Tradeoffs Between the Excess-Code-Length Exponent and the Excess-Distortion Exponent in Lossy Source Coding *

Tsachy Weissman Neri Merhav

April 25, 2001

Abstract

Lossy compression of a discrete memoryless source (DMS) with respect to a single-letter distortion measure, is considered. We study the best attainable tradeoff between the exponential rates of the probabilities that the codeword length and that the cumulative distortion exceed respective thresholds for two main cases. The first scenario examined is that where the source is corrupted by a discrete memoryless channel prior to reaching the coder. In the second part of this work we examine the universal setting, where the (noise-free) source is an unknown member, P_θ, of a given family, $\{P_\theta, \theta \in \Theta\}$. Here, inspired by an approach which was proven fruitful recently in the context of composite hypothesis testing, we allow the constraint on the excess-code-length exponent to be θ-dependent. Corollaries are derived for some special cases of interest, including Marton's classical source coding exponent and its generalization to the case where the constraint on the rate of the code is relaxed from an almost sure constraint to a constraint on the excess-code-length exponent.

1 Introduction

Let $X_1, X_2, \ldots$ be a source sequence with components in a finite alphabet $\mathcal{X}$ and let $\hat{\mathcal{X}}$ be a reconstruction alphabet. Let further $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be a given single-letter distortion measure and define the distortion between sequences $X^n \in \mathcal{X}^n$ and $\hat{X}^n \in \hat{\mathcal{X}}^n$ by $\rho_n(X^n, \hat{X}^n) = \sum_{i=1}^n \rho(X_i, \hat{X}_i)$. The problem of lossy source coding, or, rate-distortion theory, is typically the following: Given a rate R and a block length n, one seeks a codebook, $B_n \subseteq \hat{\mathcal{X}}^n$ of size 2^{nR} and a mapping, $\hat{X}^n = \varphi_n(X^n)$, from $\mathcal{X}^n$ into B_n, making the distortion $\rho_n(X^n, \hat{X}^n)$ as small as possible in some quantitative sense, most commonly the expected distortion sense. Alternatively, lossy source coding is often concerned with the reverse problem of fixed distortion and variable rate (see [1], [9]).

Another sensible performance criterion is the large deviations behavior of the distortion, which was first addressed, for a DMS, in Marton's 1974 paper [16] (cf. also [2, 3]). Specifically, the goal in that work was to find the best exponential decay rate of $\Pr\{\rho_n(X^n, \hat{X}^n) > nd\}$ for a given distortion level d, when the size of the codebook is limited to 2^{nR} words. This exponential decay rate was shown to be given by $F_d(R) = \min_{Q:R(Q,d)\geq R} D(Q\|P)$, where P is the marginal of the source, $R(Q, \cdot)$ denotes the rate-distortion function a source Q, and $D(\cdot\|\cdot)$ denotes

*Both authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. tsachy@tx.technion.ac.il, merhav@ee.technion.ac.il.

the Kullback-Leibler distance. Error exponents for source codes have since been studied by others (cf. [4, 5] and references therein) and have been extended to more general settings, such as that of successive refinement [13]. The excess-code-length exponent has also been studied in the context of lossless coding, cf. [11, 12, 21, 17] and the recent work [10] for the information-spectrum approach.

In the work related to error exponents of various lossy source coding scenarios, ordinarily, the rate constraint imposed is very strict. The code-book size is not allowed to exceed 2^{nR}, which means, of course, that the codeword length should not exceed nR bits, with probability one. The criteria by which the distortion is typically evaluated, on the other hand, are considerably softer and more pliable. Alternatively, the other setting typically considered is that where the constraint on the distortion is strict, while that on the rate is softer.

Our goal, in this work, is to treat the distortion and the rate in a more symmetric way, in the context of large deviations performance, in certain problems of lossy source coding. We shall measure performance in terms of the tradeoff between the probabilities that the distortion and the associated codeword length exceed respective thresholds. Specifically, we characterize the best achievable exponential rate of $\Pr\{\rho_n(X^n, \hat{X}^n) > nd\}$ (henceforth referred to as the "excess-distortion exponent"), subject to the constraint $\Pr\{\text{code-length of } \hat{X}^n > nR\} \leq e^{-n\lambda}$, for a given rate R, and a given $\lambda > 0$ (henceforth referred to as the "excess-code-length exponent").

Consider first the extension of the setting of [16] where we allow variable-length codes but restrict attention only to those for which $\Pr\{\text{code-length of } \hat{X}^n > nR\} \leq e^{-n\lambda}$. Assuming, without loss of optimality (as will be proved in Section 3), that sequences of the same type class are assigned codewords of roughly the same length, the optimal scheme allots approximately nR bits to any sequence whose empirical probability mass function (PMF) is within Kullback-Leibler distance of λ from P and $n \log |\mathcal{X}|$ bits when the Kullback-Leibler distance from P is more than λ. To get the idea of this, let us now examine two complimentary cases:

Case 1 All types Q with $D(Q\|P) \leq \lambda$ have $R(Q,d) < R$.

Case 2 There are types Q with $D(Q\|P) \leq \lambda$ such that $R(Q,d) \geq R$.

In **Case 1**, by the type-covering lemma [5], the types in which we cannot afford more than nR bits, can be covered within distortion nd. Hence, *all* sequences are covered within distortion nd under the rate constraint, thus achieving an infinite excess-distortion exponent. In **Case 2**, essentially all sequences within types Q for which $R(Q,d) > R$ are distorted by more than nd and, hence, the best excess-distortion exponent achievable will be the minimum of

$D(Q\|P)$ over all Q for which $R(Q,d) \geq R$, which is precisely $F_d(R)$. Furthermore, note that **Case 1** prevails if and only if $\lambda < F_d(R)$. To sum up, the excess-distortion exponent is a step function in λ, assuming the value ∞ for $\lambda < F_d(R)$ and the value $F_d(R)$ for $\lambda \geq F_d(R)$ (cf. Figure 1 of subsection 3.C.2). This extension of Marton's result will be shown to be obtainable as a special case of the more general setting, considered in Section 3, where the source is corrupted by noise prior to encoding. As will also be shown, in this more general case, the behavior of the excess-distortion exponent is normally more "graceful" than in the noise-free case.

In Section 3, we shall analyze source coding exponents for the case of a DMS corrupted by a discrete memoryless channel (DMC). In this setting[1], the encoder accesses a noisy version $Z_1, Z_2, \ldots$ of the clean source sequence $X_1, X_2, \ldots$ and produces a reconstruction sequence $\hat{X}_1, \hat{X}_2, \ldots$. We shall characterize the best achievable excess-distortion exponent (where the distortion measured is between $X_1, X_2, \ldots$ and $\hat{X}_1, \hat{X}_2, \ldots$) subject to the constraint that the excess-code-length exponent is at least λ. Note that the problem considered in [16] is a special case of this section, where the DMC is the clean channel (under which $X_i = Z_i$ a.s.) and $\lambda = \infty$. Staying with the case where $\lambda = \infty$, the results of this section can be considered analogous to those of [16] for the noisy case. To the best of our knowledge, the lossy source coding error exponent for noisy sources has not been previously obtained in the literature. Nor are we aware of results pertaining to the excess distortion exponent even for $R = \log|\mathcal{X}|$, that is, pure filtering.

A more detailed outline of Section 3 is as follows. Subsection 3.A presents the main result on the above-described exponent, whose proof is given in subsection 3.B. In subsection 3.C, we verify that Marton's exponent is obtained as a special case ($\lambda = \infty$) and we obtain the excess-distortion exponent for the noise-free case as a function of λ. Subsection 3.D will be dedicated to the properties of the exponent function. Finally, in subsection 3.E, we merely mention that a generalization of the model to the case where additional side information is available at both encoder and decoder is possible.

Section 4 analyzes a universal coding problem from the perspective described above. Returning to the case of encoding the clean sequence (uncorrupted by a DMC), emitted by a DMS P_θ, where θ is an unknown parameter taking values in a set Θ, we characterize the best achievable (θ-dependent) distortion exponent, such that the code-length exponent, w.r.t. P_θ, is at least as large as a given function $\lambda(\theta)$, uniformly for all $\theta \in \Theta$. A universally optimal

[1] Originally, studied under the ordinary, expected rate and distortion criteria [7, 8].

scheme will be constructed. We shall also attempt to establish a quantitative relationship between the "price" paid for universality (relative to the source-dependent coding case), the geometry of Θ, and the form of the function $\lambda(\cdot)$. This somewhat parallels a recent work [15], where a competitive Neyman-Pearson approach was developed for the composite hypothesis testing problem. In that work, the goal was to decide, upon observing $(X_1, \ldots, X_n) \sim P_\theta$, whether θ belongs to Θ_1 or to Θ_2, and the best achievable error exponent of the second kind was characterized such that the constraint that the error exponent of the first kind is lower bounded by $\lambda(\theta)$, for all θ. As it turns out, the optimal scheme of [15] is intimately related and, in some sense, analogous to the one herein. As will be elaborated on in subsection 4.B, the coding scheme which turns out to be optimal in the setting of this work uses two different strategies in two complimentary subsets of sources, and these subsets exactly correspond to the decision rule of [15].

As the proofs of the main results reveal, there is a basic structural feature, shared by the sequences of optimum schemes in all of the scenarios considered. In all cases, the schemes dichotomize between two complimentary sets of types. If the observed sequence is of a type which belongs to one set, the codeword will be about nR bits long. If the type belongs to the complimentary set, the codeword will be significantly longer. In Section 3, the dichotomy is according to whether $D(Q\|P) \leq \lambda$ or $D(Q\|P) > \lambda$. In Section 4, the dichotomy is dictated by the Kullback-Leibler distances between the type of the sequence and the sources P_θ, $\theta \in \Theta$, corresponding to the decision rule proposed in [15]. This phenomenon can be attributed to the fact that the probability of excess-code-length is the expectation of the indicator function of the excess-code-length event. Thus, all codewords shorter than nR bits incur zero loss, whereas those longer than nR bits share the value 1. Hence, in terms of distortion, one can never gain from using significantly less than nR bits and, on the other hand, having allotted more than nR bits, one may use arbitrarily many bits with no extra payment.

The remainder of the paper is organized as follows. Section 2, in which we present the notation and conventions used throughout, will be followed by the two main sections (3 and 4) described above. Finally, in Section 5, we summarize the paper and discuss a direction for future research.

2 Notation and Preliminaries

For an arbitrary pair of non negative sequences, $\{a_n\}$ and $\{b_n\}$, we write $a_n \doteq b_n$ as short-hand notation for

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0, \tag{1}$$

where we take throughout $0 \log 0 \triangleq 0$, $\log 0 \triangleq -\infty$, $\frac{0}{0} \triangleq 1$ and $\frac{a}{0} \triangleq \infty$ for $a > 0$. We shall write $a_n \stackrel{\cdot}{\geq} b_n$ for

$$\liminf_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} \geq 0, \tag{2}$$

similarly, $a_n \stackrel{\cdot}{>} b_n$ will be synonymous to the statement that (2) holds with a strict inequality. Note that the negation of $a_n \stackrel{\cdot}{\geq} b_n$ does *not* imply that $a_n \stackrel{\cdot}{<} b_n$, but it does imply that $\limsup_{n \to \infty} \frac{1}{n} \log \frac{b_n}{a_n} > 0$, or, in other words[2], the existence of a $\delta > 0$ and an increasing sequence of natural numbers $\{n_k\}_k$ such that $b_{n_k} \geq a_{n_k} e^{n_k \delta}$ for all k. For an increasing sequence of natural numbers $\{n_k\}_k$, we shall slightly abuse the notation $\doteq$ by writing $a_{n_k} \doteq b_{n_k}$ for $\lim_{k \to \infty} \frac{1}{n_k} \log \frac{a_{n_k}}{b_{n_k}} = 0$. The notation $a_{n_k} \stackrel{\cdot}{\geq} b_{n_k}$ should be understood analogously.

For finite sets $\mathcal{A}$ and $\mathcal{B}$, we let $\mathcal{M}(\mathcal{A})$ denote the set of all probability measures on $\mathcal{A}$ and $\mathcal{C}(\mathcal{A} \to \mathcal{B})$ the set of all stochastic matrices (or "channels", or "conditional distributions") from $\mathcal{A}$ to $\mathcal{B}$. We will let $\mathcal{M}_+(\mathcal{A})$ denote the subset of $\mathcal{M}(\mathcal{A})$ containing all probability measures which assign strictly positive mass to all members of $\mathcal{A}$. For $P \in \mathcal{M}(\mathcal{A})$ we let $H(P)$ denote the entropy of a random variable distributed according to P. For any $P \in \mathcal{M}(\mathcal{A})$ and $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ we will write $P \times W$ for the measure governing the pair $(A, B) \in \mathcal{A} \times \mathcal{B}$ when A is generated according to $P(\cdot)$ and then B is taken as the output of the channel W whose input is A. We also let

$$H(W|P) = \sum_{a \in \mathcal{A}} P(a) H(W(\cdot|a)) \tag{3}$$

denote the conditional entropy of B given A and $I(P; W)$ the mutual information between A and B. Alternatively, for $Q \in \mathcal{M}(\mathcal{A} \times \mathcal{B})$, we shall sometimes write $I(Q)$ to denote the mutual information between A and B when jointly distributed according to Q. For $P, Q \in \mathcal{M}(\mathcal{A})$ we will denote the Kullback-Leibler (informational) divergence by $D(P\|Q)$ and for $V, W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ we will let

$$D(V\|W|P) = \sum_{a \in \mathcal{A}} P(a) D(V(\cdot|a)\|W(\cdot|a)) \tag{4}$$

denote the conditional (informational) divergence.

We assume a finite source alphabet $\mathcal{X}$, a finite reproduction alphabet $\hat{\mathcal{X}}$, and a given single-letter distortion measure $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ satisfying[3] $\min_{\hat{X} \in \hat{\mathcal{X}}} \rho(X, \hat{X}) = 0$ for all $X \in \mathcal{X}$. For $X^n \in \mathcal{X}^n$ and $\hat{X}^n \in \hat{\mathcal{X}}^n$ we define

$$\rho_n(X^n, \hat{X}^n) = \sum_{i=1}^{n} \rho(X_i, \hat{X}_i) \tag{5}$$

[2]This is the form of the statement which will be used in a certain place in the paper, so the opportunity is taken to make it explicit.
[3]Note that there is no essential loss of generality in this assumption as, otherwise, we can look at the modified distortion measure given by $\tilde{\rho}(X, \hat{X}) = \rho(X, \hat{X}) - \min_{\hat{X}' \in \hat{\mathcal{X}}} \rho(X, \hat{X}')$.

and for any $B \subseteq \hat{\mathcal{X}}^n$ we denote $\rho_n(X^n, B) = \min_{\hat{X}^n \in B} \rho_n(X^n, \hat{X}^n)$. The subscript n will be omitted from ρ_n in the sequel when there is no room for ambiguity. We shall adopt the convention throughout that capital letters represent random variables, while the corresponding lower case letters represent specific sample values. For $P \in \mathcal{X}$ we shall let $R(P, d)$ and $D(P, R)$ denote the rate-distortion and distortion-rate functions, respectively, associated with the source P.

For any $a^n \in \mathcal{A}^n$ we let $P_{a^n} \in \mathcal{M}(\mathcal{A})$ denote the associated empirical measure. For $P \in \mathcal{M}(\mathcal{A})$ we let $T_P = \{a^n \in \mathcal{A}^n : P_{a^n} = P\}$ denote the type class of P. For $n \in \mathbb{N}$ we let $\mathcal{M}_n(\mathcal{A})$, or simply $\mathcal{M}_n$ when the alphabet is clear from the context, denote the set of all $P \in \mathcal{M}(\mathcal{A})$ for which $T_P \neq \emptyset$. For $P \in \mathcal{M}_n$ we let $\mathcal{C}_n(P)$ denote the set of all $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ for which $T_{P \times W}$ (as a subset of $(\mathcal{A} \times \mathcal{B})^n$) is not empty. Following [5], for any given $\hat{x}^n \in \hat{\mathcal{X}}^n, z^n \in \mathcal{Z}^n$ and $V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X})$, we let $T_V(\hat{x}^n, z^n)$ denote the set of sequences $x^n \in \mathcal{X}^n$ having conditional type V given $(\hat{x}^n, z^n)$.

When dealing with expectations of functions or with functionals of random variables, we shall sometimes subscript their distributions when we want to make these explicit. Thus, for example, for any $f : \mathcal{A} \to \mathbb{R}$ and $P \in \mathcal{M}(\mathcal{A})$ we shall write $E_P f(A)$ for the expectation of $f(A)$ when A is distributed according to P. Similarly, we shall write, for example, $I_Q(X; Y|Z)$, to denote the conditional mutual information between X and Y given Z, when the triple (X, Y, Z) is distributed according to $Q \in \mathcal{M}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. Also, if $Q \in \mathcal{M}(\mathcal{Y} \times \mathcal{Z})$ and $V \in \mathcal{C}(\mathcal{Z} \to \mathcal{X})$, we shall sometimes slightly abuse the notation by writing $Q \times V$ to denote the distribution on $(Y, Z, X) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ where (Y, Z) are generated according to Q and then X is taken as the output of the channel V whose input is Z (note that in this case, $X \ominus Z \ominus Y$ form a Markov chain). Also, for $V \in \mathcal{C}(\mathcal{Y} \times \mathcal{Z} \to \mathcal{X})$, $U \in \mathcal{C}(\mathcal{Z} \to \mathcal{X})$ and $Q \in \mathcal{M}(\mathcal{Y} \times \mathcal{Z})$, we shall sometimes slightly abuse the notation by writing $D(V\|U|Q)$ for $D(V\|\tilde{U}|Q)$, where $\tilde{U} \in \mathcal{C}(\mathcal{Y} \times \mathcal{Z} \to \mathcal{X})$ is the channel which coincides with U (i.e., the output of the channel $\tilde{U}$ is independent of the $\mathcal{Y}$-valued component of the input).

Since we will only be dealing with finite alphabets, there will be no technicalities associated with convergence of probability measures. For any sequence of probability measures $\{P_n\}_n$ and a P, all of which are members of, say, $\mathcal{M}(\mathcal{A})$, we will write $P_n \to P$ for $\|P_n - P\| \to 0$ where, in the second limit, we regard P_n and P as elements of $\mathbb{R}^{|\mathcal{A}|}$.

For any function I, we use the standard notation $I(R \pm 0)$ to denote the respective limits $\lim_{\varepsilon \to 0^\pm} I(R + \varepsilon)$, whenever these limits exist. We shall use analogous notation with multivariate functions as well. Thus, for example, the notation $I(R+0,d,\lambda-0)$ will stand for $\lim_{\varepsilon\to 0^+,\delta\to 0^+} I(R+\varepsilon,d,\lambda-\delta)$ and, similarly, $I(R-0,d,\lambda+0)$ for $\lim_{\varepsilon\to 0^+,\delta\to 0^+} I(R-\varepsilon,d,\lambda+\delta)$, whenever the limits exist.

Finally, the infimum or the minimum over the empty set is defined throughout as ∞ and $\backslash$ will denote subtraction of sets, i.e., for two sets A and B, $\mathrm{A}\backslash\mathrm{B} = \mathrm{A}\cap\mathrm{B}^c$.

3 Encoding a Noisy Source

In this section, we assume a DMS with finite input alphabet $\mathcal{X}$ and probability distribution P_X, corrupted by a noisy channel $P_{Z|X}$ with output alphabet $\mathcal{Z}$. We let further P_Z, $P_{X,Z}$, and $P_{X|Z}$ denote the marginal on $\mathcal{Z}$, the joint distribution on $\mathcal{X}\times\mathcal{Z}$, and the backward channel, respectively, induced by the source P_X and the DMC $P_{Z|X}$. To avoid trivial technicalities, we assume that $P_{X,Z}\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Z})$. $P^n_{X,Z}$ will denote the product distribution on $(\mathcal{X}\times\mathcal{Z})^n$ induced by $P_{X,Z}$.

A variable-length scheme D_n for combined filtering and compression of noisy input vectors of length n, $Z^n\in\mathcal{Z}^n$, consists of a triple $D_n = (B_n, \hat{X}^n(\cdot), C_n)$ with the following ingredients: a reconstruction codebook $B_n\subseteq\hat{\mathcal{X}}^n$, a mapping $\hat{X}^n : \mathcal{Z}^n\to B_n$ so that $\hat{X}^n(Z^n)\in\hat{\mathcal{X}}^n$ is the reconstruction vector, and a uniquely decodable code $C_n : B_n\to\{0,1\}^*$. For the noisy vector Z^n, let $L_n(Z^n)$ denote the length of the codeword $C_n(\hat{X}^n(Z^n))$. We shall write $L_n(Z^n, D_n)$ when we want to make the dependence on the particular code D_n explicit. Let $\mathcal{D}_n$ denote the set of all such schemes for combined filtering and compression of noisy input vectors of length n.

In the classical setting of [16], which is noise-free, and there is a strict constraint on the codeword length ($\lambda=\infty$), the compression scheme is completely determined by the codebook B_n as, clearly, the optimal way to use a given codebook is to apply the nearest neighbor rule w.r.t. ρ. In the noisy setting, given Z^n, it is not obvious how to encode, as distortion is measured with respect to the unobserved X^n. Moreover, even when there is no noise, but the excess-code-length constraint is relaxed from the stringent assumption $\lambda=\infty$ of [16], to the milder one where λ is finite, the compression scheme is not immediately determined by B_n, because there might exist X^n which may better be represented by a reproduction word which is not the nearest-neighbor, but which has a shorter codeword.

For $R\geq 0$ and $\lambda\in[0,\infty]$ we define now

$$A_n(R,\lambda) = \left\{D_n\in\mathcal{D}_n : P^n_{X,Z}\{L_n(Z^n,D_n) > nR\}\leq e^{-\lambda n}\right\} \tag{6}$$

and for $d \geq 0$ we let

$$G_n(R,d,\lambda) = \min_{D_n \in A_n(R,\lambda)} P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\}, \tag{7}$$

where the $\hat{X}^n(\cdot)$ in the right side of (7) is the mapping from $\mathcal{Z}^n$ into B_n associated with the scheme D_n.

A Statement of the Main Result

The main result of this section is the following.

Theorem 1. *For all $d \geq 0$, $R \geq 0$, and $\lambda \in [0,\infty]$ we have*

$$I(R-0,d,\lambda+0) \leq \liminf_{n\to\infty} \left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] \tag{8}$$

$$\leq \limsup_{n\to\infty} \left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] \leq I(R+0,d,\lambda-0), \tag{9}$$

where

$$I(R,d,\lambda) = \min\left\{\inf_{P: D(P\|P_Z)\geq\lambda} a(P,\infty,d), \inf_{P: D(P\|P_Z)<\lambda} a(P,R,d)\right\}, \tag{10}$$

$a(P,R,d)$ is defined, for $P \in \mathcal{M}(\mathcal{Z})$, by

$$a(P,R,d) = D(P\|P_Z) + \sup_{\substack{W \in \mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}}): \\ I(P;W)\leq R}} F(P\times W, d), \tag{11}$$

and $F(\cdot,d)$ is defined, for $Q \in \mathcal{M}(\hat{\mathcal{X}}\times\mathcal{Z})$, by

$$F(Q,d) = \inf_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}}\times\mathcal{Z}\to\mathcal{X}): \\ E_{Q\times V}\rho(X,\hat{X})>d}} D\left(V\|P_{X|Z}|Q\right). \tag{12}$$

Remark: The monotonicity of $I(R,d,\lambda)$ in both R (increasing) and λ (decreasing) is shown in Appendix B.1 to imply that the limits defining $I(R-0,d,\lambda+0)$ and $I(R+0,d,\lambda-0)$ both exist. Furthermore, in Appendix B.2 we show that Theorem 1 implies that for all pairs (R,λ) outside a set $\mathcal{S}$ of zero Lebesgue measure

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] = I(R,d,\lambda). \tag{13}$$

Note that the case of pure filtering is obtained by taking $R=\infty$ or $\lambda=0$, where Theorem 1 gives the best achievable excess-distortion exponent for the filtering problem. The case of "pure coding", on the other hand, when the codebook is confined to a size no greater than 2^{nR} words, is obtained by taking $\lambda=\infty$. A further discussion of the exponent, $I(R,d,\lambda)$, and of special cases of Theorem 1 of particular interest, is deferred to subsection D.

B Proof of Theorem 1

The outline of the proof is the following. We start by obtaining an exponentially tight estimate for $P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd|Z^n = z^n\}$, for an arbitrary $z^n \in \mathcal{Z}^n$ and scheme D_n (which defines the mapping $\hat{X}^n(\cdot) : \mathcal{Z}^n \to \hat{\mathcal{X}}^n$). This estimate will be a functional of the empirical distribution $P_{\hat{X}^n(z^n),z^n}$. So, in effect, we are looking for that scheme in $A_n(R, \lambda)$ with empirical distributions $P_{\hat{X}^n(z^n),z^n}$ that minimize this functional, if possible, for all $z^n \in \mathcal{Z}^n$. In the converse part, we show that any member of $A_n(R, \lambda)$ must be such that $I(P_{\hat{X}^n(z^n),z^n}) \leq R$ for essentially all $z^n \in \mathcal{Z}^n$ for which $D(P_{z^n}\|P_Z) < \lambda$. Therefore, the exponent achievable for $P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\}$ must clearly be upper bounded by the exponent obtained when maximizing the above functional with respect to $P_{\hat{X}^n(z^n),z^n}$, for each z^n, under the constraint that $I(P_{\hat{X}^n(z^n),z^n}) \leq R$ whenever $D(P_{z^n}\|P_Z) < \lambda$. The direct part is then established by a construction of a scheme in $A_n(R, \lambda)$, which achieves the upper bound on the exponent established in the converse part. This scheme is one under which $P_{\hat{X}^n(z^n),z^n}$ maximizes the above functional, subject to the constraint that $I(P_{\hat{X}^n(z^n),z^n}) \leq R$ whenever $D(P_{z^n}\|P_Z) < \lambda$. The existence of such a scheme in $A_n(R, \lambda)$ is essentially guaranteed because the fact that $I(P_{\hat{X}^n(z^n),z^n}) \leq R$ for all $z^n \in T_P$ guarantees the existence of a scheme that will need no more than nR bits to convey $\hat{X}^n(z^n)$ for $z^n \in T_P$. Hence, whenever $z^n \in T_P$ for P with $D(P\|P_Z) < \lambda$, the codeword will be no more than nR bits long, guaranteeing the membership of the associated scheme in $A_n(R, \lambda)$. We now turn to making the above line of argumentation precise.

For any scheme D_n we have

$$P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\} = \sum_{(x^n,z^n)\in(\mathcal{X}\times\mathcal{Z})^n:\rho(x^n,\hat{X}^n(z^n))>nd} P^n_Z(z^n) \cdot P^n_{X|Z}(x^n|z^n) \tag{14}$$

$$= \sum_{z^n\in\mathcal{Z}^n} P^n_Z(z^n) \left(\sum_{x^n\in\mathcal{X}^n:\rho(x^n,\hat{X}^n(z^n))>nd} P^n_{X|Z}(x^n|z^n) \right). \tag{15}$$

Considering the inner sum in (15), we have for any $z^n \in \mathcal{Z}^n$ and $\hat{x}^n \in \hat{\mathcal{X}}^n$

$$
\begin{aligned}
& \sum_{x^n \in \mathcal{X}^n : \rho(x^n, \hat{x}^n) > nd} P^n_{X|Z}(x^n|z^n) && (16)\\
= & \sum_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X}) : \\ E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d}} \sum_{x^n \in T_V(\hat{x}^n, z^n)} P^n_{X|Z}(x^n|z^n) && (17)\\
= & \sum_{V : E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d} |T_V(\hat{x}^n, z^n)| \exp\left\{-n E_{P_{\hat{x}^n, z^n} \times V}\left[-\log P_{X|Z}(X|Z)\right]\right\} && (18)\\
= & \sum_{V : E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d} |T_V(\hat{x}^n, z^n)| \exp\left\{-n H(V|P_{\hat{x}^n, z^n})\right\} && (19)\\
& \times \exp\left\{-n\left(-H(V|P_{\hat{x}^n, z^n}) + E_{P_{\hat{x}^n, z^n} \times V}\left[-\log P_{X|Z}(X|Z)\right]\right)\right\} && (20)\\
= & \sum_{V : E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d} |T_V(\hat{x}^n, z^n)| \exp\left\{-n H(V|P_{\hat{x}^n, z^n})\right\} \exp\left\{-n D\left(V \| P_{X|Z} | P_{\hat{x}^n, z^n}\right)\right\} && (21)
\end{aligned}
$$

where equality (18) follows standardly by a rewriting of $P^n_{X|Z}(x^n|z^n)$ as

$$
P^n_{X|Z}(x^n|z^n) = \prod_{i=1}^{n} P_{X|Z}(x_i|z_i) = \prod_{(a,b) \in \mathcal{X} \times \mathcal{Z}} P_{X|Z}(a|b)^{N((a,b)|(x^n, z^n))}, \tag{22}
$$

where $N((a,b)|(x^n, z^n))$ is the count of the pair (a, b) along (x^n, z^n). To get to equality (21), we have used the easily verifiable relationship

$$
D\left(V \| P_{X|Z} | P_{\hat{x}^n, z^n}\right) = -H(V|P_{\hat{x}^n, z^n}) + E_{P_{\hat{x}^n, z^n} \times V}\left[-\log P_{X|Z}(X|Z)\right]. \tag{23}
$$

For an upper and a lower bound on (21), we now let $F_n(\cdot, d)$, for $Q \in \mathcal{M}_n(\hat{\mathcal{X}} \times \mathcal{Z})$, be defined by

$$
F_n(Q, d) = \min_{\{V : E_{Q \times V} \rho(X, \hat{X}) > d\} \cap \mathcal{C}_n(Q)} D\left(V \| P_{X|Z} | Q\right). \tag{24}
$$

Recall further (cf., e.g., [5, Lemma 2.5]) that for all $V, \hat{x}^n, z^n$

$$
(n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \leq |T_V(\hat{x}^n, z^n)| \exp\left\{-n H(V|P_{\hat{x}^n, z^n})\right\} \leq 1. \tag{25}
$$

Therefore (21) is upper bounded by

$$
\begin{aligned}
& \sum_{V : T_V(\hat{x}^n, z^n) \neq \emptyset, E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d} \exp\left\{-n D\left(V \| P_{X|Z} | P_{\hat{x}^n, z^n}\right)\right\} && (26)\\
\leq \quad & (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \exp\{-n F_n(P_{\hat{x}^n, z^n}, d)\} && (27)
\end{aligned}
$$

and lower bounded by

$$
\begin{aligned}
& (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{V: T_V(\hat{x}^n, z^n) \neq \emptyset, E_{P_{\hat{x}^n, z^n} \times V} \rho(X, \hat{X}) > d} \exp\left\{-nD\left(V \| P_{X|Z} | P_{\hat{x}^n, z^n}\right)\right\} && (28)\\
\geq \quad & (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \exp\{-nF_n(P_{\hat{x}^n, z^n}, d)\}. && (29)
\end{aligned}
$$

We turn to the proof of (9) first. Fix arbitrary $\varepsilon, \delta, \eta > 0$ and a sequence of coding schemes $\{D_n\}$ such that $D_n \in A_n(R-\varepsilon, \lambda)$. Combining (15) with (29), we have the existence of $n_0(\varepsilon, \delta, \eta)$ (independent on the sequence $\{D_n\}$) such that for all $n \geq n_0(\varepsilon, \delta, \eta)$,

$$
\begin{aligned}
& P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\} && (30)\\
\geq \quad & (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{z^n \in \mathcal{Z}^n} P^n_Z(z^n) \exp\left\{-nF_n(P_{\hat{X}^n(z^n), z^n}, d)\right\} && (31)\\
= \quad & (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \left[\sum_{\{P: D(P\|P_Z) \geq \lambda-\delta\} \cap \mathcal{M}_n} \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n), z^n}, d)\right]\right\}\right. && (32)\\
& \left. + \sum_{\{P: D(P\|P_Z) < \lambda-\delta\} \cap \mathcal{M}_n} \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n), z^n}, d)\right]\right\}\right] && (33)\\
\geq \quad & (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \left[\sum_{\{P: D(P\|P_Z) \geq \lambda-\delta\} \cap \mathcal{M}_n} \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P)} F_n(P \times W, d)\right]\right\}\right. \\
& \left. + \sum_{\{P: D(P\|P_Z) < \lambda-\delta\} \cap \mathcal{M}_n} \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P): \\ I(P;W) \leq R}} F_n(P \times W, d) + \eta\right]\right\}\right], && (34)
\end{aligned}
$$

where we only need to justify the last inequality. To this end, observe first that (32) upper bounds the first line of (34) as there clearly exists a $W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P_{z^n})$ such that $P_{z^n} \times W = P_{z^n, \hat{X}^n(z^n)}$. So we have left to argue why (33) upper bounds the second line of (34). For this it will suffice to establish the existence of $n_0(\varepsilon, \delta, \eta)$ (independent on the sequence $\{D_n\}$ and on P) such that for all $n \geq n_0(\varepsilon, \delta, \eta)$ and any $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$ with $D(P\|P_Z) < \lambda - \delta$

$$
\begin{aligned}
& \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n), z^n}, d)\right]\right\} && (35)\\
\geq \quad & \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{\{W: I(P;W) \leq R\} \cap \mathcal{C}_n(P)} F_n(P \times W, d) + \eta\right]\right\}. && (36)
\end{aligned}
$$

Inequality (36) is proved in Appendix C. The independence of $n_0(\varepsilon, \delta, \eta)$ on the sequence $\{D_n\}$ guarantees that (30) can, in fact, be replaced by $\min_{D_n \in A_n(R-\varepsilon, \lambda)} P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\} = G_n(R-\varepsilon, d, \lambda)$. Furthermore, since the sums in (36) are over expressions which do not depend on $z^n \in T_P$, (34) can be further lower bounded by

$$(n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|}\left[\sum_{\{P:D(P\|P_Z)\geq\lambda-\delta\}\cap\mathcal{M}_n}|T_P|\exp\left\{-n\left[D(P\|P_Z)+H(P)+\max_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)}F_n(P\times W,d)\right]\right\}\right.$$

$$\left.+\sum_{\{P:D(P\|P_Z)<\lambda-\delta\}\cap\mathcal{M}_n}|T_P|\exp\left\{-n\left[D(P\|P_Z)+H(P)+\max_{\{W:I(P;W)\leq R\}\cap\mathcal{C}_n(P)}F_n(P\times W,d)+\eta\right]\right\}\right],$$

$$\geq (n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\left[\sum_{\{P:D(P\|P_Z)\geq\lambda-\delta\}\cap\mathcal{M}_n}\exp\left\{-n\left[D(P\|P_Z)+\max_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)}F_n(P\times W,d)\right]\right\}\right.$$

$$\left.+\sum_{\{P:D(P\|P_Z)<\lambda-\delta\}\cap\mathcal{M}_n}\exp\left\{-n\left[D(P\|P_Z)+\max_{\{W:I(P;W)\leq R\}\cap\mathcal{C}_n(P)}F_n(P\times W,d)+\eta\right]\right\}\right] \tag{37}$$

$$\geq (n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\left(-n\min\left\{\min_{\{P:D(P\|P_Z)\geq\lambda-\delta\}\cap\mathcal{M}_n}\left[D(P\|P_Z)+\max_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)}F_n(P\times W,d)\right],\right.\right.$$

$$\left.\left.\min_{\{P:D(P\|P_Z)<\lambda-\delta\}\cap\mathcal{M}_n}\left[D(P\|P_Z)+\max_{\{W:I(P;W)\leq R\}\cap\mathcal{C}_n(P)}F_n(P\times W,d)+\eta\right]\right\}\right), \tag{38}$$

we have for all $n \geq n_0(\eta,\varepsilon,\delta)$

$$-\frac{1}{n}\log G_n(R-\varepsilon,d,\lambda) \tag{39}$$

$$\leq \frac{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|}{n}\log(n+1)+\min\left\{\min_{\{P:D(P\|P_Z)\geq\lambda-\delta\}\cap\mathcal{M}_n}\left[D(P\|P_Z)+\max_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)}F_n(P\times W,d)\right],\right.$$

$$\left.\min_{\{P:D(P\|P_Z)<\lambda-\delta\}\cap\mathcal{M}_n}\left[D(P\|P_Z)+\max_{\{W:I(P;W)\leq R\}\cap\mathcal{C}_n(P)}F_n(P\times W,d)+\eta\right]\right\}. \tag{40}$$

Combining the result of Appendix D (Claim 2) with the definition of $I(R,d,\lambda-\delta)$ (recall (10)) and the arbitrariness of $\eta>0$ gives

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log G_n(R-\varepsilon,d,\lambda)\right]\leq I(R,d,\lambda-\delta), \tag{41}$$

which, by the arbitrariness of $\varepsilon,\delta>0$, finally implies (9).

Turning to establish (8), we fix arbitrary $\varepsilon,\delta>0$ and construct a $D_n\in A_n(R+\varepsilon,\lambda)$ as follows. We first construct the codebook B_n:

- For $z^n\in T_P$ with $D(P\|P_Z)\geq\lambda+\delta$ we take

$$\hat{X}^n(z^n)=\arg\max_{\hat{x}^n\in\hat{\mathcal{X}}^n}F_n(P_{\hat{x}^n,z^n},d). \tag{42}$$

- For $z^n\in T_P$ with $D(P\|P_Z)<\lambda+\delta$ let $W^*\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)$ achieve

$$\max_{\substack{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P):\\ I(P;W)\leq R}}F_n(P\times W,d). \tag{43}$$

Then, as is well-known [5], there exists, for sufficiently large n (dependent on ε yet not on P), a set $B(P) \subseteq \hat{\mathcal{X}}^n$ of size $|B(P)| \leq 2^{n(I(P;W^*)+\varepsilon/2)} \leq 2^{n(R+\varepsilon/2)}$ and a map $\hat{X}^n(\cdot) : T_P \to B(P)$ such that $(z^n, \hat{X}^n(z^n)) \in T_{P\times W^*}$ for all $z^n \in T_P$. Let, for each $z^n \in T_P$, $\hat{X}^n(z^n)$ be given by such a map.

The above two items completely specify the mapping F_n and the codebook B_n comprising D_n. We have left to specify the codewords, or, equivalently, the uniquely decodable map $C_n : B_n \to \{0,1\}^*$. We construct the codewords as follows: use no more than $|\mathcal{Z}|\log(n+1)$ bits to convey the type to which z^n belongs. Now, if $z^n \in T_P$ with $D(P\|P_Z) \geq \lambda + \delta$, use any number of additional bits to convey $\hat{X}^n(z^n)$. Otherwise, use no more than $R + \varepsilon/2$ bits to convey $\hat{X}^n(z^n)$. This can clearly be done as, once the type to which z^n belongs is known, $\hat{X}^n(z^n)$ is known to belong to $B(P)$, whose size is no more than $2^{n(R+\varepsilon/2)}$. Using this scheme, for sufficiently large n, the fact that $L_n(Z^n, D_n) > n(R+\varepsilon)$ implies that $Z^n \in T_P$ for some P such that $D(P\|P_Z) \geq \lambda + \delta$. Consequently, for sufficiently large n (dependent on ε and δ), we have

$$\begin{aligned} P^n_{X,Z}\{L_n(Z^n, D_n) > n(R+\varepsilon)\} &\leq P^n_{X,Z}\left\{ \bigcup_{P\in\mathcal{M}(\mathcal{Z}):D(P\|P_Z)\geq\lambda+\delta} \{z^n \in T_P\} \right\} && (44) \\ &< e^{-n\lambda}. && (45) \end{aligned}$$

Hence, our scheme D_n is a bona fide member of $A_n(R+\varepsilon, \lambda)$ for n sufficiently large. Now, for the schemes $\{D_n\}$, we have

$$\begin{aligned} & P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\} && (46) \\ \leq\ & (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \Bigg[\sum_{\{P:D(P\|P_Z)\geq\lambda+\delta\}\cap\mathcal{M}_n} \sum_{z^n\in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n),z^n}, d)\right]\right\} && (47) \\ & + \sum_{\{P:D(P\|P_Z)<\lambda+\delta\}\cap\mathcal{M}_n} \sum_{z^n\in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n),z^n}, d)\right]\right\}\Bigg] && (48) \\ =\ & (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \Bigg[\sum_{\{P:D(P\|P_Z)\geq\lambda+\delta\}\cap\mathcal{M}_n} \sum_{z^n\in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})\cap\mathcal{C}_n(P)} F_n(P\times W, d)\right]\right\} \\ & + \sum_{\{P:D(P\|P_Z)<\lambda+\delta\}\cap\mathcal{M}_n} \sum_{z^n\in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{\{W:I(P;W)\leq R\}\cap\mathcal{C}_n(P)} F_n(P\times W, d)\right]\right\}\Bigg], \end{aligned}$$

where the inequality follows by combining (15) with (27) and the equality by the construction of $\hat{X}^n(\cdot)$. Since $P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\}$ clearly upper bounds $G_n(R+\varepsilon, d, \lambda)$, we have the inequality

$$\limsup_{n\to\infty} \left[\frac{1}{n}\log G_n(R+\varepsilon, d, \lambda)\right] \leq -I(R, d, \lambda+\delta), \tag{49}$$

which follows when considering the normalized logarithmic limit of the two sides of the last chain of inequalities and invoking Claim 2 of Appendix D. Finally, the arbitrariness of $\varepsilon, \delta > 0$ implies (8). □

Remark: Note that the optimal scheme constructed in the direct part of the proof is dependent on d, since the maximizers in (42) and (43) will, in general, depend on d. This is in contrast to Marton's setting (cf. [16]), for which the scheme that essentially represents the sequences of each type with the lowest distortion possible at the given rate is optimal for all values of d.

C A Special Case: The Noise-Free Setting

We dedicate this subsection to verifying that $I(R,d,\lambda)$ and Theorem 1 coincide with Marton's classical result when the source is not corrupted by noise and $\lambda = \infty$. We then look at the explicit form that $I(R,d,\lambda)$ assumes when Marton's setting is relaxed to $\lambda < \infty$.

C.1 Marton's Exponent

Let us verify that for the noise-free case and $\lambda = \infty$, Theorem 1 coincides with Marton's classical result [16]. To comply with the general formulation, we assume that in this case $\mathcal{Z} = \mathcal{X}$ and $Z = X$ $P_{X,Z}$-a.s., so that $P_{X|Z} = \delta_{X|Z}$, where $\delta_{X|Z}$ denotes the "clean" channel from $\mathcal{Z}$ into $\mathcal{X}$.

Taking $\lambda = \infty$ in (10) we have

$$I(R,d,\infty) = \inf_{P \in \mathcal{M}(\mathcal{Z})} \left[D(P\|P_Z) + \sup_{W \in \mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}}): I(P;W) \le R} F(P \times W, d) \right]. \tag{50}$$

It is easy to see that for any $Q \in \mathcal{M}(\hat{\mathcal{X}} \times \mathcal{Z})$

$$\begin{aligned} D\left(V\|P_{X|Z}|Q\right) &= D\left(Q \times V \| Q \times P_{X|Z}\right) &(51)\\ &= D\left(Q \times V \| Q \times \delta_{X|Z}\right) &(52)\\ &= \begin{cases} 0 & \text{if } Q \times V = Q \times \delta_{X|Z} \\ \infty & \text{otherwise.} \end{cases} &(53) \end{aligned}$$

Consequently,

$$\begin{aligned} F(Q,d) &= \begin{cases} 0 & \text{if } E_{Q \times \delta_{X|Z}} \rho(X,\hat{X}) > d \\ \infty & \text{otherwise} \end{cases} &(54)\\ &= \begin{cases} 0 & \text{if } E_Q \rho(Z,\hat{X}) > d \\ \infty & \text{otherwise.} \end{cases} &(55) \end{aligned}$$

Therefore

$$
\begin{aligned}
&\sup_{W:I(P;W)\leq R} F(P\times W,d) && (56)\\
=\ &\begin{cases} 0 & \text{if } \forall W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}}) \text{ with } I(P;W)\leq R,\ E_{P\times W}\rho(Z,\hat{X})>d. \\ \infty & \text{otherwise} \end{cases} && (57)\\
=\ &\begin{cases} 0 & \text{if } R(P,d)\geq R \\ \infty & \text{otherwise.} \end{cases} && (58)
\end{aligned}
$$

Finally, this implies

$$
\begin{aligned}
I(R,d,\infty) &= \inf_{P\in\mathcal{M}(\mathcal{Z}):R(P,d)\geq R} D(P\|P_Z) && (59)\\
&\triangleq F_d(R), && (60)
\end{aligned}
$$

where $F_d(R)$ is precisely the exponent from [16].

C.2 A Slight Extension

Staying in the noise-free setting, we now generalize the result of [16], where $\lambda=\infty$, to a general $\lambda\in[0,\infty]$. It is straightforward to show, similarly as in the derivation of the preceding subsection, that for all $P\in\mathcal{M}(\mathcal{Z})$, we have in the noise-free case

$$
\sup_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})} F(P\times W,d)=\infty. \qquad (61)
$$

Plugging into (10) gives

$$
\begin{aligned}
I(R,d,\lambda) &= \inf_{P\in\mathcal{M}(\mathcal{Z}):D(P\|P_Z)<\lambda}\left[D(P\|P_Z)+\sup_{W\in\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}}):I(P;W)\leq R} F(P\times W,d)\right] && (62)\\
&= \inf_{P\in\mathcal{M}(\mathcal{Z}):D(P\|P_Z)<\lambda,R(P,d)\geq R} D(P\|P_Z) && (63)\\
&= \begin{cases} \infty & \text{if } \forall P \text{ with } D(P\|P_Z)<\lambda, R(P,d)<R \\ F_d(R) & \text{otherwise} \end{cases} && (64)\\
&= \inf_{P\in\mathcal{M}(\mathcal{Z}):D(P\|P_Z)<\lambda,R(P,d)\geq R} D(P\|P_Z) && (65)\\
&= \begin{cases} \infty & \lambda<F_d(R) \\ F_d(R) & \text{otherwise,} \end{cases} && (66)
\end{aligned}
$$

where (63) follows from (58) and the rest from the definition of $F_d(R)$. Theorem 1 now gives us the following generalization of the result of [16].

Corollary 2. *For all $d\geq 0$, $R\geq 0$ and all $\lambda\in(F_d(R),\infty]$ we have*

$$
F_d(R-0)\leq\liminf_{n\to\infty}\left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] \qquad (67)
$$

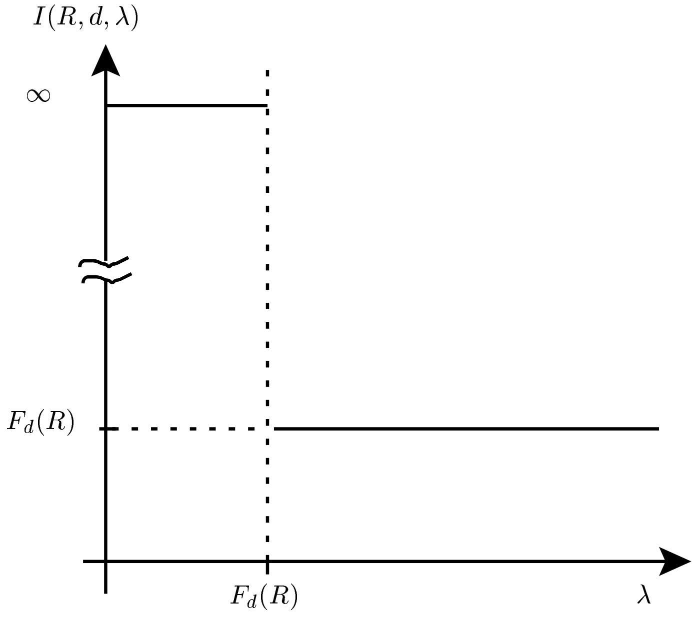

Figure 1: $I(R,d,\lambda)$ for the extension of Marton's setting to the case of a general excess-code-length exponent.

$$\leq \limsup_{n\to\infty} \left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] \leq F_d(R+0). \tag{68}$$

For $\lambda \in [0, F_d(R))$ *we have*

$$\lim_{n\to\infty} \left[-\frac{1}{n}\log G_n(R,d,\lambda)\right] = \infty. \tag{69}$$

Corollary 2 tells us two things:

- Even when the very strict requirement $\lambda = \infty$ of [16] is considerably relaxed to and $\lambda > F_d(R)$, the excess-distortion exponent is not improved.

- $P^n_{X,Z}\{\rho(X^n, \hat{X}^n(Z^n)) > nd\}$ may decay at a super-exponential rate, while keeping the excess-codelength exponent arbitrarily close to $F_d(R)$. A more detailed analysis shows that, in fact, for $\lambda \in [0, F_d(R))$, we have $G_n(R,d,\lambda) = 0$ for large enough n.

A qualitative explanation for this behavior has been given in Section 1. A more thorough discussion of the dichotomic phenomenon exhibited in Corollary 2 is deferred to Section 4, where the setting of [16] will be further generalized to the case of universal coding.

D Properties of the Function $I(R, d, \lambda)$

Unfortunately, $I(R, d, \lambda)$ turns out to be quite arduous for analysis. Convexity in R, for example, has no hope of reigning, as even $F_d(R)$ is not necessarily convex. Also, the explicit analytic form of $I(R, d, \lambda)$ for even the simplest cases, such as the binary memoryless source, corrupted by a BSC with Hamming distortion measure, turns out to be far too intricate to shed any insight on the problem.

We dedicate this subsection to a qualitative investigation of the function $I(R, d, \lambda)$ and to a verification that the form of $I(R, d, \lambda)$ adds up with what is known about rate-distortion coding of noisy sources (cf. [7]).

D.1 Qualitative Investigation of the Function $I(R, d, \lambda)$

We recall here, for convenience, that

$$I(R, d, \lambda) = \min \left\{ \inf_{P: D(P \| P_Z) \geq \lambda} a(P, \infty, d), \inf_{P: D(P \| P_Z) < \lambda} a(P, R, d) \right\}. \tag{70}$$

Clearly, we expect $I(R, d, \lambda)$ to be non-increasing in λ. As this may not be obvious from (70), let us verify that this is indeed the case.

Claim 1. *$I(R, d, \lambda)$ is a monotone non-increasing function of λ, for fixed R, d.*

Proof: For R, d fixed and $\lambda \in [0, \infty]$, consider the function $f_\lambda : \mathcal{M}(\mathcal{Z}) \to \mathbb{R} \cup \infty$ defined by

$$f_\lambda(P) = \begin{cases} a(P, \infty, d) & \text{if } D(P \| P_Z) \geq \lambda \\ a(P, R, d) & \text{if } D(P \| P_Z) < \lambda. \end{cases} \tag{71}$$

Since the function $a(\cdot, \infty, d)$ lies above the function $a(\cdot, R, d)$, it is clear that $\{f_\lambda(\cdot)\}_{\lambda \in [0, \infty]}$ is a non-increasing family of functions (with increasing λ). Consequently, the respective minima of these functions is a non-increasing function of λ. But, for each λ, $\inf_{P \in \mathcal{M}(\mathcal{Z})} f_\lambda(P) = I(R, d, \lambda)$. $\square$

To get a feel for the qualitative behavior of the function $I(R, d, \cdot)$ we further let

$$I_R = \inf_{P \in \mathcal{M}(\mathcal{Z})} a(P, R, d), \tag{72}$$

we let $P_R^* \in \mathcal{M}(\mathcal{Z})$ denote the achiever in the right side of (72), and we let $\lambda_R = D(P_R^* \| P_Z)$. Consider now one of two possibilities:

Possibility 1: $I_\infty \leq a(P_Z, R, d)$. In this case, the minimum value of the function $a(\cdot, \infty, d)$ lies below the value of $a(P_Z, R, d)$ and, hence, for all $P \in \mathcal{M}(\mathcal{Z})$ with $D(P \| P_Z)$ sufficiently small, we also have $I_\infty < a(P, R, d)$. This means that there is a whole interval $[0, \tilde{\lambda})$ in which $\inf_{P: D(P\|P_Z)<\lambda} a(P,R,d) > I_\infty$ for all $\lambda \in [0, \tilde{L})$, where $\tilde{\lambda} = \inf\{D(P\|P_Z) : a(P,R,d) \geq I_\infty\}$. Now, since $\tilde{\lambda} \leq \lambda_\infty$ (as $a(P_\infty^*, R, d) \leq a(P_\infty^*, \infty, d) = I_\infty$ and therefore clearly, for any $\lambda > \lambda_\infty$, $\inf_{P: D(P\|P_Z)<\lambda} a(P,R,d) \leq \inf_{P: D(P\|P_Z)\leq\lambda_\infty} a(P,R,d) = \inf_{P: D(P\|P_Z)\leq D(P_\infty^*\|P_Z)} a(P,R,d) \leq a(P_\infty^*, R, d) \leq I_\infty$), this means that for any $\lambda \in [0, \tilde{\lambda})$ we have $\inf_{P: D(P\|P_Z)\geq\lambda} a(P,\infty,d) = I_\infty$. Consequently, for all $\lambda \in [0, \tilde{\lambda})$, we have $I(R,d,\lambda) = \min\left\{\inf_{P: D(P\|P_Z)\geq\lambda} a(P,\infty,d), \inf_{P: D(P\|P_Z)<\lambda} a(P,R,d)\right\} = \inf_{P: D(P\|P_Z)\geq\lambda} a(P,\infty,d) = I_\infty$. By the definition of $\tilde{\lambda}$ it is also clear that for all $\lambda \geq \tilde{\lambda}$, the right branch in (70) "kicks in" and we have $I(R,d,\lambda) = \inf_{P: D(P\|P_Z)<\lambda} a(P,R,d)$. In particular, for all $\lambda \geq \lambda_R$, we clearly have $I(R,d,\lambda) = I_R$.

To sum up, the graph of the function $I(R,d,\cdot)$ starts with a plateau at level I_∞ until the right branch in (70) "kicks in". From the point λ_R and on, the function is at a plateau again at level I_R (see Figure 2 for a qualitative illustration). Note that the heights of the two plateaus mentioned are, respectively, I_∞ and I_R, the exponents associated with the "pure filtering" and "pure coding" cases. Hence we see that, in this case, there is nothing to loose in the excess distortion exponent relative to the "pure filtering" regime when the constraint on the excess code-length exponent is mild enough. It is also seen that when the excess code-length exponent is confined to be above a certain rate, there is nothing to gain relative to the "pure coding" regime where the codewords are confined to a length less than nR with probability one.

Possibility 2: $I_\infty > a(P_Z, R, d)$. In this case the filtering regime, the left branch in the right side of (70), does not "kick in" at all except for $\lambda = 0$. The reason is that, in this case, we have for any $\lambda > 0$, $\inf_{P: D(P\|P_Z)<\lambda} a(P,R,d) \leq a(P_Z,R,d) < I_\infty \leq \inf_{P: D(P\|P_Z)\geq\lambda} a(P,\infty,d)$ so that $I(R,d,\lambda) = \inf_{P: D(P\|P_Z)<\lambda} a(P,R,d)$. On the other hand, when $\lambda = 0$, the set $\{P : D(P\|P_Z) < \lambda\}$ is empty, in which case the right branch in the right side of (70) is infinite and we have $I(R,d,0) = \inf_{P: D(P\|P_Z)\geq 0} a(P,\infty,d) = I_\infty$.

Hence, in this case, we have an interesting discontinuity phenomenon at $\lambda = 0$. A qualitative explanation for this phenomenon is that, in this case, where $I_\infty > a(P_Z,R,d)$, P_Z-typical sequences are such that the optimal filtering scheme on these sequences needs more than nR bits, and, even when confined to nR bits only on P_Z, the optimal scheme will have an excess-distortion exponent no higher than $a(P_Z,R,d)$. Hence, when $\lambda = 0$, this is a purely filtering regime, whose exponent is I_∞. For any $\lambda > 0$, however, one cannot afford to allot the P_Z-typical sequences more than nR bits, in which case the associated exponent cannot be higher than

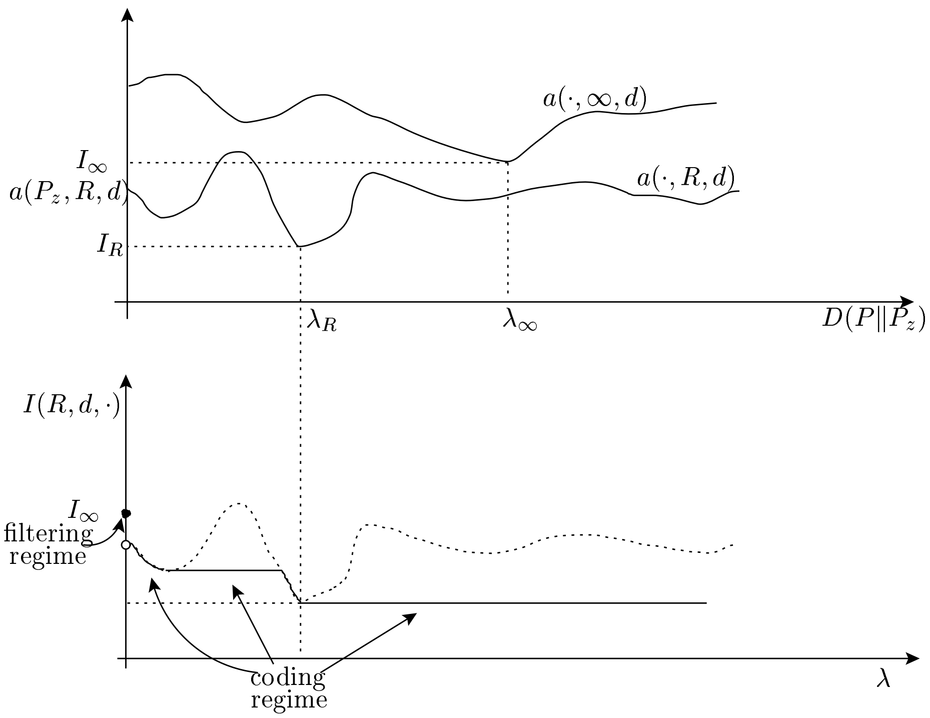

Figure 2: Possible behavior of $a(\cdot, R, d)$, $a(\cdot, \infty, d)$, and $I(R, d, \cdot)$ in the case $I_\infty \leq a(P_Z, R, d)$.

$a(P_Z, R, d)$ (see Figure 2 for a qualitative illustration). Note that the regime of textbfPossibility 2 will prevail when R is sufficiently small.

D.2 The Region Where $I(R, d, \lambda) = 0$

Clearly, if the excess-distortion exponent is positive, we must have $E\rho(X^n, \hat{X}^n(Z^n)) < nd$ for large n. We therefore expect, for any $\lambda > 0$, to have $I(R, d, \lambda) = 0$ for all $R < R^*(P_{X,Z}, d)$, where $R^*(P_{X,Z}, \cdot)$ is the rate-distortion function associated with the noisy source coding problem which was introduced in [7]. Specifically, it was shown in [7] that $R^*(P_{X,Z}, \cdot)$ is no more than the rate-distortion function of the source P_Z, with respect to the modified distortion measure $\rho^* : \mathcal{Z} \times \hat{\mathcal{X}} \to [0, \infty)$, given by:

$$\rho^*(z, \hat{x}) = E_{P_{X,Z}}[\rho(X, \hat{x}) | Z = z]. \tag{73}$$

Let us now show that our belief was justified.

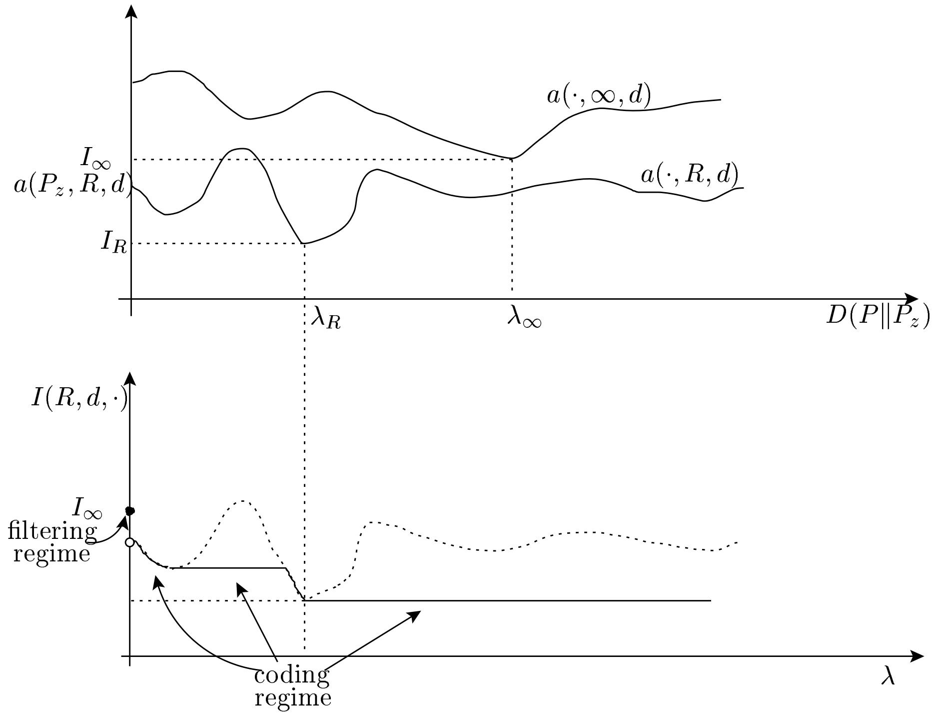

Figure 3: Possible behavior of $a(\cdot, R, d)$, $a(\cdot, \infty, d)$, and $I(R, d, \cdot)$ in the case $I_\infty > a(P_Z, R, d)$ (note the discontinuity at the origin).

Proposition 1. *For any $\lambda > 0$, we have*

$$I(R, d, \lambda) = 0 \quad \forall R < R^*(P_{X,Z}, d). \tag{74}$$

Proof: Fix a $\lambda > 0$. From the definition of $I(R, d, \lambda)$, it clearly follows that

$$I(R, d, \lambda) \leq \inf_{\substack{P \in \mathcal{M}(\mathcal{Z}) : \\ D(P\|P_Z) < \lambda}} \left[D(P\|P_Z) + \sup_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) : \\ I(P;W) \leq R}} F(P \times W, d) \right]. \tag{75}$$

Since $I(\cdot, d, \lambda)$ is non decreasing, it will suffice to show that

$$\inf_{\substack{P \in \mathcal{M}(\mathcal{Z}) : \\ D(P\|P_Z) < \lambda}} \left[D(P\|P_Z) + \sup_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) : \\ I(P;W) < R^*(P_{X,Z}, d)}} F(P \times W, d) \right] = 0. \tag{76}$$

Taking $P = P_Z$, the left side of (76), is upper bounded by

$$\sup_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) : \\ I(P_Z;W) < R^*(P_{X,Z}, d)}} F(P_Z \times W, d). \tag{77}$$

It will therefore suffice to show that the expression in (77) equals zero. But this will be the case if and only if

$$F(P_Z \times W, d) = 0 \quad \forall W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \text{ s.t. } I(P_Z; W) < R^*(P_{X,Z}, d). \tag{78}$$

Consequently, since $F(P_Z \times W, d) = 0$ if and only if $E_{(P_Z \times W) \times P_{X|Z}} \rho(X, \hat{X}) \geq d$, we will be done upon showing that

$$E_{(P_Z \times W) \times P_{X|Z}} \rho(X, \hat{X}) \geq d \quad \forall W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \text{ s.t. } I(P_Z; W) < R^*(P_{X,Z}, d). \tag{79}$$

But

$$\begin{aligned} R^*(P_{X,Z}, d) &= \inf_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}): \\ E_{P_Z \times W} \rho^*(Z, \hat{X}) \leq d}} I(P_Z; W) && (80) \\ &= \inf_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}): \\ E_{(z,\hat{x}) \sim P_Z \times W} \left[E_{P_{X,Z}} [\rho(X, \hat{x}) | Z = z] \right] \leq d}} I(P_Z; W) && (81) \\ &= \inf_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}): \\ E_{(P_Z \times W) \times P_{X|Z}} \rho(X, \hat{X}) \leq d}} I(P_Z; W). && (82) \end{aligned}$$

Consequently, if $I(P_Z; W) < R^*(P_{X,Z}, d)$ then $E_{(P_Z \times W) \times P_{X|Z}} \rho(X, \hat{X}) > d$, which implies (79). □

E Noisy Source Coding with Side Information

The setting considered thus far can be generalized to the case where side information is available at both encoder and decoder. Specifically, one can assume that the source $\{(X_k, Z_k, Y_k)\}_{k=1}^{\infty}$ is a sequence of independent drawings of a triple of dependent random variables X, Y, Z, taking values in the finite sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, respectively, and distributed according to $P_{X,Y,Z}$. As before, it may be desired to encode the sequence $\{X_k\}$ in blocks of length n, based on its noisy version $\{Z_k\}$, while complying with rate limitations of the type considered in previous subsections. In this setting we assume the availability of side information. Namely, we assume that both encoder *and* decoder have access to the side information $\{Y_k\}$. Clearly, this setting contains that considered thus far as a special case. A result generalizing and analogous to Theorem 1 can be obtained for this setting. The principles underlying the treatment of this setting and the proof of such a generalized result are the same as those underlying the proof of Theorem 1, with the only added ingredient of conditioning everything on Y^n. The interested reader is referred to [20, Section 3.E] for an elaborate treatment of this setting and for results which generalize those of the previous subsections to the case where side information is present.

4 Error Exponent for Universal Lossy Coding

In this section, we consider the case of lossy coding for a memoryless source (in the absence of noise and of side information), which, rather than being completely known, is only known to belong to a given parametric family. Note that the first step in this direction has been taken by Marton in [16], as the scheme achieving the optimal exponent in that setting did not depend on the source. As we saw in previous sections, however, when the constraint on the code-length overflow exponent is relaxed from infinity to a finite value, the optimal scheme becomes source-dependent. The basic reason for this is that, as opposed to the setting of [16], where sequences of all types could be allotted no more than nR bits, here the optimal scheme is one under which the only types that are allotted no more than nR bits are those that are sufficiently close (in the Kullback-Leibler sense) to the source. Consequently, in general, such a scheme would clearly be source-dependent. Thus, when the excess-code length constraint is relaxed to a finite value of the exponent, the issue of universality is more involved than it was in the setting of [16].

A The Error Exponent and the Universally Optimal Scheme

Let $\mathbf{P}_\Theta \stackrel{\triangle}{=} \{P_\theta : \theta \in \Theta\}$ denote a parametric family of memoryless sources with the finite alphabet $\mathcal{X}$ and the marginal $P_\theta \in \mathcal{M}(\mathcal{X})$. Considering the standard correspondence between elements of $\mathcal{M}(\mathcal{X})$ and $[0,1]^{|\mathcal{X}|}$, we may think of Θ as a subset of (the simplex in) $[0,1]^{|\mathcal{X}|}$. Furthermore, to avoid technical nuances, we assume throughout that Θ is a closed (and, hence, compact) subset of $(0,1)^{|\mathcal{X}|}$. In particular, this guarantees the continuity of $D(\cdot\|P_\theta)$ in $\mathcal{M}_+(\mathcal{X})$, for all $\theta \in \Theta$. This also implies that $\min_{x\in\mathcal{X}} \min_{\theta\in\Theta} P_\theta(x) > 0$, a fact that we will rely on in the sequel. Note that the consideration of a general subset Θ of $\mathcal{M}(\mathcal{X})$, rather than restricting attention only to the whole of $\mathcal{M}(\mathcal{X})$, is motivated by more than a desire to maximize the mathematical generality. As will be seen in the sequel, there is a clear trade-off between the size and structure of the uncertainty set Θ and the deterioration of the performance relative to the non-universal setting, so that when the uncertainty set is smaller than the whole of $\mathcal{M}(\mathcal{X})$ it is advantageous to consider coding schemes which are tailored for this smaller set. Furthermore, in many situations of practical interest, the source is most naturally modelled by a family of distributions which is a proper subset of $\mathcal{M}(\mathcal{X})$: e.g., the family of symmetric distributions or, for a large alphabet, families of one-parameter exponential distributions (cf. [19, 18]).

Given a sequence $D = \{D_n\}_{n\geq 1}$ of compression schemes, we associate with every $\theta \in \Theta$ an exponent for distortion level d defined by

$$e_d(D|\theta) \stackrel{\triangle}{=} \liminf_{n\to\infty} -\frac{1}{n}\log P_\theta^n\{\rho(X^n, \hat{X}^n) > nd\}, \tag{83}$$

where $\hat{X}^n = \hat{X}^n(X^n)$, $\hat{X}^n(\cdot)$ being the mapping from $\mathcal{X}^n$ into B_n associated with the scheme D_n (recall Section 3 for the complete definition of a coding scheme). Our goal is to find a sequence D that is independent of θ and which maximizes $e_d(D|\theta)$, uniformly for all θ if possible, under the requirement that

$$e_R(D|\theta) \stackrel{\triangle}{=} \liminf_{n\to\infty} -\frac{1}{n}\log P_\theta^n\{L_n(X^n, D_n) > nR\} \geq \lambda(\theta) \quad \forall\theta \in \Theta, \tag{84}$$

for a continuous $\lambda : \Theta \to (0, \infty]$ which is assumed given and fixed throughout this section. Note that a requirement which is similar to that in (84) can be given as follows. Letting $A_n^\theta(R, \lambda) = \left\{D_n \in \mathcal{D}_n : P_\theta^n\{L_n(Z^n, D_n) > nR\} \leq e^{-\lambda n}\right\}$, one might require that, for all n sufficiently large

$$D_n \in \bigcap_{\theta\in\Theta} A_n^\theta(R, \lambda(\theta)). \tag{85}$$

Clearly, the non-asymptotic requirement in (85) is slightly stronger than the asymptotic one of (84), though they are similar in spirit. The idea in letting $\lambda(\theta)$ vary with θ is motivated by the fact that certain sources in the parameter space Θ may be "easier" to code than others, so it may make sense to require a larger excess-code-length exponent for such sources, and a smaller one for the more "difficult" sources. More specifically, the set $\bigcap_{\theta\in\Theta} A_n^\theta(R, \lambda(\theta))$ of equation (85) can, in general, be made considerably larger by considering θ-dependent $\lambda(\cdot)$s, rather than some value of λ which would be constant all across the parameter space. This point will be elaborated on in subsection B. A similar approach was recently proven quite fruitful in the context of composite hypothesis testing (cf. [15]). Indeed, there is an intimate relationship between the decision rule proposed in [15] and our coding scheme.

As we show next, for a given $R > 0$ there exists a sequence D^u (where the superscript u stands for "universal") whose elements comply with (85) while maximizing $e_d(D|\theta)$ universally in Θ. The idea in the construction of D^u is the following. Assuming without loss of optimality (what will be justified formally in the proof of Theorem 4) that sequences of the same type are assigned codewords of essentially the same length, if the source sequence $x^n \in T_P$ and $D(P\|P_\theta) \leq \lambda(\theta)$, for some $\theta \in \Theta$, then we cannot afford more than nR bits, otherwise, under P_θ, the probability of excess code-length would be more than the probability of T_P, which would, in turn, be essentially lower bounded by $\exp\{-nD(P\|P_\theta)\} > \exp\{-n\lambda(\theta)\}$, violating (84). Consequently, if $D(P\|P_\theta) < \lambda(\theta)$ for some $\theta \in \Theta$ or, in other words, $u(P) = u(P, \lambda(\cdot)) \stackrel{\triangle}{=} \inf_{\theta \in \Theta}[D(P \| P_\theta) - \lambda(\theta)] \leq 0$, the best we can do is to allot sequences in T_P just about nR bits. For types with $u(P) > 0$, we have no limitation. This is the rationale behind the construction of D^u which follows.

For a fixed $\varepsilon > 0$, we construct the sequence $D^\varepsilon = \{D_n^\varepsilon\}_{n \geq 1}$ as follows. For the associated codebook, B_n^ε, we take

- For $x^n \in T_P$ with $u(P) > 0$, we take $\hat{X}^n(x^n) = x^n$.
- For $x^n \in T_P$ with $u(P) \leq 0$, the type covering lemma assures us of the existence of a set $B(P) \subseteq \hat{\mathcal{X}}^n$ with size $|B(P)| \leq 2^{n(R-\varepsilon)}$ and such that $\frac{1}{n}\rho(x^n, B(P)) \leq D(P, R-\varepsilon) + \delta_n(\varepsilon)$ for all $x^n \in T_P$, where $\delta_n(\varepsilon) \to 0$ as $n \to \infty$ and the sequence $\{\delta_n(\varepsilon)\}$ is independent of P. Let, for each $x^n \in T_P$, $\hat{X}^n(x^n)$ be the ρ-nearest neighbor of x^n in $B(P)$.

The above two items completely specify the mapping $\hat{X}^n(\cdot)$ and the codebook B_n comprising D_n. We construct the uniquely decodable map $C_n : B_n \to \{0,1\}^*$ as follows: use no more than $|\mathcal{X}| \log(n+1)$ bits to convey the type of x^n. Now, if $x^n \in T_P$ with $u(P) \geq 0$, use $n \log |\mathcal{X}|$ additional bits to convey $\hat{X}^n(x^n) = x^n$. Otherwise, use $n(R-\varepsilon)$ bits to convey $\hat{X}^n(x^n) \in B(P)$. Using this scheme, for sufficiently large n (such that $|\mathcal{X}|(\log(n+1))/n < \varepsilon$), $\{L_n(X^n, D_n^\varepsilon) > nR\} \subseteq T_P$ for some P with $u(P) \geq 0$. Consequently, there exists $N(\varepsilon)$, independent of $\theta \in \Theta$, such that for all $n > N(\varepsilon)$ and all $\theta \in \Theta$

$$
\begin{aligned}
P_\theta^n\{L_n(X^n, D_n^\varepsilon) > nR\} &\leq P_\theta^n\left\{\bigcup_{P: u(P) \geq 0} T_P\right\} && (86)\\
&\leq P_\theta^n\left\{\bigcup_{P: [D(P\|P_\theta) - \lambda(\theta)] \geq 0} T_P\right\} && (87)\\
&\leq (n+1)^{|\mathcal{X}|} \exp\{-n\lambda(\theta)\}. && (88)
\end{aligned}
$$

We now have a sequence of compression schemes D^ε for any $\varepsilon > 0$. We construct our scheme, D^u, as follows. Fix any decreasing sequence $\{\varepsilon_m\}_{m \geq 1}$ of positive reals satisfying $\varepsilon_m \to 0$. By the above construction and, in particular, by the uniform (in Θ) and non asymptotic nature of (88), we can readily construct a sequence of integers $\{m_n\}_{n \geq 1}$ satisfying

$$
m_n \stackrel{n \to \infty}{\longrightarrow} \infty \qquad (89)
$$

yet sufficiently slowly such that

$$\delta_n(\varepsilon_{m_n}) \stackrel{n\to\infty}{\longrightarrow} 0 \tag{90}$$

(recall the second item in the construction of the code book for the definition of $\delta_n(\varepsilon)$) and such that for all sufficiently large n

$$P_\theta^n\{L_n(X^n, D_n^{\varepsilon_{m_n}}) > nR\} \leq \exp\{-n[\lambda(\theta) - \varepsilon_{m_n}]\} \quad \forall \theta \in \Theta. \tag{91}$$

Finally, we define $D^u = \{D_n^{\varepsilon_{m_n}}\}_{n\geq 1}$. We shall write $D^{u(R)}$ when we want to make the dependence of D^u on R explicit. We begin by assessing the performance of D^u. In what follows, we let $\hat{X}_u^n(\cdot)$ denote the mapping from $\mathcal{X}^n$ into B_n associated with D_n^u.

Theorem 3. *(a) D^u satisfies, for all n*

$$D_n^u \in \bigcap_{\theta\in\Theta} A_n^\theta(R, \lambda(\theta) - \varepsilon_{m_n}). \tag{92}$$

(b) For each $\theta \in \Theta$

$$\begin{aligned} e_d(D^u|\theta) &\stackrel{\triangle}{=} \liminf_{n\to\infty} -\frac{1}{n}\log P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} && (93)\\ &\geq I_u(R, d-0, \theta) && (94)\end{aligned}$$

and

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} \leq I_u(R, d+0, \theta), \tag{95}$$

where

$$I_u(R, d, \theta) \stackrel{\triangle}{=} \inf_{B(R,d,\lambda(\cdot))} D(P\|P_\theta) \tag{96}$$

and we define

$$B(R, d, \lambda(\cdot)) \stackrel{\triangle}{=} \{P \in \mathcal{M}(\mathcal{X}) : u(P) \leq 0, D(P, R) \geq d\}. \tag{97}$$

Remark: Note the dependence of $I_u(R, d, \theta)$ on the threshold function $\lambda(\cdot)$. This dependence is suppressed in order to avoid cumbersome notation. Note also that, since $I_u(R, d, \theta)$ is a non decreasing function of d, it can only be discontinuous at a countable number of values of d. Thus, Theorem 3 gives

$$e_d(D^u|\theta) = \lim_{n\to\infty} -\frac{1}{n}\log P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} = I_u(R, d, \theta), \tag{98}$$

except, possibly, for a countable number of values of d.

Proof: Item (a) follows from the construction of D^u and, in particular, (91). Turning to the proof of (b), we note that, by our construction of D^u, we have for any $\xi > 0$ and sufficiently large n

$$\begin{aligned} P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} &\leq P_\theta^n\left\{ \bigcup_{\{P: u(P)<0, D(P,R-\varepsilon_{m_n})+\delta_n(\varepsilon_{m_n})>d\}\cap\mathcal{M}_n} T_P \right\} && (99)\\ &\leq P_\theta^n\left\{ \bigcup_{\{P: u(P)<0, D(P,R)>d-\xi\}\cap\mathcal{M}_n} T_P \right\} && (100)\\ &\leq (n+1)^{|\mathcal{X}|} \exp\left\{ -n \inf_{\{P: u(P)\leq 0, D(P,R)\geq d-\xi\}} D(P\|P_\theta) \right\}. \end{aligned}$$

Inequality (94) now follows by considering the normalized logarithm of the two ends of the above chain, letting $n \to \infty$, and then $\xi \searrow 0$.

For (95), we fix $\xi > 0$ and note that, by construction of D^u, we have for all sufficiently large n

$$\begin{aligned} P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} &\geq P_\theta^n\{\rho(X^n, \hat{X}_u^n) \geq n(d+\xi)\} && (101)\\ &\geq P_\theta^n\left\{ \bigcup_{\{P: u(P)<0, D(P,R)\geq d+2\xi\}\cap\mathcal{M}_n} T_P \right\} && (102)\\ &\geq (n+1)^{-|\mathcal{X}|} \exp\left\{ -n \min_{\{P: u(P)<0, D(P,R)\geq d+2\xi\}\cap\mathcal{M}_n} D(P\|P_\theta) \right\}. \end{aligned}$$

Considering the normalized logarithm of the two ends of the last chain of inequalities, letting $n \to \infty$, and invoking the fact that

$$\lim_{n\to\infty} \left(\min_{\{P: u(P)<0, D(P,R)\geq d\}\cap\mathcal{M}_n} D(P\|P_\theta) \right) = I_u(R, d, \theta), \tag{103}$$

which is a consequence of the definition of $I_u(R, d, \theta)$, the continuity of $D(\cdot\|P_\theta)$, the continuity of $\lambda(\cdot)$, and the continuity of $D(P, R)$ in P (cf. [5, Lemma 2.2]), we arrive at

$$\limsup_{n\to\infty} -\frac{1}{n} \log P_\theta^n\{\rho(X^n, \hat{X}_u^n) > nd\} \leq I_u(R, d+2\xi, \theta). \tag{104}$$

Finally, we take $\xi \searrow 0$ and establish (95). □

Theorem 3 assessed the performance of D^u. The following theorem tells us that no scheme can do better.

Theorem 4. *Let $\mathcal{C}(R, d, \Theta)$ denote the set of all $\theta \in \Theta$ for which $I_u(R, \cdot, \theta)$ is continuous at d, and $I_u(\cdot, d, \theta)$ is continuous at R. For any sequence of schemes D satisfying*

$$e_R(D|\theta) > \lambda(\theta) \quad \forall \theta \in \Theta, \tag{105}$$

we have

$$e_d(D|\theta) \le e_d(D^u|\theta) = I_u(R,d,\theta) \quad \forall\theta \in \mathcal{C}(R,d,\Theta). \tag{106}$$

Remark: Theorem 4 establishes the universal optimality of the scheme D^u, with respect to the class of all schemes complying with the code-word overflow probability constraint dictated by $\lambda(\cdot)$. Note that item (a) of Theorem 3 implies, in particular, that D^u satisfies (84) *uniformly* in Θ. I.e.,

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \left[-\frac{1}{n}\log P_\theta^n\{L_n(X^n, D_n^u) > nR\} - \lambda(\theta)\right] \ge 0. \tag{107}$$

Theorem 4 tells us that if a scheme has an excess-code-length exponent which is better than that of D^u, it will necessarily have an excess-distortion exponent upper bounded by that of D^u, uniformly for essentially[4] *all* values of the parameter space.

Proof: Let $D = \{D_n\}$ be any sequence of schemes satisfying (105) and assume, conversely, that there exists a $\hat\theta \in \mathcal{C}(R,d,\Theta)$ for which

$$e_d(D|\hat\theta) > e_d(D^u|\hat\theta) = I_u(R,d,\hat\theta). \tag{108}$$

Before plunging into formalities, let us outline the idea. Inequality (108) essentially implies, for large n, the existence of a type T_P on which D gives distortion $\le nd$ while D^u gives distortion $> nd$. The latter fact essentially implies that $R(P,d) > R$ *and* that P satisfies $u(P) < 0$. This is because, by the construction of D^u, the only types on which this (sequence of) scheme(s) suffers distortion $> nd$ are those for which $u(P) < 0$ and $D(P,R) > d$. Now since $R(P,d) > R$ and since D suffers distortion $\le nd$ on T_P, essentially all $x^n \in T_P$ have $L_n(x^n, D_n) > nR$. Furthermore, the fact that $u(P) < 0$ implies the existence of a $\tilde\theta \in \Theta$ for which $D(P\|P_{\tilde\theta}) \le \lambda(\tilde\theta)$. Consequently, for large n,

$$\exp\{-ne_R(D|\tilde\theta)\} \approx P_{\tilde\theta}^n\{L_n(X^n, D_n) > nR\} \gtrsim P_{\tilde\theta}^n\{T_P\} \approx \exp\{-nD(P\|P_{\tilde\theta})\} \ge \exp\{-n\lambda(\tilde\theta)\}, \tag{109}$$

which contradicts (105) for $\tilde\theta$.

Turning to the formal proof, the fact that $\hat\theta \in \mathcal{C}(R,d,\Theta)$ implies that (98) holds for $\theta = \hat\theta$. It thus follows from (108) that we can find an $\eta = \eta(\hat\theta) > 0$ such that for all sufficiently large[5] n

$$-\frac{1}{n}\log P_{\hat\theta}^n\{\rho(X^n, \hat X_u^n) > nd\} + \eta \le -\frac{1}{n}\log P_{\hat\theta}^n\{\rho(X^n, \hat X^n) > nd\}, \tag{110}$$

[4]It is tedious but straightforward to show, using the continuity of $D(\cdot,\cdot)$ in $\mathcal{M}(\mathcal{X}) \times [0,\infty)$, that when Θ and $\lambda(\cdot)$ are sufficiently well-behaved, $\mathcal{C}(R,d,\Theta) = \Theta$.

[5]Throughout this proof, the "sufficiently large n" may be dependent on $\hat\theta$.

where $\hat{X}^n = \hat{X}^n(X^n)$, $\hat{X}^n(\cdot)$ being the mapping associated with the scheme D_n. Therefore, for all sufficiently large n

$$\begin{aligned} P_{\hat{\theta}}^n\{\rho(X^n,\hat{X}^n) > nd\} &\le P_{\hat{\theta}}^n\{\rho(X^n,\hat{X}_u^n) > nd\}\cdot e^{-n\eta} && (111)\\ &\le \exp\{-n[I_u(R,d,\hat{\theta})+\eta/2]\}, && (112)\end{aligned}$$

where the first inequality is a rewriting of (110) and the second inequality follows from a reuse of the fact that (98) holds for $\theta = \hat{\theta}$. The continuity of $I_u(\cdot, d, \hat{\theta})$ at R (recall that $\hat{\theta} \in \mathcal{C}(R, d, \Theta)$) guarantees the existence of some $\alpha = \alpha(\hat{\theta}) > 0$ such that for all sufficiently large n

$$P_{\hat{\theta}}^n\{\rho(X^n,\hat{X}^n) > nd\} \le \exp\{-n[I_u(R+\alpha,d,\hat{\theta})+\eta/3]\}. \tag{113}$$

Now, by the definition of $I_u(R+\alpha, d, \hat{\theta})$ (recall (96)), there exists a $Q \in \mathcal{M}(\mathcal{X})$ and a sequence $\{P_n\}$, where $P_n \in \mathcal{M}(\mathcal{X}) \cap \mathcal{M}_n$ such that $u(P_n) \le 0$, $D(P_n, R+\alpha) \ge d$, $P_n \to Q$, and $D(Q\|P_{\hat{\theta}}) = I_u(R+\alpha, d, \hat{\theta})$. Hence, by the continuity of $D(\cdot\|P_{\hat{\theta}})$, we must have for all sufficiently large n

$$D(P_n\|P_{\hat{\theta}}) \le I_u(R+\alpha,d,\hat{\theta}) + \eta/4. \tag{114}$$

Also, the fact that $u(P_n) \le 0$ (recall the continuity of P_θ in θ, the continuity of $\lambda(\cdot)$, and the compactness of Θ) implies the existence of $\tilde{\theta}_n \in \Theta$ with

$$D(P_n\|P_{\tilde{\theta}_n}) \le \lambda(\tilde{\theta}_n), \tag{115}$$

such that $\theta_n \to \theta$ and

$$D(Q\|P_\theta) \le \lambda(\theta). \tag{116}$$

On the other hand, since P_n satisfies (114), there must exist a $B(P_n) \subseteq T_{P_n}$ with, say, $|B(P_n)| \ge \frac{1}{2}|T_{P_n}|$ such that

$$\rho(x^n, \hat{X}^n(x^n)) \le nd \quad \forall x^n \in B(P_n) \tag{117}$$

since, otherwise, (113) would be violated. By the converse to lossy coding (or type covering), (117) necessarily implies that, for all sufficiently large n (dependent on α yet not on the particular sequence $\{P_n\}$ chosen), $|\{U_n(x^n) : x^n \in B(P_n)\}| \ge 2^{n[R(P_n,d)-\alpha/2]}$ which, since $D(P_n, R+\alpha/2) \ge d$, leads to

$$|\{U_n(x^n) : x^n \in B(P_n)\}| \ge 2^{n(R+\alpha/2)}. \tag{118}$$

Inequality (118) inevitably implies, for sufficiently large n (again, independent of the particular sequence $\{P_n\}$ chosen), the existence of $\overline{B}(P_n) \subseteq B(P_n)$ with, say, $|\overline{B}(P_n)| \geq \frac{1}{2}|B(P_n)| \geq \frac{1}{4}|T_{P_n}|$ such that

$$L_n(x^n, D_n) \geq n(R + \alpha/4) \quad \forall x^n \in \overline{B}(P_n) \tag{119}$$

(this is established easily like in the proof of Theorem 1, cf. (A.36) in particular). Consequently, for sufficiently large n,

$$\begin{aligned} P^n_{\tilde{\theta}_n}\{L_n(x^n, D_n) > nR\} &\geq P^n_{\tilde{\theta}_n}\{L_n(x^n, D_n) \geq n(R + \alpha/4)\} & (120)\\ &\geq P^n_{\tilde{\theta}_n}\{\overline{B}(P_n)\} & (121)\\ &\geq \frac{1}{4}(n+1)^{-|\mathcal{X}|}\exp\{-nD(P\|P_{\tilde{\theta}_n})\} & (122)\\ &\geq \frac{1}{4}(n+1)^{-|\mathcal{X}|}\exp\{-n\lambda(\tilde{\theta}_n)\}, & (123) \end{aligned}$$

where the last inequality follows from (115). Therefore

$$\liminf_{n\to\infty} -\frac{1}{n}\log P^n_{\tilde{\theta}_n}\{L_n(x^n, D_n) > nR\} \leq \liminf_{n\to\infty} \lambda(\tilde{\theta}_n) = \lambda(\theta), \tag{124}$$

where the equality follows from the continuity of $\lambda(\cdot)$ and the fact that $\tilde{\theta}_n \to \tilde{\theta}$. To conclude, it is shown in the Appendix that for any sequence $\{A_n\}$, where $A_n \subseteq \mathcal{X}^n$ and $A_n \neq \emptyset$, we have

$$\lim_{n\to\infty} \frac{1}{n}\log \frac{P^n_{\tilde{\theta}_n}\{A_n\}}{P^n_{\tilde{\theta}}\{A_n\}} = 0, \tag{125}$$

which implies

$$\liminf_{n\to\infty} -\frac{1}{n}\log P^n_{\tilde{\theta}_n}\{L_n(x^n, D_n) > nR\} = \liminf_{n\to\infty} -\frac{1}{n}\log P^n_{\tilde{\theta}}\{L_n(x^n, D_n) > nR\} = e_R(D|\tilde{\theta}). \tag{126}$$

Consequently, we have established

$$e_R(D|\tilde{\theta}) \leq \lambda(\tilde{\theta}), \tag{127}$$

contradicting (105). $\square$

B Discussion

Working with the rate-distortion rather than the distortion-rate function, equation (96) implies that

$$I_u(R, d, \theta) = \inf_{\{P: u(P) \leq 0, R(P,d) \geq R\}} D(P\|P_\theta). \tag{128}$$

Let further $F_d^\theta(R)$ denote Marton's exponent [16] corresponding to the source P^θ:

$$F_d^\theta(R) = \inf_{\{P: R(P,d) \geq R\}} D(P \| P_\theta). \tag{129}$$

Note first that when Θ is a singleton we have

$$I_u(R, d, \theta) = \inf_{\{P: D(P\|P_\theta) \leq \lambda(\theta), R(P,d) \geq R\}} D(P \| P_\theta) = \begin{cases} F_d^\theta(R) & \text{if } \lambda(\theta) \geq F_d^\theta(R) \\ \infty & \text{otherwise,} \end{cases} \tag{130}$$

where the ∞ branch on the right side of (130) follows from the fact that when $\lambda(\theta) < F_d^\theta(R)$, the set $\{P : D(P\|P_\theta) \leq \lambda(\theta), R(P,d) \geq R\}$ is empty. Hence, as required, when Θ is a singleton, $I_u(R,d,\theta)$ coincides with the exponent function derived in subsection 3.C.2.

The next simple observation we make is that when $\lambda(\theta) \geq F_d^\theta(R)$ then the P achieving the minimum in (129) satisfies $D(P\|P_\theta) \leq \lambda(\theta)$ and, a fortiori, satisfies $u(P) \leq 0$ so that we have

$$I_u(R, d, \theta) = F_d^\theta(R) \quad \forall \theta \in \Theta \text{ s.t. } \lambda(\theta) \geq F_d^\theta(R). \tag{131}$$

Equation (131) can be given both an optimistic and a pessimistic interpretation. On the one hand, as we have seen in Corollary 2, even in the non universal setting, the best achievable exponent for the distortion when subject to an overflow exponent larger than $F_d(R)$ is $F_d(R)$ itself. Hence, equation (131) tells us that at all points $\theta \in \Theta$ in which $\lambda(\theta) \geq F_d^\theta(R)$, we are not paying any price for universality, or, more explicitly, for the fact that our scheme has to comply with the overflow exponent constraint dictated by $\lambda(\cdot)$ for *all* $\theta' \in \Theta$. For a pessimistic view, note that, in particular, the regime of equation (131) holds for $\lambda(\cdot)$ given by $\lambda(\theta) = \infty$ $\forall \theta \in \Theta$. In this case, we are back to Marton's setting [16], where the code word length is restricted to nR bits with probability one. In this context it is worthwhile to note that the optimal (sequence of) scheme(s) in Marton's setting for achieving the best exponent, which essentially covers each type to within the lowest distortion achievable with nR bits, is universal. Equation (131) tells us that any relaxation of the overflow exponent constraint to a value $\lambda(\theta)$ greater than $F_d^\theta(R)$ will not lead to a better distortion exponent.

A more significant divergence from the non-universal setting is observed for values of θ for which $\lambda(\theta) < F_d^\theta(R)$. While in the source-dependent setting, Corollary 2 gave a rate of ∞, $I_u(R,d,\theta)$ can well be finite. Technically, this follows from the fact that in the source-dependent setting, where $\Theta = \{\theta\}$, the minimizing set in (128), namely $\{P \in \mathcal{M}(\mathcal{X}) : D(P\|P_\theta) - \lambda(\theta) \leq 0, R(P,d) \geq R\}$ is empty whenever $\lambda(\theta) < F_d^\theta(R)$ (cf. subsection 3.C.2). In the universal setting, however, even when the set $\{P \in \mathcal{M}(\mathcal{X}) : D(P\|P_\theta) - \lambda(\theta) \le 0, R(P,d) \ge R\}$ is empty, the set $\{P \in \mathcal{M}(\mathcal{X}) : u(P) \le 0, R(P,d) \ge R\}$ may not be empty and, hence, the exponent $I(R,d,\theta)$ may be finite. The rationale behind this phenomenon is the following: The fact that $\lambda(\theta) < F_d^\theta(R)$ implies that all types P with $R(P,d) > R$ are such that $D(P\|P_\theta) > \lambda(\theta)$. Therefore, in the non universal setting, one could allot to each type, P, the $\approx nR(P,d)$ bits necessary to cover it with distortion no more than nd, thereby essentially annihilating the probability for distortion exceeding nd, while complying with the requirement on the exponent associated with the code word length. In the universal setting, on the other hand, the fact that $\lambda(\theta) < F_d^\theta(R)$ does *not* imply that all types P with $R(P,d) > R$ can be covered with distortion less than nd. It only means, as in the non universal case, that all types P with $R(P,d) > R$ are such that $D(P\|P_\theta) > \lambda(\theta)$. There might exist some other $\theta' \in \Theta$ for which $D(P\|P_\theta) < \lambda(\theta')$ and, if this is the case, complying with the requirement on the code word exponent forces one to allot no more than nR bits to the type P and, hence, to suffer distortion greater than nd on this type.

Thus, unlike the non universal setting, $\lambda(\theta) < F_d^\theta(R)$ does not automatically imply $I_u(R,d,\theta) = \infty$. What it does imply, however, is that if all P_θ^*-s achieving $F_d^\theta(R)$ are such that $u(P_\theta^*) > 0$ then $I_u(R,d,\theta)$ is strictly greater than $F_d^\theta(R)$. The situation in these cases is similar to that in the non universal setting in the dichotomy exhibited between the regime $\lambda(\theta) > F_d^\theta(R)$, where the best achievable exponent is $F_d^\theta(R)$, and the regime $\lambda(\theta) > F_d^\theta(R)$, where a better exponent is achievable.

Notably, the discussion above tells us that if $\lambda(\cdot)$ is such that the set $\{P \in \mathcal{M}(\mathcal{X}) : u(P) \le 0, R(P,d) \ge R\}$ is not empty then, at all $\theta \in \Theta$ for which $\lambda(\theta) < F_d^\theta(R)$, the exponential price of universality is infinite. This is because, in this case and for such a $\theta \in \Theta$, the θ-dependent scheme complying only with the code word overflow exponent constraint for the source P_θ is infinite, while the universal one is finite. A natural question arising in this context is whether there exists a $\lambda(\cdot) > 0$ such that the $I_u(R,d,\theta) = \infty$ regime is attained. For this we would need the set $\{P \in \mathcal{M}(\mathcal{X}) : u(P) \le 0, R(P,d) \ge R\}$ to be empty or, in other words, we need $u(P) > 0$ for all P with $R(P,d) \ge R$. This implies that $I_u(R,d,\theta) = \infty$ if and only if $\inf_{\Psi(R,d)} u(P) = \inf_{\Psi(R,d)} \inf_{\theta' \in \Theta}[D(P\|P_{\theta'}) - \lambda(\theta')] = \min_{\Psi(R,d)} \min_{\theta' \in \Theta}[D(P\|P_{\theta'}) - \lambda(\theta')] > 0$, where $\Psi(R,d) \stackrel{\triangle}{=} \{P \in \mathcal{M}(\mathcal{X}) : R(P,d) \ge R\}$ and the equality holds by our compactness assumption on Θ, the continuity of $\lambda(\cdot)$, and the compactness of $\Psi(R,d)$ (which follows from the continuity of $R(\cdot,\cdot)$ in $\mathcal{M}(\mathcal{X}) \times [0,\infty)$). The fact that

$$\inf_{\Psi(R,d)} u(P) = \min_{\Psi(R,d)} \min_{\theta' \in \Theta}[D(P\|P_{\theta'}) - \lambda(\theta')] = \min_{\theta' \in \Theta}[\min_{\Psi(R,d)} D(P\|P_{\theta'}) - \lambda(\theta')] \tag{132}$$

implies that a necessary and sufficient condition for $I_u(R,d,\theta)=\infty$ is

$$\min_{\Psi(R,d)} D(P\|P_{\theta'}) > \lambda(\theta'), \quad \forall \theta' \in \Theta. \tag{133}$$

Finally, equation (133) (again, by the compactness of $\Psi(R,d)$ and Θ) implies that a necessary and sufficient condition for the existence of a strictly positive $\lambda(\cdot)$ which will satisfy (133) is that $\min_{\Psi(R,d)} D(P\|P_{\theta'}) > 0$ for all $\theta' \in \Theta$ or, in other words, that

$$\Psi(R,d) \cap P_\Theta = \emptyset. \tag{134}$$

This condition can be qualitatively explained as follows. Since (134) implies that $R(P_\theta,d) < R$ for every $\theta \in \Theta$, no more than nR bits are needed to represent sequences of types sufficiently close to P_θ with distortion less than nd. Hence, one can afford to cover all types with a sufficient rate to guarantee that the distortion does not exceed nd for *all* sequences and, at the same time, guarantee that the probability of the event that more than nR bits would be needed will decay exponentially. Conversely, when (134) is not satisfied, there exists a $\theta \in \Theta$ with $R(P_\theta,d) \geq R$. Since, in order to maintain exponential decay of the probability of codeword length overflow, sequences whose types are close to P_θ cannot be allotted more than nR bits, the distortion on essentially all such sequences will have to exceed nd and, consequently, under P_θ^n, the probability of exceeding distortion nd cannot be exponentially negligible.

We note the dichotomy established in the above discussion between the regime $I_u(R,d,\theta)=\infty$ and that where $I_u(R,d,\theta)<\infty$. Specifically, note that the condition (134) is independent of $\theta \in \Theta$. In other words, we have either $I_u(R,d,\theta)=\infty$ for all $\theta \in \Theta$ or $I_u(R,d,\theta)<\infty$ for all $\theta \in \Theta$. The qualitative reason for this is that, as was discussed above, the regime $I_u(R,d,\theta)=\infty$ is reached when a scheme whose distortion on *all* sequences does not exceed nd is employed. In such a case, the probability of exceeding distortion nd is annihilated under any source P_θ^n. In the case where $I_u(R,d,\theta)<\infty$, on the other hand, the scheme employed is such that the distortion suffered on some types exceeds nd. The probability of these types cannot be exponentially negligible under any source P_θ^n and, thus, we will have $I_u(R,d,\theta)<\infty$ for all $\theta \in \Theta$.

From the above discussion it follows that when condition (134) does hold, a logical choice of the function $\lambda(\cdot)$ should be based on the left side of (133). One example for such a choice is given by $\lambda(\theta)=(1-\varepsilon)\cdot\min_{\Psi(R,d)} D(P\|P_{\theta'})$, for some small $\varepsilon>0$, which clearly satisfies (133) whenever condition (134) holds, while also satisfying $\lambda(\theta)/\lambda'(\theta) > (1-\varepsilon)$ for all $\theta \in \Theta$ and any other $\lambda'(\cdot)$ satisfying (133).

A discussion of the analogy between the optimal decision rule in [15] and that of the present work now seems in order. The setting in [15] is one of composite hypothesis testing. The problem is that of deciding, based on observing an i.i.d. sample y^n of marginal P_θ whether $\theta \in \Theta_1$ or $\theta \in \Theta_2$. The goal is to find a decision rule which will maximize the mis-detection exponent, subject to a constraint on the false-alarm exponent, which may be θ-dependent. More specifically, the goal is to find a decision rule which maximizes the second kind error exponent uniformly over θ_2, subject to the condition that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_{e_1}(\Omega^n|\theta_1) \geq \lambda(\theta_1), \quad \forall\theta_1 \in \Theta_1, \tag{135}$$

where $P_{e_1}(\Omega^n|\theta_1)$ is the first kind error probability associated with the decision rule Ω^n, when the active source is $\theta_1 \in \Theta_1$. The optimal decision rule Ω^n for this setting was shown to be one which decides on the first hypothesis when

$$\inf_{\theta_1\in\Theta_1}[D(P_{y^n}\|P_{\theta_1}) - \lambda(\theta_1)] \leq 0, \tag{136}$$

and on the second hypothesis otherwise. Note the similarity between this optimal decision rule and the optimal coding scheme constructed for the setting of the present work in subsection 4.A. The dichotomy between the sequences on which the first hypothesis will be chosen and those on which the second hypothesis will be chosen, in the former case, and that between the sequences that will be allotted just about nR bits and those that will be allotted an unlimited number of bits, in the latter case, is essentially identical. This similarity is not due to mere chance. The reasons for this dichotomy, in both settings, are analogous: In the setting of [15], assuming that sequences of equal types receive equal treatment (an assumption which is justified by the fact that P_{y^n} is a sufficient statistic for this problem, cf., e.g., [22, Lemma 1]), the fact that (136) holds implies the existence of $\theta_1 \in \Theta_1$ for which $D(P_{y^n}\|P_{\theta_1}) \leq \lambda(\theta_1)$. Consequently, the first hypothesis must be chosen whenever (136) is satisfied since, otherwise, (135) would be violated. Analogously, in the setting of the present work, as discussed in the previous subsection, assuming that sequences of equal types are allotted code-words of approximately equal lengths (an assumption which is justified by part B of Theorem 4), the fact that the source sequence belongs to a type P satisfying $u(P) \leq 0$ implies the existence of $\theta \in \Theta$ for which $D(P\|P_\theta) \leq \lambda(\theta)$. Hence, such source sequences must be allotted no more that nR bits as, otherwise, the excess-code-length requirement (84) would be violated. Another analogy with [15] is in the motivation for taking a θ-dependent $\lambda(\cdot)$ and in the considerations guiding the search of a sensible $\lambda(\cdot)$: In [15], a judicious choice of $\lambda(\cdot)$ can make the difference between the existence of a decision rule with exponential decay of $P_{e_2}(\Omega^n|\theta_2)$ for all $\theta_2 \in \Theta_2$ and the non-existence of such a rule. In the present setting, as discussed above, a sensible choice of $\lambda(\cdot)$ could mean the difference between not having to pay a price for universality and having to pay such a price. Furthermore, when there is an inevitable price to be paid, the right choice of $\lambda(\cdot)$ will mean the difference between paying essentially this price and paying significantly more in the excess-code-length exponent.

B.1 Special Cases

The Case where $P_\Theta = \mathcal{M}(\mathcal{X})$

As an extreme example, note that when[6] $P_\Theta = \mathcal{M}(\mathcal{X})$, we trivially have, for any $\lambda(\cdot)$, $u(P) \leq 0$ for all $P \in \mathcal{M}(\mathcal{X})$. Therefore, in this case, $I_u(R,d,\theta) = F_d^\theta(R)$. Qualitatively, the reason is that, in this case, one can clearly not afford to allot sequences of any type more than nR bits. Under this restriction, the best scheme is that of Marton's, which is already universal for the case $\lambda(\cdot) = \infty$. Hence, for this case, the fact that $\lambda(\cdot) < \infty$ does not result in a better exponent. Let us now examine a slightly less trivial case, which concretely illustrates some of the points discussed above.

The Case $\Theta = \{\theta_1, \theta_2\}$

Consider the case where there are only two possible sources, P_{θ_1} and P_{θ_2}. In this case we have

$$I_u(R,d,\theta_1) = \inf_{\left\{P: \left\{\begin{array}{c} D(P\|P_{\theta_1}) \leq \lambda(\theta_1) \text{ or} \\ D(P\|P_{\theta_2}) \leq \lambda(\theta_2) \end{array}\right\}, R(P,d) \geq R\right\}} D(P\|P_{\theta_1}) \tag{137}$$

$$= \begin{cases} F_d^{\theta_1}(R) & \text{if} \quad \lambda(\theta_1) \geq F_d^{\theta_1}(R) \\ \infty & \text{if} \quad \left\{\begin{array}{c} \lambda(\theta_1) < F_d^{\theta_1}(R) \text{ and} \\ \lambda(\theta_2) < F_d^{\theta_2}(R) \end{array}\right\} \\ \left(\inf_{\{P: D(P\|P_{\theta_2}) \leq \lambda(\theta_2), R(P,d) \geq R\}} D(P\|P_{\theta_1})\right) & \text{if} \quad \left\{\begin{array}{c} \lambda(\theta_1) < F_d^{\theta_1}(R) \text{ and} \\ \lambda(\theta_2) \geq F_d^{\theta_2}(R) \end{array}\right\} \end{cases} \tag{138}$$

Note that, by symmetry, $I_u(R,d,\theta_2)$ looks the same with $1 \to 2$ and $2 \to 1$. Hence, we see that, as discussed above, we either have $I_u(R,d,\theta_1) = I_u(R,d,\theta_2) = \infty$ or both $I_u(R,d,\theta_1)$ and $I_u(R,d,\theta_2)$ finite. These two regimes are reached, respectively, according to whether both $\lambda(\theta_1) < F_d^{\theta_1}(R)$ and $\lambda(\theta_2) < F_d^{\theta_2}(R)$ or not. It is also evident from (138) that if we have both $\lambda(\theta_1) < F_d^{\theta_1}(R)$ and $\lambda(\theta_2) \geq F_d^{\theta_2}(R)$, then $F_d^{\theta_1}(R) < I_u(R,d,\theta_1) < \infty$. To see why the inequality $F_d^{\theta_1}(R) < I_u(R,d,\theta_1)$ is strict note that in this case, where we assume $\lambda(\theta_1) < F_d^{\theta_1}(R)$, if P^* achieves $F_d^{\theta_1}(R)$ then, by the definition of $F_d^{\theta_1}(R)$, $D(P^*\|P_{\theta_1}) > \lambda(\theta_1)$ and, hence, P^* cannot belong to the minimizing set

[6] Note that, formally, this case does not conform to our standing assumption that $P_\Theta \subseteq \mathcal{M}_+(\mathcal{X})$. We will not be pedantic about that as the discussion of this trivial case is only a qualitative one.

in (137). Thus we see that in the case where $\lambda(\theta_1) < F_d^{\theta_1}(R)$ yet $\lambda(\theta_2) \geq F_d^{\theta_2}(R)$, as in the source-dependent case (cf. subsection 3.D), a better exponent than $F_d^{\theta_1}(R)$ is achievable. Unfortunately, unlike the source-dependent case, it is finite.

5 Conclusion and Future Directions

The study of error exponents for lossy coding under a constrained probability of code word overflow has been initiated in this work. A single-letter expression was obtained first for the case of lossy coding of noise-corrupted memoryless sources, and then generalized for the case where side information is available at both encoder and decoder. Finally, the case of a clean (uncorrupted) sequence, generated by an unknown member of a family of memoryless sources, was considered.

It would be interesting to study the natural extension of the universal setting of Section 4 to the noisy case. This is not a trivial extension of the noise-free universal setting. For this case, one can show that the discrimination between the types on which z^n will be allotted no more than nR bits and those on which it will not be limited in codeword length is quite similar to that made in the noise-free case. The difficulty comes in at the stage where one needs to find the best scheme within each type. While in the noise-free case, the optimal scheme within each type was independent of the particular active source, P_θ, in the noisy setting, as was seen in Section 3, this is no longer the case. There would, therefore, no longer be hope of finding a single universal scheme which will be optimal for all sources in the class, in contrast to what was shown to be the case in the noise-free setting. Furthermore, in universal coding of noisy sources, under the expected distortion criterion, there, generally, do not exist schemes which are optimal for all sources in the reference class. Thus, in that setting, one is led to consider somewhat less ambitious optimality criteria, such as the minimax criterion (cf. [14, 6]). This will probably also be the case in the study of error-exponents for the noisy universal setting, which is under current investigation.

Appendix

A Proof of (125)

Note that by our assumption on Θ, $P_{\tilde{\theta}} \in \mathcal{M}_+(\mathcal{X})$ so that $a(\tilde{\theta}) \stackrel{\triangle}{=} \min_{x \in \mathcal{X}} P_{\tilde{\theta}}(x) > 0$. Let $\varepsilon_n = \|P_{\tilde{\theta}} - P_{\tilde{\theta}_n}\|$ and recall that, by our choice of $\{\tilde{\theta}_n\}$, $\varepsilon_n \to 0$. Now, for each n

$$\frac{P^n_{\tilde{\theta}_n}\{A_n\}}{P^n_{\tilde{\theta}}\{A_n\}} = \frac{\sum_{x^n \in A_n} P^n_{\tilde{\theta}_n}\{x^n\}}{\sum_{x^n \in A_n} P^n_{\tilde{\theta}}\{x^n\}} \tag{A.1}$$

$$= \frac{\sum_{x^n \in A_n} \prod_{i=1}^n P_{\tilde{\theta}_n}(x_i)}{\sum_{x^n \in A_n} \prod_{i=1}^n P^n_{\tilde{\theta}}(x_i)} \tag{A.2}$$

$$\leq \frac{\sum_{x^n \in A_n} \prod_{i=1}^n [P_{\tilde{\theta}}(x_i) + \varepsilon_n]}{\sum_{x^n \in A_n} \prod_{i=1}^n P_{\tilde{\theta}}(x_i)} \tag{A.3}$$

$$= \frac{\sum_{x^n \in A_n} \prod_{i=1}^n P_{\tilde{\theta}}(x_i) \left[1 + \frac{\varepsilon_n}{P_{\tilde{\theta}}(x_i)}\right]}{\sum_{x^n \in A_n} \prod_{i=1}^n P_{\tilde{\theta}}(x_i)} \tag{A.4}$$

$$\leq \left[1 + \frac{\varepsilon_n}{a(\tilde{\theta})}\right]^n. \tag{A.5}$$

Consequently

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{P^n_{\tilde{\theta}_n}\{A_n\}}{P^n_{\tilde{\theta}}\{A_n\}} \leq \limsup_{n\to\infty} \log \left[1 + \frac{\varepsilon_n}{a(\tilde{\theta})}\right] = 0. \tag{A.6}$$

Analogously to (A.5), we can show that

$$\frac{P^n_{\tilde{\theta}_n}\{A_n\}}{P^n_{\tilde{\theta}}\{A_n\}} \geq \left[1 - \frac{\varepsilon_n}{a(\tilde{\theta})}\right]^n, \tag{A.7}$$

which, upon taking $\liminf_{n\to\infty} \frac{1}{n}$ log's, completes the proof. $\square$

B

B.1 Proof that $I(R-0, d, \lambda+0)$ and $I(R+0, d, \lambda-0)$ are Well-Defined

We need to prove that the limits

$$\lim_{\varepsilon\to 0^+, \delta\to 0^+} I(R+\varepsilon, d, \lambda-\delta) \tag{A.8}$$

and

$$\lim_{\varepsilon\to 0^+, \delta\to 0^+} I(R-\varepsilon, d, \lambda+\delta) \tag{A.9}$$

exist. To establish the existence of the first limit, let $\{(R_1^{(n)}, \lambda_1^{(n)})\}$ and $\{(R_2^{(n)}, \lambda_2^{(n)})\}$ be two arbitrary sequences of pairs converging to the pair (R, λ) and satisfying $R_i^{(n)} > R$ and $\lambda_i^{(n)} < \lambda$ for all n and $i = 1, 2$. We will be done upon showing that the limits $\lim_{n\to\infty} I(R_1^{(n)}, d, \lambda_1^{(n)})$ and $\lim_{n\to\infty} I(R_2^{(n)}, d, \lambda_2^{(n)})$ exist and are equal. The existence of these limits is guaranteed by the monotonicity of $I(R, d, \lambda)$ in both R (increasing) and λ (decreasing). More specifically, this implies that for $i = 1, 2$ and any n_0 there exists $m_i(n_0)$ (namely, one for which $R_i^{(n)} \leq R_i^{(n_o)}$ and $\lambda_i^{(n)} \geq \lambda_i^{(n_0)}$ for all $n \geq m_i(n_0)$) such that $I(R_i^{(n_0)}, d, \lambda_i^{(n_0)}) \geq I(R_i^{(n)}, d, \lambda_i^{(n)})$ for all $n \geq m_i(n_0)$, which clearly implies the existence of the limits $\lim_{n\to\infty} I(R_i^{(n)}, d, \lambda_i^{(n)})$. To establish the equality between the limits it suffices to show that

$$\lim_{n\to\infty} I(R_1^{(n)}, d, \lambda_1^{(n)}) \geq \lim_{n\to\infty} I(R_2^{(n)}, d, \lambda_2^{(n)}) \tag{A.10}$$

(since the arbitrary labelling of the indices $i = 1, 2$ will imply the reverse inequality as well). Inequality (A.10), however, follows easily from the monotonicity of $I(R, d, \lambda)$ as well by observing that for any n_0 there exists $m_(n_0)$ (namely, one for which $R_2^{(n)} \leq R_1^{(n_o)}$ and $\lambda_2^{(n)} \geq \lambda_1^{(n_0)}$ for all $n \geq m(n_0)$) such that $I(R_1^{(n_0)}, d, \lambda_1^{(n_0)}) \geq I(R_2^{(n)}, d, \lambda_2^{(n)})$ for all $n \geq m(n_0)$. Thus, the existence of the limit in (A.8) is established. The existence of the limit in (A.9) is proven similarly. □

B.2 Proof that Equation (13) Holds Outside a Set of Pairs (R, λ) of Zero Lebesgue Measure

Since the monotonicity of $I(R, d, \lambda)$ implies, for all pairs (R, λ), $I(R - 0, d, \lambda + 0) \leq I(R, d, \lambda) \leq I(R + 0, d, \lambda - 0)$, it follows from Theorem 1 that it will suffice to show that the set $\mathcal{S} \subseteq \mathbb{R}^2$ defined by

$$\mathcal{S} = \{(R, \lambda) \in [0, \infty)^2 : I(R - 0, d, \lambda + 0) < I(R + 0, d, \lambda - 0)\} \tag{A.11}$$

has zero Lebesgue measure. To this end we note first that the fact that $I(R - 0, d, \lambda + 0)$ and $I(R + 0, d, \lambda - 0)$ are well-defined, i.e. that the limits which define these respective quantities exist (cf. subsection B.1), implies that for each pair (R, λ) we have both

$$I(R - 0, d, \lambda + 0) = \lim_{\varepsilon\to 0^+} I(R - \varepsilon, d, \lambda + \varepsilon) \tag{A.12}$$

and

$$I(R + 0, d, \lambda - 0) = \lim_{\varepsilon\to 0^+} I(R + \varepsilon, d, \lambda - \varepsilon). \tag{A.13}$$

Consider now the family of univariate monotone functions $\{f_c\}_{c\geq 0}$, where $f_c : [0, c] \to [0, \infty]$ is defined by

$$f_c(R) = I(R, d, c - R) \quad 0 \leq R \leq c. \tag{A.14}$$

We note that, by the definition of $f_c(\cdot)$ and equations (A.12) - (A.13), for each $c \geq 0$ and $0 \leq R \leq c$ we have

$$f_c(R+0) = I(R+0, d, c-R-0) \quad \text{and} \quad f_c(R-0) = I(R-0, d, c-R+0). \tag{A.15}$$

Letting

$$\mathcal{S}_c = \{R \in [0, c] : f_c(R+0) > f_c(R-0)\} \tag{A.16}$$

denote the set of discontinuity points of f_c, it follows from the definitions of $\mathcal{S}$ and $\mathcal{S}_c$ and from equation (A.15) that

$$\mathcal{S} = \bigcup_{c \geq 0} \{(R, c-R) : R \in \mathcal{S}_c\}, \tag{A.17}$$

where the union on the right side of (A.17) is clearly a disjoint one. Letting μ_1 and μ_2 denote the Lebesgue measures on $\mathbb{R}^1$ and $\mathbb{R}^2$, respectively, equation (A.17) gives

$$\mu_2(\mathcal{S}) = \int_{c \geq 0} \mu_1(\mathcal{S}_c) d\mu_1(c). \tag{A.18}$$

Finally, we note that for all $c \geq 0$, $f_c(\cdot)$ is a monotone (non-decreasing) function which, therefore, has, at most, a countable number of points of discontinuity. In particular, we have $\mu_1(\mathcal{S}_c) = 0$ for all $c \geq 0$ which annihilates the right side of (A.18) and completes the proof. □

C Proof of Inequality (36)

Assume, conversely, the existence of $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$ such that $D(P\|P_Z) < \lambda - \delta$ for which

$$\sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + F_n(P_{\hat{X}^n(z^n), z^n}, d)\right]\right\} \tag{A.19}$$

$$< \sum_{z^n \in T_P} \exp\left\{-n\left[D(P\|P_Z) + H(P) + \max_{\{W: I(P;W) \leq R\} \cap \mathcal{C}_n(P)} F_n(P \times W, d) + \eta\right]\right\}, \tag{A.20}$$

or, equivalently,

$$\sum_{z^n \in T_P} \exp\left\{-nF_n(P_{\hat{X}^n(z^n), z^n}, d)\right\} \tag{A.21}$$

$$< |T_P| \cdot \exp\left\{-n\left[\max_{\{W: I(P;W) \leq R\} \cap \mathcal{C}_n(P)} F_n(P \times W, d) + \eta\right]\right\}. \tag{A.22}$$

Letting $\tilde{s}(P) \triangleq \{z^n \in T_P : I(P_{\hat{X}^n(z^n),z^n}) > R\}$, we have

$$\sum_{z^n \in T_P} \exp\left\{-nF_n(P_{\hat{X}^n(z^n),z^n}, d)\right\} \tag{A.23}$$

$$= \sum_{z^n \in T_P \setminus \tilde{s}(P)} \exp\left\{-nF_n(P_{\hat{X}^n(z^n),z^n}, d)\right\} + \sum_{z^n \in \tilde{s}(P)} \exp\left\{-nF_n(P_{\hat{X}^n(z^n),z^n}, d)\right\} \tag{A.24}$$

$$\geq \sum_{z^n \in T_P \setminus \tilde{s}(P)} \exp\left\{-n\left[\max_{\{W: I(P;W) \leq R\} \cap \mathcal{C}_n(P)} F_n(P \times W, d)\right]\right\} \tag{A.25}$$

$$= |T_P \setminus \tilde{s}(P)| \exp\left\{-n\left[\max_{\{W: I(P;W) \leq R\} \cap \mathcal{C}_n(P)} F_n(P \times W, d)\right]\right\}. \tag{A.26}$$

Combining (A.26) with (A.22) gives

$$e^{-n\eta}|T_P| > |T_P \setminus \tilde{s}(P)|, \tag{A.27}$$

or

$$|\tilde{s}(P)| > \left(1 - e^{-n\eta}\right)|T_P| \geq \frac{1}{2}|T_P|, \tag{A.28}$$

where the right inequality holds for all sufficiently large n (dependent on η yet independent of $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$). Combining (A.28) with the fact that $|\mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P)| \leq (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|}$ implies the existence of $W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P)$ with $I(P \times W) > R$ such that the set $s(P) \triangleq \{z^n : P_{z^n,\hat{X}^n(z^n)} = P \times W\}$ (which is a subset of $\tilde{s}(P)$) satisfies

$$|s(P)| \geq \frac{1}{2}(n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|}|T_P| \geq \frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\{nH(P)\}. \tag{A.29}$$

On the other hand, letting $q(P) \triangleq \{\hat{X}^n(z^n) : z^n \in s(P)\}$, we have

$$|s(P)| = \sum_{\hat{x}^n \in q(P)} |\{z^n \in s(P) : \hat{X}^n(z^n) = \hat{x}^n\}| \tag{A.30}$$

$$\leq \sum_{\hat{x}^n \in q(P)} \exp\{nH_{P \times W}(Z|\hat{X})\} \tag{A.31}$$

$$= |q(P)|\exp\{nH_{P \times W}(Z|\hat{X})\}, \tag{A.32}$$

which, combined with (A.29), implies

$$|q(P)| \geq \frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\{n[H(P) - H_{P \times W}(Z|\hat{X})]\} \tag{A.33}$$

$$= \frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\{n[I(P \times W)]\} \tag{A.34}$$

$$\geq \frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\{nR\}. \tag{A.35}$$

Now, since the number of members of $q(P)$ that will be mapped into a binary vector of length not exceeding $n(R-\varepsilon)$ is upper bounded by $\exp\{n(R-\varepsilon+1/n)\}$, and since there can be no more than $\exp\{nH_{P\times W}(Z|\hat{X})\}$ members of $s(P)$ that are mapped into one member of $q(P)$, we have

$$\begin{aligned} |\{z^n \in s(P) : L_n(z^n, D_n) \le n(R-\varepsilon)\}| &\le \exp\{n(R-\varepsilon+1/n+H_{P\times W}(Z|\hat{X}))\} && \text{(A.36)} \\ &\le \exp\{n(I(P\times W)-\varepsilon+1/n+H_{P\times W}(Z|\hat{X}))\} && \text{(A.37)} \\ &= \exp\{n(H(P)-\varepsilon+1/n)\}. && \text{(A.38)} \end{aligned}$$

Consequently,

$$\begin{aligned} & P^n_{X,Z}\{L_n(z^n, D_n) > n(R-\varepsilon)\} && \text{(A.39)} \\ \ge\ & P^n_{X,Z}\{z^n \in s(P), L_n(z^n, D_n) > n(R-\varepsilon)\} && \text{(A.40)} \\ =\ & (|s(P)| - |\{z^n \in s(P) : L_n(z^n, D_n) \le n(R-\varepsilon)\}|)\exp\{-n(D(P\|P_Z)+H(P))\} && \text{(A.41)} \\ \ge\ & \left(\frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)}\exp\{nH(P)\} - \exp\{n(H(P)-\varepsilon+1/n)\}\right)\exp\{-n(D(P\|P_Z)+H(P))\} && \text{(A.42)} \\ =\ & \left(\frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)} - \exp\{-n(\varepsilon-1/n)\}\right)\exp\{-nD(P\|P_Z)\} && \text{(A.43)} \\ \ge\ & \left(\frac{1}{2}(n+1)^{-(|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|+|\mathcal{Z}|)} - \exp\{-n(\varepsilon-1/n)\}\right)\exp\{-n(\lambda-\delta)\} && \text{(A.44)} \\ >\ & \exp\{-n\lambda\}, && \text{(A.45)} \end{aligned}$$

where the last (strict) inequality holds for all sufficiently large n (dependent on $\eta, \varepsilon, \delta$, yet independent of the $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$ assumed to satisfy (A.20)). Hence, for all $n \ge n_0(\eta, \varepsilon, \delta)$, the assumption of the existence of $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$ such that $D(P\|P_Z) < \lambda - \delta$ and for which (A.20) holds leads to a contradiction, as (A.45) clearly violates the fact that $\{D_n\}$ were taken such that $D_n \in A_n(R-\varepsilon, \lambda)$. Thus, we have established the fact that whenever $n \ge n_0(\eta, \varepsilon, \delta)$, for any $P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n$ such that $D(P\|P_Z) < \lambda - \delta$, (36) holds for all $D_n \in A_n(R-\varepsilon, \lambda)$. □

D Technical Claim in Proof of Theorem 1

In this subsection, we aim to prove the following.

Claim 2. *For any* $P_{X,Z} \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$ *we have*

$$\lim_{n\to\infty} \min_{\substack{P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n : \\ D(P\|P_Z) < \lambda}} \left[D(P\|P_Z) + \max_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P) : \\ I(P;W) \le R}} F_n(P \times W, d) \right]$$

$$= \inf_{\substack{P \in \mathcal{M}(\mathcal{Z}) : \\ D(P\|P_Z) < \lambda}} \left[D(P\|P_Z) + \sup_{\substack{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) : \\ I(P;W) \le R}} F(P \times W, d) \right] \tag{A.46}$$

and

$$\lim_{n\to\infty} \min_{\substack{P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n : \\ D(P\|P_Z) \ge \lambda}} \left[D(P\|P_Z) + \max_{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P)} F_n(P \times W, d) \right]$$

$$= \inf_{\substack{P \in \mathcal{M}(\mathcal{Z}) : \\ D(P\|P_Z) \ge \lambda}} \left[D(P\|P_Z) + \sup_{W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})} F(P \times W, d) \right], \tag{A.47}$$

where $F_n(\cdot,\cdot)$ *and* $F(\cdot,\cdot)$ *are defined by (24) and (12), respectively.*

Proof: We start by recalling that

$$F(Q,d) = \inf_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X}) : \\ E_{Q \times V} \rho(X, \hat{X}) > d}} D\left(V \| P_{X|Z} | Q\right) \tag{A.48}$$

and that

$$F_n(Q,d) = \min_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X}) \cap \mathcal{C}_n(Q) : \\ E_{Q \times V} \rho(X, \hat{X}) > d}} D\left(V \| P_{X|Z} | Q\right). \tag{A.49}$$

Since, by our standing assumption, $P_{X,Z} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Z})$, it follows that, in particular, $P_{X|Z}(x|z) > 0$ for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. It therefore follows that $D\left(V\|P_{X|Z}|Q\right)$, which is ultimately given by a finite convex combination of functions (of the form $\log\left(V(x|\hat{x},z)/P_{X|Z}(x|z)\right)$) that are all uniformly continuous in V when $P_{X,Z} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Z})$, is uniformly continuous in the pair (V,Q). This, combined with the clear fact that

$$\sup_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X}) : \\ E_{Q \times V} \rho(X, \hat{X}) > d}} \min_{\substack{V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X}) \cap \mathcal{C}_n(Q) : \\ E_{Q \times V} \rho(X, \hat{X}) > d}} \|P - P'\| \to 0$$

as $n \to \infty$, implies that $F(Q,d)$ and $F_n(Q,d)$ are uniformly continuous in Q and, further, that

$$|F(Q,d) - F_n(Q,d)| \to 0 \tag{A.50}$$

uniformly in $Q \in \mathcal{M}(\hat{\mathcal{X}} \times \mathcal{Z})$. The proof of (A.46) is completed by combining (A.50) with the easily verifiable fact that

$$\sup_{Q \in \mathcal{Q}} \min_{Q' \in \mathcal{Q}_n} \|Q' - Q\| \to 0 \quad \text{as } n \to \infty, \tag{A.51}$$

where

$$\mathcal{Q}_n = \left\{ Q = P \times W : P \in \mathcal{M}(\mathcal{Z}) \cap \mathcal{M}_n, D(P \| P_Z) < \lambda, W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}) \cap \mathcal{C}_n(P), I(P; W) \le R \right\}$$

and

$$\mathcal{Q} = \left\{ Q = P \times W : P \in \mathcal{M}(\mathcal{Z}), D(P \| P_Z) < \lambda, W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}}), I(P; W) \le R \right\}.$$

Equality (A.47) is established similarly. □

References

[1] T. Berger. *Rate-Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall, Englewood Cliffs, N.J., 1971.

[2] R. E. Blahut. "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, IT-20:405–417, 1974.

[3] R. E. Blahut. "Information bounds of the Fano-Kullback type," *IEEE Trans. Inform. Theory*, IT-22:410–421, 1976.

[4] R. E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, 1987.

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

[6] A. Dembo and T. Weissman. "On universal coding of noisy sources," *preprint*, April 2001.

[7] Y. Ephraim and R. M. Gray. "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, 34(4):826–834, July 1988.

[8] E. Erkip and T. M. Cover. "The efficiency of investment information," *IEEE Trans. Inform. Theory*, 44(3):1026–1040, May 1998.

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

[10] T. S. Han. "The reliability functions of the general source with fixed-length coding," *IEEE Trans. Inform. Theory*, IT-46:2117–2132, September 2000.

[11] P. A. Humblet. "Generalization of Huffman coding to minimize the probability of buffer overflow," *IEEE Trans. Inform. Theory*, IT-27(2):230–237, March 1981.

[12] F. Jelinek. "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, IT-14(3):490–501, May 1968.

[13] A. Kanlis and P. Narayan. "Error exponents for succesive refinement by partitioning," *IEEE Trans. Inform. Theory*, 42(1):275–282, January 1996.

[14] I. Lev. "Universal signal enhacement by coding," Master's thesis, Technion, Haifa, Israel, April 1996.

[15] E. Levitan and N. Merhav. "A competitive Neyman - Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, December 2000. Submitted.

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

[17] N. Merhav. "Universal coding with minimum probability of codeword length overflow," *IEEE Trans. Inform. Theory*, 37:556–563, 1991.

[18] N. Merhav, G. Seroussi, and M. J. Weinberger. "Coding of sources with two-sided geometric distributions and unknown parameters," *IEEE Trans. Inform. Theory*, IT-46:229–236, January 2000.

[19] N. Merhav, G. Seroussi, and M. J. Weinberger. "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Trans. Inform. Theory*, IT-46:121–135, January 2000.

[20] T. Weissman and N. Merhav. "Tradeoffs between the excess-code-length exponent and the excess-distortion exponent in lossy source coding," CCIT 341 EE 1275, Technion-I.I.T., EE Dept., April 2001.

[21] A. D. Wyner. "On the probability of buffer overflow under an arbitrary bounded input-output distribution," *SIAM J. Appl. Math.*, 27:544–570, 1974.

[22] O. Zeitouni and M. Gutman. "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, 37:285–290, March 1991.