Information-Theoretic Applications of the Logarithmic Probability Comparison Bound

Rami Atar[*] Neri Merhav[†]

Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 3200003, ISRAEL
{atar, merhav}@ee.technion.ac.il

June 4, 2015

Abstract

A well-known technique in estimating probabilities of rare events in general and in information theory in particular (used, e.g., in the sphere–packing bound), is that of finding a reference probability measure under which the event of interest has probability of order one and estimating the probability in question by means of the Kullback-Leibler divergence. A method has recently been proposed in [2], that can be viewed as an extension of this idea in which the probability under the reference measure may itself be decaying exponentially, and the Rényi divergence is used instead. The purpose of this paper is to demonstrate the usefulness of this approach in various information–theoretic settings. For the problem of channel coding, we provide a general methodology for obtaining matched, mismatched and robust error exponent bounds, as well as new results in a variety of particular channel models. Other applications we address include rate-distortion coding and the problem of guessing.

Index Terms: change-of-measure, error exponent, mismatch, Rényi divergence.

1 Introduction

A key approach to obtaining lower bounds on probabilities of rare events is based on the idea of a change of measure. In this approach, the underlying probability measure is replaced by a reference probability measure under which the probability of the event in question does not decay exponentially, and the exponent of the bound is given by the Kullback–Leibler (KL) divergence between the two probability measures. One then optimizes the estimate over all reference measures having the property alluded to above. This idea is standard for deriving lower bounds in large deviations theory (see, e.g., [5, p. 32]), where it is sometimes referred to as tilting. In the context of information theory it has been used in applications including (i)

[*] Research supported in part by the Israel Science Foundation (ISF) Grant 1315/12.

[†] Research supported in part by the Israel Science Foundation (ISF) Grant 412/12.

the derivation of the sphere–packing bound for discrete memoryless channels (DMC's), using Csiszár and Körner's method [4, Theorem 5.3]; (ii) Marton's converse theorem on the source coding exponent [13]. In the former, the resulting exponential error bound is tight at least in some range of high coding rates. In the latter, it is virtually always tight (for finite–alphabet memoryless sources), as there exists a matching upper bound.

In [2], Atar, Chowdhary and Dupuis presented what may be viewed as an extension of this approach to situations where the probability of the event of interest may also decay exponentially under the reference measure. The estimate is then given in terms of the corresponding Rényi divergence. At the heart of the approach lies the *logarithmic probability comparison bounds* (LPCB) that compare the probability of an event under two measures at a logarithmic scale in terms of the respective Rényi divergence. Specifically, if P and Q are probability measures on a measurable space and $\mathcal{A}$ is an event on it then

$$\frac{1}{\alpha-1}\ln P(\mathcal{A}) \le \frac{1}{\alpha}\ln Q(\mathcal{A}) + D_\alpha(P\|Q) \tag{1}$$

for $\alpha > 1$, where D_α denotes Rényi divergence of order α (see definition and details in Section 2). This bound is tight in the sense that, given P and $\mathcal{A}$, one can find Q for which it holds as equality. Thus if $\{P_n\}$ and $\{Q_n\}$ are sequences of probability measures and we denote by $E_P = -\limsup_{n\to\infty} n^{-1}\ln P_n(\mathcal{A})$ the exponential decay rate of the probability under P_n and by E_Q that under Q_n, then with $\bar{D}_\alpha(P\|Q) = \limsup_{n\to\infty} n^{-1}D_\alpha(P_n\|Q_n)$, one obtains

$$E_P \ge \frac{\alpha-1}{\alpha}E_Q - (\alpha-1)\bar{D}_\alpha(P\|Q), \qquad \alpha > 1. \tag{2}$$

This gives a lower bound on the decay rate E_P under a sequence of measures of interest in terms of that under reference measures, E_Q. By switching the roles of $\{P_n\}$ and $\{Q_n\}$ one obtains an analogous upper bound. One natural use of (2) is when Q is a model for which we have information on the decay rate (exactly or as a bound), whereas P is harder to analyze. In this case, a key step is to provide a useful estimate of the divergence term $\bar{D}_\alpha(P\|Q)$. Another way to view (2) is as what is often called a *robust bound*, where one attempts to obtain performance bounds on a whole family of true models P, and Q serves in defining this family. For example, the family of true models might consist of all P for which the divergence from Q does not exceed a certain bound, in the sense that $\bar{D}_\alpha(P\|Q) \le \epsilon(\alpha)$, some $\epsilon(\cdot)$. Then it is immediate from (2) that for all P in the family,

$$E_P \ge \frac{\alpha-1}{\alpha}E_Q - (\alpha-1)\epsilon(\alpha). \tag{3}$$

While the latter has been the main motivation in [2], both viewpoints will be addressed in this paper. Some benefits of the approach include: (i) the ability to compare, not only probabilities of a given event, but also expectations of a given function under the two (sequences of) measures (this relies on a more general inequality than (1); see Section 2), (ii) the presence of the free parameter α, that can be optimized over in order to tighten the bound, and (iii) the possibility to derive both upper and lower bounds by the same method.

The objective of this paper is to present an approach based on the LPCB and to demonstrate its usefulness for problems in Information Theory, as a tool for deriving upper and lower bounds in a variety of applications, including both source coding and channel coding scenarios. Because it compares two probability measures, the bound is especially natural to apply in situations of mismatch between the true underlying model and the one to which the coding–decoding schemes are tailored. Also, as will be seen in the sequel, in most of these applications, the setting is sufficiently general that no alternative bounds are available to the best knowledge of the authors, such as, for example, coding for channels with additive interference of unlimited memory and mismatch. In some of these scenarios, the exponential bounds obtained are tight in the sense that they are attained at least for some instance of the problem.

Our main contributions are summarized as follows.

- Highlighting the relevance of the approach to information theory;
- Developing general upper and lower bounds on channel coding error exponents for the matched, mismatched and robust settings;
- Using the approach to derive new bounds on error exponents for a host of particular channel models including Gaussian channels with long memory interference, the intersymbol interference channel, the fading channel, and the binary erasure channel;
- Obtaining new bounds for source coding and the problem of guessing.

The outline of the remaining part of this paper is as follows. In Section 2, we present the LPCB. In Section 3, we explain its use in estimating probabilities of rare events. We also present a corollary regarding small perturbations between reference and true models. Section 4 is devoted to the channel coding framework. Finally, Section 5 provides further application examples.

Notation. A vector (deterministic or random) of the form $(x_1, x_2, \ldots, x_n)$ will be written as x^n. When the dimension n is understood from the context, the vector will sometimes be written as the corresponding bold font letter, $\boldsymbol{x}$. The probability law of a random variable X under a probability measure P is denoted by P_X, and the conditional law of Y given X under P by $P_{Y|X}(\cdot|\cdot)$. When there is no room for ambiguity, these subscripts will be omitted. Expectation with respect to a probability measure P will be denoted by $\boldsymbol{E}_P\{\cdot\}$. Again, the subscript will be omitted if the underlying probability distribution is clear from the context. The entropy of a distribution Q will be denoted by $H(Q)$.

2 Rényi divergence and the LPCB

Let a measurable space $(\mathcal{S}, \mathcal{F})$ be given, and denote by $\mathcal{P}$ the set of all probability measures on it. For $\alpha > 1$ and $P, Q \in \mathcal{P}$, the Rényi divergence of degree α of Q from P is defined by[1]

$$D_\alpha(Q\|P) = \begin{cases} \dfrac{1}{\alpha(\alpha-1)} \ln \displaystyle\int \left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^\alpha \mathrm{d}P & \text{if } Q \ll P, \\ +\infty & \text{otherwise,} \end{cases} \tag{4}$$

[1] Some authors use the factor $\frac{1}{\alpha-1}$ rather than $\frac{1}{\alpha(\alpha-1)}$. By choosing the latter we follow the notation used in [12].

where $Q \ll P$ denotes absolute continuity of Q with respect to P, and $\frac{\mathrm{d}Q}{\mathrm{d}P}$ denotes the Radon-Nikodym derivative. For $\alpha = 1$ one extends this definition by letting $D_1 = D$ be the KL divergence, namely

$$D(Q\|P) = \begin{cases} \int \left(\ln \frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q & \text{if } Q \ll P, \\ +\infty & \text{otherwise.} \end{cases} \tag{5}$$

For Q and P fixed, $\alpha \mapsto \alpha D_\alpha(Q\|P)$ is nondecreasing as a map from $[1, \infty)$ to $[0, \infty]$. Moreover, if $\bar{\alpha} := \sup\{\alpha : D_\alpha(Q\|P) < \infty\}$ and $\bar{\alpha} > 1$ then $D_\alpha(Q\|P)$ is finite and continuous on $[1, \bar{\alpha})$. For extension to $\alpha \in \mathbb{R}$ and many other useful properties of the divergence, see [9], [12], [18] and [19].

The well-known convex duality between exponential integrals and KL divergence [6] states that for any bounded measurable function $g : \mathcal{S} \to \mathbb{R}$, and every $Q \in \mathcal{P}$,

$$\ln \int e^g \mathrm{d}Q = \sup_{P \in \mathcal{P}} \Big[\int g \mathrm{d}P - D(P\|Q) \Big]. \tag{6}$$

It has recently been shown (in [2]; earlier related calculations appeared in [7]) that

$$\frac{1}{\alpha} \ln \int e^{\alpha g} \mathrm{d}Q = \sup_{P \in \mathcal{P}} \Big[\frac{1}{\alpha - 1} \ln \int e^{(\alpha-1)g} \mathrm{d}P - D_\alpha(P\|Q) \Big], \qquad \alpha > 1. \tag{7}$$

Formally, one can recover (6) from (7) by taking the limit $\alpha \to 1$ and using D_1 in place of the limit of D_α as $\alpha \to 1$. Now, as a consequence of (7) one obtains for $\alpha > 1$ and $P, Q \in \mathcal{P}$ the bound

$$\frac{1}{\alpha - 1} \ln \int e^{(\alpha-1)g} \mathrm{d}P \le \frac{1}{\alpha} \ln \int e^{\alpha g} \mathrm{d}Q + D_\alpha(P\|Q). \tag{8}$$

Given an event $\mathcal{A} \in \mathcal{F}$, one can take g to assume the values 0 and $-M$ on $\mathcal{A}$ and its complement, respectively, and on taking the limit $M \to \infty$, deduce from the above that

$$\frac{1}{\alpha - 1} \ln P(\mathcal{A}) \le \frac{1}{\alpha} \ln Q(\mathcal{A}) + D_\alpha(P\|Q), \tag{9}$$

provided $P(\mathcal{A}) > 0$ and $Q(\mathcal{A}) > 0$ (see [2] for the details). Inequalities (8) and (9) are referred to in [2] as the *risk-sensitive functionals comparison bound* and *logarithmic probability comparison bound*, respectively. It is important to mention that both inequalities are tight in the sense that given Q and g, there exists a (unique) measure, namely $dP = e^{\alpha g} dQ/Z$, $Z = \int e^{\alpha g} \mathrm{d}Q$, for which (8) holds with equality. And given Q [resp., P] and $\mathcal{A}$ for which $Q(\mathcal{A}) > 0$ [resp., $P(\mathcal{A}) > 0$], there exists a (unique) measure, namely $Q(\cdot|\mathcal{A})$ [resp., $P(\cdot|\mathcal{A})$] for which (9) holds with equality. Another way of obtaining inequality (9) (but not (8)) is via the the data processing inequality for the Rényi divergence, as illustrated in Appendix A.1. A further useful fact is that both also give a lower bound, in addition to an upper bound, by interchanging the roles of the measures. Thus

$$\frac{1}{\alpha} \ln P(\mathcal{A}) \ge \frac{1}{\alpha - 1} \ln Q(\mathcal{A}) - D_\alpha(Q\|P). \tag{10}$$

3 Implications on exponential rate of decay

It is well known that (6) can be used to obtain estimates on probabilities of rare events (see [6]). By an approach developed in [2], the representation (7) also leads to such estimates, by appealing to (8) and (9). We now present this approach. Consider first the simple case where a sequence of real valued random variables $X_1, X_2, \ldots$ defined on the given measurable space is i.i.d. under both probability measures P and Q. Denote by P_n and Q_n the respective probability laws of the vector $X^n = (X_1, \ldots, X_n)$. It is a simple fact that the Rényi divergence scales as $D_\alpha(P_n\|Q_n) = nD_\alpha(P_1\|Q_1)$. Thus for $n \in \mathbb{N}$ and any event $\mathcal{A}_n$ measurable on the sigma-field generated by $(X_1, \ldots, X_n)$, that is, for some Borel subset B_n of $\mathbb{R}^n$, $\mathcal{A}_n = \{X^n \in B_n\}$, one has

$$\limsup_{n\to\infty} \frac{1}{n} \ln P(\mathcal{A}_n) \leq \frac{\alpha}{\alpha-1} \limsup_{n\to\infty} \frac{1}{n} \ln Q(\mathcal{A}_n) + \alpha D_\alpha(P_1\|Q_1). \tag{11}$$

This gives a comparison of the exponential rates involving only the Rényi divergence between the two marginals. In greater generality, when X_n are not necessarily i.i.d. under the measures P and Q, with P_n and Q_n still denoting the respective probability laws of $(X_1, \ldots, X_n)$, for $n \in \mathbb{N}$, let G_n and $\mathcal{A}_n$ be a bounded, measurable function and an event, that are both measurable on the sigma-field generated by that vector. Then again, from (8) and (9),

$$\frac{1}{(\alpha-1)n} \ln \boldsymbol{E}_P[e^{(\alpha-1)G_n(X^n)}] \leq \frac{1}{\alpha n} \ln \boldsymbol{E}_Q[e^{\alpha G_n(X^n)}] + \frac{1}{n} D_\alpha(P_n\|Q_n), \tag{12}$$

$$\frac{1}{(\alpha-1)n} \ln P(\mathcal{A}_n) \leq \frac{1}{\alpha n} \ln Q(\mathcal{A}_n) + \frac{1}{n} D_\alpha(P_n\|Q_n). \tag{13}$$

Denote

$$E_*(P) = -\limsup_{n\to\infty} \frac{1}{n} \ln P(\mathcal{A}_n), \qquad E^*(P) = -\liminf_{n\to\infty} \frac{1}{n} \ln P(\mathcal{A}_n), \qquad \text{for } P \in \mathcal{P}, \tag{14}$$

and

$$\bar{D}_\alpha(P\|Q) = \limsup_{n\to\infty} \frac{1}{n} D_\alpha(P_n\|Q_n) \qquad \text{for } (P,Q) \in \mathcal{P}^2. \tag{15}$$

Then

$$\frac{1}{\alpha-1} E_*(P) \geq \frac{1}{\alpha} E_*(Q) - \bar{D}_\alpha(P\|Q). \tag{16}$$

Combining this with the bound obtained by interchanging P and Q, one obtains the two-sided bound on the exponential decay rate under P in terms of that under Q:

$$\frac{\alpha-1}{\alpha} E_*(Q) - (\alpha-1)\bar{D}_\alpha(P\|Q) \leq E_*(P) \leq E^*(P) \leq \frac{\alpha}{\alpha-1} E^*(Q) + \alpha \bar{D}_\alpha(Q\|P). \tag{17}$$

Note that upper and a lower bounds analogous to (17) can be deduced from (12) for limits of the left-hand side of (12). In the sequel, when the limits exist, we write $E^*(\cdot)$ and $E_*(\cdot)$ as $E(\cdot)$. We will usually take P to be the model of interest, or the 'true' model, and Q will be the reference model.

It is instructive to note that inequalities (12) and (13), that are valid for each n, provide some information that is lost when passing to the limit, as for example in the i.i.d. case alluded to above, where the divergence term $n^{-1}D_\alpha(P_n\|Q_n) = D_\alpha(P_1\|Q_1)$ is given explicitly. This viewpoint of the approach has been further developed in [2]. However, in this paper, we will use the bounds exclusively in their limit forms, given by (17).

To relate (17) to the standard change of measure technique, consider the upper bound on $E^*(P)$ (which corresponds to a lower bound on probabilities) in the case where the probabilities of the event of interest are order 1 at the logarithmic scale, namely $E^*(Q) = 0$. Then one can take $\alpha \to 1$. Since the divergence term converges (formally) to that given in terms of the KL divergence, the standard change of measure method recovers.

The bounds (17) are useful when for a given model of interest P, one can find a reference model Q for which the exponents are known or can be bounded, and at the same time, one can efficiently estimate the divergence term. This is demonstrated in this article several times. Whereas the case alluded to above, in which both P and Q have i.i.d. structure, is most instructive, we will apply the bounds (17) in scenarios that go far beyond that. In fact, the bound we develop are more effective in situations where the model of interest P has long memory properties (such as, in the setting of channel coding, models that have interference, fading or erasure with long range correlations).

Second moment bounds

A useful framework is when the true model consists of a small perturbation of the reference model. Here we analyze a simple case where the alphabet is finite, and obtain a bound involving the second moment of the perturbation size. While the proof of the result is simple, it is an archetype of the argument used several times in the sequel for more complicated models in which the noise is dominant. These include the very noisy channel (see p. 155, eq. (3.4.23) of [21]) $P(y|x) = Q(y)[1 + \epsilon(x, y)]$ and, in the same spirit, the weak interference channel $P(y_t|x^{n-1}, y^{t-1}) = Q(y_t|x_t)[1 + \epsilon(x^n, y^t)]$.

Let a vector X^n take values in $\mathcal{X}^n$, where $\mathcal{X}$ is a finite set, and assume that the vector is i.i.d. under both the measures P and Q. Denote by P_n and Q_n the respective probability laws of the vector. Denoting $p = P_1$ and $q = Q_1$, assume that $p(x) = q(x)[1 + \epsilon(x)]$ for all $x \in \mathcal{X}$, where $\sum_x q(x)\epsilon(x) = 0$. Assume that the support of both P and Q is the entire alphabet $\mathcal{X}$, provided that $\|\epsilon\| := \max_x |\epsilon(x)|$ is small. Let $\mathcal{A}_n$ be any sequence of events of the form $\mathcal{A}_n = \{X^n \in B_n\}$, where B_n is a Borel subset of $\mathbb{R}^n$ and use the notation (14) for $E^*(P)$ and $E^*(Q)$.

Proposition 3.1 *Denote $\overline{\epsilon^2} = \sum_{x\in\mathcal{X}} q(x)\epsilon^2(x)$. Then*

$$E^*(P) \le \left(\sqrt{E^*(Q)} + \sqrt{\frac{\overline{\epsilon^2}}{2}}\right)^2 + o(\|\epsilon\|^2). \tag{18}$$

Proof: One has

$$\begin{aligned}
D_\alpha(q\|p) &= \frac{1}{\alpha(\alpha-1)}\ln\left[\sum_x q^\alpha(x)p^{1-\alpha}(x)\right] && (19)\\
&= \frac{1}{\alpha(\alpha-1)}\ln\left[\sum_x q(x)[1+\epsilon(x)]^{1-\alpha}\right] && (20)\\
&\leq \frac{1}{\alpha(\alpha-1)}\ln\left[\sum_x q(x)[1+(1-\alpha)\epsilon(x)+\frac{1}{2}\alpha(\alpha-1)\epsilon^2(x)]+c(\alpha)\|\epsilon\|^3\right] && (21)
\end{aligned}$$

for a suitable constant $c(\alpha)$, where we used Taylor's expansion of $z\mapsto(1+z)^{1-\alpha}$. Using $\ln(1+z)\leq z$, for $z>0$,

$$\begin{aligned}
D_\alpha(q\|p) &\leq \frac{1}{2}\sum_x q(x)\epsilon^2(x)+\tilde{c}(\alpha)\|\epsilon\|^3 && (22)\\
&= \frac{1}{2}\overline{\epsilon^2}+\tilde{c}(\alpha)\|\epsilon\|^3, && (23)
\end{aligned}$$

for a suitable constant $\tilde{c}(\alpha)$. Now, by the assumed i.i.d. structure, $D_\alpha(Q_n\|P_n)=nD_\alpha(q\|p)$. Thus by (17), for every $\alpha>1$,

$$E^*(P)\leq\frac{\alpha}{\alpha-1}E^*(Q)+\alpha\bar{D}_\alpha(Q\|P)\leq\frac{\alpha}{\alpha-1}E^*(Q)+\alpha\Big[\frac{1}{2}\overline{\epsilon^2}+\tilde{c}(\alpha)\|\epsilon\|^3\Big]. \tag{24}$$

The function $\alpha\mapsto\alpha(\alpha-1)^{-1}u+\alpha v$, $\alpha\in(1,\infty)$, $u,v>0$ attains minimum at $\alpha^*=\sqrt{u/v}+1$ and the minimum is given by $(\sqrt{u}+\sqrt{v})^2$. Therefore

$$E^*(P)\leq\left(\sqrt{E^*(Q)}+\sqrt{\frac{\overline{\epsilon^2}}{2}}\right)^2+\alpha^*\tilde{c}(\alpha^*)\|\epsilon\|^3=\left(\sqrt{E^*(Q)}+\sqrt{\frac{\overline{\epsilon^2}}{2}}\right)^2+O(\|\epsilon\|^3). \tag{25}$$

□

Remark 3.1 *A more general setting is discussed in a recent paper [20] (specifically eq. (50) therein) where for a parametric family $\{P_\theta,\ \theta\in\Theta\}$, one has*

$$\lim_{\theta'\to\theta}\frac{D_\alpha(P_\theta\|P_{\theta'})}{(\theta'-\theta)^2}=\frac{J(\theta)}{2}, \tag{26}$$

where $J(\theta)$ is the Fisher information. In this case, the bound (18) *is valid with $\sqrt{\frac{\overline{\epsilon^2}}{2}}$ replaced by $\sqrt{J(\theta)/2}\cdot|\theta'-\theta|$.*

4 Applications to channel coding

This section addresses the use of the lower and upper bounds (17) in the context of channel coding. We begin by considering, in Subsection 4.1, a general framework where we describe the relevance of the bounds in three contexts: (1) Bounds on performance for a given channel in terms of a reference channel; (2) Bounds for mismatched decoding; (3) Robust bounds. In Subsections 4.2–4.5, we consider several specific channel models of interest, where our methods yield new bounds. These include interference with long range dependence, discrete and continuous time Gaussian (and non-Gaussian) channels with fading, and the binary channel with erasure.

4.1 Generalities

Setting and main estimates

In channel coding, messages are encoded, transmitted over a noisy channel and decoded. The precise setting that we shall use is as follows. A message m from a set of $M = e^{nR}$ messages, $\mathcal{M} = \{0, 1, \ldots, M-1\}$, is encoded into a codeword $\boldsymbol{x}_m = (x_{m,1}, \ldots, x_{m,n})$ of length n, whose coordinates all take on values in a space $\mathcal{X}$, that for the purposes of this paper may be either finite or a Euclidean space $\mathbb{R}^k$ (some $k \geq 1$). Here, $R > 0$ is the coding rate in nats per channel use. We let $\mathcal{C}_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\}$ denote the codebook. When a codeword $\boldsymbol{x}_m \in \mathcal{C}_n$ is transmitted over a channel, a channel output $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ is produced, where again $\mathcal{Y}$ is either a given finite set or $\mathbb{R}^\ell$ (some $\ell \geq 1$). The decoder observes the vector $\boldsymbol{y}$ and produces an estimate $\hat{m} \in \mathcal{M}$ using a metric decoder, i.e.,

$$\hat{m} = \mathrm{argmin}_{m \in \mathcal{M}} d_n(\boldsymbol{x}_m, \boldsymbol{y}) \tag{27}$$

where ties are broken by an arbitrary deterministic rule, and $d_n(\boldsymbol{x}, \boldsymbol{y})$ is an additive *decoding metric* function, that is, it takes the form

$$d_n(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} d(x_i, y_i), \tag{28}$$

where $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is a given Borel measurable function. To describe the model probabilistically (both the channel transmission and the coding-decoding process), we consider now the input and output of the channel as random variables, and write them as $\boldsymbol{X}$ and $\boldsymbol{Y}$. Our analysis allows for the codebook to be either deterministic or random, settings which we refer to as deterministic and random coding, respectively. The message and the estimated message, m and $\hat{m}$, are also random now. The collection of these random variables (for all values of n), as well as the codebooks (in the case of random coding) are defined on a probability space $(\Omega, \mathcal{F}, P)$. In the case of random coding, $\mathcal{C}_n$ is a random variable with values in $(\mathcal{X}^n)^M$. The distribution of the other random elements, namely m, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\hat{m}$, is specified conditionally on $\mathcal{C}_n$. Further probabilistic assumptions of the model are as follows:
(i) m is uniformly distributed over $\mathcal{M}$. Consequently (assuming throughout that all codewords are distinct),

$$\Pi(\boldsymbol{x}) \stackrel{\triangle}{=} P(\boldsymbol{X} = \boldsymbol{x}) = \begin{cases} \frac{1}{M} & \boldsymbol{x} \in \mathcal{C}_n \\ 0 & \text{elsewhere,} \end{cases} \tag{29}$$

for deterministic coding, and

$$\Pi(\boldsymbol{x}) \stackrel{\triangle}{=} P(\boldsymbol{X} = \boldsymbol{x} | \, \mathcal{C}_n) = \begin{cases} \frac{1}{M} & \boldsymbol{x} \in \mathcal{C}_n \\ 0 & \text{elsewhere,} \end{cases} \tag{30}$$

in the case of random coding.
(ii) The model for the channel is described by the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. This conditional distribution is denoted by

$$P(\boldsymbol{y}|\boldsymbol{x}) \stackrel{\triangle}{=} P\{\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}\}. \tag{31}$$

If we denote by $\mathcal{E}_n = \{\hat{m} \neq m\}$ the error event then the error probability is given by $P(\mathcal{E}_n)$. In the case of random coding, this can be written as $P(\mathcal{E}_n) = \boldsymbol{E}_P[P(\mathcal{E}_n|\mathcal{C}_n)]$, which is interpreted as the mean probability of error when averaged over codes. We will say that a decoding metric is *matched* to a given channel if, under this metric, $\hat{m}$ is precisely the maximum likelihood estimator of m given $\boldsymbol{Y}$. The decoding metric d_n is *not* assumed to be matched to the channel (as is the case, for example, when $P(\boldsymbol{y}|\boldsymbol{x})$ takes the product form $\prod_{i=1}^n p(y_i|x_i)$ and $d(x,y)$ is proportional to $-\ln p(y|x)$). We will sometimes assume (without essential loss of generality) that the given code $\mathcal{C}_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\}$ is a constant composition code (CCC), that is, all codewords have the same empirical distribution, which converges to a given probability distribution $\mu = \{\mu(x),\ x \in \mathcal{X}\}$ as $n \to \infty$.

A *reference channel* is another probability measure, Q, on $(\Omega, \mathcal{F})$, which models a (possibly) different channel. In this work, we will always assume that, under Q, the distribution of the codes (in the case of random coding) as well as the probability of each codeword, is the same as under P; specifically, (29) and (30) are valid with P replaced by Q.

For deterministic coding, let P_n and Q_n denote the joint distribution of the two n-vectors $(\boldsymbol{X}, \boldsymbol{Y})$ under P and Q, respectively. It will be assumed that, given n, the correspondence between m and $\boldsymbol{X}$ is one-to-one. As a result, the error event is measurable with respect to the σ-field generated by $(\boldsymbol{X}, \boldsymbol{Y})$. In the case of random coding, it is not natural to assume that the correspondence alluded to above is always one-to-one. In this case we use the same notation, P_n and Q_n, to denote the respective distributions of the quadruple $(\mathcal{C}_n, m, \boldsymbol{X}, \boldsymbol{Y})$. The error event is then measurable with respect to the σ-field of this quadruple. Thus by (13), we have for every n and every $\alpha > 1$, the lower bound

$$\frac{1}{n} \ln P(\mathcal{E}_n) \geq \frac{\alpha}{n(\alpha-1)} \ln Q(\mathcal{E}_n) - \frac{\alpha}{n} D_\alpha(Q_n \| P_n), \tag{32}$$

and the upper bound

$$\frac{1}{n} \ln P(\mathcal{E}_n) \leq \frac{\alpha-1}{n\alpha} \ln Q(\mathcal{E}_n) + \frac{\alpha-1}{n} D_\alpha(P_n \| Q_n). \tag{33}$$

Adapting the notation (14) to the present setting, we write

$$E_*(R, \hat{P}, d) = -\limsup_{n\to\infty} \frac{1}{n} \ln \hat{P}(\mathcal{E}_n), \qquad E^*(R, \hat{P}, d) = -\liminf_{n\to\infty} \frac{1}{n} \ln \hat{P}(\mathcal{E}_n), \tag{34}$$

where $\hat{P} \in \mathcal{P}$ is any channel model, and we emphasize the dependence on the rate R and on the metric d (however, in the sequel, we sometimes suppress the dependence on R and d when there is no room for confusion). The notation $\bar{D}_\alpha$ from (15) will be used here with $\hat{P}_n$ and $\hat{Q}_n$ again denoting the respective joint distribution of the n-vectors $(\boldsymbol{X}, \boldsymbol{Y})$. We thus obtain from (17), for every $\alpha > 1$, the bounds

$$\begin{aligned} \frac{\alpha-1}{\alpha} E_*(R, Q, d) - (\alpha-1)\bar{D}_\alpha(P\|Q) &\leq E_*(R, P, d) \\ &\leq E^*(R, P, d) \leq \frac{\alpha}{\alpha-1} E^*(R, Q, d) + \alpha \bar{D}_\alpha(Q\|P). \end{aligned} \tag{35}$$

Remark 4.2 *In [15, Theorem 26], the so called metaconverse theorem provides a general lower bound on the probability of error under channel P in terms of the one under channel Q. The main idea in the proof of that theorem is to relate these two error probabilities to those of a certain binary hypothesis testing problem for discriminating between P and Q. Therefore the aforementioned data processing theorem for the Rényi divergence (see Appendix A.1, eq.* (A.8)*) applied to this binary hypothesis testing problem, may serve as an alternative route to obtain a result equivalent to lower and upper bounds* (32) *and* (33)*. However, to the best of our knowledge, this direction has not been pursued before.*

Three interpretations of the bounds

We identify three ways in which the above bounds can be used. In all cases, we think of P as the true channel model and Q as a reference.

(i) *Bounds on performance of the true channel in terms of a reference channel.*

One can obtain lower [upper] bounds on error exponents for true channel models by means of a lower [resp., upper] bound for a reference model. Suppose that d and a reference channel Q are given, where d is matched to Q. More generally, suppose that a parametric family $\{Q_\theta\}$ is given such that a given, fixed metric d is matched to each member of the family. Assume further that one knows a lower bound, $E_L(R, Q_\theta, d)$ on the error exponent $E(R, Q_\theta, d)$. Then for a metric d_P that is matched to P, we obtain from (35),

$$E_*(R, P, d_P) \geq E_*(R, P, d) \geq \sup_{\alpha>1} \sup_{\theta} \Big[\frac{\alpha-1}{\alpha} E_L(R, Q_\theta, d) - (\alpha-1)\bar{D}_\alpha(P\|Q_\theta)\Big]. \tag{36}$$

Similarly, an upper bound is possible for given P and d, when for reference channels Q one knows an upper bound $E_U(R, Q, d)$ on $E(R, Q, d)$ (when d is not necessarily matched to Q) and then

$$E^*(R, P, d) \leq \inf_{\alpha>1} \inf_{\theta} \Big[\frac{\alpha}{\alpha-1} E_U(R, Q, d) + \alpha\bar{D}_\alpha(Q_\theta\|P)\Big]. \tag{37}$$

(ii) *Bounds on performance of mismatched decoding.*

When d is matched to a reference channel Q, or a parametric family thereof, the second inequality in (36) serves as an upper bound on the mismatched error exponent (of using d with the true channel P) in terms of matched error exponent bounds (of using d with the reference channels Q_θ to which it is matched). A similar statement is valid for the upper bound (37). To recapitulate, the above inequalities give bounds on the error exponents under the true channel, operating with a decoder that is matched to another channel in terms of error exponents of the latter.

(iii) *Robust bounds.*

Consider a family F of true channels. As a performance criterion for the decoder, it is of interest to study the minimum error exponent within the family, namely

$$\mathcal{E}(R, F, d) := \inf_{P\in F} E(R, P, d). \tag{38}$$

Optimizing over decoders gives

$$\mathcal{E}(R, F) := \sup_{d} \mathcal{E}(R, F, d). \tag{39}$$

Thus $\mathcal{E}(R, F)$ is the best possible guarantee on the performance of all channels within the family when the communication system operates with a single decoder d (where 'best' refers to the selection of d). We can take advantage of the fact that the aforementioned bounds for a fixed channel model, P, are independent of P for $P \in F$, in order to obtain information on $\mathcal{E}(R, F)$. To this end, fix a reference channel Q, and assume that it is a member of the family F. Then automatically, $\mathcal{E}(R, F) \leq E(R, Q, d_Q)$, where d_Q is matched to Q. As far as a lower bound is concerned, recall that for $P \in F$, and fixed α,

$$E(R, P, d) \geq \frac{\alpha - 1}{\alpha} E(R, Q, d) - (\alpha - 1)\bar{D}_\alpha(P\|Q). \tag{40}$$

Let $r(\alpha) = (\alpha - 1)\sup_{P \in F} \bar{D}_\alpha(P\|Q)$. Then, for $\alpha > 1$,

$$\mathcal{E}(R, F) \geq \frac{\alpha - 1}{\alpha} \sup_d \mathcal{E}(R, Q, d) - r(\alpha). \tag{41}$$

Whereas the max-min problem posed by (39) is typically notoriously hard, the optimization problem that now appears in the bound is easy to handle, since the optimal decoder for Q is the one matched to it. Thus we have

$$\sup_{\alpha > 1} \left[\frac{\alpha - 1}{\alpha} E(R, Q, d_Q) - r(\alpha)\right] \leq \mathcal{E}(R, F) \leq E(R, Q, d_Q). \tag{42}$$

The points of view (i)–(iii) presented above will be further explored and demonstrated for the specific models to be considered.

Implications on general memoryless channels

Here we consider the mismatched channel problem where both P and Q are memoryless. For simplicity, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are discrete. Given a metric d, it is natural to consider as a parametric family of reference channels $Q = Q_{\theta,\psi}$ given by

$$Q(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} q(y_i|x_i), \tag{43}$$

where

$$q(y|x) = q_{\theta,\psi}(y|x) = \frac{\psi(y) \cdot e^{-\theta d(x,y)}}{\sum_{y' \in \mathcal{Y}} \psi(y') e^{-\theta d(x,y')}}, \qquad x \in \mathcal{X}, y \in \mathcal{Y}, \tag{44}$$

and $\theta \geq 0$ and $\psi(y) \geq 0$, $y \in \mathcal{Y}$, are the parameters of the channel. Then the decoding metric d is matched to each of these channels, namely d is the maximum likelihood (ML) decoding metric for $Q_{\theta,\psi}$ for each θ and ψ. It is instructive to note that, as $\theta \to 0$, the channel becomes "noisier", i.e., the output becomes proportional to $\psi(y)$, independently of the input. Assume a constant composition code. Then, for $Q = Q_{\theta,\psi}$, we can calculate the divergence term as

$$\frac{\alpha}{n} D_\alpha(Q_n \| P_n) = \frac{1}{n(\alpha-1)} \ln\left(\sum_{\boldsymbol{x}\in\mathcal{C}_n} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} [\Pi(\boldsymbol{x})Q(\boldsymbol{y}|\boldsymbol{x})]^\alpha [\Pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})]^{1-\alpha} \right) \tag{45}$$

$$= \frac{1}{n(\alpha-1)} \ln\left(\sum_{\boldsymbol{x}\in\mathcal{C}_n} \Pi(\boldsymbol{x}) \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \prod_{i=1}^n [q^\alpha(y_i|x_i) p^{1-\alpha}(y_i|x_i)] \right) \tag{46}$$

$$= \frac{1}{n(\alpha-1)} \ln\left(\sum_{\boldsymbol{x}\in\mathcal{C}_n} \Pi(\boldsymbol{x}) \prod_{i=1}^n \left[\sum_{y\in\mathcal{Y}} q^\alpha(y|x_i) p^{1-\alpha}(y|x_i) \right] \right). \tag{47}$$

Using the constant composition code assumption, the empirical distributions of all codewords are equal, hence

$$\frac{\alpha}{n} D_\alpha(Q_n \| P_n) = \frac{1}{n(\alpha-1)} \ln\left(\sum_{\boldsymbol{x}\in\mathcal{C}_n} \Pi(\boldsymbol{x}) \prod_{\bar{x}\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} q^\alpha(y|\bar{x}) p^{1-\alpha}(y|\bar{x}) \right]^{n\mu(\bar{x})} \right) \tag{48}$$

$$= \frac{1}{(\alpha-1)} \sum_{x\in\mathcal{X}} \mu(x) \ln\left(\sum_{y\in\mathcal{Y}} q^\alpha(y|x) p^{1-\alpha}(y|x) \right) \tag{49}$$

$$= \alpha \sum_{x\in\mathcal{X}} \mu(x) D_\alpha(q(\cdot|x)\|p(\cdot|x)). \tag{50}$$

Thus the term reduces to one that involves Rényi divergence at the single-letter conditional marginals. Substituting in (35), we obtain

$$E(R,P,d) \le \frac{\alpha}{\alpha-1} E(R,Q,d) + \alpha \sum_{x\in\mathcal{X}} \mu(x) D_\alpha(q(\cdot|x)\|p(\cdot|x)). \tag{51}$$

Now, $E(R,Q,d)$ is an error exponent for *matched decoding* for the channel Q, and is therefore upper bounded by any upper bound on the reliability function, such as the well-known straight–line bound $E_{\text{sl}}(R,Q)$ (cf. Sections 3.6–3.8 of [21]). Thus, we have

$$E(R,P,d) \le \inf_{\psi,\theta} \inf_{\alpha>1} \left[\frac{\alpha}{\alpha-1} E_{\text{sl}}(R, Q_{\theta,\psi}) + \alpha \sum_{x\in\mathcal{X}} \mu(x) D_\alpha(q_{\theta,\psi}(\cdot|x)\|p(\cdot|x)) \right]. \tag{52}$$

Remark 4.3 *To put* (52) *in the context of known results, let $I(\mu,Q)$ denote the single–letter mutual information between X and Y, induced by the joint distribution $\mu\times Q$, that is,*

$$I(\mu,Q) = \sum_{x\in\mathcal{X}} \mu(x) \sum_{y\in\mathcal{X}} q(y|x) \ln\left[\frac{q(y|x)}{\sum_{x'\in\mathcal{X}} \mu(x') q(y|x')} \right]. \tag{53}$$

In is known [4] that

$$E(R,P) \le \sup_{\mu} \inf_{Q: I(\mu,Q)\le R} D(Q\|P|\mu). \tag{54}$$

Let us show that (52) *if fact reduces to* (54). *Given R, and a random coding distribution μ, consider Q for which $I(\mu, Q) \leq R$. Then $E_{\rm sl}(R, Q) = 0$, and so eq. (52) is further upper bounded by*

$$E(R, P, d) \leq \inf_{\alpha>1} \alpha \sum_{x\in\mathcal{X}} \mu(x) D_\alpha(q(\cdot|x)\|p(\cdot|x)). \tag{55}$$

Now, $\alpha D_\alpha(q(\cdot|x)\|p(\cdot|x))$ is a monotonically non–decreasing as a function of α, and taking the limit $\alpha \downarrow 1$ results in $\sum_{x\in\mathcal{X}} \mu(x) D(q(\cdot|x)\|p(\cdot|x))$, which recovers (54) *by minimizing over Q and maximizing over μ.*

Iterated use of the LPCB

Recall the general upper bound (33)

$$\frac{1}{n}\ln P(\mathcal{E}_n) \leq \frac{\alpha-1}{n\alpha}\ln Q(\mathcal{E}_n) + \frac{\alpha-1}{n} D_\alpha(P_n\|Q_n), \tag{56}$$

which holds for any pair of channel models P and Q and every $\alpha > 1$. We can iterate this estimate so as to compare another model, $\hat{P}$, to Q by relating it first to P and P to Q. This may be useful in situations when estimating the divergence of P from Q and that of $\hat{P}$ from P is easier than estimating the divergence of $\hat{P}$ from Q. Indeed, expressing relation (56) for $\hat{Q}$ and P gives, for any $\beta > 1$,

$$\frac{1}{n}\ln \hat{P}(\mathcal{E}_n) \leq \frac{\beta-1}{n\beta}\ln P(\mathcal{E}_n) + \frac{\beta-1}{n} D_\beta(\hat{P}_n\|P_n). \tag{57}$$

Consequently, for any $\alpha > 1$ and $\beta > 1$,

$$\frac{1}{n}\ln \hat{P}(\mathcal{E}_n) \leq \frac{(\alpha-1)(\beta-1)}{n\alpha\beta}\ln Q(\mathcal{E}_n) + \frac{(\alpha-1)(\beta-1)}{n\beta} D_\alpha(P_n\|Q_n) + \frac{\beta-1}{n} D_\beta(\hat{P}_n\|P_n). \tag{58}$$

We use this approach in one of the results of Subsection 4.2.

4.2 Interference with long range dependence

In this section we are interested in a channel of the form

$$Y_t = X_t + g_t(X^n, Y^{t-1}) + W_t, \tag{59}$$

for a generic sequence of functions $g_t : \mathbb{R}^n \times \mathbb{R}^{t-1} \to \mathbb{R}$. Here $\{W_t\}$ is an i.i.d. $\mathcal{N}(0, \sigma^2)$ noise (although we will also address a more general i.i.d. noise in the sequel). The main assumption is that the interference functions g_t are bounded. However, no assumption is made about g_t that limits the range of correlations of the interference signal. We let P be a probability measure under which $\{W_t\}$ is as described above, and $\{X_t\}$ and $\{W_t\}$ are mutually independent. This model will be studied via the reference channel Q, under which

$$Y_t = X_t + \tilde{W}_t, \tag{60}$$

where $\{\tilde{W}_t\}$ are i.i.d. $\mathcal{N}(0, \sigma^2/s)$ ($s > 0$ being a parameter), independent of $\{X_t\}$. Thus under the true channel,

$$P(\boldsymbol{y}|\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}[y_t - x_t - g_t(x^n, y^{t-1})]^2\right\}, \tag{61}$$

while under $Q = Q_s$,

$$Q(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{2\pi\sigma^2}{s}\right)^{-n/2} \exp\left\{-\frac{s}{2\sigma^2}\sum_{t=1}^{n}(y_t - x_t)^2\right\}. \tag{62}$$

Theorem 4.1 *Denote by $E_{\mathrm{sl}}(R, Q_s)$ the straight–line (upper) bound on $E(R, Q_s)$. Assume that for every n, and t,*

$$\max_{x^n, y^{t-1}} |g_t(x^n, y^{t-1})| \leq \mathit{\Gamma}_t, \tag{63}$$

and $\sum_{t=1}^{n} \mathit{\Gamma}_t^2 \leq n\mathit{\Gamma}^2$ for constants $\{\mathit{\Gamma}_t\}$ and $\mathit{\Gamma}$. Then, for any sequence of codes and any decoder,

$$\begin{aligned} E(R, P) \leq \inf_{\alpha>1} \inf_{s>1-1/\alpha} \bigg\{ &\frac{\alpha E_{sl}(R, Q_s)}{\alpha - 1} + \frac{\alpha \ln s}{2(\alpha - 1)} \\ &- \frac{\ln[1 + \alpha(s-1)]}{2(\alpha - 1)} + \frac{\alpha s \mathit{\Gamma}^2}{2\sigma^2[1 + \alpha(s-1)]} \bigg\}. \end{aligned} \tag{64}$$

The proof of this result, that appears in the appendix, uses the following identity in estimating the Rényi divergence. For any real u, v, a and b such that $a + b > 0$,

$$\int_{-\infty}^{+\infty} \mathrm{d}y \exp\{-a(y-u)^2 - b(y-v)^2\} = \sqrt{\frac{\pi}{a+b}} \cdot \exp\left\{-\frac{ab(u-v)^2}{a+b}\right\}. \tag{65}$$

This identity is used, in addition, in several other proofs in the sequel.

We emphasize that for the model under consideration, the authors are not aware of any other alternative bound on the error exponents.

Consider now the choice $s = 1$ in (64). In this case, the expression simplifies to

$$E(R, P) \leq \inf_{\alpha>1} \left\{\frac{\alpha E_{\mathrm{sl}}(R, Q_1)}{\alpha - 1} + \frac{\alpha \mathit{\Gamma}^2}{2\sigma^2}\right\}. \tag{66}$$

The optimal α is easily found to be

$$\alpha^* = 1 + \frac{\sigma\sqrt{2E_{\mathrm{sl}}(R, Q_1)}}{\mathit{\Gamma}}, \tag{67}$$

which yields

$$E(R, P) \leq E_U(R) \triangleq \left(\sqrt{E_{\mathrm{sl}}(R, Q_1)} + \frac{\mathit{\Gamma}}{\sqrt{2}\sigma}\right)^2. \tag{68}$$

The structure of the above bound is reminiscent of the bound from Proposition 3.1. However, the bound above is valid not only for weak interference. Results of similar structure appear several times in the sequel.

Note that the above bound has a clear weakness of having a floor of $\mathit{\Gamma}^2/2\sigma^2$ independent of the rate R. This is an inherent limitation stemming from the way the we apply the bound. However, one may apply additional considerations to address this difficulty. Specifically, one can use the idea of the straight–line bound (c.f. Theorem 3.8.1 in [21][2]), to improve the bound using the smallest straight–line function that touches the curve $E_U(R)$, passing through the point $(C,0)$, where C is the capacity of the true channel. The latter is upper bounded by $C \le \frac{1}{2}\ln[1+(\sqrt{S}+\mathit{\Gamma})^2/\sigma^2]$, where S is an upper bound on the average power of $\boldsymbol{X}$. In what follows we will denote this improved bound by $E_1(R)$.

Very noisy channel

We now focus on the case of a very noisy channel, where bounds can be computed explicitly and insight can be obtained. We thus study the implication of Theorem 4.1, specifically of (68), to the case where $\sigma^2 \gg S+\mathit{\Gamma}^2$, where $\{X_t\}$ satisfies $\sum_{t=1}^n X_t^2 \le nS$ a.s., for a given power limitation $S>0$. In this case, the capacity of the reference channel (with $s=1$) is about $C_Q = S/2\sigma^2$ and the capacity of the true channel is (upper bounded by) $C = (\sqrt{S}+\mathit{\Gamma})^2/2\sigma^2$. The error exponent is given by (see p. 157, eq. (3.4.33) of [21])

$$E(R,Q) = \begin{cases} C_Q/2 - R & R < C_Q/4 \\ \left(\sqrt{C_Q}-\sqrt{R}\right)^2 & C_Q/4 \le R < C_Q \\ 0 & R > C_Q. \end{cases} \tag{69}$$

Now, accordingly,

$$\begin{aligned} E_U(R) &= \begin{cases} \left[\sqrt{C_Q/2-R}+\mathit{\Gamma}/(\sqrt{2}\sigma)\right]^2 & R < C_Q/4 \\ \left[\sqrt{C_Q}+\mathit{\Gamma}/(\sqrt{2}\sigma)-\sqrt{R}\right]^2 & C_Q/4 \le R < C_Q \\ \mathit{\Gamma}^2/(2\sigma^2) & R > C_Q \end{cases} \\ &= \begin{cases} \left[\sqrt{C_Q/2-R}+\mathit{\Gamma}/(\sqrt{2}\sigma)\right]^2 & R < C_Q/4 \\ \left(\sqrt{C}-\sqrt{R}\right)^2 & C_Q/4 \le R < C_Q \\ \mathit{\Gamma}^2/(2\sigma^2) & R > C_Q. \end{cases} \end{aligned} \tag{70}$$

Note that at least in the intermediate range, between $C_Q/4$ and C_Q, the bound is tight in the sense that there exists an interference signal that achieves it. It corresponds to the coherent sum of the desired signal and the interference, which is the case when $g_t(x^n, y^{t-1})$ is proportional to x_t.

The improvement at high rates is provided by the straight–line that passes through the points $(C_Q, \mathit{\Gamma}^2/2\sigma^2)$ and $(C,0)$. The result we thus obtain for the very noisy channel is

$$E(R,P) \le E_1(R) \stackrel{\triangle}{=} \begin{cases} \left[\sqrt{C_Q/2-R}+\mathit{\Gamma}/(\sqrt{2}\sigma)\right]^2 & R < C_Q/4 \\ \left(\sqrt{C}-\sqrt{R}\right)^2 & C_Q/4 \le R < C_Q \\ \frac{\mathit{\Gamma}^2(C-R)}{2\sigma^2(C-C_Q)} & C_Q \le R < C. \end{cases} \tag{71}$$

[2] This theorem requires, in principle, the sphere–packing bound for list decoders, and for such a general channel, we don't know the sphere–packing bound. Nonetheless, one can still use the theorem when the higher rate is the capacity since the probability of list error is bounded away from zero, for any codebook of size $M = e^{n(C+\lambda+\epsilon)}$ and list of size $e^{\lambda n}$, as can easily be shown by a simple extension of Fano's inequality for list decoding. This is done by using the fact that $H(\boldsymbol{X}|\boldsymbol{Y},\text{ no list error}) \le n\lambda$ (unlike the case of ordinary decoding where $H(\boldsymbol{X}|\boldsymbol{Y},\text{ no error}) = 0$).

Remark 4.4 *At rate zero (and general SNR), the bound one obtains from the discussion above, by selecting $s = 1$, is*

$$E_1(0) = \Big(\sqrt{E_{ex}(0,Q)} + \frac{\Gamma}{\sqrt{2}\sigma}\Big)^2 = \Big(\sqrt{\frac{C_Q}{2}} + \frac{\Gamma}{\sqrt{2}\sigma}\Big)^2 = \frac{(\sqrt{S} + \Gamma\sqrt{2})^2}{4\sigma^2}. \tag{72}$$

It turns out that for $R = 0$ one can solve the full optimization problem (64)*, including the minimization over the parameter s. In fact, one can even solve an extended problem, in which the reference model has one additional free parameter, namely a gain factor ϕ: Instead of* (62)*, one considers $Q = Q_{\theta,\phi}$ of the form $\prod_{t=1}^{n} Q(y_t|x_t)$, where*

$$Q(y|x) = \Big(\frac{s}{2\pi\sigma^2}\Big)^{1/2} \exp\Big\{-\frac{s}{2\sigma^2}(y - \phi x)^2\Big\}, \quad s > 0,\ \phi > 0. \tag{73}$$

However, the bound one obtains is exactly (72)*.*

Lower bound on the exponent

We can also derive a lower bound by appealing to (37). In this context, it is more natural to consider the setting of random coding because existing bounds for reference models are of this type. Denoting the sequence of random codes by $\{\mathcal{C}_n\}$, the relevant divergence term for using (37) is

$$D_\alpha(P_n\|Q_n) = \frac{1}{\alpha(\alpha-1)} \ln \boldsymbol{E}_Q\Big[\Big(\frac{P(\mathcal{C}_n, m, \boldsymbol{X}, \boldsymbol{Y})}{Q(\mathcal{C}_n, m, \boldsymbol{X}, \boldsymbol{Y})}\Big)^\alpha\Big]. \tag{74}$$

Recalling our assumption that under the true model P and under the reference model Q the distribution of the codes is equal, we have

$$D_\alpha(P_n\|Q_n) = \frac{1}{\alpha(\alpha-1)} \ln \boldsymbol{E}_Q\Big[\Big(\frac{P(\boldsymbol{X}, \boldsymbol{Y}|\mathcal{C}_n, m)}{Q(\boldsymbol{X}, \boldsymbol{Y}|\mathcal{C}_n, m)}\Big)^\alpha\Big]. \tag{75}$$

Now, the estimate on the divergence term appearing in the proof of Theorem 4.1 can be carried out for the above in a similar manner, and one obtains the same bound (A.21) regardless of the code $\mathcal{C}_n$. For simplicity, we can specialize to $s = 1$, which gives

$$E(R,P) \quad \geq \quad E_L(R,P) \stackrel{\triangle}{=} \sup_{\alpha>1}\left[\frac{\alpha-1}{\alpha}E(R,Q_1) - \frac{(\alpha-1)\Gamma^2}{2\sigma^2}\right], \tag{76}$$

the solution of which is

$$E_L(R,P) = \begin{cases} \Big(\sqrt{E(R,Q_1)} - \frac{\Gamma}{\sqrt{2}\sigma}\Big)^2 & E(R,Q_1) \geq \Gamma^2/2\sigma^2 \\ 0 & \text{elsewhere.} \end{cases} \tag{77}$$

One can use the above bound to estimate the capacity of the the channel P. It is bounded below by the rate R at which $E(R,Q_1) = \frac{\Gamma^2}{2\sigma^2}$. For the example of the very noisy channel, this gives $C_P \geq (\sqrt{S} - \Gamma)^2/2\sigma^2$. This bound is attained by the interference signal that is anti–coherent with the desired signal, i.e., $g_t(x^t, y^{t-1}) = -\Gamma x_t/\sqrt{S}$.

Robust bound interpretation

All three interpretations mentioned in Subsection 4.1 are relevant for the results of this section. Specifically, the bounds of Theorem 4.1 and (77) are valid whether d is matched to P or not. Next, to demonstrate the robust bounds interpretation in the context of these results, let $Q = Q_1$ denote the reference channel (with $s = 1$) and for a fixed $\mathit{\Gamma}$, denote by F the family of true channels P for which g_t are all bounded by $\mathit{\Gamma}$. Then by (42) and the bound $r(\alpha) = (\alpha - 1)\sup_{P\in F} \bar{D}_\alpha(P\|Q) \leq (\alpha-1)\mathit{\Gamma}^2/(2\sigma^2)$ that follows from the previous paragraph, we have

$$\left(\sqrt{E(R,Q_1)} - \frac{\mathit{\Gamma}}{\sqrt{2}\sigma}\right)^2 \leq \sup_d \inf_{P\in F} E(R,P,d) \leq E(R,Q_1). \tag{78}$$

Specifically, the performance of a single decoder, namely the one matched to Q_1, is bounded by the above two bounds whenever the interference signal is bounded by the constant $\mathit{\Gamma}$.

Non Gaussian noise

Here we use the idea of iterating the bound, as presented in Subsection 4.1, in order address non-Gaussian i.i.d. noise. Going back to the general setting of Theorem 4.1, recall from (59) and (60) that under P and Q, respectively, we have the models

$$Y_t = X_t + g_t(X^n, Y^{t-1}) + W_t, \tag{79}$$

$$Y_t = X_t + \tilde{W}_t, \tag{80}$$

where $\{W_t\}$ and $\{\tilde{W}_t\}$ are i.i.d. $\mathcal{N}(0,\sigma^2)$, independent of $\{X_t\}$. Consider now an additional model $\hat{P}$ described by

$$Y_t = X_t + g_t(X^n, Y^{t-1}) + \hat{W}_t, \tag{81}$$

where $\{\hat{W}_t\}$ are i.i.d. but need not be Gaussian. The main point is that estimating the divergence of $\hat{P}$ from P is simple, whereas the estimates on the divergence of P from Q have already been established, thus by appealing to (58), one can relate $\hat{P}$ to Q by combining the two estimates.

Denote by $\delta(\beta) = D_\beta(\mathcal{L}_{\hat{P}}(\hat{W}_1)\|\mathcal{L}_P(W_1))$ the single–letter Rényi divergence, where for a measure μ and r.v. U, $\mathcal{L}_\mu(U)$ denotes the probability law of U under μ. Since both $\{\hat{W}_t\}$ and $\{W_t\}$ are i.i.d., we can make use of the simple fact that

$$D_\beta(\mathcal{L}_{\hat{P}}(\hat{W}^n)\|\mathcal{L}_P(W^n)) = nD_\beta(\mathcal{L}_{\hat{P}}(\hat{W}_1)\|\mathcal{L}_P(W_1)) = n\delta(\beta). \tag{82}$$

Moreover, since under both $\hat{P}$ and P, the noise sequence is independent of the signal $\{X_t\}$ and the latter has the same law, it follows that

$$D_\beta(\mathcal{L}_{\hat{P}}(X^n,\hat{W}^n)\|\mathcal{L}_P(X^n,W^n)) = n\delta(\beta). \tag{83}$$

Now, denote by $\hat{P}_n$ and P_n the respective laws of (X^n,Y^n) under $\hat{P}$ and P. Note by (79) and (81) that $(X^n,Y^n) = F_n(X^n,W^n)$ and $(X^n,Y^n) = G_n(X^n,\hat{W}^n)$ for suitable deterministic functions F_n and G_n. As a result, the data processing inequality (see [19, Section II]) gives $D_\beta(\hat{P}_n\|P_n) \leq D_\beta(\mathcal{L}_{\hat{P}}(X^n,\hat{W}^n)\|\mathcal{L}_P(X^n,W^n))$. Hence

$$D_\beta(\hat{P}_n\|P_n) \leq n\delta(\beta). \tag{84}$$

Using (84) in (58) gives

$$E(R,\hat{P}) \geq \frac{(\alpha-1)(\beta-1)}{\alpha\beta}E(R,Q) - \frac{(\alpha-1)(\beta-1)}{\beta}\bar{D}_\alpha(P\|Q) - (\beta-1)\delta(\beta). \tag{85}$$

We use our previous results that estimate $\bar{D}_\alpha(P\|Q)$ and optimize over α. With E_L given by (77), we have

$$E(R,\hat{P}) \geq \frac{(\beta-1)}{\beta}E_L(R,P) - (\beta-1)\delta(\beta). \tag{86}$$

An analogous estimate can be established for an upper bound on the exponent, as well as for all other channel models that we treat in the sequel.

Example 4.1 *Consider truncated Gaussian noise distribution for $\hat{W}_1$, namely, for a given constant u, assume $f_{\hat{W}_1}(w) = z^{-1}f(w)1_{[-u,u]}(w)$, where $f(w) = (2\pi)^{-1/2}e^{-w^2/2}$ is the standard normal density, and $z = \int_{-u}^{u} f(w)\mathrm{d}w$. Assume W_1 is standard normal. It is easy to see that*

$$\delta(\beta) = D_\beta(\hat{W}_1\|W_1) = \frac{\ln(1/z)}{\beta}. \tag{87}$$

Thus using (86) *and taking the limit $\beta \to \infty$,*

$$E(R,\hat{P}) \geq E_L(R,P) + \ln z. \tag{88}$$

Robust bounds for the ISI channel

We next study the Gaussian intersymbol interference (ISI) channel model, denoted by P, given by

$$Y_t = X_t + \sum_{i=1}^{k} h_i X_{t-i} + W_t, \tag{89}$$

where $\{W_t\}$ is i.i.d. $\mathcal{N}(0,\sigma^2)$, independent of $\{X_t\}$, and $\boldsymbol{h} = (h_1,\ldots,h_k)^T$ is given. While the proposed method yields new results for interference with unlimited correlation length (Theorem 4.1), an analogous treatment of the model (89) turns out not to be useful, as it leads to bounds that are inferior to existing bounds, for both matched and mismatched decoding. However, as we now demonstrate, the robust bound interpretation discussed in Subsection 4.1 gives rise to new results for this model.

Note that the model is a special case of the main model studied in this section. Because of the special structure of the interference (89) and some further assumptions we make regarding the correlation structure, the bounds that we are able to provide are much more explicit than those given by Theorem 4.1.

Aiming at bounds that may depend on the codeword energy and correlation function, but only on these important parameters, we consider a family thereof, where these are kept fixed. Namely, we assume that all codewords have energy

$$\sum_{t=1}^{n} x_t^2 = nS \tag{90}$$

and a fixed empirical autocorrelation structure

$$\sum_{t=i+1}^{n} x_t x_{t-i} = nc_i S. \quad i = 1, 2, \ldots, k. \tag{91}$$

Moreover, the rate is $R = 0$, and the channel is very noisy, that is, $\sigma \gg \sqrt{S}$. The decoder uses the mismatched decoding metric $d(x, y) = (x - y)^2$. Denote $\boldsymbol{c} = (c_1, \ldots, c_k)^T$ and $\boldsymbol{C} = [c_{|i-j|}]_{i,j=1}^{k}$ and let $r_1 = \boldsymbol{h}^T\boldsymbol{c}$ and $r_2 = \boldsymbol{h}^T\boldsymbol{C}\boldsymbol{h}$. Then r_1 and r_2 are related to the empirical correlation between signal and interference $g_t := \sum_{i=1}^{k} h_i x_{t-i}$ and interference power, respectively. Specifically,

$$\sum_{t=1}^{n} x_t g_t = nSr_1, \qquad \sum_{t=1}^{n} g_t^2 = nSr_2. \tag{92}$$

Note that always $r_2 \geq r_1^2$.

Theorem 4.2 *Consider a sequence of codes satisfying* (90) *and* (91) *for a specific vector* $\boldsymbol{c}$. *Denote by F the family of true models P of the form* (89)*, where h varies over all vectors having fixed r_1 and r_2. Denote $a = r_2 - r_1^2$ and $b = (1 + r_1)^2$. If $a < b$ then*

$$\frac{S}{4\sigma^2}(\sqrt{b} - \sqrt{a})^2 \leq \sup_{d} \inf_{P \in F} E(0, P, d) \leq \frac{S}{4\sigma^2}(\sqrt{b} + \sqrt{a})^2. \tag{93}$$

The proof appears in the appendix.

4.3 Discrete time Gaussian channel with fading

We consider the channel

$$Y_t = (1 + \theta_t)X_t + W_t, \tag{94}$$

where $\{W_t\}$ is an additive noise process and $\{\theta_t\}$ is a fading process. We let P be a probability measure under which the processes $\{\theta_t\}$, $\{W_t\}$ and $\{X_t\}$ are mutually independent, and $\{W_t\}$ is i.i.d. $\mathcal{N}(0, \sigma^2)$. Also, $\{X_t\}$ is assumed to satisfy the constraint $|X_t| \leq A$ for all t. As a reference, consider a channel with no fading. That is, consider a probability measure Q under which

$$Y_t = X_t + \tilde{W}_t, \tag{95}$$

where the law of triplet $(X, \theta, \tilde{W})$ under Q is the same as that of (X, θ, W) under P. In particular, under Q, $\{\tilde{W}_t\}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, and the three processes $\{X_t\}$, $\{\theta_t\}$ and $\{\tilde{W}_t\}$ are mutually independent.

We assume that $\{\theta_t\}$ is a stationary, zero-mean Gaussian process and that $r_k = E[\theta_0\theta_k]$ are absolutely summable. Let $\varSigma_\theta$ denote the spectral density of θ, namely

$$\varSigma_\theta(\omega) = \sum_{k=-\infty}^{\infty} r_k e^{-ik\omega}. \tag{96}$$

Theorem 4.3 *Let P and Q stand for the discrete-time Gaussian noise channel with and, respectively, without fading, described above. Denote $c = c(\alpha) = \alpha(\alpha-1)A^2/(2\sigma^2)$. Then for any $\alpha > 1$ such that $2c\sup_\omega \varSigma_\theta(\omega) < 1$,*

$$E(P) \leq \frac{\alpha}{\alpha-1}E(Q) - \frac{1}{4\pi(\alpha-1)}\int_0^{2\pi} \ln[1-2c\varSigma_\theta(\omega)]\mathrm{d}\omega, \tag{97}$$

$$E(P) \geq \frac{\alpha-1}{\alpha}E(Q) + \frac{1}{4\pi\alpha}\int_0^{2\pi} \ln[1-2c\varSigma_\theta(\omega)]\mathrm{d}\omega. \tag{98}$$

See the appendix for a proof.

Note that for fixed α, the gap between the upper and lower bound increases with $\varSigma_\theta$. This occurs due to the fact that the distance between the model P and the reference model Q, as measured in terms of the divergence, increases by strengthening the fading. When $\varSigma_\theta \equiv 0$, the models P and Q agree, and then so do the upper and lower bounds (upon optimizing over α).

While it is difficult to optimize over α in general, in the next paragraph we consider special cases where the results are more explicit.

AR fading model

Consider the case of $\{\theta_t\}$ given by the autoregressive (AR) model

$$\theta_t = a\theta_{t-1} + b\hat{W}_t, \tag{99}$$

where $\{\hat{W}_t\}$ are i.i.d. $\mathcal{N}(0,1)$, $|a| < 1$ and $\{\theta_t\}$ is stationary. We have $r_k = r_0 a^{|k|}$, $r_0 = b^2/(1-a^2)$, and

$$\varSigma_\theta(\omega) = \frac{b^2}{1-2a\cos(\omega)+a^2}. \tag{100}$$

Thus $f(\omega) \overset{\triangle}{=} 1 - 2c(\alpha)\varSigma_\theta(\omega)$ gives

$$f(\omega) = \frac{1-2a\cos(\omega)+a^2-2cb^2}{1-2a\cos(\omega)+a^2}, \tag{101}$$

and so, whenever f is bounded away from zero, which holds iff

$$(1-|a|)^2 > 2c(\alpha)b^2, \tag{102}$$

one has

$$E(P) \leq \frac{\alpha}{\alpha-1}E(Q) - \frac{1}{4\pi(\alpha-1)}\int_0^{2\pi} \ln f(\omega)\mathrm{d}\omega. \tag{103}$$

We next further develop (103) based on the residue theorem, by which one has $\int_0^{2\pi}\ln(1+re^{i\omega})\mathrm{d}\omega = 0$ whenever $|r| < 1$ for the complex logarithmic function $\ln(\cdot)$. Specifically, if we express f as

$$f(\omega) = k\frac{(1-re^{-i\omega})(1-re^{i\omega})}{(1-ae^{-i\omega})(1-ae^{i\omega})}, \tag{104}$$

for a real r with $|r| < 1$, and $k > 0$, then $\int_0^{2\pi} \ln f(\omega)\mathrm{d}\omega = 2\pi \ln k$. To calculate r and k, write for $z \in \mathbb{C}$,

$$(1 - az)(1 - az^{-1}) - 2cb^2 = k(1 - rz)(1 - rz^{-1}). \tag{105}$$

The solution to this is $k = a/r$,

$$r_{1,2} = \frac{\xi(\alpha) \pm \sqrt{\xi(\alpha)^2 - 4a^2}}{2a}. \tag{106}$$

where

$$\xi(\alpha) = 1 - 2c(\alpha)b^2 + a^2. \tag{107}$$

Under (102), the discriminant is positive, and therefore $r_{1,2}$ and k are real numbers. Also, one checks that under (102), $r_1 > 1$ hence not to be considered. As for r_2, we have $|r_2| < 1$ under that condition. Thus $k = a/r_2$, and we have

$$E(P) \leq \frac{\alpha}{\alpha - 1}E(Q) - \frac{2\pi \ln k}{4\pi(\alpha - 1)} \tag{108}$$

$$= \frac{\alpha}{\alpha - 1}E(Q) + \frac{1}{2(\alpha - 1)} \ln \frac{\xi(\alpha) - \sqrt{\xi^2(\alpha) - 4a^2}}{2a^2}. \tag{109}$$

The limit case $b \to 0$: This is when the fading amplitude goes to zero, we have the bound converging to $\alpha/(\alpha - 1)E(Q)$, and optimizing over α gives $E(Q)$, that is the best possible bound under the circumstances.

Using the lower bound gives the following bound, complementing (103), namely

$$E(P) \geq \frac{\alpha - 1}{\alpha}E(Q) + \frac{1}{4\pi\alpha}\int_0^{2\pi} \ln[1 - 2c\Sigma_\theta(\omega)]\mathrm{d}\omega \tag{110}$$

$$= \frac{\alpha - 1}{\alpha}E(Q) - \frac{1}{2\alpha} \ln \frac{\xi(\alpha) - \sqrt{\xi^2(\alpha) - 4a^2}}{2a^2}, \tag{111}$$

for all $\alpha > 1$ satisfying (102).

Figure 1 depicts the above bound as a function of α for various values of $E(Q)$. Note that the range of the parameter α is of the form $(1, \alpha^*)$, where α^* is the smallest α which violates condition (102). The right end of the graphs in Figures 1(a) and 1(b) correspond to α^*.

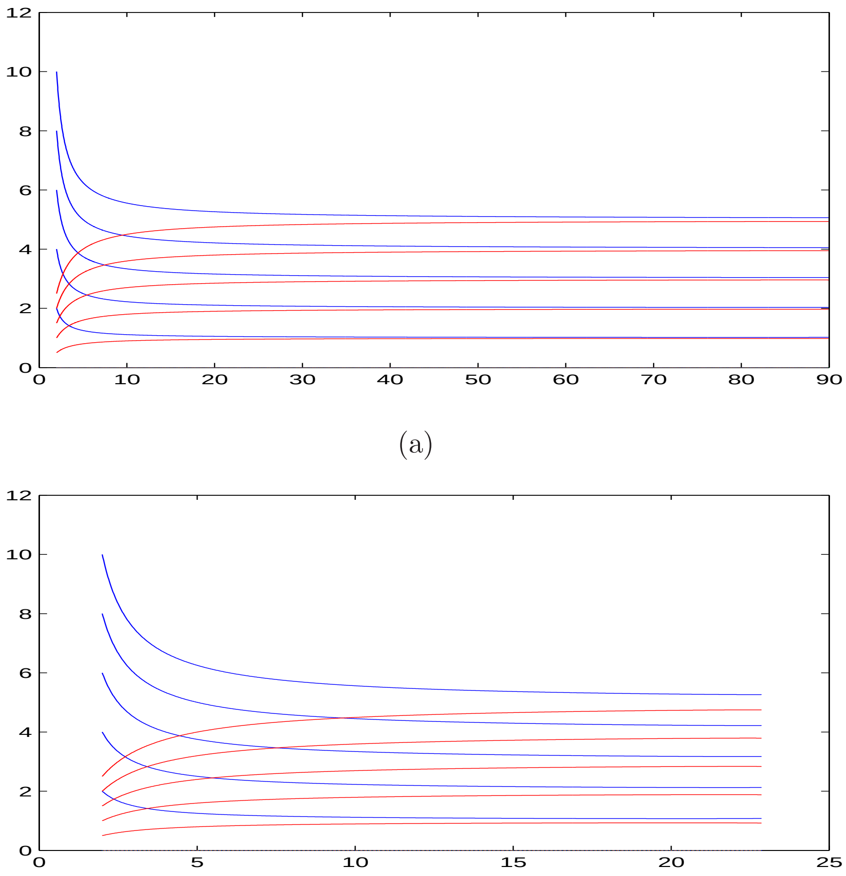
We comment that the bounds are tight in the small fading limit. Namely, as the amplitude of the fading perturbation goes to zero, the optimal bounds (obtained by choosing α suitably) converge to $E(Q)$. Indeed, as $b \to 0$, the argument of the logarithmic function converges to 1, by which that follows.

Note that one can treat the small fading limit in greater generality (beyond the AR process). Denote $\Sigma_{\max} = \sup_\omega \Sigma_\theta(\omega)$. Fix $0 < \delta < 1$ and assume $2c\Sigma_{\max} \leq \delta$. Denote $\kappa = \frac{1}{2(1-\delta)^2}$. Using the bound $\ln(1 + x) \geq x - \kappa x^2$ for all x s.t. $|x| < \delta$ in Theorem 4.3 gives, for every fixed $\alpha > 1$,

$$E(P) \leq \frac{\alpha}{\alpha - 1}E(Q) + \frac{1}{4\pi(\alpha - 1)}\int_0^{2\pi} [2c\Sigma_\theta(\omega) + \kappa\delta^2]\mathrm{d}\omega \tag{112}$$

$$= \frac{\alpha}{\alpha - 1}E(Q) + \frac{\alpha A^2 r_0}{2\sigma^2} + \frac{\kappa\delta^2}{2(\alpha - 1)}. \tag{113}$$

Optimizing over the parameter α in the range $\{\alpha > 1\} \cap \{2c(\alpha)\Sigma_{\max} \leq \delta\}$ can now be carried out easily (in a manner similar to that in Proposition 3.1).



(a)

(b)

Figure 1: *Plots of the upper bound* (109) *(blue) and lower bound* (111) *(red) on* $E(P)$ *as function of* α*. The five graphs correspond to* $E(Q) = 1, 2, 3, 4$ *and* 5*, with* $a = 0.2$ *and* $A^2/(2\sigma^2) = 0.1$*, where in plot (a),* $b = 0.02$*, and in plot (b),* $b = 0.08$*.*

4.4 Continuous–time white noise with fading

A standard model for a white Gaussian channel in continuous time is given by

$$Y_t = \int_0^t X_s \mathrm{d}s + \sigma W_t, \tag{114}$$

where $\{W_t\}$ is a Brownian motion. Let Q be a probability measure under which $\{W_t\}$ is a standard Brownian motion, and let $\{X_t\}$ and $\{\theta_t\}$ be real-valued processes such that the three processes $\{W_t\}$, $\{X_t\}$ and $\{\theta_t\}$ are mutually independent. Assume that $\{X_t\}$ satisfies the amplitude constraint $|X_t| \le A$ for all t, Q-a.s., where A is a constant. One can obtain from Q a model for a channel with fading, in which θ is the fading process, by means of a change of measure. To this end, consider the filtration

$$\mathcal{F}_t = \sigma\{X_s, \theta_s, W_s : s \in [0,t]\}, \tag{115}$$

and let

$$Z_t = \exp\Big[\frac{1}{\sigma}\int_0^t \theta_s X_s \mathrm{d}W_s - \frac{1}{2\sigma^2}\int_0^t (\theta_s X_s)^2 \mathrm{d}s\Big], \qquad t \ge 0. \tag{116}$$

It is assumed throughout that, for every $T > 0$,

$$\boldsymbol{E}_Q \exp\Big\{\frac{A^2}{2\sigma^2}\int_0^T \theta_s^2 \mathrm{d}s\Big\} < \infty. \tag{117}$$

We later provide a sufficient condition for this to hold. Note that, as a result, one has $\boldsymbol{E}_Q \exp\Big\{\frac{1}{2\sigma^2}\int_0^T (X_s\theta_s)^2 \mathrm{d}s\Big\} < \infty$, and so Novikov's condition for $\{Z_t\}$ to be an $\{\mathcal{F}_t\}$-martingale under Q is satisfied (see Corollary 3.5.13 of [11]). For $T > 0$, let Q_T and P_T be probability measures on $\mathcal{F}_T$, defined by

$$Q_T(\mathcal{A}) = Q(\mathcal{A}), \qquad P_T(\mathcal{A}) = \boldsymbol{E}_Q[1_{\mathcal{A}} Z_T], \qquad \mathcal{A} \in \mathcal{F}_T, \tag{118}$$

where $1_{\mathcal{A}}$ denotes the indicator function of $\mathcal{A}$. Then $\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = Z_T$, and by Girsanov's theorem (Theorem 3.5.1 of [11]) one has

$$Y_t = \int_0^t (1+\theta_s) X_s \mathrm{d}s + \sigma \tilde{W}_t, \tag{119}$$

where, under P_T, the triplet $(\theta_t, X_t, \tilde{W}_t, t \in [0,T])$ has the same law as that of $(\theta_t, X_t, W_t, t \in [0,T])$ under Q_T (thus under Q). In particular, under the measure P_T, $\{\tilde{W}_t\}$ is a standard Brownian motion, and the three processes $\{W_t\}$, $\{X_t\}$ and $\{\theta_t\}$ are mutually independent. As a result, P_T is a model for an additive white Gaussian noise channel with a fading process $\{\theta_t\}$.

It is assumed that $\{\theta_t\}$ is a separable, zero-mean stationary Gaussian process (under Q_T; equivalently under P_T). The spectral density of $\{\theta_t\}$, that is, the function $\varSigma_\theta$ for which $\boldsymbol{E}[\theta_0\theta_t] = \int_{-\infty}^{\infty} e^{it\omega}\varSigma_\theta(\omega)\mathrm{d}\omega$, is assumed to satisfy $\varSigma_{\max} := \operatorname{ess\,sup} \varSigma_\theta < \infty$.

The following, that can be seen as a continuous-time analogue of Theorem 4.3, is the main result of this subsection.

Theorem 4.4 *Let P and Q stand for the continuous time white noise channel models with and without fading, described above. Assume $p := A^2/(2\sigma^2) < 1/(4\pi\varSigma_{\max})$. Then* (117) *holds. Moreover, with $c(\alpha) = \alpha(\alpha-1)p$, for any $\alpha > 1$ such that $c(\alpha) < 1/(4\pi\varSigma_{\max})$,*

$$E(P) \le \frac{\alpha}{\alpha-1}E(Q) - \frac{1}{4\pi(\alpha-1)}\int_{-\infty}^{\infty}\ln[1-4\pi c(\alpha)\varSigma_\theta(\omega)]\mathrm{d}\omega \tag{120}$$

$$E(P) \ge \frac{\alpha-1}{\alpha}E(Q) + \frac{1}{4\pi\alpha}\int_{-\infty}^{\infty}\ln[1-4\pi c(\alpha)\varSigma_\theta(\omega)]\mathrm{d}\omega. \tag{121}$$

See the appendix for a proof.

For an encoder/decoder optimized for Q, an expression for $E(R,Q)$ is well known (see Section 8.2 of [8]), namely, with $C = A^2/(2\sigma^2)$,

$$E(R,Q) = \begin{cases} C/2 - R & R < C/4 \\ (\sqrt{C}-\sqrt{R})^2 & C/4 \le R < C \\ 0 & R \ge C. \end{cases} \tag{122}$$

As a result, (120) and (121) give bounds on the mismatched error exponents for the model with fading, when the encoder and decoder are matched to Q. The lower bound (121) appears to be new even for the matched channel exponent, that is, when the right-hand side of (121) serves as a lower bound on the error exponent for an encoder/decoder that are matched to P.

Low frequency fading

The expression in (A.68) is simple when $\varSigma_\theta$ is constant on its support. Specifically, consider the case $\varSigma_\theta(\omega) = \varSigma_0$ on the interval $[-B, B]$. Then

$$\frac{\alpha-1}{\alpha}E(Q) - \frac{1}{\alpha}r(\alpha) \le E(P) \le \frac{\alpha}{\alpha-1}E(Q) + \frac{1}{\alpha-1}r(\alpha), \tag{123}$$

where

$$r(\alpha) = -\frac{2B}{4\pi}\ln[1-4\pi\alpha(\alpha-1)p\varSigma_0], \tag{124}$$

provided $\max\{\alpha(\alpha-1), 1\} < 1/(4\pi p\varSigma_0)$.

Ornstein-Uhlenbeck fading

Next consider a model where the fading process takes the form of a stationary Ornstein-Uhlenbeck process, namely

$$\mathrm{d}\theta_t = -a\theta_t\mathrm{d}t + b\,\mathrm{d}\hat{W}_t, \tag{125}$$

where $\hat{W}$ is a standard Brownian motion and $a > 0$ and $b > 0$ are constants. Then the spectral density is given by $\varSigma_\theta(\omega) = (1/\pi)b^2/(a^2+\omega^2)$, and by a calculation from p. 130 of [3], one has

$$-\frac{1}{4\pi}\int_{-\infty}^{\infty}\ln[1-4\pi c\varSigma_\theta(\omega)]\mathrm{d}\omega = \frac{1}{2}a - \frac{1}{2}\sqrt{a^2-4b^2c} \tag{126}$$

provided $c < a^2/(4b^2)$. Thus

$$\frac{\alpha-1}{\alpha}E(Q) - \frac{1}{\alpha}r(\alpha) \le E(P) \le \frac{\alpha}{\alpha-1}E(Q) + \frac{1}{\alpha-1}r(\alpha) \tag{127}$$

where

$$r(\alpha) = \frac{1}{2}a - \frac{1}{2}\sqrt{a^2 - 4b^2 p\alpha(\alpha-1)}, \tag{128}$$

provided $c(\alpha) = p\alpha(\alpha-1) < a^2/(4b^2)$ and $p < a^2/(4b^2)$.

While it is hard to optimize over α, it is possible to do so if we bound r from above by

$$\bar{r}(\alpha) = \frac{1}{2}a - \frac{1}{2}\sqrt{a^2 - 4b^2 p\alpha^2} \tag{129}$$

and assume $p\alpha^2 \le a^2/(4b^2)$. That is, $\alpha \in (1, a/(2b\sqrt{p}))$. In particular, we must assume $a > 2b\sqrt{p}$. We therefore have from (127)

$$E(P) \le E_U(\alpha) := \frac{\alpha}{\alpha-1}E(Q) + \frac{1}{\alpha-1}\bar{r}(\alpha). \tag{130}$$

The minimum of this upper bound over all α in that range can be computed. Indeed, note that, as $\alpha \to 1$ from the right, $E_U(\alpha) \to \infty$. Moreover, the derivative of E_U, that is given by

$$E'_U(\alpha) = -\frac{E(Q)}{(\alpha-1)^2} - \frac{1}{(\alpha-1)^2}\left[\frac{1}{2}a - \frac{1}{2}\sqrt{a^2 - 4b^2p\alpha^2}\right] + \frac{1}{\alpha-1}\frac{2b^2p\alpha}{\sqrt{a^2 - 4b^2p\alpha^2}}, \tag{131}$$

tends to ∞ as $\alpha \to a/(2b\sqrt{p})$ from the left. As a result, and since the equation $E'_U(\alpha) = 0$ turns out to have a unique root α^* in that range, the minimizing α must be equal to α^*. This unique root is given by

$$\alpha^* = \frac{a^2\gamma_2 + a\gamma_1\sqrt{\gamma_1^2 + \gamma_2^2 - a^2}}{\gamma_2(\gamma_1^2 + \gamma_2^2)}, \tag{132}$$

where

$$\gamma_1 = a + 2E(Q), \quad \gamma_2 = 2b\sqrt{p}. \tag{133}$$

With this notation, the optimal upper bound of the form (130) is given by

$$E_U = E_U(\alpha^*) = \frac{2\alpha^* E(Q) + a - \sqrt{a^2 - \gamma_2^2(\alpha^*)^2}}{2(\alpha^*-1)}. \tag{134}$$

As $b \to 0$, we have $\alpha^* \to \infty$ and as a consequence $E_U \to E(Q)$. That is, we recover the exponent $E(Q)$ as the fading intensity tends to zero.

As for a corresponding lower bound, we have

$$E(P) \ge E_L(\alpha) := \frac{\alpha-1}{\alpha}E(Q) - \frac{1}{\alpha}\bar{r}(\alpha). \tag{135}$$

A calculation shows that the maximizing α is

$$\hat{\alpha} = \sqrt{\left(\frac{a}{\gamma_2}\right)^2 - \left(\frac{a^2}{\gamma_1\gamma_2}\right)^2} \tag{136}$$

and so

$$E_L = E_L(\hat{\alpha}) = \frac{2\gamma_1(\hat{\alpha}-1)E(Q) - a\gamma_1 + a^2}{2\gamma_1\hat{\alpha}}. \tag{137}$$

As $b \to 0$ we have $\hat{\alpha} \to \infty$ and so $E_L \to E(Q)$.

4.5 Binary channel with erasure

We next consider the channel

$$Y_t = (a_t X_t) \oplus N_t, \tag{138}$$

where N_t is i.i.d. noise whereas a_t is an erasure process. Here, a_t, X_t and N_t take values in $\{0,1\}$ and $\oplus$ denotes addition modulo 2. It is assumed that $\{a_t\}$ and $\{N_t\}$ are mutually independent. We let $p = P(N_1 = 1)$ and assume $p \le 1/2$. The first model we examine for $\{a_t\}$ is a hidden Markov model (an additional model appears afterwards). Specifically, we let $\{A_t\}$ be a stationary Markov process on the state space $\{1,\dots,d\}$ (independent of $(\{X_t\},\{N_t\})$) with a given transition probability matrix $\mathit{\Pi}$, assumed to be irreducible. For a given function $f\{1,\dots,d\} \to \{0,1\}$, a is given by $a_t = f(A_t)$, $t = 1,\dots,n$. Denote by P the probability measure induced by the above processes. Let Q denote a reference probability measure, under which

$$Y_t = X_t \oplus \tilde{N}_t, \tag{139}$$

where, for each n, the law of the triplet $(\boldsymbol{X}, \boldsymbol{a}, \tilde{\boldsymbol{N}})$ is the same as that of $(\boldsymbol{X}, \boldsymbol{a}, \boldsymbol{N})$ under P (in particular, the three are mutually independent under Q).

To calculate the Rényi divergence, note that

$$P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{A}) = P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{A})P(\boldsymbol{X})P(\boldsymbol{A}) = \Big[\prod_{t=1}^{n} P(Y_t|a_t X_t)\Big]P(\boldsymbol{X})P(\boldsymbol{A}), \tag{140}$$

where for $(x,y) \in \{0,1\}^2$, $P(y|x) = p$ if $y \ne x$ and $P(y|x) = 1-p$ if $y = x$. Also,

$$Q(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{A}) = \Big[\prod_{t=1}^{n} P(Y_t|X_t)\Big]P(\boldsymbol{X})P(\boldsymbol{A}). \tag{141}$$

Denoting by P_n and Q_n the respective laws of $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{A})$, we have

$$\begin{aligned}
\frac{\alpha}{n} D_\alpha(Q_n \| P_n) &= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\Big(\prod_{t=1}^{n} \frac{P(Y_t|X_t)}{P(Y_t|a_t X_t)}\Big)^{\alpha}\Big] && (142)\\
&= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\prod_{t:a_t=0,X_t=1} \Big(\frac{P(Y_t|X_t)}{P(Y_t|a_t X_t)}\Big)^{\alpha}\Big] && (143)\\
&= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P \prod_{t:a_t=0,X_t=1} \Big[p\Big(\frac{1-p}{p}\Big)^{\alpha} + (1-p)\Big(\frac{p}{1-p}\Big)^{\alpha}\Big] && (144)\\
&\le \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P \prod_{t:a_t=0} \delta(\alpha) && (145)\\
&= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\delta(\alpha)^{n-\sum_{t=1}^{n} a_t}\Big], && (146)
\end{aligned}$$

where we use the fact that $\delta(\alpha) := p(\frac{1-p}{p})^{\alpha} + (1-p)(\frac{p}{1-p})^{\alpha} \ge 1$. Let $\bar{f}(i) = 1 - f(i)$, $i = 1,\dots,d$, and denote by

$$Z_n = \frac{1}{n}\sum_{t=1}^{n} 1_{\{a_t=0\}} = \frac{1}{n}\sum_{t=1}^{n} \bar{f}(A_t) \tag{147}$$

the frequency of times t when $a_t = 0$. Then we can write the above as

$$\frac{\alpha}{n} D_\alpha(Q_n \| P_n) \leq \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P[e^{nZ_n \ln \delta(\alpha)}]. \tag{148}$$

By similar considerations one obtains

$$\frac{\alpha-1}{n} D_\alpha(P_n \| Q_n) \leq \frac{1}{n\alpha} \ln \boldsymbol{E}_P[e^{nZ_n \ln \delta(\alpha)}]. \tag{149}$$

For $\lambda \in \mathbb{R}$, let $\Pi_\lambda = (\pi_\lambda(i,j))$, where

$$\pi_\lambda(i,j) = \pi(i,j) e^{\lambda \bar{f}(j)}, \qquad i,j \in \{1,\ldots,d\}. \tag{150}$$

Then Π_λ is an irreducible matrix for every λ and, by the Perron-Frobenius theorem, has a real positive eigenvalue, denoted by $\rho(\Pi_\lambda)$, that dominates all eigenvalues in absolute value. It is known that the random variables Z_n satisfy the large deviation principle with the good rate function $I : \mathbb{R} \to [0,\infty]$, defined as

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \ln \rho(\Pi_\lambda)\} \tag{151}$$

(for the terminology see [5]; for the above result see Theorem 3.1.2 therein). Thus by Varadhan's lemma (Theorem 4.3.1 of [5]), it follows that

$$\lim_{n\to\infty} \frac{1}{n} \ln \boldsymbol{E}_P[e^{nZ_n \ln \delta(\alpha)}] = \sup_{x \in \mathbb{R}} [x \ln \delta(\alpha) - I(x)]. \tag{152}$$

We thus have

Theorem 4.5 *For Q the binary channel and P the binary channel with erasure described above, for every $\alpha > 1$,*

$$E(P) \leq \frac{\alpha}{\alpha-1} E(Q) + \frac{1}{\alpha-1} \sup_{x \in \mathbb{R}} [x \ln \delta(\alpha) - I(x)], \tag{153}$$

and

$$E(P) \geq \frac{\alpha-1}{\alpha} E(Q) - \frac{1}{\alpha} \sup_{x \in \mathbb{R}} [x \ln \delta(\alpha) - I(x)]. \tag{154}$$

Bounded fraction of erasures

We now examine another model for the erasure process $\{a_t\}$. In this model, the erasure process satisfies a single hard constraint, namely that the relative number of erasures $Z_n = \frac{1}{n}\sum_{t=1}^n 1_{\{a_t=0\}}$ is a.s.-bounded. Specifically, for some constant $z \in [0,1]$, it is assumed that $Z_n \leq z$ a.s., for every n. To relate this to the previous model, note that this may occur when the (stationary, Markov) process A_t taking values in $\{1,\ldots,d\}$ is cyclic, and where the subset $S \subset \{1,\ldots,d\}$ of states corresponding to erasure has cardinality k with $k/d \leq z$. Of course, the class of processes a satisfying the current assumption is much broader.

Note that (148) and (149) are valid. As a result, we obtain in this case, for $\alpha > 1$,

$$\frac{\alpha-1}{\alpha} E(Q) - \frac{1}{\alpha} z \ln \delta(\alpha) \leq E(P) \leq \frac{\alpha}{\alpha-1} E(Q) + \frac{1}{\alpha-1} z \ln \delta(\alpha). \tag{155}$$

Clearly, this model has a property analogous to that established for the channel with fading, namely that as $z \to 0$, both bounds converge (upon optimization with respect to α) to $E(Q)$.

5 Other applications

5.1 Rate–distortion coding

Consider the problem of rate–distortion coding of a source sequence $Y_1, Y_2, \ldots$ given by

$$Y_t = X_t + Z_t, \tag{156}$$

where, under the probability measure P, $\{X_t\}$ is an i.i.d., $\mathcal{N}(0, \sigma^2)$ process and $\{Z_t\}$ is a process that is independent of $\{X_t\}$. For simplicity, assume the random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ has density, denoted f_Z. Each source sequence $\boldsymbol{y} = (y_1, \ldots, y_n)$ is compressed to a string of nR nats, from which the decoder reconstructs an approximated sequence $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$. We are interested in a lower bound on

$$P(\mathcal{E}_{n,d}), \qquad \mathcal{E}_{n,d} := \Big\{ \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 > nd \Big\}, \tag{157}$$

where d is large enough so that this probability decays exponentially.

The joint density of $(\boldsymbol{Y}, \boldsymbol{Z})$ under P is thus given by $g(\boldsymbol{y} - \boldsymbol{z}) f_Z(\boldsymbol{z})$, where g is the i.i.d. $\mathcal{N}(0, \sigma^2))$ density. We consider a reference measure Q, under which the joint density of $(\boldsymbol{Y}, \boldsymbol{Z})$ is $g(\boldsymbol{y}) f_Z(\boldsymbol{z})$. Since the event $\mathcal{E}_{n,d}$ is measurable on the sigma-field of $\boldsymbol{Y}$, and under Q, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are mutually, independent, the law of $\boldsymbol{Z}$ is irrelevant for the estimation of $Q(\mathcal{E}_{n,d})$, in the sense that $Q(\mathcal{E}_{n,d}) = G(\mathcal{E}_{n,d})$, where we denote by G the law of $\boldsymbol{Y}$ under Q (equivalently, that of $\boldsymbol{X}$ under P). In the appendix, we show that

$$\liminf_{n \to \infty} \frac{\ln G(\mathcal{E}_{n,d})}{n} \geq -\Phi[R - R_G(d)], \tag{158}$$

where $R_G(d) = \frac{1}{2} \ln \frac{\sigma^2}{d}$ is the rate–distortion function of the Gaussian source $\{X_t\}$ and

$$\Phi(u) \stackrel{\triangle}{=} \frac{e^{2u} - 1}{2} - u. \tag{159}$$

We now calculate the divergence term. With P_n and Q_n denoting the respective laws of $(\boldsymbol{Y}, \boldsymbol{Z})$,

$$\begin{aligned} \frac{\alpha}{n} D_\alpha(Q_n \| P_n) &= \frac{1}{n(\alpha - 1)} \ln \left[\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{z} f_Z(\boldsymbol{z}) \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{y} \cdot g^\alpha(\boldsymbol{y}) g^{1-\alpha}(\boldsymbol{y} - \boldsymbol{z}) \right] \\ &= \frac{1}{n(\alpha - 1)} \ln \left[\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{z} f_Z(\boldsymbol{z}) \exp \left\{ \frac{\alpha(\alpha - 1) \|\boldsymbol{z}\|^2}{2\sigma^2} \right\} \right], \end{aligned} \tag{160}$$

where the second step follows by appealing to identity (65). The usefulness of the bound will now depend on estimating the last expression. Obviously, for this expression to be finite, the tails of f_Z must decay faster than those of a Gaussian.

Consider the, for example, the case where $\sum_{t=1}^n Z_t^2 \leq nA^2$ almost surely. In this case, the right-hand side of (160) is bounded by $\frac{\alpha A^2}{2\sigma^2}$. Using this bound together with (158) in (17) gives

$$E(R, P) \leq \inf_{\alpha \geq 1} \left[\frac{\alpha \Phi[R - R_G(d)]}{\alpha - 1} + \frac{\alpha A^2}{2\sigma^2} \right] = \left(\sqrt{\Phi[R - R_G(d)]} + \frac{A}{\sqrt{2}\sigma} \right)^2. \tag{161}$$

In a similar way, one obtains

$$E(R,P) \geq \left(\sqrt{\Phi[R-R_G(d)]} - \frac{A}{\sqrt{2}\sigma}\right)^2. \tag{162}$$

An analogous derivation can be made for the case where $\{X_t\}$ is a binary memoryless source with parameter p, $\{Z_t\}$ is a binary interference with normalized Hamming weight limited by A, and $Y_t = X_t \oplus Z_t$. We then end up with

$$E(R) \leq \inf_{\alpha>1}\left[\frac{\alpha F(R,D)}{\alpha-1} + \frac{A\ln[p^\alpha(1-p)^{1-\alpha}+(1-p)^\alpha p^{1-\alpha}]}{\alpha-1}\right], \tag{163}$$

where $F(R,D)$ is the source coding error exponent [13] associated with $\{X_t\}$.

5.2 Extension to a pair of sources

A possible extension of this example is associated with the problem of separate encodings and joint decoding of correlated sources. Let $\{(X_i,Y_i)\}_{i=1}^n$ be n independent copies of a random pair (X,Y) distributed according to $P_{XY}(x,y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. The sequences $\{X_i\}$ and $\{Y_i\}$ are compressed separately by two encoders (that do not cooperate) at rates R_x and R_y, respectively. The respective compressed bit–streams are both fed into a joint decoder that produces reconstructions $\{\hat{X}_i\}$ and $\{\hat{Y}_i\}$, whose components take on values in alphabets $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$, respectively. Let $\rho_x : \mathcal{X}\times\hat{\mathcal{X}} \to \mathrm{I\!R}$ and $\rho_y : \mathcal{Y}\times\hat{\mathcal{Y}} \to \mathrm{I\!R}$ be given distortion functions. We are interested in a lower bound on

$$P\left\{\sum_{i=1}^n \rho_x(X_i,\hat{X}_i) \geq nd_x,\ \sum_{i=1}^n \rho_y(Y_i,\hat{Y}_i) \geq nd_y\right\} \tag{164}$$

for some prescribed distortion levels d_x and d_y. We wish to pass to a reference source for which $\{X_i\}$ and $\{Y_i\}$ are statistically independent, that is, $Q_{XY}(x,y) = Q_X(x)Q_Y(y)$. Under Q, the probability of the above event decays exponentially at rate $F_x^Q(R_x,d_x) + F_y^Q(R_y,d_y)$, where F_x^Q and F_y^Q are the source coding exponents of the separate reference sources, Q_X and Q_Y, respectively. Thus, our upper bound on the exponent is given by

$$\begin{aligned}&E(R_x,R_y,d_x,d_y)\\ &\quad\leq \inf_{\alpha>1}\inf_{Q_X,Q_Y}\left\{\frac{\alpha}{\alpha-1}[F_x^Q(R_x,d_x)+F_y^Q(R_y,d_y)] + \alpha D_\alpha(Q_X\times Q_Y\|P_{XY})\right\}.\end{aligned} \tag{165}$$

In this setting, to the best of our knowledge, there does not exist any competing bound in the literature.

5.3 The problem of guessing

Let $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$ be a random vector with a given distribution. Let $\hat{\boldsymbol{Y}}_1, \hat{\boldsymbol{Y}}_2,\ldots$ be a sequence of 'guesses' of the random vector $\boldsymbol{Y}$ that is generated without observing $\boldsymbol{Y}$. within distortion d from $\boldsymbol{y}$, Denoting by ρ the Hamming distance and fixing a distortion level $d \geq 0$, let $\mathit{\Gamma}(\boldsymbol{Y})$ denote the number of trials it takes to correctly guess $\boldsymbol{Y}$ within distortion level d, i.e.,

$$\mathit{\Gamma}(\boldsymbol{Y}) = \min\{i: \ \rho(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_i) \le nd\}. \tag{166}$$

In [1], it was shown that for a given discrete memoryless source Q and a given parameter $\lambda > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \ln \boldsymbol{E}_Q\{\mathit{\Gamma}(\boldsymbol{Y})^\lambda\} \ge \sup_{\hat{Q}_1}[\lambda R(d, \hat{Q}_1) - D(\hat{Q}_1 \| Q_1)], \tag{167}$$

where Q_1 denotes the marginal of Q, $R(\cdot, \hat{Q}_1)$ denotes the rate–distortion function of the source $\hat{Q}_1$, and the supremum is over $\hat{Q}$ in the set of probability measures over the alphabet of Y_1.

Using the comparison bounds, we can estimate this quantity for a more general model. Specifically, consider the model discussed at the end of Subsection 5.1. Namely, Y_t is binary and takes the form $Y_t = X_t \oplus Z_t$, where, under a probability measure P, $\{X_t\}$ and $\{Z_t\}$ are mutually independent, and X_t are i.i.d. with parameter p. Assuming that the normalized number of times t when $Z_t = 1$ is bounded by a constant A, the Rényi divergence term is bounded by

$$\frac{\alpha}{n} D_\alpha(Q_n \| P_n) \le \frac{A}{\alpha - 1} \ln[p^\alpha(1-p)^{1-\alpha} + (1-p)^\alpha p^{1-\alpha}], \tag{168}$$

where as before, P_n and Q_n are the respective laws of $(\boldsymbol{Y}, \boldsymbol{Z})$. We can now appeal to (12). Using this inequality (with the roles of P and Q interchanged), we have for arbitrary $\alpha > 1$ and denoting $s = (\alpha - 1)/\alpha$,

$$\frac{1}{n} \ln \boldsymbol{E}_P\{\mathit{\Gamma}(\boldsymbol{Y})^\lambda\} \ge \frac{\alpha}{n(\alpha-1)} \ln \boldsymbol{E}_Q\{\mathit{\Gamma}(\boldsymbol{Y})^{s\lambda}\} - \frac{\alpha}{n} D_\alpha(Q_n \| P_n). \tag{169}$$

Using (167) and (168) in (169) gives

$$\begin{aligned} \liminf_{n\to\infty} & \frac{1}{n} \ln \boldsymbol{E}_P\{\mathit{\Gamma}(\boldsymbol{Y})^\lambda\} \\ & \ge \sup_{\alpha > 1} \Big\{ \sup_{\hat{Q}_1} \Big[\lambda R(d, \hat{Q}_1) - \frac{\alpha}{\alpha - 1} D(\hat{Q}_1 \| Q_1) \Big] \\ & \qquad - \frac{A}{\alpha - 1} \ln[p^\alpha(1-p)^{1-\alpha} + (1-p)^\alpha p^{1-\alpha}] \Big\}. \end{aligned} \tag{170}$$

Acknowledgment. We are grateful to an anonymous referee for thoughtful comments, and to another anonymous referee for various valuable comments, most notably, for pointing out the relation of the LPCB to the data processing inequality, as well as for kindly providing the proof that appears in the second part of Appendix A.1.

A Appendix

A.1 Two proofs of the LPCB

We provide two proofs of the LPCB. The first is for the version that allows expectations with respect to a general function as in (8), but is limited to the case where the support is a finite set (the reader is referred to [2] for the general setting). The goal of presenting it here is to show that the argument is very simple in this setting.

The second proof is of (9) (i.e., for comparing probabilities). Our goal here is to show that this inequality is an immediate consequence of the *data processing inequality* for the Rényi divergence.

Let $\mathcal{X}$ be a finite set, let $\{P_i\}_{i\in\mathcal{X}}$ and $\{Q_i\}_{i\in\mathcal{X}}$ be two probability distributions defined on it and let $G:\mathcal{X}\to[0,\infty)$ be a given function.

Proposition A.1 *Assume $\sum_{i\in\mathcal{X}} G_iP_i>0$ and $\sum_{i\in\mathcal{X}} G_iQ_i>0$. Then for all $\alpha>1$*

$$\frac{1}{\alpha-1}\ln\sum_i G_i^{\alpha-1}P_i\le\frac{1}{\alpha}\ln\sum_i G_i^{\alpha}Q_i+D_\alpha(P\|Q). \tag{A.1}$$

Moreover, given P, G and α as above, there exists Q for which (A.1) *holds with equality.*

Proof: When one does not have $P\ll Q$, the divergence term above equals $+\infty$ by definition, and there is nothing to prove. Hence assume $P\ll Q$. Denote by S_P, S_Q and S_G the support of P, Q and G, respectively. Let $S=S_Q\cap S_G$. Using Hölder's inequality with the exponents α and $\alpha/(\alpha-1)$ and measure $\{Q_i\}_{i\in S}$,

$$\begin{aligned}\sum_S G_i^{\alpha-1}P_i &= \sum_S\frac{P_i}{Q_i}G_i^{\alpha-1}Q_i && \text{(A.2)}\\ &\le\Big(\sum_S\Big(\frac{P_i}{Q_i}\Big)^{\alpha}Q_i\Big)^{1/\alpha}\Big(\sum_S\Big(G_i^{\alpha-1}\Big)^{\alpha/(\alpha-1)}Q_i\Big)^{(\alpha-1)/\alpha} && \text{(A.3)}\\ &=\Big(\sum_S\Big(\frac{P_i}{Q_i}\Big)^{\alpha}Q_i\Big)^{1/\alpha}\Big(\sum_S G_i^{\alpha}Q_i\Big)^{(\alpha-1)/\alpha}. && \text{(A.4)}\end{aligned}$$

Thus

$$\begin{aligned}\Big(\sum_S G_i^{\alpha-1}P_i\Big)^{\alpha}\Big(\sum_S G_i^{\alpha}Q_i\Big)^{(1-\alpha)} &\le\sum_S\Big(\frac{P_i}{Q_i}\Big)^{\alpha}Q_i && \text{(A.5)}\\ &\le\sum_{S_Q}\Big(\frac{P_i}{Q_i}\Big)^{\alpha}Q_i. && \text{(A.6)}\end{aligned}$$

For i not in S, $G_i^\alpha Q_i=0$, and because $P\ll Q$, also $G_i^{\alpha-1}P_i=0$. Thus, on the left-hand side, the summation can be performed over all of $\mathcal{X}$. As a result, taking logarithms and dividing by $\alpha(\alpha-1)$, using the definition of the divergence (4) gives the inequality (A.1). To show the final assertion set $Q_i=G_i^{-1}P_i/Z$ for $i\in S_P\cap S_G$ and 0 off of that set. Here, $Z=\sum_{i\in S_P\cap S_G}G_i^{-1}P_i>0$ by assumption. Substituting in (A.1) gives equality by a direct calculation. □

Proof of the LPCB via the data processing inequality

Let

$$d_\alpha(p\|q)=\frac{1}{\alpha(\alpha-1)}\ln\Big[q\Big(\frac{p}{q}\Big)^{\alpha}+(1-q)\Big(\frac{1-p}{1-q}\Big)^{\alpha}\Big], \tag{A.7}$$

for $p \in [0,1]$ and $q \in (0,1)$. Then, given an event $\mathcal{A} \in \mathcal{F}$, denoting $p = P(\mathcal{A}) > 0$ and $q = Q(\mathcal{A}) > 0$, we have

$$D_\alpha(P\|Q) \geq d_\alpha(p\|q) \geq \frac{1}{\alpha(\alpha-1)} \ln \frac{p^\alpha}{q^{\alpha-1}} = \frac{1}{\alpha-1} \ln P(\mathcal{A}) - \frac{1}{\alpha} \ln Q(\mathcal{A}), \tag{A.8}$$

where the first inequality is due to the data processing inequality for the Rényi divergence (see eg. [19, Section II], for a proof see [12, Theorem 1.24 and Corollary 1.29]) and the second uses the monotonicity of the logarithmic function (recall $\alpha > 1$). This given (9).

A.2 Proof of Theorem 4.1

A bound on the divergence between any two univariate Gaussians is deduced from identity (65) as follows. Given $x \in \mathbb{R}$, $\xi \in \mathbb{R}$ such that $|\xi| \leq \mathit{\Gamma}$, and any $\alpha > 1$ and $s > 1 - 1/\alpha$,

$$\begin{aligned}
&D_\alpha(\mathcal{N}(x, \sigma^2/s)\|\mathcal{N}(x+\xi, \sigma^2)) \\
&= \frac{1}{\alpha(\alpha-1)} \ln \left[\frac{s^{\alpha/2}}{\sqrt{1+\alpha(s-1)}} \cdot \exp\left\{ -\frac{\alpha(1-\alpha)s\xi^2}{2\sigma^2[1+\alpha(s-1)]} \right\} \right] \\
&= \frac{1}{\alpha(\alpha-1)} \left\{ \frac{\alpha \ln s}{2} - \frac{\ln[1+\alpha(s-1)]}{2} + \frac{\alpha(\alpha-1)s\xi^2}{2\sigma^2[1+\alpha(s-1)]} \right\} \\
&= \frac{\ln s}{2(\alpha-1)} - \frac{\ln[1+\alpha(s-1)]}{2\alpha(\alpha-1)} + \frac{s\xi^2}{2\sigma^2[1+\alpha(s-1)]} \\
&\leq \frac{\ln s}{2(\alpha-1)} - \frac{\ln[1+\alpha(s-1)]}{2\alpha(\alpha-1)} + \frac{s\mathit{\Gamma}^2}{2\sigma^2[1+\alpha(s-1)]}.
\end{aligned} \tag{A.9}$$

Let P_n and Q_n denote the respective probability laws of $(\boldsymbol{X}, \boldsymbol{Y})$. Then

$$\begin{aligned}
D_\alpha(Q_n\|P_n) &= \frac{1}{\alpha(\alpha-1)} \ln \boldsymbol{E}_P \Big[\Big(\frac{Q(\boldsymbol{X},\boldsymbol{Y})}{P(\boldsymbol{X},\boldsymbol{Y})} \Big)^\alpha \Big] && \text{(A.10)} \\
&= \frac{1}{\alpha(\alpha-1)} \ln \sum_{\boldsymbol{x}} \int \mathrm{d}\boldsymbol{y} \Big(\frac{Q(\boldsymbol{y}|\boldsymbol{x})}{P(\boldsymbol{y}|\boldsymbol{x})} \Big)^\alpha P(\boldsymbol{y}|\boldsymbol{x}) \Pi(\boldsymbol{x}) && \text{(A.11)} \\
&= \frac{1}{\alpha(\alpha-1)} \ln \sum_{\boldsymbol{x}} \Pi(\boldsymbol{x}) \left[\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{y} (2\pi\sigma^2/s)^{-\alpha n/2} (2\pi\sigma^2)^{-(1-\alpha)n/2} \times \right. \\
&\quad \left. \exp\left\{ -\frac{s\alpha}{2\sigma^2} \sum_{t=1}^n (y_t - x_t)^2 \right\} \cdot \exp\left\{ -\frac{1-\alpha}{2\sigma^2} \sum_{t=1}^n [y_t - x_t - g_t(x^n, y^{t-1})]^2 \right\} \right] \\
&= \frac{n \ln s}{2(\alpha-1)} - \frac{n \ln(2\pi\sigma^2)}{2\alpha(\alpha-1)} + \frac{1}{\alpha(\alpha-1)} \ln \sum_{\boldsymbol{x}} \Pi(\boldsymbol{x}) \left[\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{y} \times \right. \\
&\quad \left. \exp\left\{ -\sum_{t=1}^n \left(\frac{s\alpha}{2\sigma^2} [y_t - x_t]^2 + \frac{1-\alpha}{2\sigma^2} [y_t - x_t - g_t(x^n, y^{t-1})]^2 \right) \right\} \right] && \text{(A.12)} \\
&\stackrel{\triangle}{=} \frac{n \ln s}{2(\alpha-1)} - \frac{n \ln(2\pi\sigma^2)}{2\alpha(\alpha-1)} + \frac{1}{\alpha(\alpha-1)} \cdot Z_n && \text{(A.13)}
\end{aligned}$$

Let us focus on the expression of Z_n. For $t \in \{1,\ldots,n\}$ let $\Pi(\xi^t) = \sum_{x^n : x^t = \xi^t} \Pi(x^n)$. Then

$$\begin{aligned}
Z_n &= \ln \sum_{\boldsymbol{x}} \Pi(\boldsymbol{x}) \left[\int_{\mathrm{I\!R}^{n-1}} \mathrm{d}y^{n-1} \times \right. \\
&\quad \exp\left\{ -\sum_{t=1}^{n-1} \left(\frac{s\alpha}{2\sigma^2}[y_t - x_t]^2 + \frac{1-\alpha}{2\sigma^2}[y_t - x_t - g_t(x^n, y^{t-1})]^2 \right) \right\} \times \\
&\quad \left. \int_{\mathrm{I\!R}} \mathrm{d}y_n \exp\left\{ -\left(\frac{s\alpha}{2\sigma^2}(y_n - x_n)^2 + \frac{1-\alpha}{2\sigma^2}[y_n - x_n - g_n(x^n, y^{n-1})]^2 \right) \right\} \right] \qquad \text{(A.14)} \\
&= \ln \sum_{\boldsymbol{x}} \Pi(\boldsymbol{x}) \left[\int_{\mathrm{I\!R}^{n-1}} \mathrm{d}y^{n-1} \times \right. \\
&\quad \exp\left\{ -\sum_{t=1}^{n-1} \left(\frac{s\alpha}{2\sigma^2}[y_t - x_t]^2 + \frac{1-\alpha}{2\sigma^2}[y_t - x_t - g_t(x^n, y^{t-1})]^2 \right) \right\} \times \\
&\quad \left. \sqrt{\frac{2\pi\sigma^2}{1+\alpha(s-1)}} \exp\left\{ \frac{s\alpha(\alpha-1) g_n^2(x^n, y^{n-1})}{2\sigma^2[1+\alpha(s-1)]} \right\} \right] \qquad \text{(A.15)} \\
&\le \ln \sum_{x^{n-1}} \Pi(x^{n-1}) \left[\int_{\mathrm{I\!R}^{n-1}} \mathrm{d}y^{n-1} \times \right. \\
&\quad \left. \exp\left\{ -\sum_{t=1}^{n-1} \left(\frac{s\alpha}{2\sigma^2}[y_t - x_t]^2 + \frac{1-\alpha}{2\sigma^2}[y_t - x_t - g_t(x^n, y^{t-1})]^2 \right) \right\} \right] + \\
&\quad \frac{1}{2} \ln \left[\frac{2\pi\sigma^2}{1+\alpha(s-1)} \right] + \frac{s\alpha(\alpha-1) \max_{x^n, y^{n-1}} g_n^2(x^n, y^{n-1})}{2\sigma^2[1+\alpha(s-1)]} \qquad \text{(A.16)} \\
&\le Z_{n-1} + \frac{1}{2} \ln \left[\frac{2\pi\sigma^2}{1+\alpha(s-1)} \right] + \frac{s\alpha(\alpha-1)\mathit{\Gamma}_n^2}{2\sigma^2[1+\alpha(s-1)]}. \qquad \text{(A.17)}
\end{aligned}$$

From this recursion on Z_n, we have

$$\begin{aligned}
Z_n &\le \frac{n}{2} \ln \left[\frac{2\pi\sigma^2}{1+\alpha(s-1)} \right] + \frac{s\alpha(\alpha-1)\sum_{t=1}^n \mathit{\Gamma}_t^2}{2\sigma^2[1+\alpha(s-1)]} \qquad \text{(A.18)} \\
&\le \frac{n}{2} \ln \left[\frac{2\pi\sigma^2}{1+\alpha(s-1)} \right] + \frac{ns\alpha(\alpha-1)\mathit{\Gamma}^2}{2\sigma^2[1+\alpha(s-1)]}. \qquad \text{(A.19)}
\end{aligned}$$

Therefore

$$\begin{aligned}
&D_\alpha(Q_n \| P_n) \qquad \text{(A.20)} \\
&= \frac{n \ln s}{2(\alpha-1)} - \frac{n \ln(2\pi\sigma^2)}{2\alpha(\alpha-1)} + \frac{1}{\alpha(\alpha-1)} \cdot Z_n \\
&\le \frac{n \ln s}{2(\alpha-1)} - \frac{n \ln(2\pi\sigma^2)}{2\alpha(\alpha-1)} + \frac{1}{\alpha(\alpha-1)} \left\{ \frac{n}{2} \ln \left[\frac{2\pi\sigma^2}{1+\alpha(s-1)} \right] + \frac{ns\alpha(\alpha-1)\mathit{\Gamma}^2}{2\sigma^2[1+\alpha(s-1)]} \right\} \\
&= \frac{n \ln s}{2(\alpha-1)} - \frac{n \ln[1+\alpha(s-1)]}{2\alpha(\alpha-1)} + \frac{ns\mathit{\Gamma}^2}{2\sigma^2[1+\alpha(s-1)]}. \qquad \text{(A.21)}
\end{aligned}$$

Substituting in (35), using the bound $E(R, Q_s, d) \le E_{\mathrm{sl}}(R, Q_s)$ for every d, and finally optimizing over s and α, yields (64). □

A.3 Proof of Theorem 4.2

As a reference, we will use the models $Q = Q_{\phi,\theta}$, under which

$$Y_t = \phi X_t + \tilde{W}_t, \tag{A.22}$$

where $\{\tilde{W}_t\}$ are i.i.d. $\mathcal{N}(0, \theta)$, independent of $\{X_t\}$. Here, $\phi > 0$ and $\theta > 0$ are parameters. Note that, for each of the models Q, d is the optimal decoding metric. One has

$$P(\boldsymbol{y}|\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \prod_{t=1}^{n} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_t - x_t - \sum_{i=1}^{k} h_i x_{t-1} \right)^2 \right\}, \tag{A.23}$$

and $Q(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} Q(y_t|x_t)$, where $Q(y|x)$ is given by

$$Q(y|x) = \sqrt{\frac{\theta}{\pi}} \cdot \exp\{-\theta(y - \phi x)^2\}, \quad \theta > 0, \ \phi > 0. \tag{A.24}$$

In order to calculate the Rényi divergence, we use the identity (65) with the assignments: $a = \alpha/(2\sigma^2)$, $b = (1-\alpha)\theta$, $u = x_t + \sum_{i=1}^{k} h_k x_{t-i}$ and $v = \phi x_t$, to get, under the assumption

$$a + b = \frac{\alpha}{2\sigma^2} + (1-\alpha)\theta > 0, \tag{A.25}$$

$$\begin{aligned}
& \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{y} \cdot \prod_{t=1}^{n} \exp\left\{ -\frac{\alpha}{2\sigma^2} \left(y_t - x_t - \sum_{i=1}^{k} h_i x_{t-i} \right)^2 - (1-\alpha)\theta(y_t - \phi x_t)^2 \right\} && \text{(A.26)} \\
= & \left[\frac{\pi}{(1-\alpha)\theta + \alpha/2\sigma^2} \right]^{n/2} \cdot \exp\left\{ -\frac{\alpha(1-\alpha)\theta \sum_t \left[(1-\phi)x_t + \sum_{i=1}^{k} h_i x_{t-i} \right]^2}{\alpha + 2(1-\alpha)\theta\sigma^2} \right\} && \text{(A.27)} \\
= & \left[\frac{2\pi\sigma^2}{\alpha + 2(1-\alpha)\theta\sigma^2} \right]^{n/2} \cdot \exp\left\{ -\frac{n\alpha(1-\alpha)\theta S[(1-\phi)^2 + 2(1-\phi)r_1 + r_2]}{\alpha + 2(1-\alpha)\theta\sigma^2} \right\}. && \text{(A.28)}
\end{aligned}$$

Therefore

$$\begin{aligned}
& \frac{1}{n} D_\alpha(P_n \| Q_n) \\
= & \frac{1}{n\alpha(\alpha-1)} \ln \left[\left(\frac{\theta}{\pi} \right)^{n(1-\alpha)/2} (2\pi\sigma^2)^{-n\alpha/2} \left(\frac{2\pi\sigma^2}{\alpha + 2(1-\alpha)\theta\sigma^2} \right)^{n/2} \right. \\
& \quad \left. \times \exp\left\{ -\frac{n\alpha(1-\alpha)\theta S[(1-\phi)^2 + 2(1-\phi)r_1 + r_2]}{\alpha + 2(1-\alpha)\theta\sigma^2} \right\} \right] \\
= & \frac{1}{n\alpha(\alpha-1)} \ln \left[\frac{(2\theta\sigma^2)^{n(1-\alpha)/2}}{(\alpha + 2(1-\alpha)\theta\sigma^2)^{n/2}} \cdot \exp\left\{ -\frac{n\alpha(1-\alpha)\theta S[(1-\phi)^2 + 2(1-\phi)r_1 + r_2]}{\alpha + 2(1-\alpha)\theta\sigma^2} \right\} \right] \\
= & -\frac{\ln(2\theta\sigma^2)}{2\alpha} - \frac{\ln(\alpha + 2(1-\alpha)\theta\sigma^2)}{2\alpha(\alpha-1)} + \frac{\theta S[(1-\phi)^2 + 2(1-\phi)r_1 + r_2]}{\alpha + 2(1-\alpha)\theta\sigma^2}. && \text{(A.29)}
\end{aligned}$$

For a code of rate zero operating over the reference channel Q, the best achievable exponent is known to be

$$E(0, Q, d) = \frac{S\theta\phi^2}{2}, \tag{A.30}$$

where we have used an extension of the zero–rate lower bound of [16], [17] that applies to codes with a given composition μ (see Sections 2 and 4 of [14]). Then we have a lower bound from (35), for $\alpha > 1$,

$$E(0, P, d) \geq \frac{\alpha - 1}{\alpha} E(0, Q, d) - (\alpha - 1)\bar{D}_\alpha(P\|Q). \tag{A.31}$$

Thus

$$\begin{aligned} E(0, P, d) \quad \geq \quad & \sup_{(\alpha,\theta,\phi)\in\mathcal{S}} \left[\frac{(\alpha - 1)S\theta\phi^2}{2\alpha} + \frac{(\alpha - 1)\ln(2\theta\sigma^2) + \ln(\alpha + 2(1 - \alpha)\theta\sigma^2)}{2\alpha} \right. \\ & \left. - \frac{(\alpha - 1)\theta S[(1 - \phi)^2 + 2(1 - \phi)r_1 + r_2]}{\alpha + 2(1 - \alpha)\theta\sigma^2} \right], \end{aligned} \tag{A.32}$$

where

$$\mathcal{S} = \left\{ (\alpha, \theta, \phi) : \ \alpha > 1, \ \theta < \frac{\alpha}{2(\alpha - 1)\sigma^2}, \ \phi > 0 \right\}.$$

The maximization over ϕ is simple since the objective is quadratic in ϕ. In particular, the part that depends on ϕ is of the form $A\phi^2 + B\phi$, where

$$A = -\frac{\theta S(\alpha - 1)(\alpha + 2(\alpha - 1)\theta\sigma^2)}{2\alpha(\alpha + 2(1 - \alpha)\theta\sigma^2)} < 0, \tag{A.33}$$

and

$$B = \frac{2\theta S(\alpha - 1)(1 + r_1)}{\alpha + 2(1 - \alpha)\theta\sigma^2}. \tag{A.34}$$

The maximum of $A\phi^2 + B\phi$ is

$$-\frac{B^2}{4A} = -\frac{2\theta\alpha(\alpha - 1)S(1 + r_1)^2}{4(\alpha - 1)^2\theta^2\sigma^4 - \alpha^2}, \tag{A.35}$$

and our lower bound becomes,

$$\begin{aligned} & \frac{(\alpha - 1)\ln(2\theta\sigma^2) + \ln(\alpha + 2(1 - \alpha)\theta\sigma^2)}{2\alpha} - \frac{(\alpha - 1)\theta S[1 + 2r_1 + r_2]}{\alpha + 2(1 - \alpha)\theta\sigma^2} \\ & \quad - \frac{2\theta\alpha(\alpha - 1)S(1 + r_1)^2}{4(\alpha - 1)^2\theta^2\sigma^4 - \alpha^2}. \end{aligned} \tag{A.36}$$

It would be more convenient to define $\theta = \tau\alpha/[2(\alpha - 1)\sigma^2]$, $\tau \in (0, 1)$, and to transform the parameter set from (α, θ) to (α, τ). Denoting

$$\varPhi(\alpha, \tau) := \frac{(\alpha - 1)\ln[\tau\alpha/(\alpha - 1)] + \ln[\alpha(1 - \tau)]}{2\alpha}, \tag{A.37}$$

the expression is then

$$\begin{aligned} &\varPhi(\alpha,\tau)+\frac{S\tau}{2\sigma^2}\Big[-\frac{1+2r_1+r_2}{1-\tau}+\frac{2(1+r_1)^2}{1-\tau^2}\Big] \\ &=\varPhi(\alpha,\tau)+\frac{S}{2\sigma^2}\Big[-\frac{\tau}{1-\tau}(r_2-r_1^2)+\frac{\tau}{1+\tau}(1+r_1)^2\Big], \end{aligned} \tag{A.38}$$

to be maximized over $(\tau,\alpha)\in(0,1)\times(1,\infty)$. Now the function $\varPhi(\alpha,\tau)$ is always non–positive (the maximum over $\tau\in(0,1)$ for a given α is zero) and it vanishes for $\alpha=1/(1-\tau)$ (hence this is the optimum choice of α). Thus, we are left with maximizing the second term of (A.38) over τ. Recall that $r_2\geq r_1^2$, and that $a=r_2-r_1^2\geq 0$ and $b=(1+r_1)^2$. The maximum is given by

$$E(0,P)\geq\begin{cases}\frac{S}{4\sigma^2}(\sqrt{b}-\sqrt{a})^2=\frac{S}{4\sigma^2}[|1+r_1|-\sqrt{r_2-r_1^2}]^2 & a\leq b,\\ 0 & \text{otherwise.}\end{cases} \tag{A.39}$$

This establishes the first inequality in (93).

In the case $a<b$, the maximizing τ is given by $(\sqrt{b}-\sqrt{a})^2/(b-a)\in(0,1)$. If we use this in the expression for the optimal α and θ, we obtain that the optimal θ is $1/(2\sigma^2)$ and the optimal ϕ^2 is given by $\phi^2=(\sqrt{b}+\sqrt{a})^2$. Thus under the selected reference model,

$$E(0,Q,d)=\frac{S\theta\phi^2}{2}=\frac{S}{4\sigma^2}(\sqrt{b}+\sqrt{a})^2. \tag{A.40}$$

By virtue of (42), this gives namely

$$\sup_d\inf_{P\in F}E(0,P,d)\leq\frac{S}{4\sigma^2}(\sqrt{b}+\sqrt{a})^2. \tag{A.41}$$

□

A.4 Proof of Theorem 4.3

To work with the upper bound, we compute the Rényi divergence term,

$$\frac{\alpha}{n}D_\alpha(Q_n\|P_n)=\frac{1}{n(\alpha-1)}\ln\boldsymbol{E}_P\Big[\Big(\frac{Q(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})}{P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})}\Big)^\alpha\Big]. \tag{A.42}$$

We have

$$P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})=P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})P(\boldsymbol{X})P(\boldsymbol{\theta})=\Big[\prod_{t=1}^n g(Y_t|(1+\theta_t)X_t)\Big]P(\boldsymbol{X})P(\boldsymbol{\theta}), \tag{A.43}$$

where $g(y|x)=(2\pi\sigma^2)^{-1/2}e^{-(y-x)^2/(2\sigma^2)}$, and

$$Q(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})=\Big[\prod_{t=1}^n g(Y_t|X_t)\Big]P(\boldsymbol{X})P(\boldsymbol{\theta}). \tag{A.44}$$

Thus

$$\frac{\alpha}{n} D_\alpha(Q_n \| P_n) = \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\Big(\prod_{t=1}^{n} \frac{g(Y_t|X_t)}{g(Y_t|(1+\theta_t)X_t)}\Big)^{\alpha}\Big] \tag{A.45}$$

$$= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\exp\Big\{\frac{\alpha}{2\sigma^2}\sum_{t=1}^{n}[\theta_t^2 X_t^2 - 2\theta_t X_t (Y_t - X_t)]\Big\}\Big] \tag{A.46}$$

$$= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\exp\Big\{\frac{\alpha}{2\sigma^2}\sum_{t=1}^{n}[-\theta_t^2 X_t^2 - 2\theta_t X_t W_t]\Big\}\Big] \tag{A.47}$$

$$= \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\exp\Big\{\frac{\alpha(\alpha-1)}{2\sigma^2}\sum_{t=1}^{n}\theta_t^2 X_t^2\Big\}\Big] \tag{A.48}$$

$$\leq \frac{1}{n(\alpha-1)} \ln \boldsymbol{E}_P\Big[\exp\Big\{\frac{\alpha(\alpha-1)A^2}{2\sigma^2}\sum_{t=1}^{n}\theta_t^2\Big\}\Big], \tag{A.49}$$

where in the last line we have used the assumption $|X_t| \leq A$.

We have assumed that θ is a stationary, zero-mean Gaussian process. Thus the limit

$$\lim_{n\to\infty} \frac{1}{n} \ln \boldsymbol{E}_P\Big\{\exp\Big(c\sum_{t=1}^{n}\theta_t^2\Big)\Big\} \tag{A.50}$$

can be computed using Szego's theorem (see [10]). To this end, note first that the exponential moment is given by

$$\boldsymbol{E}_P\Big\{\exp\Big(c\sum_{t=1}^{n}\theta_t^2\Big)\Big\} = \det(I - 2cV_n)^{-1/2}, \tag{A.51}$$

where V_n is the covariance matrix of θ_t, $t = 1, \ldots, n$. Next, if T_n is a sequence of Hermitian Toeplitz matrices of the form $T_n = [t_{k-j}; k, j = 0, 1, 2, \ldots, n-1]$, where t_k are absolutely summable, and their spectral density $f(\omega) = \sum_{k=-\infty}^{\infty} t_k e^{-ik\omega}$, $\omega \in \mathbb{R}$, satisfies $f(\omega) \geq m > 0$, $\omega \in \mathbb{R}$, one has by Theorem 13 of [10], that

$$\lim_{n\to\infty} \frac{1}{n} \ln\det(T_n) = \frac{1}{2\pi}\int_0^{2\pi} \ln f(\omega) \mathrm{d}\omega. \tag{A.52}$$

Recall that we assume that r_k are absolutely summable. Then, with $c = c(\alpha) = \alpha(\alpha - 1)A^2/(2\sigma^2)$, we obtain the bound

$$\limsup_{n\to\infty} \frac{\alpha}{n} D_\alpha(Q_n\|P_n) \leq \limsup_{n\to\infty} -\frac{1}{2(\alpha-1)n} \ln\det(I - 2cV_n) \tag{A.53}$$

$$= -\frac{1}{4\pi(\alpha-1)}\int_0^{2\pi} \ln(1 - 2c\Sigma_\theta(\omega))\mathrm{d}\omega, \tag{A.54}$$

assuming $2c\sup_\omega \Sigma_\theta(\omega) < 1$. As for the lower bound, a calculation similar to that of (A.45)–(A.49) gives

$$\frac{\alpha-1}{n} D_\alpha(P_n\|Q_n) = \frac{1}{n\alpha} \ln E_Q\Big[\exp\frac{\alpha(\alpha-1)}{2\sigma^2}\sum_{t=1}^{n}\theta_t^2 X_t^2\Big] \tag{A.55}$$

$$\leq \frac{1}{n\alpha} \ln E_Q\Big[\exp\frac{\alpha(\alpha-1)A^2}{2\sigma^2}\sum_{t=1}^{n}\theta_t^2\Big]. \tag{A.56}$$

Using the same considerations as before gives

$$\limsup_{n\to\infty} \frac{\alpha-1}{n} D_\alpha(P_n\|Q_n) \le -\frac{1}{4\pi\alpha}\int_0^{2\pi} \ln(1-2c\Sigma_\theta(\omega))\mathrm{d}\omega, \tag{A.57}$$

where again we assume that $\{\Sigma_\theta\}$ satisfies $2c\sup_\omega \Sigma_\theta(\omega) < 1$. □

A.5 Proof of Theorem 4.4

Our estimates of the Rényi divergence are based on large deviation results from [3]. We first note that the divergence is given by

$$\begin{aligned} D_\alpha(Q_T\|P_T) &= \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{P_T}[(Z_T)^{-\alpha}] && \text{(A.58)}\\ &= \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{P_T}\exp\Big[-\frac{\alpha}{\sigma}\int_0^t \theta_s X_s \mathrm{d}W_s + \frac{\alpha}{2\sigma^2}\int_0^t (\theta_s X_s)^2\mathrm{d}s\Big]. && \text{(A.59)}\end{aligned}$$

Note by (114) and (119) that

$$\sigma W_t = \int_0^t \theta_s X_s \mathrm{d}s + \sigma\tilde{W}_t, \tag{A.60}$$

and so

$$\int_0^t \theta_s X_s \mathrm{d}W_s = \sigma^{-1}\int_0^t (\theta_s X_s)^2 \mathrm{d}s + \int_0^t \theta_s X_s \mathrm{d}\tilde{W}_s. \tag{A.61}$$

Thus

$$D_\alpha(Q_T\|P_T) = \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{P_T}\exp\Big[-\frac{\alpha}{\sigma}\int_0^t \theta_s X_s \mathrm{d}\tilde{W}_s - \frac{\alpha}{2\sigma^2}\int_0^t (\theta_s X_s)^2\mathrm{d}s\Big]. \tag{A.62}$$

Under P_T, conditioned on $(\theta_s, X_s, s\in[0,t])$, the integral $\int_0^t \theta_s X_s \mathrm{d}\tilde{W}_s$ is a Gaussian random variable with mean zero and variance $\int_0^t (\theta_s X_s)^2 \mathrm{d}s$. Thus

$$D_\alpha(Q_T\|P_T) = \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{P_T}\exp\Big[\frac{\alpha(\alpha-1)}{2\sigma^2}\int_0^t (\theta_s X_s)^2\mathrm{d}s\Big]. \tag{A.63}$$

A similar calculation for

$$\begin{aligned} D_\alpha(P_T\|Q_T) &= \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{Q_T}[(Z_T)^{\alpha}] && \text{(A.64)}\\ &= \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{Q_T}\exp\Big[\frac{\alpha}{\sigma}\int_0^t \theta_s X_s \mathrm{d}W_s - \frac{\alpha}{2\sigma^2}\int_0^t (\theta_s X_s)^2\mathrm{d}s\Big] && \text{(A.65)}\end{aligned}$$

gives

$$D_\alpha(P_T\|Q_T) = \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{Q_T}\exp\Big[\frac{\alpha(\alpha-1)}{2\sigma^2}\int_0^t (\theta_s X_s)^2\mathrm{d}s\Big]. \tag{A.66}$$

As a result, the two divergences are equal, and using $|X_t|\le A$, we can bound them as follows:

$$D_\alpha(Q_T\|P_T) = D_\alpha(P_T\|Q_T) \le \frac{1}{\alpha(\alpha-1)}\ln \boldsymbol{E}_{Q_T}\exp\Big[\frac{\alpha(\alpha-1)A^2}{2\sigma^2}\int_0^T \theta_t^2\mathrm{d}t\Big]. \tag{A.67}$$

It is shown in Lemma 3 of [3] that

$$\lim_{T\to\infty}\frac{1}{T}\ln \boldsymbol{E}_{Q_T}\exp\left[c\int_0^T \theta_t^2 \mathrm{d}t\right] = -\frac{1}{4\pi}\int_{-\infty}^{\infty}\ln[1-4\pi c\varSigma_\theta(\omega)]\mathrm{d}\omega, \tag{A.68}$$

provided that $c < 1/(4\pi M)$. Specifically, (117) holds since we have assumed that $p := A^2/(2\sigma^2) < 1/(4\pi M)$. Moreover, with $c(\alpha) = \alpha(\alpha-1)p$, for any $\alpha > 1$ such that $c(\alpha) < 1/(4\pi M)$, we obtain from (9) and (10) (by a derivation analogous to that of (17))

$$E(P) \le \frac{\alpha}{\alpha-1}E(Q) - \frac{1}{4\pi(\alpha-1)}\int_{-\infty}^{\infty}\ln[1-4\pi c(\alpha)\varSigma_\theta(\omega)]\mathrm{d}\omega \tag{A.69}$$

$$E(P) \ge \frac{\alpha-1}{\alpha}E(Q) + \frac{1}{4\pi\alpha}\int_{-\infty}^{\infty}\ln[1-4\pi c(\alpha)\varSigma_\theta(\omega)]\mathrm{d}\omega. \tag{A.70}$$

□

A.6 Proof of (158)

To prove (158), we follow the main steps of [13], with a little twist since in our case the distortion measure (which is quadratic) is unbounded. Consider an arbitrary rate–distortion code $\mathcal{C} = \{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_M\}$, $M = e^{nR}$, R being the coding rate. Let us denote the event under discussion by

$$\mathcal{E} = \left\{\boldsymbol{y}:\ \min_m \sum_{t=1}^n (y_t - \hat{y}_{m,t})^2 > nd\right\}, \tag{A.71}$$

where $\hat{y}_{m,t}$ is the t–th component of the reproduction word $\hat{\boldsymbol{y}}_m$. Let $R_G(d,\tilde{\sigma}^2) = \frac{1}{2}\ln\frac{\tilde{\sigma}^2}{d}$ denote the rate–distortion function of the Gaussian memoryless source $\tilde{G}$ with variance $\tilde{\sigma}^2$. We first show that under the assumption that $R_G(d,\tilde{\sigma}^2) > R$, there exists a constant $\alpha(\tilde{\sigma}^2,d,R) > 0$ such that $\tilde{G}(\mathcal{E}) \ge \alpha(\tilde{\sigma}^2,d,R)$ for all sufficiently large n. Let

$$\tilde{d}(\mathcal{C}) \stackrel{\triangle}{=} \frac{1}{n}\tilde{\boldsymbol{E}}\{\min_m \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_m\|^2\}, \tag{A.72}$$

where $\tilde{\boldsymbol{E}}$ denotes expectation under $\tilde{G}$. Let $d_1 = \tilde{\sigma}^2 e^{-2R}$ denote the optimum distortion of $\tilde{G}$ at rate R. Then, obviously,

$$R_G(d,\tilde{\sigma}^2) > R = R_G(d_1,\tilde{\sigma}^2) \ge R_G(\tilde{d}(\mathcal{C}),\tilde{\sigma}^2), \tag{A.73}$$

where the first inequality is by our assumption. the equality is by definition of d_1 and the second inequality is due to the fact that $\mathcal{C}$ may not be optimal for $\tilde{G}$. Since $R_G(\cdot,\tilde{\sigma}^2)$ is monotonically decreasing, then

$$d < d_1 \le \tilde{d}(\mathcal{C}). \tag{A.74}$$

Now, let us denote $\delta(\boldsymbol{y}) = \min_m \|\boldsymbol{y} - \hat{\boldsymbol{y}}_m\|^2/n$ and let $d_0 > d_1$ be an arbitrary large distortion level. Then, assuming, without loss of generality, that the zero–vector belongs to $\mathcal{C}$, and so,

$\delta(\boldsymbol{y}) \le \|\boldsymbol{y}\|^2/n$, we have:

$$\begin{aligned}
\tilde{d}(\mathcal{C}) &\le [1-\tilde{G}(\mathcal{E})]\cdot d + \int_{\boldsymbol{y}:\ \delta(\boldsymbol{y})\ge d} \tilde{G}(\boldsymbol{y})\delta(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \\
&= [1-\tilde{G}(\mathcal{E})]\cdot d + \int_{\boldsymbol{y}:\ d\le\delta(\boldsymbol{y})\le d_0} \tilde{G}(\boldsymbol{y})\delta(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + + \int_{\boldsymbol{y}:\ \delta(\boldsymbol{y})\ge d_0} \tilde{G}(\boldsymbol{y})\delta(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \\
&\le [1-\tilde{G}(\mathcal{E})]\cdot d + \tilde{G}(\mathcal{E})\cdot d_0 + \frac{1}{n}\int_{\boldsymbol{y}:\ \|\boldsymbol{y}\|^2\ge nd_0} \tilde{G}(\boldsymbol{y})\cdot\|\boldsymbol{y}\|^2\mathrm{d}\boldsymbol{y} \qquad \text{(A.75)}
\end{aligned}$$

Now, the last term, which is

$$\delta_n \stackrel{\triangle}{=} \frac{1}{n}\cdot(2\pi\tilde{\sigma}^2)^{-n/2}\int_{\|\boldsymbol{y}\|^2\ge nd_0} \|\boldsymbol{y}\|^2\cdot\exp\{-\|\boldsymbol{y}\|^2/2\tilde{\sigma}^2\}\mathrm{d}\boldsymbol{y}, \qquad \text{(A.76)}$$

is easily shown[3] to decrease exponentially provided that $d_0 > \tilde{\sigma}^2$. Thus, we have

$$\tilde{G}(\mathcal{E}) \ge \frac{\tilde{d}(\mathcal{C}) - d - \delta_n}{d_0 - d} \ge \frac{d_1 - d - \delta_n}{d_0 - d}, \qquad \text{(A.77)}$$

which is positive for n large enough. For example, beyond a certain n_0, it exceeds $\frac{d_1-d}{2(d_0-d)}$, which we take to be $\alpha(\tilde{\sigma}^2, d, R)$. Now, for a given $\epsilon > 0$, let $\mathcal{T}_\epsilon = \{\boldsymbol{y} :\ |\ln\frac{\tilde{G}(\boldsymbol{y})}{G(\boldsymbol{y})} - nD(\tilde{G}\|G)| \le n\epsilon\}$. Then, by the weak law of large numbers, $\tilde{G}(\mathcal{T}_\epsilon) \ge 1 - \alpha(\tilde{\sigma}^2, d, R)/2$ for all large n. Thus,

$$\begin{aligned}
G(\mathcal{E}) &\ge \int_{\mathcal{E}\cap\mathcal{T}_\epsilon} G(\boldsymbol{y})\mathrm{d}\boldsymbol{y} && \text{(A.78)}\\
&= \int_{\mathcal{E}\cap\mathcal{T}_\epsilon} \tilde{G}(\boldsymbol{y})e^{-\ln[\tilde{G}(\boldsymbol{y})/G(\boldsymbol{y})]}\mathrm{d}\boldsymbol{y} && \text{(A.79)}\\
&\ge \tilde{G}(\mathcal{E}\cap\mathcal{T}_\epsilon)\cdot\exp\{-n[D(\tilde{G}\|G)+\epsilon]\} && \text{(A.80)}\\
&\ge [\tilde{G}(\mathcal{E}) - \tilde{G}(\mathcal{T}_\epsilon^c)]\cdot\exp\{-n[D(\tilde{G}\|G)+\epsilon]\} && \text{(A.81)}\\
&\ge [\alpha(\tilde{\sigma}^2,d,R) - \frac{1}{2}\alpha(\tilde{\sigma}^2,d,R)]\cdot\exp\{-n[D(\tilde{G}\|G)+\epsilon]\} && \text{(A.82)}\\
&= \frac{1}{2}\alpha(\tilde{\sigma}^2,d,R)\cdot\exp\{-n[D(\tilde{G}\|G)+\epsilon]\}. && \text{(A.83)}
\end{aligned}$$

Since this is true for all $\tilde{\sigma}^2$ with $R_G(d,\tilde{\sigma}^2) > R$, the tightest bound is obtained by minimizing

$$D(\tilde{G}\|G) = \frac{1}{2}\left[\frac{\tilde{\sigma}^2}{\sigma^2} - \ln\frac{\tilde{\sigma}^2}{\sigma^2} - 1\right], \qquad \text{(A.84)}$$

in the range $\tilde{\sigma}^2 \ge de^{2R}$, which is attained at $\tilde{\sigma}^2 = de^{2R}$, yielding the following upper bound on the exponent:

$$E(R) \le \frac{1}{2}\left[\frac{de^{2R}}{\sigma^2} - \ln\frac{de^{2R}}{\sigma^2} - 1\right] = \Phi[R - R_G(d)]. \qquad \text{(A.85)}$$

□

[3] Apply the Chernoff bound and use the fact that $\|\boldsymbol{y}\|^2e^{-s\|\boldsymbol{y}\|^2}$ is the negative derivative of $e^{-s\|\boldsymbol{y}\|^2}$ w.r.t. s.

References

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

[2] R. Atar, K. Chowdhary and P. Dupuis. Robust bounds on risk–sensitive functionals via Rényi divergence. *SIAM J. Uncertainty Quant.*, to appear, 2015, `arXiv:1310.6391 [math.PR]`.

[3] W. Bryc and A. Dembo. Large deviations for quadratic functionals of Gaussian processes. *J. Theoret. Probab.*, 10(2):307–332, 1997.

[4] I. Csiszár and J. Körner. *Information Theory. Coding Theorems for Discrete Memoryless Systems.* Cambridge University Press, Cambridge, second edition, 2011.

[5] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics (New York).* Springer-Verlag, New York, second edition, 1998.

[6] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, 1997.

[7] K. Dvijotham and E. Todorov. A unified theory of linearly solvable optimal control. *Artificial Intelligence (UAI)*, page 1, 2011.

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

[9] L. Golshani, E. Pasha, and G. Yari. Some properties of Rényi entropy and Rényi entropy rate. *Inform. Sci.*, 179(14):2426–2433, 2009.

[10] R. M. Gray. *Toeplitz and circulant matrices: A review.* now publishers inc, 2006.

[11] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, second edition, 1991.

[12] F. Liese and I. Vajda. *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics].* BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[13] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Trans. Information Theory*, IT-20:197–199, 1974.

[14] N. Merhav. On zero-rate error exponents of finite-state channels with input-dependent states. *IEEE Trans. Inform. Theory*, to appear, 2015, `arXiv:1406.7092 [cs.IT]`.

[15] Y. Polyanskiy, H. V. Poor and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory*, 56, 2307–2359, 2010.

[16] 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*, 10:65–103, 1967.

[17] 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*, 10:522–552, 1967.

[18] I. Vajda. Distances and discrimination rates for stochastic processes. *Stochastic Process. Appl.*, 35(1):47–57, 1990.

[19] T. van Erven and P. Harremoës. Rényi divergence and majorization. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1335–1339. IEEE, 2010.

[20] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inform. Theory*, 60(7):3797–3820, 2014.

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