evenly spaced points along $\mathcal{U}$, denoted $\{u_1, u_2, \ldots, u_M\}$, where $u_1 = -1/2 + e^{-nR(\rho)}$ and $u_M = 1/2 - e^{-nR(\rho)}$ (see also [7, Theorem 2]). If $\rho > \rho_-$, assign to

each point u_i a codeword of a code of rate $R(\rho)$ that achieves the expurgated exponent $E_{ex}[R(\rho)]$. Otherwise, do the same with a code that achieves $E_r[R(\rho)]$ (see [4, p. 139, Corollary 1] or [13, Theorem 3.2.1]). Given u, let $f_n(u)$ be the codeword $\boldsymbol{x}_i$ that is assigned to the grid point u_i, which is closest to u. Given $\boldsymbol{y}$, let $g_n(\boldsymbol{y})$ be the grid point u_j that corresponds to the codeword $\boldsymbol{x}_j$ that has been decoded based on $\boldsymbol{y}$ using the ML decoder for the given DMC. For every $R \geq 0$, we have:

$$\begin{aligned}
\boldsymbol{E}\{|\hat{U}-U|^{\rho}\} &= \boldsymbol{E}\left\{|\hat{U}-U|^{\rho}\Big||\hat{U}-U| \leq e^{-nR}\right\} \cdot \Pr\{|\hat{U}-U| \leq e^{-nR}\} + \\
&\quad \boldsymbol{E}\left\{|\hat{U}-U|^{\rho}\Big||\hat{U}-U| > e^{-nR}\right\} \cdot \Pr\{|\hat{U}-U| > e^{-nR}\} \\
&\leq [e^{-nR}]^{\rho} \cdot 1 + 1^{\rho} \cdot \Pr\{|\hat{U}-U| > e^{-nR}\} \\
&= e^{-n\rho R} + \Pr\{|\hat{U}-U| > e^{-nR}\}. \qquad \text{(B.2)}
\end{aligned}$$

Now, it follows from the construction of the proposed scheme that if R is the coding rate and the spacing between each two consecutive grid points is $2e^{-nR}$, then the event $\{|\hat{U}-U| > e^{-nR}\}$ occurs iff the ML decoder errs. Thus, $\Pr\{|\hat{U}-U| > e^{-nR}\}$ is exactly the probability of decoding error. Considering the case $\rho > \rho_-$, this code is assumed to achieve the expurgated exponent, and so, this probability of error is upper bounded by $e^{-n\{E_{ex}(R)]-o(n)\}}$. Since ρR is an increasing function of R and $E_{ex}(R)$ is a decreasing function, the best choice of R is the solution to the equation

$$\rho R = E_{ex}(R) \qquad \text{(B.3)}$$

or, equivalently

$$\rho R = \sup_{\varrho \geq 1}[E_x(\varrho) - \varrho R]. \qquad \text{(B.4)}$$

Below we show that the solution to this equation is given by

$$R = R(\rho) \stackrel{\Delta}{=} \sup_{\varrho \geq 1} \frac{E_x(\varrho)}{\varrho + \rho} \qquad \text{(B.5)}$$

and for this choice of R, both exponents in the last line of (B.2) are given by

$$\rho R(\rho) = \sup_{\varrho \geq 1} \frac{\rho E_x(\varrho)}{\varrho + \rho} \qquad \text{(B.6)}$$

which is exactly the expression of $\underline{E}(\rho)$ in the range $\rho > \rho_-$. In the range $\rho < \rho_+$, exactly the same arguments hold, except that $E_{ex}(R)$ and $E_x(\varrho)$ and $\sup_{\varrho \geq 1}$ are replaced by $E_r(R)$, $E_0(\varrho)$, and $\sup_{0 \leq \varrho \leq 1}$, respectively. In the intermediate range, the same line of arguments hold once again, with $\varrho = 1$ and $E_x(1) \equiv E_0(1)$.

It remains to show that $R(\rho)$ in (B.5) solves equation (B.4) for $\rho > \rho_-$, and then similar arguments will follow for the two other ranges. Let $R(\rho)$ be defined as in (B.5) and let $R'(\rho)$ be defined as the solution to (B.4). We wish to prove that $R(\rho) = R'(\rho)$. To this end, we will prove that both $R(\rho) \geq R'(\rho)$ and $R(\rho) \leq R'(\rho)$. To prove the first inequality, let $\varrho(R)$ denote the achiever of $E_{ex}(R) = \sup_{\varrho \geq 1}[E_x(\varrho) - \varrho R]$. Then, by definition of $R'(\rho)$, we obviously have

$$\rho R'(\rho) = E_x[\varrho(R'(\rho))] - \varrho[R'(\rho)]R'(\rho) \tag{B.7}$$

i.e.,

$$R'(\rho) = \frac{E_x[\varrho(R'(\rho))]}{\varrho[R'(\rho)] + \rho} \leq \sup_{\varrho \geq 1} \frac{E_x(\varrho)}{\varrho + \rho} \equiv R(\rho). \tag{B.8}$$

To prove the second (opposite) inequality, let $\varrho(\rho)$ be the achiever of $R(\rho)$, that is,

$$R(\rho) = \frac{E_x[\varrho(\rho)]}{\varrho(\rho) + \rho}, \tag{B.9}$$

or, equivalently,

$$\rho R(\rho) = E_x[\varrho(\rho)] - \varrho(\rho)R(\rho). \tag{B.10}$$

But the l.h.s. cannot exceed $\sup_{\varrho \geq 1}[E_x(\varrho) - \varrho R(\rho)] = E_{ex}[R(\rho)]$, and so,

$$\rho R(\rho) \leq E_{ex}[R(\rho)]. \tag{B.11}$$

Now, as mentioned earlier, the function ρR is increasing in R whereas the function $E_{ex}(R)$ is decreasing. Thus, the value of R for which there is equality $\rho R = E_{ex}(R)$, which is $R'(\rho)$, cannot be smaller than any value of R, for which $\rho R \leq E_{ex}(R)$, like $R(\rho)$. Hence, $R(\rho) \leq R'(\rho)$. This completex the proof of Theorem 2.

Appendix C

Derivation of a lower bound on the generalized rate–distortion function. Consider the minimization of the generalized mutual information

$$\tilde{I}(U;\hat{U}) = -\boldsymbol{E}\left\{\int_{\mathcal{U}} \mathrm{d}\hat{u} \prod_{i=1}^{k} p(\hat{u}|U_i)^{\alpha_i}\right\} = -\int_{\mathcal{U}} \mathrm{d}\hat{u} \prod_{i=1}^{k} \int_{\mathcal{U}} \mathrm{d}u_i p(u_i) p(\hat{u}|u_i)^{\alpha_i}. \tag{C.1}$$

Similarly as in [18, Sect. IV, Example 2] and [6], since we are dealing with an exponentially small estimation error level (small distortion), then for reasons of convenience, we approximate our distortion measure $d(u,\hat{u}) = |\hat{u} - u|^\rho$ $(u,\ \hat{u} \in \mathcal{U})$ by

$$d'(u,\hat{u}) = |(\hat{u} - u) \bmod 1|^\rho. \tag{C.2}$$

where

$$t \bmod 1 \stackrel{\Delta}{=} \left\langle t + \frac{1}{2} \right\rangle - \frac{1}{2} \tag{C.3}$$

$\langle r \rangle$ being the fractional part of r, that is, $\langle r \rangle = r - \lfloor r \rfloor$. The justification is that for very small distortion (the high–resolution limit), the modulo 1 operation has a negligible effect, and hence $d'(u, \hat{u})$ becomes essentially equivalent to the original distortion measure $d(u, \hat{u}) = |\hat{u} - u|^\rho$. Using the same reasoning as in [18, Sect. IV, Example 2] and [6], there is no loss of optimality by confining attention to channels $p(\hat{u}|u)$ of the form $f(w)$ with $w = \hat{u} - u \bmod 1$. Thus, the minimization of $\tilde{I}(U; \hat{U})$ reduces to the maximization of

$$U(f) = \prod_{i=1}^{k} \mathrm{d}w_i \int_{-1/2}^{+1/2} \mathrm{d}w_i [f(w_i)]^{\alpha_i} \tag{C.4}$$

subject to the constraints

$$\int_{-1/2}^{+1/2} \mathrm{d}w \cdot f(w) = 1 \tag{C.5}$$

$$\int_{-1/2}^{+1/2} \mathrm{d}w \cdot |w|^\rho f(w) = D. \tag{C.6}$$

This optimization problem is not trivial, but we can find an upper bound on $U(f)$ in terms of D for small D. We begin with the following bound for each one of the factors of $U(f)$:

$$\int_{-1/2}^{+1/2} \mathrm{d}w \cdot [f(w)]^{\alpha_i} = \int_{-1/2}^{+1/2} \mathrm{d}w \cdot [f(w)]^{\alpha_i} \cdot \left(\frac{|w|^\rho + D}{|w|^\rho + D} \right)^{\alpha_i} \tag{C.7}$$

$$= \int_{-1/2}^{+1/2} \mathrm{d}w \cdot [f(w)(|w|^\rho + D)]^{\alpha_i} \cdot \left[\frac{1}{(|w|^\rho + D)^{\theta_i}} \right]^{1-\alpha_i} \tag{C.8}$$

$$\leq \left[\int_{-1/2}^{+1/2} \mathrm{d}w \cdot f(w)(|w|^\rho + D) \right]^{\alpha_i} \cdot \left[\int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(|w|^\rho + D)^{\theta_i}} \right]^{1-\alpha_i} \tag{C.9}$$

$$= (2D)^{\alpha_i} \cdot \left[\int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(|w|^\rho + D)^{\theta_i}} \right]^{1-\alpha_i}. \tag{C.10}$$

where $\theta_i = \alpha_i/(1 - \alpha_i)$ and the third line follows from Hölder's inequality. It remains to evaluate the integral

$$I = \int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(|w|^\rho + D)^{\theta_i}}. \tag{C.11}$$

To this end, we have to distinguish between the cases $\theta_i > 1/\rho$ and $\theta_i < 1/\rho$ (the case $\theta_i = 1/\rho$ can be solved separately or approached as a limit of $\theta_i \to 1/\rho$ from either side). For the case $\theta_i > 1/\rho$,

letting

$$c_i = \int_{-\infty}^{+\infty} \frac{\mathrm{d}t}{(|t|^\rho + 1)^{\theta_i}}, \tag{C.12}$$

we can easily bound I as follows:

$$\begin{aligned} I &= D^{-\theta_i} \int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(|w/D^{1/\rho}|^\rho + 1)^{\theta_i}} && \text{(C.13)} \\ &\leq D^{1/\rho-\theta_i} \int_{-\infty}^{+\infty} \frac{\mathrm{d}(w/D^{1/\rho})}{(|w/D^{1/\rho}|^\rho + 1)^{\theta_i}} && \text{(C.14)} \\ &\leq c_i D^{1/\rho-\theta_i}. && \text{(C.15)} \end{aligned}$$

For $\theta_i < 1/\rho$, we proceed as follows:

$$\begin{aligned} I &= D^{1/\rho-\theta_i} \int_{-1/(2D^{1/\rho})}^{+1/(2D^{1/\rho})} \frac{\mathrm{d}t}{(|t|^\rho + 1)^{\theta_i}} && \text{(C.16)} \\ &= 2D^{1/\rho-\theta_i} \int_{0}^{+1/(2D^{1/\rho})} \frac{\mathrm{d}t}{(t^\rho + 1)^{\theta_i}} && \text{(C.17)} \\ &\leq 2D^{1/\rho-\theta_i} \int_{0}^{+1/(2D^{1/\rho})} \frac{\mathrm{d}t}{(\max\{t^\rho, 1\})^{\theta_i}} && \text{(C.18)} \\ &= 2D^{1/\rho-\theta_i} \int_{0}^{+1/(2D^{1/\rho})} \frac{\mathrm{d}t}{\max\{t^{\rho\theta_i}, 1\}} && \text{(C.19)} \\ &= 2D^{1/\rho-\theta_i} \left[\int_0^1 \frac{\mathrm{d}t}{1} + \int_1^{+1/(2D^{1/\rho})} \frac{\mathrm{d}t}{t^{\rho\theta_i}} \right] && \text{(C.20)} \\ &= 2D^{1/\rho-\theta_i} \left[1 + \frac{t^{1-\rho\theta_i}}{1-\rho\theta_i} \Big|_1^{1/(2D^{1/\rho})} \right] && \text{(C.21)} \\ &= 2D^{1/\rho-\theta_i} \left[1 + \frac{2^{\rho\theta_i-1} D^{\theta_i-1/\rho} - 1}{1-\rho\theta_i} \right] && \text{(C.22)} \\ &\leq \frac{2^{\rho\theta_i}}{1-\rho\theta_i}. && \text{(C.23)} \end{aligned}$$

Thus, defining $c_i' = 2^{\alpha_i} \max\{c_i, 2^{\rho\theta_i}/(1-\rho\theta_i)\}$, we have

$$\begin{aligned} \int_{-1/2}^{+1/2} \mathrm{d}w \cdot [f(w)]^{\alpha_i} &\leq (2D)^{\alpha_i} I^{1-\alpha_i} && \text{(C.24)} \\ &\leq c_i' \cdot D^{\zeta_\rho(\alpha_i)}, && \text{(C.25)} \end{aligned}$$

where the function $\zeta_\rho(\cdot)$ is defined as in (40). Thus,

$$U(f) \leq c \cdot D^{\sum_{i=1}^k \zeta_\rho(\alpha_i)} \tag{C.26}$$

where $c = \prod_{i=1}^{k} c_i'$. Finally, it follows that

$$\tilde{R}(D) \geq -c \cdot D^{\sum_{i=1}^{k} \zeta_\rho(\alpha_i)} \tag{C.27}$$

as claimed.

References

[1] K. L. Bell, Y. Steinberg, Y. Ephraim, and H. L. van Trees, "Extended Ziv—Zakai lower bounds for vector parameter estimation," *IEEE Trans. Inform. Theory*, vol. 43, no. 2, pp. 626–637, March 1997.

[2] Z. Ben–Haim and Y. C. Eldar, "A lower bound on the Bayesian MSE based on the optimal bias function," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5179–5196, November 2009.

[3] D. Chazan, M. Zakai, and J. Ziv, "Improved lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 1, pp. 90–93, January 1975.

[4] R. G. Gallager, *Information Theory and Reliable Communication*, New York, Wiley 1968.

[5] I. Leibowitz and R. Zamir, "A Ziv–Zakai–Rényi lower bound on distortion at high resolution," *Proc. 2008 IEEE Information Theory Workshop*, Porto, Portugal, May 2008.

[6] N. Merhav, "Data processing inequalities based on a certain structured class of information measures with application to estimation theory," *IEEE Trans. Inform. Theory*, vol. 58, no. 8, pp. 5287–5301, August 2012.

[7] N. Merhav, "On optimum parameter modulation–estimation from a large deviations perspective," *IEEE Trans. Inform. Theory*, vol. 58, no. 12, pp. 7215–7225, December 2012.

[8] A. No, K. Venkat, and T. Weissman, "Joint source–channel coding of one random variable over the Poisson channel," *Proc. ISIT 2012*, July 2012.

[9] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, " Lower bounds to error probability for coding on discrete memoryless channels. I" *Information and Control*, vol. 10, pp. 65–103, January 1967.

[10] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, " Lower bounds to error probability for coding on discrete memoryless channels. II" *Information and Control*, vol. 10, pp. 522–552, May 1967.

[11] S. Tridenski and R. Zamir, "Bounds for joint source–channel coding at high SNR," *Proc. ISIT 2011*, pp. 874–878, St. Petersburg, Russia, August 2011.

[12] H. L. Van Trees and K. L. Bell (Eds), *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*, IEEE Press, Published by John Wiley & Sons, 2007.

[13] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw–Hill, New York, 1979.

[14] A. J. Weiss, *Fundamental Bounds in Parameter Estimation*, Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, June 1985.

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

[16] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, 1975, pp. 87–123.

[17] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT–15, no. 3, pp. 386–391, May 1969.

[18] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 3, pp. 275–283, May 1973.