IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Data Processing Inequalities Based on a Certain Structured Class of Information Measures with Application to Estimation Theory

Neri Merhav

CCIT Report #796
September 2011

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

CCIT REPORT #796 September 2011

Data Processing Inequalities Based on a Certain Structured Class of Information Measures with Application to Estimation Theory *

Neri Merhav

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

Abstract

We study data processing inequalities that are derived from a certain class of generalized information measures, where a series of convex functions and multiplicative likelihood ratios are nested alternately. While these information measures can be viewed as a special case of the most general Zakai–Ziv generalized information measure, this special nested structure calls for attention and motivates our study. Specifically, a certain choice of the convex functions leads to an information measure that extends the notion of the Bhattacharyya distance (or the Chernoff divergence): While the ordinary Bhattacharyya distance is based on the (weighted) geometric mean of two replicas of the channel's conditional distribution, the more general information measure allows an arbitrary number of such replicas. We apply the data processing inequality induced by this information measure to a detailed study of lower bounds of parameter estimation under additive white Gaussian noise (AWGN) and show that in certain cases, tighter bounds can be obtained by using more than two replicas. While the resulting lower bound may not compete favorably with the best bounds available for the ordinary AWGN channel, the advantage of the new lower bound, relative to the other bounds, becomes significant in the presence of channel uncertainty, like unknown fading. This different behavior in the presence of channel uncertainty is explained by the convexity property of the information measure.

Index Terms: Data processing inequality, Chernoff divergence, Bhattacharyya distance, Gallager function, parameter estimation, fading.

*This research was supported by the Israeli Science Foundation (ISF), grant no. 208/08.

1 Introduction

In classical Shannon theory, data processing inequalities (in various forms) are frequently used to prove converses to coding theorems and to establish fundamental properties of information measures, like the entropy, the mutual information, and the Kullback–Leibler divergence [5]. A very well–known example is the converse to the joint source–channel coding theorem, which sets the stage for the separation theorem of Information Theory: When a source with rate–distortion function $R(D)$ is encoded and transmitted across a channel with capacity C, the distortion of the reconstruction at the decoder must obey the inequality $R(D) \leq C$, or equivalently, $D \geq R^{-1}(C)$. This lower bound is achievable (e,g., by separate source coding and channel coding) in the limit of large block length.

Ziv and Zakai [24] (see also Csiszár [6], [7], [8] for related work) have observed that in order to obtain a wider class of data processing inequalities, the (negative) logarithm function, that plays a role in the classical mutual information, can be replaced by an arbitrary convex function Q, provided that it obeys certain regularity conditions. This generalized mutual information, $I_Q(X;Y)$, was further generalized in [22] to be based on multivariate convex functions, as opposed to the univariate convex functions in [24]. In analogy to the classical converse to the joint source–channel coding theorem, one can then define a generalized rate–distortion function $R_Q(D)$ (as the minimum of the generalized mutual information between the source and the reproduction, s.t. some distortion constraint) and a generalized channel capacity C_Q (as the maximum generalized mutual information between the channel input and output) and establish another lower bound on the distortion via the inequality $R_Q(D) \leq C_Q$ that stems from the data processing inequality of I_Q. While this lower bound obviously cannot be tighter than its classical counterpart in the limit of long blocks (which is asymptotically achievable), Ziv and Zakai have demonstrated that for short block codes (e.g., codes of block length 1), sharper lower bounds can certainly be obtained (see also [14] for more recent developments).

Gurantz, in his M.Sc. work [10] (supervised by Ziv and Zakai), continued the work in [24] at a specific direction: He constructed a special class of generalized information functionals defined by iteratively alternating between applications of convex functions and multiplications by likelihood

ratios[1] (or more generally, Radon–Nykodim derivatives). After proving that this functional obeys a data processing inequality, Gurantz demonstrated how it can be used to improve on the Arimoto bound for coding above capacity [2] and on the Gallager upper bound of random coding [9] by a pre-factor of 1/2.

Motivated by the belief that the interesting nested structure of Gurantz' information functional can be further exploited, we continue, in this work, to investigate this information measure and we further study its properties and potential.

We begin by putting the Gurantz' functional in the broader perspective of the other information measures due to Ziv and Zakai [22], [24] (Section 2). Specifically, we first discuss two possible methods to define a generalized mutual information from the Gurantz' functional, each one with its advantages and disadvantages. We then show that both of these generalized mutual informations can be viewed as special cases of the generalized mutual information of [22], which is based on multivariate convex functions. The proof of this fact then naturally suggests a way to broaden the scope and define a family of information measures with a tree structure of convex functions and likelihood ratios.

We then focus on a concrete choice of the convex functions (Section 3) in the Gurantz' information measure (in particular, power functions), which turn out to yield an information measure that extends the notion of the Bhattacharyya distance (or the Chernoff divergence): While the ordinary Bhattacharyya distance is based on the (weighted) geometric mean of two replicas of the channel's conditional distribution (see, e.g., [17, eq. (2.3.15)]), the more general information measure considered here, allows an arbitrary number of such replicas. This generalized Bhattacharyya distance is also intimately related to the Gallager function $E_0(\rho, Q)$ [9], [17], which is indeed another information measure obeying a data processing inequality [13, Proposition 2], since it is yet another special case of the information measures in [22].

Finally, we apply the data processing inequality, induced by the above described generalized Bhattacharyya distance, to a detailed study of lower bounds on parameter estimation under additive white Gaussian noise (AWGN) and show that in certain cases, tighter bounds can be obtained by using more than two replicas (Section 4). In this particular case, it turns out that three is the

[1]The exact form of this will be given in the sequel.

optimum number of replicas in the high SNR regime. While the resulting lower bound may still not compete favorably with the best available bounds for the ordinary AWGN channel, the advantage of the new lower bound, relative to the other bounds, becomes apparent in the presence of channel uncertainty, like in the case of an AWGN channel with unknown fading. This different behavior, in the presence of channel uncertainty, is explained by the convexity property of the information measure.

2 Preliminaries and Basic Observations

In [10], a generalized information functional was defined in the following manner: Let X and Y be random variables taking on values in alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, where here and throughout the sequel, all alphabets may either be finite, countably infinite, or uncountably infinite, like intervals or the entire real line. Let $x_1, x_2, \ldots, x_k$ be a given list of symbols (possibly with repetitions) from $\mathcal{X}$. Let $Q_1, Q_2, \ldots, Q_k$ be a collection of univariate functions, defined on the positive reals, with the following properties, holding for all i:

1. $\lim_{t\to 0} tQ_i(1/t) = 0$.
2. $|Q_i(0)| < \infty$.
3. Either the function $\hat{Q}_i \stackrel{\Delta}{=} Q_1 \circ Q_2 \circ \ldots \circ Q_i$ is monotonically non-decreasing and Q_{i+1} is convex, or $\hat{Q}_i$ is monotonically non–increasing and Q_{i+1} is concave (here, the notation $\circ$ means function composition).

Now, define the *Gurantz' functional* as

$$G(Y|x, x_1, \ldots, x_k) = \int_{\mathcal{Y}} \mathrm{d}y \cdot P_{Y|X}(y|x) \times Q_1\left(\frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x)} \cdot Q_2\left(\frac{P_{Y|X}(y|x_2)}{P_{Y|X}(y|x_1)} \cdot Q_3\left(\ldots Q_k\left(\frac{P_{Y|X}(y|x_k)}{P_{Y|X}(y|x_{k-1})}\right)\ldots\right)\right)\right),$$

where here and throughout, it is understood that integrals and probability density functions should be replaced, in the countable alphabet case, by summations and probability mass functions, respectively.

The data processing inequality associated with the Gurantz' functional is the following: Let $X \to Y \to Z$ be a Markov chain and let Q_1 be a convex function which, together with $Q_2, \ldots, Q_k$, complies with rules 1–3 above. Then,

$$G(Y|x, x_1, \ldots, x_k) \geq G(Z|x, x_1, \ldots, x_k). \tag{1}$$

The direct proof of this inequality is fairly straightforward [10]: First, observe that

$$G(Y|x, x_1, \ldots, x_k) = G(Y, Z|x, x_1, \ldots, x_k) \tag{2}$$

due to the Markov property. Then, one can easily obtain a sequence of lower bounds on the right–hand–side (r.h.s.) of eq. (2) by successive applications of Jensen's inequality, where at each stage, the expectation with respect to (w.r.t.) $P_{Y|X_i,Z}$ propagates into the next convex function and then partially cancels out with the factor $P_{Y,Z|X_i}(y, z|x_i)$ at the denominator of the likelihood ratio.

Note that according to the definition of $G(Y|x, x_1, \ldots, x_k)$, x is the random variable that controls the distribution of Y (as the averaging is w.r.t. $P_{Y|X}(\cdot|x)$), whereas $x_1, \ldots, x_k$ can be viewed as 'dummy' variables. One way to define a generalized mutual information based on G, which is a functional of $\{P_{XY}(x, y)\}$, is by assigning a certain probability distribution to $(x, x_1, \ldots, x_k)$. Let $P(x, x_1, \ldots, x_k) = P_X(x)P(x_1, \ldots, x_k|x)$, where $P_X(\cdot)$ is the actual distribution of the random variable X and $P(x_1, \ldots, x_k|x)$ is an arbitrary conditional distribution of $(X_1, \ldots, X_k)$ given $X = x$, for example, $P(x_1, \ldots, x_k|x) = \prod_{i=1}^k P_X(x_i)$ or $P(x_1, \ldots, x_k|x) = \prod_{i=1}^k \delta(x_i - f_i(x))$ for some deterministic functions $\{f_i\}$. Now, for a given choice of $\{P(x_1, \ldots, x_k|x)\}$, the *Gurantz' mutual information* $I_G(X; Y)$ can be defined as

$$I_G(X; Y) = \boldsymbol{E}G(Y|X, X_1, \ldots, X_k) \tag{3}$$

where the expectation is w.r.t. the above defined joint distribution of the random variables X, X_1,..., X_k. This generalized mutual information is now a well–defined functional of $P_{XY} = P_X \times P_{Y|X}$. In principle, one may apply the generalized data processing inequality $I_G(X; Y) \geq I_G(X; Z)$ for any given choice of $\{P(x_1, \ldots, x_k|x)\}$ (consider these as parameters) and then optimize the resulting distortion bound w.r.t. the choice of these parameters.

Our first observation is that $I_G(X; Y)$ is a special case of the Zakai–Ziv generalized mutual information [22], defined as

$$I_{ZZ}(X; Y) = \boldsymbol{E}Q\left(\frac{\mu_1(X, Y)}{P_{XY}(X, Y)}, \ldots, \frac{\mu_k(X, Y)}{P_{XY}(X, Y)}\right), \tag{4}$$

where Q is a multivariate convex function of k variables and $\mu_i(\cdot,\cdot)$, $i = 1, 2, \ldots, k$, are arbitrary measures on $\mathcal{X} \times \mathcal{Y}$.

To see why this is true, consider the following: For each convex (resp., concave) function $Q_i(t)$, define the bivariate *perspective function* $\tilde{Q}_i(s,t) = s \cdot Q_i(t/s)$, where $s > 0$, which is a convex (resp., concave) function as well, and jointly in both variables [3, Subsection 3.2.6]. Thus,

$$\begin{aligned}
& G(Y|x_1, \ldots, x_k) \\
= & \int_{\mathcal{Y}} \mathrm{d}y P_{Y|X}(y|x) Q_1 \left(\frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x)} Q_2(\ldots) \right) \\
= & \int_{\mathcal{Y}} \mathrm{d}y \cdot P_{Y|X}(y|x') \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')} Q_1 \left(\frac{P_{Y|X}(y|x_1)/P_{Y|X}(y|x')}{P_{Y|X}(y|x)/P_{Y|X}(y|x')} Q_2(\ldots) \right) \\
= & \int_{\mathcal{Y}} \mathrm{d}y \cdot P_{Y|X}(y|x') \tilde{Q}_1 \left(\frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')}, \frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x')} Q_2(\ldots) \right) \\
= & \int_{\mathcal{Y}} \mathrm{d}y \cdot P_{Y|X}(y|x') \tilde{Q}_1 \left(\frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')}, \tilde{Q}_2 \left(\frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x')}, \frac{P_{Y|X}(y|x_2)}{P_{Y|X}(y|x')} Q_3(\ldots) \right) \right) \\
= & \ldots \\
= & \int_{\mathcal{Y}} \mathrm{d}y \cdot P_{Y|X}(y|x') \tilde{Q}_1 \left(\frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')}, \tilde{Q}_2 \left(\ldots \tilde{Q}_k \left(\frac{P_{Y|X}(y|x_{k-1})}{P_{Y|X}(y|x')}, \frac{P_{Y|X}(y|x_k)}{P_{Y|X}(y|x')} \right) \ldots \right) \right) \qquad (5)
\end{aligned}$$

Now, under the assumed properties of the functions $\{Q_i\}$, it is easy to see that

$$\hat{Q}(t_0, t_1, \ldots, t_k) \stackrel{\Delta}{=} \tilde{Q}_1(t_0, \tilde{Q}_2(t_1, \tilde{Q}_3(t_2, \ldots \tilde{Q}_k(t_{k-1}, t_k) \ldots))) \qquad (6)$$

is jointly convex in $(t_0, t_1, \ldots, t_k)$. Thus, upon taking the expectation of the last line of (5) w.r.t. $P_X(x')$, we have (after multiplying the numerator and the denominator of each likelihood ratio by $P_X(x')$) that $\boldsymbol{E}G(Y|X, x_1, \ldots, x_k)$ is an instance of $I_{ZZ}(X;Y)$ for every given $(x_1, \ldots, x_k)$, with the assignments $\mu_i(x,y) = P_X(x)P_{Y|X}(y|x_i)$, $i = 1, 2, \ldots, k$.

We can represent the general structure of information functionals, such as I_G and I_{ZZ}, as well as the forms in the different lines of eq. (5), graphically, in terms of *factor trees* (i.e., factor graphs which are trees) that obey the following rules.

1. There are two types of nodes, variable nodes and function nodes, and each edge of the tree connects a variable node and a function node.

2. The root of the tree is a function node whereas the leaves are variable nodes.

3. Each function node is represented by a convex function Q_i and each variable node is represented by a likelihood ratio $p(y|x_k)/p(y|x_l)$, whose shorthand notation here will be $L_{k,l}$.

4. There is a directed edge from function node Q_i to variable node $L_{j,k}$ (denoted $Q_i \to L_{j,k}$) if the information measure includes a product of the form $Q_i(\cdot) \cdot L_{j,k}$.

5. There is a directed edge from variable node $L_{i,j}$ to function node Q_k (denoted $L_{i,j} \to Q_k$) if $L_{i,j}$ multiplies an argument of Q_k.

6. For every path $L_{i,j} \to Q_k \to L_{l,m}$, j must be equal to l (namely, $x_j = x_l$).

7. For all direct offsprings of the root, $\{L_{i,j}\}$, the second subscript j is the same.

Now observe that I_G and I_{ZZ} correspond to two extreme cases: While I_{ZZ} corresponds to a factor tree where all k leaves are connected directly to the root, I_G corresponds to a simple chain (i.e., every node has one offspring and there is only one leaf), which alternates between variable nodes and function nodes. The form that appears in the last line of (5) corresponds to a binary tree with a comb structure, i.e., every node that is not a leaf has two offsprings, one of which is a leaf. More generally, every factor graph with a tree structure, that complies with the above rules, corresponds to a valid information measure that satisfies a data processing inequality. For example, the factor graph of Fig. 1 corresponds to the information measure

$$\int_{\mathcal{Y}} \mathrm{d}y \cdot p(y|x_a) Q_1 \left(\frac{p(y|x_b)}{p(y|x_a)} Q_2 \left(\frac{p(y|x_d)}{p(y|x_b)}, \frac{p(y|x_e)}{p(y|x_b)} \right), \frac{p(y|x_c)}{p(y|x_a)} Q_3 \left(\frac{p(y|x_f)}{p(y|x_c)} \right) \right). \tag{7}$$

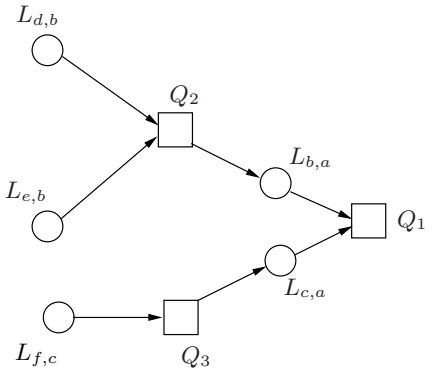

Figure 1: The factor graph that represents the generalized mutual information of eq. (7).

In view of the observation that $\boldsymbol{EG}(Y|X, x_1, \ldots, x_k)$ a special case of the $I_{ZZ}(X;Y)$, there is another way to use it to obtain data processing inequalities for communication systems. According

to [22, Theorems 3.1 and 5.1], the following is true: Let $U \to X \to Y$ be a Markov chain and let $V = g(Y)$ where g is a deterministic function. Let $\mu_i(x,y)$, $i = 1, 2, \ldots, k$, be arbitrary measures and define $\mu_i(u,y) = P_U(u) \sum_x P_{X|U}(x|u)\mu_i(x,y)/P_X(x)$, $\mu_i(u,v) = \sum_{y:\ g(y)=v} \mu_i(u,y)$, $i = 1, \ldots, k$. Then,

$$I_{ZZ}(X;Y) \geq I_{ZZ}(U;V). \tag{8}$$

As described informally in the Introduction, the maximum of the left–hand side (l.h.s.) over P_X and the minimum of the r.h.s. over $P_{V|U}$ (subject to some distortion constraint) can be thought of as generalized channel capacity and generalized rate–distortion function, respectively, as in [22]. Now, consider the special case where I_{ZZ} is based on a multivariate convex function $\hat{Q}$ as defined in (6), where each bivariate convex function $\tilde{Q}_i$ is the perspective of a certain univariate convex function, i.e., $\tilde{Q}_i(s,t) = s \cdot Q_i(t/s)$. Then by a similar argument as above (going the other direction), we get another information measure in the spirit of Gurantz:

$$I_G(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y) Q_1\left(\frac{\mu_1(x,y)}{P_{XY}(x,y)} Q_2\left(\frac{\mu_2(x,y)}{\mu_1(x,y)} \ldots Q_k\left(\frac{\mu_k(x,y)}{\mu_{k-1}(x,y)}\right)\ldots\right)\right). \tag{9}$$

Since it is a special case of $I_{ZZ}(X;Y)$, then it obviously satisfies a strong[2] data processing inequality $I_G(X;Y) \geq I_G(U;V)$. Assuming, in addition, that the encoder is given by a deterministic function $x = f(u)$, we can choose $\mu_i(x,y) = P_X(x)P_{Y|X}(y|x_i)$, where $x_i = f(u_i)$ is a specific member in $\mathcal{X}$ and then $\mu(y|u_i) = P_{Y|X}(y|f(u_i))$. We then obtain

$$\begin{aligned} &\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}x\mathrm{d}y \cdot P_{XY}(x,y) Q_1\left(\frac{P_{Y|X}(y|f(u_1))}{P_{Y|X}(y|x)} Q_2\left(\ldots Q_k\left(\frac{P_{Y|X}(y|f(u_k))}{P_{Y|X}(y|f(u_{k-1}))}\right)\ldots\right)\right) \\ \geq\ &\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}u\mathrm{d}v \cdot P_{UV}(u,v) Q_1\left(\frac{P_{V|U}(v|u_1)}{P_{V|U}(v|u)} Q_2\left(\ldots Q_k\left(\frac{P_{V|U}(v|u_k)}{P_{V|U}(v|u_{k-1})}\right)\ldots\right)\right). \end{aligned} \tag{10}$$

Multiplying both sides by $\prod_i P_U(u_i)$ and integrating over $\{u_i\}$, we get

$$\begin{aligned} &\boldsymbol{E} Q_1\left(\frac{P_{Y|X}(Y|X_1)}{P_{Y|X}(Y|X)} Q_2\left(\ldots Q_k\left(\frac{P_{Y|X}(Y|X_k)}{P_{Y|X}(Y|X_{k-1}))}\right)\ldots\right)\right) \\ \geq\ &\boldsymbol{E} Q_1\left(\frac{P_{V|U}(V|U_1)}{P_{V|U}(V|U)} Q_2\left(\ldots Q_k\left(\frac{P_{V|U}(V|U_k)}{P_{V|U}(V|U_{k-1})}\right)\ldots\right)\right). \end{aligned} \tag{11}$$

where the expectation on the l.h.s. is w.r.t. $P_{XY}(x,y)\prod_i P_X(x_i)$, and the expectation on the r.h.s. is w.r.t. $P_{UV}(u,v)\prod_i P_U(u_i)$. This is different from the data processing theorem in [10], because it allows 'moving' in both directions of the Markov chain and not only to the right.

[2]By "strong data processing inequality" we use the terminology of [22], meaning that for a Markov chain $U \to X \to Y$ and $V = g(Y)$, we have $I_G(X;Y) \geq I_G(U;Y) \geq I_G(U;V)$.

To summarize, we have seen two approaches to derive data processing inequalities from the inequality $G(Y|u,u_1,\ldots,u_k) \geq G(V|u,u_1,\ldots,u_k)$ for a Markov chain $U \to Y \to V$ (where have slightly changed the notation relative to eq. (1)): According to the first approach, one allows an arbitrary distribution $P_{U_1,\ldots U_k|U}$ and averages both sides w.r.t. $P_U \times P_{U_1,\ldots U_k|U}$. This defines the $I_G(U;Y)$ and $I_G(U;V)$ as functionals of P_{UY} and P_{UV}, respectively, where $P_{U_1,\ldots U_k|U}$ serve as free parameters that can be optimized, to get the tightest distortion bound. The advantage of this approach is the free choice of $P_{U_1,\ldots U_k|U}$, which gives many degrees of freedom. The disadvantage is that $I_G(U;Y)$ depends on the source and the encoder and there is no apparent way to prove a strong data processing theorem, in general, i.e., to prove that $I_G(U;Y)$ can be further upper bounded by $I_G(X;Y)$ (whatever its definition may be) and thereby define a channel capacity, that is independent of the source (in addition to a generalized rate distortion function, which is $\min I_G(U;V)$ s.t. some distortion constraint). The inequality $I_G(U;Y) \geq I_G(U;V)$ is relevant to situations where there is no encoder to be optimized, namely, when the channel from U to Y is given and cannot be shaped by encoding. This happens, for example, in parameter estimation problems.

According to the second approach, one limits $P_{U_1,\ldots U_k|U}(u_1,\ldots,u_k|u)$ to be $\prod_{i=1}^k P_U(u_i)$. This leaves no degrees of freedom, but it admits a strong data processing theorem, and hence allows to define both a generalized rate–distortion function and a generalized channel capacity, whose calculations are completely decoupled of each other. It is also much simpler to use. This type of data processing inequality is more suitable for coded communication systems, where there is also an encoder to optimize.

From this point onward, we essentially confine ourselves to the second option, mainly for reasons of simplicity.

3 Choice of the Convex Functions

An interesting and convenient choice of the functions $\{Q_i\}$ is the following: $Q_1(t) = -t^{a_1}$, and $Q_i(t) = t^{a_i}$ for $i \geq 2$, where $0 \leq a_i \leq 1$, $i = 1,\ldots,k$. In this case, $\tilde{Q}_i(t) = -t^{\prod_{j=1}^i a_j}$ is monotonically

decreasing and Q_{i+1} is concave, so this choice complies with the rules. In this case, we have:

$$\begin{aligned} G(Y|x_0,x_1,\ldots,x_k) &= -\int_{\mathcal{Y}} \mathrm{d}y P_{Y|X}(y|x_0) \times \\ &\quad \left(\frac{P_{Y|X}(y|x_1)}{P_{Y|X}(y|x_0)}\left(\frac{P_{Y|X}(y|x_2)}{P_{Y|X}(y|x_1)}\left(\ldots\left(\frac{P_{Y|X}(y|x_k)}{P_{Y|X}(y|x_{k-1})}\right)^{a_k}\right)^{a_{k-1}}\ldots\right)^{a_2}\right)^{a_1} \\ &= -\int_{\mathcal{Y}} \mathrm{d}y \prod_{i=0}^{k} P_{Y|X}^{b_i}(y|x_i) \end{aligned} \tag{12}$$

where $\{b_i\}$ are given by:

$$\begin{aligned} b_0 &= 1-a_1 \\ b_1 &= (1-a_2)a_1 \\ b_2 &= (1-a_3)a_1a_2 \\ &\ldots \\ b_{k-1} &= (1-a_k)\prod_{i=1}^{k-1} a_i \\ b_k &= \prod_{i=1}^{k} a_i \end{aligned} \tag{13}$$

Note that the coefficients $b_0,\ldots,b_k$ are all non–negative and their sum is equal to 1. Conversely, for every set of coefficients $\{b_i\}$ with these properties, one can find $a_1,\ldots,a_k$, all in $[0,1]$, using the following inverse transformation:

$$\begin{aligned} a_1 &= 1-b_0 \\ a_2 &= 1-\frac{b_1}{1-b_0} \\ &\ldots \\ a_k &= 1-\frac{b_{k-1}}{1-\sum_{i=0}^{k-2} b_i}. \end{aligned} \tag{14}$$

This allows us parametrize the information measure directly in terms of an arbitrary set of non–negative numbers $\{b_i\}$ summing to unity, without worrying about $\{a_i\}$. The resulting information measure can then be viewed as an extension of the Chernoff divergence between two conditional densities, $P_{Y|X}(y|x_0)$ and $P_{Y|X}(y|x_1)$, to a general number of densities, where the powers of $\{P_{Y|X}(y|x_i)\}$ always sum up to unity. Specializing this to the case $b_i = 1/(k+1)$ for all $i = 0,1,\ldots,k$, eq. (12) extends the Bhattacharyya distance. Following the discussion of the second

option at the end of Section 2, if, in addition, we assign $P_{X_1,\ldots,X_k|X_0}(x_1,\ldots,x_k|x_0)=\prod_{i=1}^{k}P_X(x_i)$, then $I_G(X;Y)=\boldsymbol{E}G(Y,X,X_1,\ldots,X_k)=-e^{-E_0(k,P_X)}$, where E_0 is the Gallager function [9]

$$E_0(\rho,P_X)=-\ln\left\{\int_{\mathcal{Y}}\mathrm{d}y\left[\int_{\mathcal{X}}\mathrm{d}xP_X(x)P_{Y|X}^{1/(1+\rho)}(y|x)\right]^{1+\rho}\right\}. \tag{15}$$

Thus, $I_G(X;Y)$ extends, not only the Chernoff divergence, but also the Gallager function, albeit only at integer values of the parameter ρ. Indeed, it was shown in [13, Proposition 2] that the Gallager function (for every real $\rho\geq 0$) satisfies a data processing inequality, because it is also a special case of $I_{ZZ}(X;Y)$. In other words, the generalized Chernoff divergence can be obtained as a special case of $I_{ZZ}(X;Y)$ in two different ways: one is via I_G and the other is via the Gallager function. The advantage of working with Gallager's function for integer values of ρ, is that an integral raised to an integer power $(k+1)$ can be expressed in terms of $(k+1)$–dimensional integration over the $(k+1)$ replicas, x_0,x_1,...,x_k, that in turn can be commuted with the additional out–most integration over $\mathcal{Y}$. In some situations, this enables explicit calculations more conveniently.

4 Application to Estimation Theory

In this section, we apply the data processing inequality associated with the generalized Bhattacharyya distance to obtain a Bayesian lower bound on the estimation error of parameter estimators of a parameter u modulated in a signal $x(t,u)$ that is in turn corrupted by Gaussian white noise. As mentioned earlier, we essentially adopt the second approach discussed at the end of Section 2: Although we use the data processing inequality $I_G(U;V)\leq I_G(U;Y)$, in some of our derivations, we eventually further upper bound $I_G(U;Y)$ by a universal bound, that is independent of the modulation scheme $x(t,\cdot)$, so in a way, it conveys the notion of generalized capacity. The model we focus on is the following.

The source symbol U, which is uniformly distributed in $\mathcal{U}=[-1/2,+1/2]$, plays the role of a random parameter to be estimated. For reasons of convenience, we define the distortion measure between a realization u of the source and an estimate v (both in $\mathcal{U}$) as

$$d(u,v)=[(u-v)\bmod 1]^2. \tag{16}$$

where

$$t\bmod 1\stackrel{\Delta}{=}\left\langle t+\frac{1}{2}\right\rangle-\frac{1}{2} \tag{17}$$

$\langle r \rangle$ being the fractional part of r, that is, $\langle r \rangle = r - \lfloor r \rfloor$. Note that in the high–resolution limit (corresponding to the high signal–to–noise (SNR) limit), the modulo 1 operation has a negligible effect, and hence $d(u,v)$ becomes essentially equivalent to the ordinary quadratic distortion. Indeed, most of our results in the sequel, refer to the high SNR regime. At any rate, under the modulo 1 quadratic distortion measure, it is convenient to visualize U as being evenly distributed across the circumference of a circle of radius $1/(2\pi)$ (or as a phase parameter) and then $d(u,v)$ is the squared length of the shorter arc (or the smaller angel) between the two corresponding points on the circle.

The channel is assumed to be an AWGN channel, namely, the channel output is given by

$$y(t) = x(t,u) + n(t), \qquad 0 \le t < T, \tag{18}$$

where $x(t,u)$ is an arbitrary waveform of unlimited bandwidth, parametrized by u and $n(t)$ is AWGN with two–sided spectral density $N_0/2$. The energy

$$E = \int_O^T x^2(t,u)\mathrm{d}t \tag{19}$$

is assumed to be independent of u (for reasons of simplicity). The estimator v is assumed to be a functional of the channel output waveform $\{y(t),\ 0 \le t < T\}$.

Before deriving lower bounds on the estimation error, $\boldsymbol{E}d(U,V)$, we first need to derive the generalized rate–distortion function and the generalized channel capacity pertaining to the generalized Bhattacharyya distance. This will be done in the next two subsections.

4.1 Derivation of $R(D)$

The "rate–distortion function" $R(D)$ w.r.t. the information measure under discussion is given by the minimum of

$$I(U;V) = -\int_{-1/2}^{+1/2} \mathrm{d}v \left[\int_{-1/2}^{+1/2} \mathrm{d}u P_{V|U}^{1/(k+1)}(v|u) \right]^{k+1}$$

subject to the constraints $\boldsymbol{E}d(U,V) \le D$ and $\int_{-1/2}^{+1/2} \mathrm{d}v P_{V|U}(v|u) = 1$. As explained in [24], it is enough to consider channels of the form $P_{V|U}(v|u) = f(v-u)$. Defining $w = (v-u) \bmod 1$, the

problem is then equivalent to

$$\begin{aligned} \max & \int_{-1/2}^{+1/2} \mathrm{d}w \cdot f^{1/(k+1)}(w) \\ \text{s.t.} & \int_{-1/2}^{+1/2} \mathrm{d}w \cdot w^2 f(w) \leq D \\ & \int_{-1/2}^{+1/2} \mathrm{d}w \cdot f(w) = 1. \end{aligned} \tag{20}$$

This problem is easily solved using calculus of variations [1]. Suppose that f^* is the optimum density and let $f = f^* + \delta g$, where g satisfies

$$\int_{-1/2}^{+1/2} \mathrm{d}w \cdot g(w) = 0. \tag{21}$$

Defining the Lagrangian

$$J(f) = -\int_{-1/2}^{+1/2} \mathrm{d}w \cdot f^{1/(k+1)}(w) + \lambda \int_{-1/2}^{+1/2} \mathrm{d}w \cdot w^2 f(w) + \nu \int_{-1/2}^{+1/2} \mathrm{d}w \cdot f(w), \tag{22}$$

the condition for f^* being an extremum is $\partial J(f+\delta g)/\partial\delta|_{\delta=0} = 0$ for all g. Now,

$$\left.\frac{\partial J(f+\delta g)}{\partial \delta}\right|_{\delta=0} = \int_{-1/2}^{+1/2} \mathrm{d}w \cdot g(w) \left[-\frac{1}{(k+1) f^{k/(k+1)}(w)} + \lambda w^2 + \nu \right] = 0. \tag{23}$$

For this integral to vanish for every g, one must have

$$-\frac{1}{(k+1) f^{k/(k+1)}(w)} + \lambda w^2 + \nu = \text{const.} \tag{24}$$

This means that f^* is of the form

$$f^*(w) = \frac{C(s)}{(1+sw^2)^{1+1/k}}, \tag{25}$$

where

$$C(s) = \left[\int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(1+sw^2)^{1+1/k}} \right]^{-1}, \tag{26}$$

and the parameter s is determined such that

$$C(s) \int_{-1/2}^{+1/2} \frac{w^2 \mathrm{d}w}{(1+sw^2)^{1+1/k}} = D. \tag{27}$$

Define also

$$F(s) = \int_{-1/2}^{+1/2} \frac{w^2 \mathrm{d}w}{(1+sw^2)^{1+1/k}}. \tag{28}$$

Let us denote then $D_s = C(s)F(s)$. Then,

$$\begin{aligned} -R(D_s) &= \left[\int_{-1/2}^{+1/2} \mathrm{d}w [f^*(w)]^{1/(k+1)}\right]^{k+1} \\ &= C(s)\left[\int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(1+sw^2)^{1/k}}\right]^{k+1} \\ &= C(s)[G(s)]^{k+1}, \end{aligned} \tag{29}$$

where we have defined

$$G(s) = \int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(1+sw^2)^{1/k}}. \tag{30}$$

To summarize, we have obtained a parametric representation of $R(D)$ via the variable s:

$$D_s = C(s)F(s) \tag{31}$$

$$R(D_s) = -C(s)[G(s)]^{k+1}, \tag{32}$$

For later use, we point out that the functions $C(s)$, $F(s)$, and $G(s)$ are intimately related. First, observe that

$$\begin{aligned} G(s) &= \int_{-1/2}^{+1/2} \frac{(1+sw^2)\mathrm{d}w}{(1+sw^2)^{1+1/k}} \\ &= \frac{1}{C(s)} + sF(s). \end{aligned} \tag{33}$$

Also, using integration by parts,

$$\begin{aligned} G(s) &= w(1+sw^2)^{-1/k}\Big|_{-1/2}^{+1/2} + \frac{2s}{k}\cdot F(s) \\ &= \left(1+\frac{s}{4}\right)^{-1/k} + \frac{2s}{k}\cdot F(s). \end{aligned} \tag{34}$$

Thus,

$$\frac{1}{C(s)} + sF(s) = \left(1+\frac{s}{4}\right)^{-1/k} + \frac{2s}{k}\cdot F(s), \tag{35}$$

which gives a direct relationship between $C(s)$ and $F(s)$ whenever $k \neq 2$. For $k = 2$, the terms pertaining to $F(s)$ cancel out, but we then have an explicit formula for $C(s)$.

While in general, $R(D)$ is given only a parametric form and not directly, in the limits of very low and very high distortion, one can approximate $R(D)$ directly as an explicit function of D. In particular, it is shown in Appendix A that in the low resolution regime,

$$D(R) \approx \frac{1}{12} - \frac{1}{15}\sqrt{1+R}, \tag{36}$$

where it should be kept in mind that for this information measure, R takes on values in the interval $[-1, 0]$. Here and throughout the sequel, the notation $A \approx B$ means that A/B tends to unity as a certain parameter (in this case, R) tends to a certain limit (in this case, -1), which will always be clear from the context. Here, the term $1/12$ is the variance of U, which is uniform over $[-1/2, +1/2]$, as no useful information is available except the prior.

In the high–resolution regime ($R \to 0$), the behavior depends on whether $k = 1$, $k = 2$, or $k > 2$. In Appendix B, derivations are provided for all three cases. For $k = 1$, the rate–distortion function is approximated as

$$R(D) \approx -4c_1\sqrt{D}. \tag{37}$$

or equivalently, the distortion–rate function is

$$D(R) \approx \frac{R^2}{16c_1^2}, \tag{38}$$

where

$$c_1 = \int_{-\infty}^{+\infty} \frac{\mathrm{d}t}{(1+t^2)^2}. \tag{39}$$

For $k > 2$, we have

$$R(D) \approx -4\left(\frac{k}{k-2}\right)^k \cdot D \quad \text{or} \quad D(R) \approx -\frac{1}{4}\left(1-\frac{2}{k}\right)^k \cdot R, \tag{40}$$

The case $k = 2$ lacks an explicit closed–form direct relation between R and D, but it shows that

$$\log D \approx \log[-R(D)], \tag{41}$$

which means that the relation between R and D is essentially linear, like in the case $k > 2$, but in a slightly weaker sense. It is also easy to extend all the derivations to higher–order moments modulo 1 (see Appendix C for the high resolution analysis).

4.2 Derivation of $I_G(U;Y)$

As mentioned earlier, the channel is assumed to be an AWGN channel with unlimited bandwidth. The probability law of the channel from U to Y is given by

$$P_{Y|U}(y|u) \propto \exp\left\{-\frac{1}{N_0}\int_0^T [y(t) - x(t,u)]^2 dt\right\}, \tag{42}$$

where y in the l.h.s. designates the entire channel output waveform $\{y(t),\ 0 \leq t < T\}$, and $\propto$ means that the constant of proportionality does not depend on u. Let us denote

$$\rho(u,u') = \frac{1}{E} \cdot \int_0^T x(t,u)x(t,u')\mathrm{d}t. \tag{43}$$

Consider the integral

$$\begin{aligned}
&\int \mathrm{d}y \prod_{i=0}^{k} [P_{Y|U}(y|u_i)]^{1/(k+1)} \\
&= \boldsymbol{E}\left\{ \frac{\prod_{i=1}^{k}[P_{Y|U}(y|u_i)]^{1/(k+1)}}{P_{Y|U}(y|u_0)^{k/(k+1)}} \Big| U = u_0 \right\} \\
&= \boldsymbol{E}\left\{ \exp\left[\frac{k}{(k+1)N_0} \int_0^T [y(t) - x(t,u_0)]^2 dt - \frac{1}{(k+1)N_0} \sum_{i=1}^{k} \int_0^T [y(t) - x(t,u_k)]^2 dt \right] \Big| U = u_0 \right\} \\
&= \boldsymbol{E} \exp\left\{ \frac{2}{(k+1)N_0} \int_0^T [x(t,u_0) + n(t)] \left[\sum_{i=1}^{k} x(t,u_i) - kx(t,u_0) \right] dt \right\} \\
&= \exp\left\{ -\frac{E}{N_0} \left[1 - \frac{1}{(k+1)^2} \sum_{i=0}^{k} \sum_{j=0}^{k} \rho(u_i,u_j) \right] \right\},
\end{aligned} \tag{44}$$

where the last passage is associated with the calculation of the moment–generating function of the Gaussian random variable

$$Z = \int_0^T n(t) \left[\sum_{i=1}^{k} x(t,u_i) - kx(t,u_0) \right] \mathrm{d}t \tag{45}$$

which has zero mean and variance $\frac{N_0}{2} \int_0^T [\sum_{i=1}^{k} x(t,u_i) - kx(t,u_0)]^2 \mathrm{d}t$.

The next step, in principle, is take another expectation over the last line of (44) w.r.t. the randomness of $\{U_i\}$. This can be done explicitly for some specific classes of signals (e.g., when U is a phase parameter of a sinusoid), but in general, it is not a trivial task. As in [1] and [22], we then resort to a lower bound (hence an upper bound on $I_G(U;Y)$) based on Jensen's inequality, by raising the expectation operator to the exponent. Denoting

$$\bar{x}(t) = \boldsymbol{E}\{x(t,U)\} = \int_{-1/2}^{+1/2} \mathrm{d}u \cdot x(t,u), \tag{46}$$

it is easily observed that since $\{U_i\}$ are independent, then for all $i \neq j$:

$$\boldsymbol{E}\rho(U_i,U_j) = \frac{1}{E} \cdot \boldsymbol{E}\left\{ \int_0^T x(t,U_i)x(t,U_j)\mathrm{d}t \right\} = \frac{1}{E} \int_0^T [\bar{x}(t)]^2 \mathrm{d}t \stackrel{\Delta}{=} \varrho. \tag{47}$$

Note that the parameter ϱ is always between 0 and 1 and it depends only on the parametric family of signals.[3] Specifically, continuing from the last line of (44), we have

$$\begin{aligned}
&\boldsymbol{E}\exp\left\{-\frac{E}{N_0}\left[1-\frac{1}{(k+1)^2}\sum_{i=0}^{k}\sum_{j=0}^{k}\rho(U_i,U_j)\right]\right\}\\
&= \exp\left\{-\frac{E}{N_0}\left[1-\frac{1}{k+1}\right]\right\}\cdot\boldsymbol{E}\exp\left\{\frac{E}{N_0(k+1)^2}\sum_{i\neq j}\rho(U_i,U_j)\right\}\\
&\geq \exp\left\{-\frac{E}{N_0}\cdot\frac{k}{k+1}\right\}\cdot\exp\left\{\frac{E}{N_0(k+1)^2}\sum_{i\neq j}\boldsymbol{E}\rho(U_i,U_j)\right\}\\
&= \exp\left\{-\frac{E}{N_0}\cdot\frac{k}{(k+1)}(1-\varrho)\right\}. \qquad (48)
\end{aligned}$$

Note that the expression $E(1-\varrho)$, that appears in the exponent, is equal to $\int_0^T \mathrm{Var}\{x(t,U)\}\mathrm{d}t$, which is a measure of the variability, or the sensitivity of the $x(t,u)$ to the parameter u (in analogy the Cramér–Rao bound that depends on the energy of the derivative of the signal w.r.t. u, as another measure of sensitivity). Accordingly, classes of signals with smaller values of ϱ (or equivalently, higher values of the integrated variance of $x(t,U)$) are expected to yield higher value of $I_G(U;Y)$, and hence smaller estimation error, at least as far as our bounds predict, and since ϱ cannot be negative, the best classes of signals, in this sense, are those for which $\varrho = 0$. Note also that for Jensen's inequality to be reasonably tight, the random variables $\{\rho(U_i,U_j)\}$ should be all close to their expectation ϱ with very high probability, and if this expectation vanishes, as suggested, then $\{\rho(U_i,U_j)\}$ should all be nearly zero with very high probability. We will get back to classes of signals with this desirable rapidly vanishing correlation property later on.

4.3 Estimation Error Bounds for the AWGN Channel

We now equate $R(D)$ to $I_G(U;Y)$ in order to obtain estimation error bounds in the high SNR regime, where the high–resolution expressions of $R(D)$ are relevant. As discussed above, in this regime, we will neglect the effect of the modulo 1 operation in the definition of the distortion measure, and will refer to it hereafter as the ordinary quadratic distortion measure. The choice $k = 1$ yields $I_G(U;Y) \leq -e^{-(1-\varrho)E/(2N_0)}$ (see also [22]), and following eq. (38), this yields

$$\boldsymbol{E}(U-V)^2 \geq D\left(-e^{-(1-\varrho)E/(2N_0)}\right) = \frac{e^{-(1-\varrho)E/N_0}}{16c_1^2}, \qquad (49)$$

[3]For example, if $x(t,u) = x_0(t-u)$ is a rectangular pulse of duration Δ then $\varrho = \Delta/T$.

and so, the exponential decay of the lower bound is according to $e^{-(1-\varrho)E/N_0}$. For $k=2$, according to eq. (41), we have $\log D \approx 2(1-\varrho)E/(3N_0)$, which means an exponential decay according to $e^{-2(1-\varrho)E/(3N_0)}$, which is better. For $k \geq 3$, we use (40) and the resulting bound decays according to $\exp\{-(1-\rho)kE/[(k+1)N_0]\}$, which is better than the result of $k=1$, but not as good as the one of $k=2$. Thus, the best choice of k for the high SNR regime is $k=2$, namely, a generalized Bhattacharyya distance with $k+1=3$ replicas, rather the two replicas of the ordinary Bhattacharyya distance.

Note that since $\varrho \geq 0$, as mentioned earlier, then for any family of signals, the exponential function $e^{-2E/(3N_0)}$ is a *universal* lower bound (at high SNR) in the sense that it applies, not only to every estimator of U, but also to every parametric family of signals $\{x(t,u)\}$, i.e., to every modulation scheme without being dependent on this modulation scheme (see also [22]). This is in contrast to most of the estimation error bounds in the literature. In other words, it sets a fundamental limit on the entire communication system and not only on the receiver end for a given transmitter. Indeed, for some classes of signals, an MSE with exponential decay in E/N_0 is attainable at least in the high SNR regime, although there might be gaps in the actual exponential rates compared to the above mentioned bound. For example, in [15], it is discussed that in the case of time delay estimation ($x(t,u)=x_0(t-u)$), it is possible to achieve an MSE of the exponential order of $e^{-E/(3N_0)}$ by allowing the pulse $s_0(t)$ to have bandwidth that grows exponentially with T.[4] Thus, by improving the lower bound $\exp(-E/N_0)$ (a special case of the above with $k=1$) to $\exp[-2E/(3N_0)]$, we are halving the gap between the exponential rates of the upper bound and the lower bound, from $2E/(3N_0)$ to $E/(3N_0)$.

Our asymptotic lower bound should be compared to other lower bounds available in the literature. One natural candidate would be the Weiss–Weinstein bound (WWB) [18], [19], [20], which for the model under discussion at high SNR, reads [18, p. 66]:

$$\mathrm{WWB} = \sup_{h \neq 0} \frac{h^2 \exp\{-[1-r(h)]E/(2N_0)\}}{2(1-\exp\{-[1-r(2h)]E/(2N_0)\})}, \tag{50}$$

where $r(h)=\rho(u,u+h)=\int_0^T x(t,u)x(t,u+h)\mathrm{d}t/E$ is assumed to depend only on h and not on u. While this is an excellent bound for a given modulation scheme $\{x(t,u),\ u \in \mathcal{U}\}$, it does not seem

[4]Other examples include chirp–like signals, e.g., $x(t,u)=\sin(ue^{Rt})$ (for some given $R>0$), as well as chaotic signals parametrized by their initial condition – see [11], [12] and references therein.

to lend itself easily to the derivation of universal lower bounds, as discussed above. To this end, in principle, the WWB should be minimized over all feasible correlation functions $r(\cdot)$, which is not a trivial task. A reasonable compromise is to first minimize the WWB over $r(\cdot)$ for a given h, and then to maximize the resulting expression over h (i.e., max–min instead of min–max). Since the expression of the bound is a monotonically increasing function of both $r(h)$ and $r(2h)$, and since both $r(h)$ and $r(2h)$ cannot be smaller than -1, we end up with

$$\text{WWB} = \frac{e^{-E/N_0}}{2(1-e^{-E/N_0})} \tag{51}$$

as a modulation–independent bound. This is a faster exponential decay rate (and hence weaker asymptotically) than that of our proposed bound for $k = 2$.

It is possible, however, to obtain a universal lower bound stronger than both bounds by a simple channel–coding argument, which is in the spirit of the Ziv–Zakai bound [23]. This bound is given by (see Appendix D for the derivation):

$$\boldsymbol{E}(U-V)^2 \geq \frac{1}{8M^2} \cdot Q\left(\sqrt{\frac{E}{N_0} \cdot \frac{M}{M-2}}\right), \tag{52}$$

where

$$Q(x) \stackrel{\Delta}{=} \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} \mathrm{d}t \tag{53}$$

and where M is a free parameter, an even integer not smaller than 4, which is subjected to optimization. Throughout the sequel, we refer to this bound as the *channel–coding bound*. In the high SNR regime, the exponential order of the channel–coding bound (for fixed M) is

$$\exp\left\{-\frac{E}{2N_0} \cdot \frac{M}{M-2}\right\}, \tag{54}$$

which for large enough M becomes arbitrarily close to $e^{-E/(2N_0)}$, and hence better than the data–processing bound of $e^{-2E/(3N_0)}$. Note that the Ziv–Zakai bound [23] would be weaker in this context of universal lower bounds, since it is based on binary hypothesis testing ($M/2 = 2$), yielding an exponent of e^{-E/N_0}.

In view of this comparison, it is natural to ask then what is benefit of our data processing lower bound. The answer is that the potential of the data–processing bound is much better exploited in situations of channel uncertainty, like in channels with fading. This is the subject of the next subsection.

4.4 Estimation Error Bounds for the AWGN Channel with Fading

It turns out that the feature that makes the data–processing–theorem approach to error lower bounds more powerful, relatively to other approaches, is the convexity property of the generalized mutual information (in this case, $I_G(U;Y)$) w.r.t. the channel $P_{Y|U}$. Suppose that the channel actually depends on an additional random parameter A (independent of U), that is known to neither the transmitter nor the receiver, namely,

$$P_{Y|U}(y|u) = \int_{-\infty}^{+\infty} \mathrm{d}a \cdot P_A(a) P_{Y|U,A}(y|u,a). \tag{55}$$

where $P_A(a)$ is the density of A. If we think of $I_G(U;Y)$ as a functional of $P_{Y|U}$, denoted $\mathcal{I}(P_{Y|U}(\cdot|u))$, then it is a convex functional, namely,

$$\mathcal{I}(P_{Y|U}(\cdot|u)) = \mathcal{I}\left(\int_{-\infty}^{+\infty} \mathrm{d}a P_A(a) P_{Y|U,A}(\cdot|u,a)\right) \leq \int_{-\infty}^{+\infty} \mathrm{d}a P_A(a) \mathcal{I}(P_{Y|U,A}(\cdot|u,a)). \tag{56}$$

This is a desirable property because the r.h.s. reflects a situation where A is known to both parties, whereas the l.h.s. pertains to the situation where A is unknown, so the lower bound associated with the case where A is unknown is always tighter than the expectation of the lower bound pertaining to a known A. The WWB, on the other hand, does not have this convexity property, as we shall see.

Consider now the case where A is a fading parameter, drawn only once and kept fixed throughout the entire observation time T. More precisely, our model is the same as before except that now the signal is subjected to fading according to

$$y(t) = a \cdot x(t,u) + n(t), \quad 0 \leq t < T, \tag{57}$$

where a and u are realizations of the random variables A and U, respectively. For the sake of convenience in the analysis, we assume that A is a zero–mean Gaussian random variable with variance σ^2 (other densities are, of course, possible too).

We next compare the three corresponding bounds in this case. The overall channel from U to Y is

$$P_{Y|U}(y|u) \propto \int_{-\infty}^{+\infty} \mathrm{d}a \cdot \frac{e^{a^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{N_0}\int_0^T [y(t) - a \cdot x(t,u)]^2 \mathrm{d}t\right\}. \tag{58}$$

Carrying out the integration, we readily obtain

$$P_{Y|U}(y|u) \propto \exp\left\{\theta\left[\int_0^T y(t)x(t,u)\mathrm{d}t\right]^2\right\}, \tag{59}$$

where

$$\theta \stackrel{\Delta}{=} \frac{2\sigma^2}{N_0^2(1+2\sigma^2E/N_0)}. \tag{60}$$

Thus,

$$\begin{aligned} -I_G(U;Y) &= \boldsymbol{E}\left\{\exp\left\{\frac{\theta}{k+1}\sum_{i=1}^{k}\left[\int_0^T y(t)x(t,u_i)\mathrm{d}t\right]^2 - \right.\right. \\ &\qquad \left.\left.\frac{\theta k}{k+1}\left[\int_0^T y(t)x(t,u_0)\mathrm{d}t\right]^2\right\}\Big|U=u_0\right\}. \end{aligned} \tag{61}$$

Upon substituting $y(t) = Ax(t,u_0) + n(t)$, one obtains, after some straightforward algebra

$$\begin{aligned} -I_G(U;Y) &= \boldsymbol{E}\exp\left\{\frac{\theta}{k+1}\left(A^2E^2\sum_{i=1}^{k}\rho^2(U_0,U_i) + 2AE\sum_{i=1}^{k}\rho(U_0,U_i)Z_i + \sum_{i=1}^{k}Z_i^2\right) - \right. \\ &\qquad \left.\frac{\theta k}{k+1}(A^2E^2 + 2AEZ_0 + Z_0^2)\right\}, \end{aligned} \tag{62}$$

where

$$Z_i = \int_0^T n(t)x(t,u_i)\mathrm{d}t, \qquad i = 0,1,2,\ldots,k, \tag{63}$$

and where the expectation is w.r.t. the randomness of A, $\{U_i\}$ and $\{Z_i\}$. Obviously, given A and $\{U_i\}$, the random variables $\{Z_i\}$ are jointly Gaussian with zero–mean with covariances $\frac{N_0}{2}E\rho(U_i,U_j)$. Motivated by the discussion at the end of Subsection 4.2, we now adopt the assumption of signals with rapidly vanishing correlation. In other words, we assume that $\rho(u,u+h)$ vanishes so rapidly[5] as a function of h for every u, that it is safe to neglect $\rho(U_i,U_j)$ altogether for all $i \neq j$. This would make $\{Z_i\}$ independent and simplify the above expression to

$$-I_G(U;Y) = \boldsymbol{E}\left[\exp\left\{-\frac{\theta kE^2A^2}{k+1}\right\}\exp\left\{-\frac{\theta k}{k+1}(Z_0^2+2AEZ_0)\right\}\right]\cdot\left(\boldsymbol{E}\exp\left\{-\frac{\theta Z_1^2}{k+1}\right\}\right)^k \tag{64}$$

Upon calculating the expectation (w.r.t. both A and $\{Z_i\}$), we obtain

$$\begin{aligned} -I_G(U;Y) &= \left[\frac{(k+1)(1+2\sigma^2E/N_0)}{k+1+2k\sigma^2E/N_0}\right]^{k/2}\times \\ &\qquad \sqrt{\frac{(k+1)(1+2\sigma^2E/N_0)}{(k+1)(1+2\sigma^2E/N_0)+2k\sigma^2E/N_0}}\cdot\frac{1}{\sqrt{1+2\mu\sigma^2}}, \end{aligned} \tag{65}$$

[5]Consider an asymptotic regime under which, the signal $x(t,u)$ depends on an additional (design) parameter Δ, so that for every $h \neq 0$, $\rho(h) \to 0$ as Δ tends to a certain limit, and that this limit is taken *before* the limit $E/N_0 \to \infty$. For example, if $x(t,u) = x_0(t-u)$ is a rectangular pulse of amplitude $\sqrt{E/\Delta}$ and duration Δ, then $\rho(h) = [1-|h|/\Delta]_+$ which obviously vanishes as $\Delta \to 0$ for every $h \neq 0$.

where

$$\mu \triangleq \frac{2k\sigma^2(E/N_0)^2}{2(2k+1)\sigma^2 E/N_0 + k + 1}. \tag{66}$$

Considering the high–SNR regime ($E/N_0 \gg 1$), this is approximated as

$$-I_G(U;Y) \approx \frac{1}{\sqrt{2}}\left(1+\frac{1}{k}\right)^{(k+1)/2} \cdot \frac{1}{\sigma\sqrt{E/N_0}} \triangleq \frac{f_k}{\sigma\sqrt{E/N_0}}. \tag{67}$$

Applying the high–resolution approximation of $D(R)$ for $k \geq 3$, we get:

$$\boldsymbol{E}(U-V)^2 \geq \frac{g_k}{\sigma} \cdot \sqrt{\frac{N_0}{E}}, \tag{68}$$

where

$$g_k = \frac{1}{4\sqrt{2}}\left(1-\frac{2}{k}\right)^k \left(1+\frac{1}{k}\right)^{(k+1)/2}. \tag{69}$$

A simple numerical study indicates that $\{g_k\}$ is monotonically increasing and so the best bound is obtained for $k \to \infty$ (infinitely many replicas), where the constant is:

$$g_\infty = \lim_{k\to\infty} g_k = \frac{1}{4\sqrt{2}} \cdot e^{-2} \cdot \sqrt{e} = \frac{1}{4\sqrt{2}e^{3/2}} = 0.03944. \tag{70}$$

Thus, our asymptotic lower bound for high SNR is

$$\liminf_{E/N_0\to\infty} \sqrt{\frac{E}{N_0}} \cdot \boldsymbol{E}(U-V)^2 \geq \frac{0.03944}{\sigma}. \tag{71}$$

The WWB [18, p. 51], in its more general form, is given by

$$\text{WWB} = \sup_{h\neq 0,\ s\in[0,1]} \frac{h^2 e^{2\mu(s,h)}}{e^{\mu(2s,h)} + e^{\mu(2-2s,-h)} - 2e^{\mu(s,2h)}}, \tag{72}$$

where

$$e^{\mu(s,h)} = \boldsymbol{E}\left[\frac{P_{Y|U}(Y|U+h)}{P_{Y|U}(Y|U)}\right]^s, \qquad s \in [0,1] \tag{73}$$

which for the fading channel under the high SNR regime of rapidly vanishing correlation signals, can be shown (using similar calculations as above) to be given by

$$e^{\mu(s,h)} \approx \begin{cases} \sqrt{\frac{1+2\sigma^2 E/N_0}{(1+2s\sigma^2 E/N_0)(1+2[1-s]\sigma^2 E/N_0)}} & h \neq 0 \\ 1 & h = 0 \end{cases} \tag{74}$$

The problem is that, unless $s = 1/2$, either $2s > 1$ or $2-2s > 1$, and so correspondingly, for large enough values of E/N_0, either $e^{\mu(2s,h)}$ or $e^{\mu(2-2s,-h)}$ at the denominator diverge, and the WWB becomes useless. Thus, the only feasible choice of s is $s = 1/2$, in which case, the WWB becomes

$$\text{WWB} = \sup_{h\neq 0} \frac{h^2 e^{2\mu(1/2,h)}}{2[1-e^{\mu(1/2,2h)}]}. \tag{75}$$

But $e^{\mu(1/2,h)}$ is exactly our information measure for $k=1$, and so,

$$\text{WWB} = \frac{f_1^2 N_0/(\sigma^2 E)}{2[1 - f_1/(\sigma\sqrt{E/N_0})]}. \tag{76}$$

As can be seen, the WWB decays according to $(E/N_0)^{-1}$ rather than $(E/N_0)^{-1/2}$ and hence inferior to the data processing bound.

The channel–coding bound is based on a universal lower bound on the probability of error, which holds for every signal set. The problem is that under fading, we are not aware of such a universal lower bound. The only remaining alternative then is to use a lower bound corresponding to the case where A is known to the receiver, and then to take the expectation w.r.t. A, although one might argue that this comparison is not quite fair. Nonetheless, the derivation of this appears in Appendix E and the result is

$$\liminf_{E/N_0\to\infty} \sqrt{\frac{E}{N_0}} \cdot \boldsymbol{E}(U-V)^2 \geq \frac{1}{128\pi\sqrt{2}\sigma} = \frac{0.001758}{\sigma}. \tag{77}$$

Thus, the data processing bound is better by a factor of 22.4 (13.5dB).

Yet another comparison, perhaps more fair, can be made with a related bound, which based on binary hypothesis testing, but has the advantage of avoiding the use of the Chebychev inequality, that was used in the channel–coding bound. This is the Chazan–Zakai–Ziv bound (CZZB), an improved version of the Ziv–Zakai bound [23]. According to the CZZB, applied to our problem (see Appendix F for the derivation),

$$\liminf_{E/N_0\to\infty} \sqrt{\frac{E}{N_0}} \cdot \boldsymbol{E}(U-V)^2 \geq \frac{0.00716}{\sigma}, \tag{78}$$

which is again significantly smaller than our bound. Thus, we observe that while the WWB and the CZZB are excellent bounds for ordinary channels without fading, when it comes to channels with fading, the proposed data–processing bound has an advantage.

5 Conclusion

In this work, we have explored a certain class of information measures [10], which although being a special case of the Zakai–Ziv information measures [22], it has an interesting structure that calls for attention. We first put this class of information measures in the broader perspective, relating

it to other information measures, like those of [22], and then, by a specific choice of the convex functions, we defined a generalized notion of the Chernoff divergence that is based on an arbitrary number of replicas of the channel. Relations have be drawn between the generalized Chernoff divergence and the Gallager function, the ordinary Chernoff divergence, and even more specifically, the Bhattacharyya distance. We have also suggested a somewhat more general structured class based on factor trees. We then applied the data processing inequality, based on the generalized Chernoff divergence, and demonstrated that sometimes bounds can be improved by using more than $k+1=2$ replicas. In particular, for the AWGN three replicas is the optimum number in the AWGN model, thus improving on [22], where only two replicas were used (the ordinary Bhattacharyya distance). While this bound still falls short compared to other bounds available from estimation theory, the data processing bound seems to be more powerful than others when it comes to channels with uncertainty, like fading channels. In this case, the limit of $k \to \infty$ gives the best result.

Acknowledgment

Interesting discussions with Shlomo Shamai are acknowledged with thanks.

Appendix A

Low Resolution Analysis

Low resolution analysis corresponds to very small values of s, which can be handled by a first order Taylor series expansion of the functions $F(s)$, $C(s)$ and $G(s)$. Specifically,

$$C(s) \approx 1 + \frac{k+1}{12k} \cdot s \tag{A.1}$$

$$F(s) \approx \frac{1}{12} - \frac{k+1}{80k} \cdot s \tag{A.2}$$

$$G^{k+1}(s) \approx 1 - \frac{k+1}{12k} \cdot s. \tag{A.3}$$

Thus,

$$D_s = C(s)F(s) \approx \frac{1}{12} - \frac{k+1}{180k} \cdot s \tag{A.4}$$

and

$$-R(D_s) = C(s)[G(s)]^{k+1} \approx 1 - \left(\frac{k+1}{12k}\right)^2 \cdot s^2 \tag{A.5}$$

or

$$s \approx \frac{12k}{k+1}\sqrt{R+1}. \tag{A.6}$$

and so

$$D(R) \approx \frac{1}{12} - \frac{1}{15}\sqrt{R+1}. \tag{A.7}$$

Appendix B

High Resolution Analysis

High resolution corresponds to $s \gg 1$. In this case, we have

$$\begin{aligned}
\frac{1}{C(s)} &= \int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{(1+sw^2)^{1+1/k}} \\
&= \frac{1}{\sqrt{s}} \int_{-\sqrt{s}/2}^{+\sqrt{s}/2} \frac{\mathrm{d}(\sqrt{s}w)}{(1+(\sqrt{s}w)^2)^{1+1/k}} \\
&\approx \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} \frac{\mathrm{d}t}{(1+t^2)^{1+1/k}} \\
&\stackrel{\Delta}{=} \frac{c_k}{\sqrt{s}},
\end{aligned} \tag{B.1}$$

Now, according to the relations between the functions C, F and G, derived in Subsection 4.1, we have:

$$\begin{aligned}
G(s) &= \frac{1}{(1+s/4)^{1/k}} + \frac{2s}{k}F(s) \\
&\approx \frac{4^{1/k}}{s^{1/k}} + \frac{2s}{k}F(s)
\end{aligned} \tag{B.2}$$

and also

$$G(s) = \frac{1}{C(s)} + sF(s) \approx \frac{c_k}{\sqrt{s}} + sF(s). \tag{B.3}$$

Comparing the two expressions of $G(s)$, we get

$$\frac{c_k}{\sqrt{s}} + sF(s) \approx \frac{4^{1/k}}{s^{1/k}} + \frac{2s}{k}F(s), \tag{B.4}$$

which leads to the equation

$$sF(s)\left(1 - \frac{2}{k}\right) \approx \frac{4^{1/k}}{s^{1/k}} - \frac{c_k}{\sqrt{s}}. \tag{B.5}$$

At this stage, we have to handle separately the cases $k = 1$, $k = 2$ and $k > 2$.

Let us consider the case $k = 1$ first. In this case, the last equation reads

$$-sF(s) \approx \frac{4}{s} - \frac{c_1}{\sqrt{s}} \approx -\frac{c_1}{\sqrt{s}} \tag{B.6}$$

and so,

$$F(s) \approx \frac{c_1}{s^{3/2}}. \tag{B.7}$$

Thus, from the distortion equation,

$$D_s = C(s)F(s) \approx \frac{\sqrt{s}}{c_1} \cdot \frac{c_1}{s^{3/2}} = \frac{1}{s}, \tag{B.8}$$

or equivalently, $s = 1/D_s$. Now,

$$G(s) = \frac{1}{C(s)} + sF(s) \approx \frac{c_1}{\sqrt{s}} + s \cdot \frac{c_1}{s^{3/2}} = \frac{2c_1}{\sqrt{s}}. \tag{B.9}$$

From the rate equation, we have

$$\begin{aligned} -R(D_s) &= C(s)[G(s)]^2 \\ &\approx \frac{\sqrt{s}}{c_1} \cdot \frac{4c_1^2}{s} \\ &= \frac{4c_1}{\sqrt{s}} = 4c_1\sqrt{D_s}, \end{aligned} \tag{B.10}$$

which means

$$R(D) \approx -4c_1\sqrt{D}. \tag{B.11}$$

or equivalently, the distortion–rate function is

$$D(R) \approx \frac{R^2}{16c_1^2}, \tag{B.12}$$

where it should be kept in mind that R takes on values in the range $[-1, 0]$ in this case.

The case $k = 2$ is handled as follows:

$$G(s) = \int_{-1/2}^{+1/2} \frac{\mathrm{d}w}{\sqrt{1 + sw^2}} = \frac{1}{\sqrt{s}} \ln \frac{\sqrt{s/4 + 1} + \sqrt{s}/2}{\sqrt{s/4 + 1} - \sqrt{s}/2} = \frac{1}{\sqrt{s}} \ln \left(1 + \frac{s}{2} + \sqrt{s\left(\frac{s}{4} + 1\right)}\right), \tag{B.13}$$

and so, for s large $G(s) \approx (\ln s)/\sqrt{s}$. By comparing the two expressions for $G(s)$, we find that $C(s) = \sqrt{1 + s/4} \approx \sqrt{s}/2$. Consequently,

$$F(s) = \frac{G(s) - 1/C(s)}{s} \approx \frac{\ln s}{s^{3/2}}. \tag{B.14}$$

Thus, $D_s = F(s)C(s) \approx (\ln s)/(2s)$ and $-R(D_s) = C(s)G^3(s) \approx (\ln^3 s)/(2s)$. In the high–resolution limit, the logarithmic terms are relatively negligible and so, we can deduce that

$$\lim_{s\to\infty} \frac{\log D_s}{\log[-R(D_s)]} = \lim_{D\to 0} \frac{\log D}{\log[-R(D)]} = 1. \tag{B.15}$$

Finally, we examine the case $k > 2$. Returning to eq. (B.5), now we have:

$$sF(s)\left(1 - \frac{2}{k}\right) \approx \frac{4^{1/k}}{s^{1/k}}, \tag{B.16}$$

and so

$$F(s) \approx \frac{k4^{1/k}}{(k-2)s^{1+1/k}}. \tag{B.17}$$

and

$$G(s) \approx \frac{c_k}{\sqrt{s}} + \frac{k4^{1/k}}{(k-2)s^{1/k}} \approx \frac{k4^{1/k}}{(k-2)s^{1/k}}. \tag{B.18}$$

The distortion equation then gives

$$\begin{aligned} D_s &= C(s)F(s) = \frac{\sqrt{s}}{c_k} \cdot \frac{k4^{1/k}}{(k-2)s^{1+1/k}} \\ &= \frac{k4^{1/k}}{(k-2)c_k s^{1/2+1/k}} \end{aligned} \tag{B.19}$$

and the rate equation yields

$$\begin{aligned} -R(D_s) &= C(s)[G(s)]^{k+1} \\ &\approx \frac{\sqrt{s}}{c_k} \cdot \left[\frac{k4^{1/k}}{(k-2)s^{1/k}}\right]^{k+1} \\ &= \frac{4^{1+1/k}}{c_k s^{1/2+1/k}} \cdot \left(\frac{k}{k-2}\right)^{k+1} \\ &= 4\left(\frac{k}{k-2}\right)^{k} \cdot D_s, \end{aligned} \tag{B.20}$$

Thus, the rate–distortion function and the distortion–rate function are approximated as

$$R(D) \approx -4\left(\frac{k}{k-2}\right)^{k} \cdot D; \qquad D(R) \approx -\frac{1}{4}\left(1 - \frac{2}{k}\right)^{k} \cdot R, \tag{B.21}$$

Appendix C

Higher Order Moments

The high–resolution analysis can easily be extended to handle general moments of the estimation error, $\boldsymbol{E}|U-V|^p$, $p>0$ (p should not necessarily be integer). This gives for large s,

$$C(s) \approx \frac{s^{1/p}}{c}; \quad c \stackrel{\Delta}{=} \int_{-1/2}^{+1/2} \frac{\mathrm{d}t}{[1+|t|^p]^{1+1/k}} \tag{C.1}$$

and

$$\left(1-\frac{p}{k}\right) sF(s) \approx \frac{2^{p/k}}{s^{1/k}} - \frac{c}{s^{1/p}}. \tag{C.2}$$

Here, we have to handle separately the cases $k<p$ and $k>p$ (and the case $k=p$ will not be covered here, but since p is allowed to be non–integer, it can be approached by either $p \downarrow k$ or $p \uparrow k$). In the case $k<p$, we have

$$\left(\frac{p}{k}-1\right) sF(s) \approx \frac{c}{s^{1/p}} \tag{C.3}$$

and so

$$F(s) \approx \frac{kc}{p-k} \cdot \frac{1}{s^{1+1/p}}. \tag{C.4}$$

Thus,

$$D_s = C(s)F(s) = \frac{k}{(p-k)s}. \tag{C.5}$$

Now,

$$G(s) = \frac{1}{C(s)} + sF(s) \approx \frac{c}{s^{1/p}} + \frac{kc}{(p-k)s^{1/p}} = \frac{pc}{(p-k)s^{1/p}} \tag{C.6}$$

and so

$$-R(D_s) = C(s)[G(s)]^{k+1} \approx c^k \left(\frac{p}{p-k}\right)^{k+1} \cdot \frac{1}{s^{k/p}}. \tag{C.7}$$

Thus,

$$D(R) \approx S_1(k,p)[-R]^{p/k}. \tag{C.8}$$

where

$$S_1(k,p) = \frac{k}{c^p(p-k)} \cdot \left(1-\frac{k}{p}\right)^{p(1+1/k)}. \tag{C.9}$$

Note that in terms of the asymptotic behavior for small values of $-R$, the best choice of k is the largest integer strictly less than p. For p integer, this means $k=p-1$. As for the case $k>p$, we

get:

$$\left(1-\frac{p}{k}\right) sF(s) \approx \frac{2^{p/k}}{s^{1/k}} \tag{C.10}$$

or

$$F(s) \approx \frac{k2^{p/k}}{(k-p)s^{1+1/k}}. \tag{C.11}$$

So

$$D_s = C(s)F(s) \approx \frac{k2^{p/k}}{(k-p)cs^{1+1/k-1/p}} \tag{C.12}$$

Here,

$$G(s) = \frac{c}{s^{1/p}} + \frac{k2^{p/k}}{(k-p)s^{1/k}} \approx ck2^{p/k}(k-p)s^{1/k}. \tag{C.13}$$

Then,

$$-R(D_s) = C(s)[G(s)]^{k+1} \approx 2^p \left(\frac{k}{k-p}\right)^k D_s, \tag{C.14}$$

and we get

$$D(R) \approx -S_2(k,p)R, \tag{C.15}$$

where

$$S_2(k,p) = 2^{-p}\left(1-\frac{p}{k}\right)^k. \tag{C.16}$$

Appendix D

Derivation of the Channel–Coding Bound

For a given positive integer M, consider the following chain of inequalities:

$$\begin{aligned}
& \boldsymbol{E}(U-V)^2 \\
\geq\ & \left(\frac{1}{2M}\right)^2 \Pr\left\{|U-V| \geq \frac{1}{2M}\right\} \\
=\ & \frac{1}{4M^2} \cdot \int_{-1/2}^{+1/2} \mathrm{d}u \cdot \Pr\left\{|U-V| \geq \frac{1}{2M}\middle| U=u\right\} \\
=\ & \frac{1}{4M^2} \cdot \sum_{i=0}^{M-1} \int_{-1/(2M)}^{+1/(2M)} \mathrm{d}u \cdot \Pr\left\{|U-V| \geq \frac{1}{2M}\middle| U=\frac{2i+1}{2M}-\frac{1}{2}+u\right\} \\
=\ & \frac{1}{4M} \cdot \int_{-1/(2M)}^{+1/(2M)} \mathrm{d}u \cdot \frac{1}{M}\sum_{i=0}^{M-1} \Pr\left\{|U-V| \geq \frac{1}{2M}\middle| U=\frac{2i+1}{2M}-\frac{1}{2}+u\right\}.
\end{aligned} \tag{D.1}$$

Now, note that the integrand of the last expression has a simple interpretation: Consider the codebook of signals $\{x(t,u_i)\}$, $0 \leq t < T$, $i = 0, 1, \ldots, M-1$ where $u_i = (2i+1)/(2M) - 1/2 + u$,

and consider the (suboptimum) decoder that first estimates U by an arbitrary estimator V and then decodes the message according to the u_i that is nearest to V. The integrand in the last line above is simply the probability of error of that decoder. This probability of error is lower bounded [17, p. 174, eqs. (3.73) and (3.75)] according to

$$\begin{aligned} & \frac{1}{M}\sum_{i=0}^{M-1}\Pr\left\{|U-V|\geq\frac{1}{2M}\middle|U=\frac{2i+1}{2M}-\frac{1}{2}+u\right\} \\ \geq\ & \frac{1}{2}Q\left(\sqrt{\frac{E}{N_0}\cdot\frac{M/2}{M/2-1}}\right) \\ =\ & \frac{1}{2}Q\left(\sqrt{\frac{E}{N_0}\cdot\frac{M}{M-2}}\right), \end{aligned} \tag{D.2}$$

where now $M/2$ should be an integer at least as large as 2, namely, $M = 4, 6, 8, \ldots$, Thus,

$$\boldsymbol{E}(U-V)^2 \geq \frac{1}{4M}\cdot\int_{-1/(2M}^{+1/(2M)}\mathrm{d}u\frac{1}{2}\cdot Q\left(\sqrt{\frac{E}{N_0}\cdot\frac{M}{M-2}}\right) = \frac{1}{8M^2}\cdot Q\left(\sqrt{\frac{E}{N_0}\cdot\frac{M}{M-2}}\right). \tag{D.3}$$

Appendix E

Channel–Coding Bound for the AWGN Fading Channel

For a given value of the fading parameter $A = a$, the earlier derivation of the channel–coding bound implies

$$\boldsymbol{E}(U-V)^2 \geq \frac{1}{8M^2}\cdot Q\left(\sqrt{\frac{a^2E}{N_0}\cdot\frac{M}{M-2}}\right). \tag{E.1}$$

Averaging over A and using Craig's formula (see, e.g., [16]), we have

$$\begin{aligned}
\boldsymbol{E}(U-V)^2 &\geq \frac{1}{8M^2}\int_{-\infty}^{+\infty}\mathrm{d}a\frac{e^{-a^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}Q\left(\sqrt{a^2\frac{E}{N_0}\cdot\frac{M}{M-2}}\right)\\
&= \frac{1}{8\pi M^2}\int_{-\infty}^{+\infty}\mathrm{d}a\frac{e^{-a^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\int_0^{\pi}\mathrm{d}\theta\cdot\exp\left\{-\frac{a^2EM}{2(M-2)N_0\sin^2\theta}\right\}\\
&= \frac{1}{8\pi M^2}\int_0^{\pi}\mathrm{d}\theta\int_{-\infty}^{+\infty}\mathrm{d}a\frac{e^{-a^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\cdot\exp\left\{-\frac{a^2EM}{2(M-2)N_0\sin^2\theta}\right\}\\
&= \frac{1}{8\pi M^2}\int_0^{\pi}\frac{\mathrm{d}\theta}{\sqrt{1+\sigma^2EM/[(M-2)N_0\sin^2\theta]}}\\
&= \frac{1}{8\pi M^2}\int_0^{\pi}\frac{\mathrm{d}\theta\sin\theta}{\sqrt{\sin^2\theta+E\sigma^2M/[N_0(M-2)]}}\\
&> \frac{1}{8\pi M^2}\int_0^{\pi}\frac{\mathrm{d}\theta\sin\theta}{\sqrt{1+\sigma^2EM/[N_0(M-2)]}}\\
&= \frac{1}{8\pi M^2\sqrt{1+\sigma^2EM/[N_0(M-2)]}}.
\end{aligned}\tag{E.2}$$

For E/N_0 large, this is approximately,

$$\frac{1}{8\pi\sigma\sqrt{E/N_0}}\sqrt{\frac{M-2}{M^5}},$$

which is maximized (for even $M>2$) by $M=4$ to yield

$$\liminf_{E/N_0\to\infty}\sqrt{\frac{E}{N_0}}\cdot\boldsymbol{E}(U-V)^2\geq\frac{1}{128\pi\sqrt{2}\sigma}=\frac{0.001758}{\sigma}.\tag{E.3}$$

Appendix F

Derivation of the Chazan–Zakai–Ziv Bound

The CZZB [4] asserts that

$$\boldsymbol{E}(U-V)^2\geq\int_0^1\mathrm{d}h\cdot h\int_{-1/2}^{1/2-h}\mathrm{d}u\cdot P_e(u,u+h),\tag{F.1}$$

where $P_e(u,u+h)$ is the probability of error associated with optimum hypothesis testing between the hypotheses $y(t)=Ax(t,u)+n(t)$ and $y(t)=Ax(t,u+h)+n(t)$, assuming equal priors. Let us denote the probabilities of error of the two kinds by $P_e(u\to u+h)$ and $P_e(u+h\to h)$. Then, according to the Shannon–Gallager–Berlekamp theorem [17, p. 159, Theorem 3.5.1], for every

$s \in [0,1]$, at least one of the two following inequalities must hold:

$$P_e(u \to u+h) > \frac{1}{4}\exp[\mu(s,h) - s\mu'(s,h) - s\sqrt{2\mu''(s,h)}] \stackrel{\Delta}{=} A(s) \quad \text{(F.2)}$$

$$P_e(u+h \to u) > \frac{1}{4}\exp[\mu(s,h) + (1-s)\mu'(s,h) - (1-s)\sqrt{2\mu''(s,h)}] \stackrel{\Delta}{=} B(s) \quad \text{(F.3)}$$

where $\mu'(s,h)$ and $\mu''(s,h)$ denote the first two partial derivatives of $\mu(s,h)$ w.r.t s, and where for rapidly–vanishing–correlation signals, $\mu(s,h)$ is given by the (first line of) eq. (74). Since $\mu(1/2,h) = \ln[f_1/(\sigma\sqrt{E/N_0})$, $\mu'(1/2,h) = 0$ and $\mu''(1/2) \approx 1/4$ at the high SNR limit, this implies that

$$\begin{aligned} P_e(u,u+h) &= \frac{P_e(u \to u+h) + P_e(u+h \to u)}{2} \\ &> \sup_{0 \le s \le 1} \frac{1}{2}\min\{A(s), B(s)\} \\ &\ge \frac{1}{2}\min\{A(1/2), B(1/2)\} \\ &= \frac{1}{8}\exp\{\mu(1/2,h) - 0.5 \cdot \sqrt{2\mu''(1/2,h)}\} \\ &\approx \frac{1}{8e^{\sqrt{2}}} \cdot \frac{f_1}{\sigma\sqrt{E/N_0}} \\ &= \frac{0.042977}{\sigma\sqrt{E/N_0}}, \end{aligned} \quad \text{(F.4)}$$

and so,

$$\boldsymbol{E}(U-V)^2 \ge \frac{0.042977}{\sigma\sqrt{E/N_0}} \int_0^1 h(1-h)\mathrm{d}h = \frac{0.00716}{\sigma\sqrt{E/N_0}}, \quad \text{(F.5)}$$

References

[1] D. Andelman, *Bounds According to a Generalized Data Processing Theorem*, M.Sc. final paper (in Hebrew), Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel, October 1974.

[2] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels", *IEEE Transactions on Information Theory*, pp. 357–359, May 1973.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[4] 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.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Second Edition, Hoboken NJ, USA, 2006.

[6] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 95–108, 1963.

[7] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2 (1–4), pp. 191–213, 1972.

[8] I. Csiszár and P. Shields, "Information theory and statistics: a tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 417–528, 2004.

[9] R. G. Gallager, *Information Theory and Reliable Communication*, J. Wiley & Sons, 1968.

[10] I. Gurantz, *Application of a Generalized Data Processing Theorem*, M.Sc. final paper (in Hebrew), Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel, August 1974.

[11] I. Hen, *The Threshold Effect in the Estimation of Chaotic Sequences*, M.Sc. dissertation, Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel, February 2002.

[12] I. Hen and N. Merhav, "On the threshold effect in the estimation of chaotic sequences," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2894–2904, November 2004.

[13] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEÜ*, vol. 47, no. 4, pp. 228–239, 1993.

[14] N. Merhav, "Data processing theorems and the second law of thermodynamics," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 4926–4939, August 2011.

[15] N. Merhav, "Threshold effects in parameter estimation as phase transitions in statistical mechanics," to appear in *IEEE Trans. Inform. Theory*, October 2011.

[16] C. Tellambura and A. Annamalai, "Derivation of Craig's formua for Gaussian probability function," *Electronic Letters*, vol. 35, no. 17, pp. 1424–1425, August 19, 1999.

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

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

[19] A. J. Weiss and E. Weinstein, "A lower bound on the mean square error in random parameter estimation," *IEEE Transactions on Information Theory*, vol. IT–31, no. 5, pp. 680–682, September 1985.

[20] E. Weinstein and A. J. Weiss, "Lower bounds on the mean square estimation error," *Proc. IEEE*, vol. 73, no. 9, pp. 1433–1434, September 1985.

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

[22] 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.

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

[24] 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.