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

Tsachy Weissman, Student Member, IEEE, and Neri Merhav, Fellow, IEEE

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 (DMC) 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 of a given family \( \mathcal{P}_\Theta \) of a given family \( \{ \mathcal{P}_\Theta, \Theta \in \Theta \} \). 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 \( \Theta \)-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.

Index Terms—Excess-code-length exponent, excess-distortion exponent, noisy source coding, Marton’s exponent, universal coding.

I. INTRODUCTION

Let \( X_1, X_2, \ldots \) be a source sequence with components in a finite alphabet \( \mathcal{X} \) and let \( \hat{X} \) be a reconstruction alphabet. Let further \( \rho: \mathcal{X} \times \hat{X} \to \mathbb{R}^+ \) be a given single-letter distortion measure and define the distortion between sequences \( X^n \in \mathcal{X}^n \) and \( \hat{X}^n \in \hat{X}^n \) 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, \( B_n \subseteq \hat{X}^n \) of size \( 2^{nR} \) and a mapping, \( \hat{X}^n = \varphi_n(X^n) \), from \( X^n \) 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], [10]).

Another sensible performance criterion is the large deviations behavior of the distortion, which was first addressed, for a discrete memoryless source (DMS), in Marton’s 1974 paper [17] (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 of a source \( Q \), and \( D(\cdot||\cdot) \) denotes the Kullback–Leibler distance. Error exponents for source codes have since been studied by others (cf. [4], [6] and references therein) and have been extended to more general settings, such as that of successive refinement [14]. The excess-code-length exponent has also been studied in the context of lossless coding, cf. [12], [13], [22], [18] and the recent work [11] 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 codebook 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 [17] 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 III), that sequences of the same type class \(^1\) 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 \( \lambda \).
from $P$ and $\eta \log |X|$ bits when the Kullback–Leibler distance from $P$ is more than $\lambda$. 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 [6], 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 $\lambda$, assuming the value $\infty$ for $\lambda < F_d(R)$ and the value $F_d(R)$ for $\lambda \geq F_d(R)$ (cf. Fig. 1 of Section III-C2). This extension of Marton’s result will be shown to be obtainable as a special case of the more general setting, considered in Section III, 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 III, we shall analyze source coding exponents for the case of a DMS corrupted by a discrete memoryless channel (DMC). In this setting, 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 $\tilde{X}_1, \tilde{X}_2, \ldots$. We shall characterize the best achievable excess-distortion exponent (where the distortion measured is between $X_1, X_2, \ldots$ and $\tilde{X}_1, \tilde{X}_2, \ldots$) subject to the constraint that the excess-code-length exponent is at least $\lambda$. Note that the problem considered in [17] is a special case of this section, where the DMC is the clean channel (under which $X_i = Z_i$ almost surely (A’s.s)) and $\lambda = \infty$. Staying with the case where $\lambda = \infty$, the results of this section can be considered analogous to those of [17] 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 $R = \log |X|$, that is, pure filtering.

A more detailed outline of Section III is as follows. Section III-A presents the main result on the above-described exponent, whose proof is given in Section III-B. In Section III-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 $\lambda$. Section III-D will be dedicated to the properties of the exponent function. Finally, in Section III-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 IV 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 $P_\theta$, where $\theta$ is an unknown parameter taking values in a set $\Theta$, we characterize the best achievable ($\theta$-dependent) distortion exponent, such that the code-length exponent, with respect to (w.r.t.) $P_\theta$, is at least as large as a given function $\lambda(\theta)$, uniformly for all $\theta \in \Theta$. A universally optimal 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 $\Theta$, and the form of the function $\lambda(\cdot)$. This somewhat parallels a recent work [16], 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 $\theta$ belongs to $\Theta_1$ or to $\Theta_2$, 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 $\theta$. As it turns out, the optimal scheme of [16] is intimately related and, in some sense, analogous to the one presented here. As will be elaborated on in Section IV-B, the coding scheme which turns out to be optimal in the setting of this work uses two different strategies in two complementary subsets of sources, and these subsets exactly correspond to the decision rule of [16].

As revealed by the proofs of the main results, 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 code word will be about $nR$ bits long. If the type belongs to the complimentary set, the codeword will be significantly longer. In Section III, the dichotomy is according to whether $D(Q||P) \leq \lambda$ or $D(Q||P) > \lambda$. In Section IV, the dichotomy is dictated by the Kullback–Leibler distances between the type of the sequence and the sources $P_\theta$, $\theta \in \Theta$, corresponding to the decision rule proposed in [16]. 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 II, in which we present the notation and conventions used throughout, will be followed by the two main sections (III and IV) described above. Finally, in Section V, we summarize the paper and discuss a direction for future research.

II. Notation and Preliminaries

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

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

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

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

(1)
similarly, \(a_n \geq b_n\) will be synonymous to the statement that (1) holds with a strict inequality. Note that the negation of \(a_n \geq b_n\) does not imply that \(a_n < b_n\), but it does imply that
\[
\limsup_{n \to \infty} \frac{1}{n} \log \frac{b_n}{a_n} > 0
\]
or, in other words,\(^3\) the existence of \(\delta > 0\) 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 \(z \mapsto b_{n_k}\) for
\[
\lim_{k \to \infty} \frac{1}{n_k} \log \frac{a_{n_k}}{b_{n_k}} = 0.
\]
The notation \(a_{n_k} \geq b_{n_k}\) 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
\[
H(W|P) = \sum_{a \in \mathcal{A}} P(a) H(W(\cdot|a))
\]
be 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
\[
D(V||W|P) = \sum_{a \in \mathcal{A}} P(a) D(V(\cdot|a)||W(\cdot|a))
\]
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 \(^4\) \(\min_{\mathcal{X} \in \mathcal{X}} \rho(X, \hat{X}) = 0\) for all \(\hat{X} \in \hat{\mathcal{X}}\). For \(X^n \in \mathcal{X}^n\) and \(\hat{X}^n \in \hat{\mathcal{X}}^n\) we define
\[
\rho_n(X^n, \hat{X}^n) = \sum_{i=1}^{n} \rho(X_i, \hat{X}_i)
\]
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 \(\rho_n\) in the sequel when there is no room for ambiguity. We shall adopt the convention
\[^3\]This is the form of the statement which will be used in a certain place in the paper, so we take the opportunity here to make it explicit.
\[^4\]Note that there is no essential loss of generality in this assumption as, otherwise, we can look at the modified distortion measure given by \(\rho(X, \hat{X}) = \rho(X, \hat{X}) - \min_{\hat{X} \in \hat{X}} \rho(X, \hat{X})\).
III. 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 $P_{Z|X}$. We assume that $P_{X|Z} \in \mathcal{M}_+$($\mathcal{X} \times \mathcal{Z}$). $P_{X|Z}^n$ will denote the product distribution on $\mathcal{X} \times \mathcal{Z}$ 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 $B_n \subseteq \mathcal{X}^n$, a mapping $\hat{X}^n; \mathcal{Z}^n \rightarrow B_n$ so that $\hat{X}^n(Z^n) \in \hat{X}^n$, and a uniquely decodable code $C_n; B_n \rightarrow \{0,1\}^n$. 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 $D_n$ denote the set of all such schemes for combined filtering and compression of noisy input vectors of length $n$.

In the classical setting of [17], 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. $P$. 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 [17], to the milder one where $\lambda$ 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

$$A_n(R, \lambda) = \{D_n \in D_n: P_{X|Z}^n \{L_n(Z^n, D_n) > nR\} \leq e^{-n\lambda} \}.$$ (2)

In words, $A_n(R, \lambda)$ is the set of all variable-length schemes for input vectors of length $n$ having excess-code-length exponent for rate $R$ at least as large as $\lambda$. For $d \geq 0$ we let

$$G_n(R, d, \lambda) = \min_{D_n \in A_n(R, \lambda)} P_{X|Z}^n(\rho(\hat{X}^n, \hat{X}^n(Z^n)) > nd)$$ (3)

where the $\hat{X}^n(\cdot)$ on the right-hand side of (3) is the mapping from $Z^n$ into $B_n$ associated with the scheme $D_n$. In words, $G_n(R, d, \lambda)$ is the least achievable probability for distortion exceeding $nd$ among all variable-length schemes having excess-code-length exponent at least as large as $\lambda$.

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

$$I(R - 0, d, \lambda + 0) \leq \liminf_{n \to \infty} \left[ -\frac{1}{n} \log G_n(R, d, \lambda) \right]$$ (4)

where

$$I(R, d, \lambda) = \min \left\{ \inf_{P_{D_P}|P_Z|P_X \geq \lambda} a(P, \infty, d), \inf_{P_{D_P}|P_Z|P_X < \lambda} a(P, R, d) \right\}$$ (6)

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

$$a(P, R, d) = D(P) + \sup_{V \in \mathcal{C}(\mathcal{Z} \rightarrow \mathcal{X})} F(P \times W, d)$$ (7)

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

$$F(Q, d) = \inf_{V \in \mathcal{C}(\hat{X} \times \mathcal{Z} \rightarrow \mathcal{X})} D(V||P_{X|Z}(Q))$$ (8)

Remarks: A qualitative explanation for the origin of the expression for the exponent in (6) can be given as follows. Assuming (what will be justified in the proof) that there is no loss of optimality in restricting attention to schemes for which the (joint) type of $(\rho(X^n, \hat{X}^n(Z^n)))$ is constant all across the type $\rho(X^n) \in \mathcal{T}_P$ for all $P \in \mathcal{M}(\mathcal{Z})$, fix a type $P$ and let $W \in \mathcal{C}(\mathcal{Z} \rightarrow \hat{X})$ denote the channel induced by the joint type $(\rho(X^n, \hat{X}^n(Z^n)))$ when $\rho(X^n) \in \mathcal{T}_P$. It is then not hard to see (and will be formally established), e.g., via the method of types, that

$$\Pr\{\rho(X^n, \hat{X}^n(Z^n)) > nd|Z^n \in \mathcal{T}_P\} \approx \exp\{-nf(PXW, d)\}.$$ So the best coding scheme (in the sense of maximizing the excess-distortion exponent) would be one for which $F(P \times W, d)$ is maximized over $W$, for each $P$. However, as discussed in Section I, if the excess-code-length constraint is to be satisfied, types with $D(P|P_Z) \leq \lambda$ cannot be afforded more than $nR$ bits. This translates into the requirement that the maximization over $W$, for $P \in \mathcal{M}(\mathcal{Z})$, be restricted to $W$’s satisfying $F(P, W) \leq R$. For types with $D(P|P_Z) > \lambda$, on the other hand, the maximization over $W$ is unrestricted. Hence, since the exponential “price” of observing a (noisy) source sequence $Z^n$ of type $P$ is $D(P|P_Z)$, the best achievable exponential behavior for Pr features $\Pr\{\rho(X^n, \hat{X}(Z^n)) > nd, Z^n \in T_P\} \approx \exp\{-na(P, R, d)\}$ when $D(P|P_Z) \leq \lambda$ and $\approx \exp\{-na(P, \infty, d)\}$ when $D(P|P_Z) > \lambda$, which explains the form of (6).

The monotonicity of $I(R, d, \lambda)$ in both $R$ (increasing) and $\lambda$ (decreasing) is shown in Appendix-B1 to imply that the limits defining $I(R - 0, d, \lambda + 0)$ and $I(R + 0, d, \lambda - 0)$ both exist. Furthermore, in Appendix-B2 we show that Theorem 1 implies that for all pairs $(R, \lambda)$ outside a set $S$ of zero Lebesgue measure

$$\lim_{n \to \infty} \left[ -\frac{1}{n} \log G_n(R, d, \lambda) \right] = I(R, d, \lambda).$$ (9)

Note that the case of pure filtering is obtained by taking $R = \infty$ or $\lambda = 0$, where Theorem 1 gives the best achievable excess-dist-
tortion 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 Section III-D.

B. Proof of Theorem 1

The outline of the proof is the following. We start by obtaining an exponentially tight estimate for

$$P_{X^n, Z}^n\{p(X^n, \hat{X}^n(Z^n)) > nd | Z^n = z^n\}$$

for an arbitrary $z^n \in Z^n$ and scheme $D_n$ (which defines the mapping $\hat{X}^n: Z^n \rightarrow \hat{X}^n$). This estimate will be a functional of the empirical distribution $P_{X^n(z^n), z^n}$. So, in effect, we are looking for that scheme in $A_n(R, \lambda)$ with empirical distributions $P_{X^n(z^n), z^n}$ that minimize this functional, if possible, for all $z^n \in Z^n$. In the converse part, we show that any member of $A_n(R, \lambda)$ must be such that $I(P_{X^n(z^n), z^n}) \leq R$ for essentially all $z^n \in Z^n$ for which $D(P_{n} || P_{Z}) < \lambda$. Therefore, the exponent achievable for $P_{X^n, Z}^n\{p(X^n, \hat{X}^n(Z^n)) > nd\}$ must clearly be upper-bounded by the exponent obtained when maximizing the above functional with respect to $P_{X^n(z^n), z^n}$, for each $z^n$, under the constraint that $I(P_{X^n(z^n), z^n}) \leq R$ whenever $D(P_{n} || P_{Z}) < \lambda$. The direct part is then established by the 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 $P_{X^n(z^n), z^n}$ maximizes the above functional, subject to the constraint that $I(P_{X^n(z^n), z^n}) \leq R$ whenever $D(P_{n} || P_{Z}) < \lambda$. The existence of such a scheme in $A_n(R, \lambda)$ is essentially guaranteed because the fact that $I(P_{X^n(z^n), z^n}) \leq R$ for all $z^n \in Z^n$ guarantees the existence of a scheme that will need no more than $nR$ bits to convey $\hat{X}^n(z^n)$ for $z^n \in Z^n$. Hence, whenever $z^n \in Z^n$ for $P$ with $D(P_n || P_{Z}) < \lambda$, 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

$$P_{X^n, Z}^n\{p(X^n, \hat{X}^n(Z^n)) > nd\} \geq \sum_{Z^n \in Z^n} P_{X^n(Z^n), z^n}^n \cdot P_{X^n, Z}^n\{p(X^n, \hat{X}^n(Z^n)) > nd\}$$

$$= \sum_{z^n \in Z^n} P_{X^n(Z^n)}^n \left( \sum_{X^n, \hat{X}^n: \rho(x^n, \hat{x}^n(Z^n)) > nd} P_{X^n, Z}^n\{p(X^n, \hat{X}^n(Z^n)) > nd\} \right).$$  \hspace{1cm} (10)

Considering the inner sum in (10), we have for any $z^n \in Z^n$ and $\hat{x}^n \in \hat{X}^n$

$$\sum_{x^n \in \chi^n: \rho(x^n, \hat{x}^n(z^n)) > nd} P_{X^n, Z}^n\{p(x^n, \hat{x}^n(z^n)) > nd\} = \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(x, y)} > d} \sum_{z^n \in Z^n} P_{X^n, Z}^n\{p(x^n, \hat{x}^n(z^n)) > nd\}$$

$$= \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(x, y)} > d} \sum_{z^n \in Z^n} P_{X^n, Z}^n\{p(x^n, \hat{x}^n(z^n)) > nd\}$$  \hspace{1cm} (11)

Therefore, (15) is upper-bounded by

$$= \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(x, y)} > d} \sum_{z^n \in Z^n} \exp\{-nH(V|P^n(Z)|P^n(z^n))\}$$

$$= \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(x, y)} > d} \sum_{z^n \in Z^n} \exp\{-nH(V|P^n(z^n))\}$$  \hspace{1cm} (12)

The direct part is then established by the 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 $P_{X^n(z^n), z^n}$ maximizes the above functional, subject to the constraint that $I(P_{X^n(z^n), z^n}) \leq R$ whenever $D(P_{n} || P_{Z}) < \lambda$. Therefore, (15) is upper-bounded by

$$= \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(x, y)} > d} \sum_{z^n \in Z^n} \exp\{-nH(V|P^n(z^n))\}$$

$$\leq (n+1)^{-|\mathcal{L}|} \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(z^n)} > d} \exp\{-nH(V|P^n(z^n))\}$$  \hspace{1cm} (19)

and lower-bounded by

$$\geq (n+1)^{-|\mathcal{L}|} \sum_{V \in \mathcal{L}^n, \chi \subseteq \chi^n: \chi \subseteq \hat{x}^n, E_{P^n(z^n)} > d} \exp\{-nH(V|P^n(z^n))\}$$  \hspace{1cm} (21)

We turn to the proof of (5) 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 (10) with (22), we have the
existence of \(\eta_0(\varepsilon, \delta, \eta)\) (independent on the sequence \(\{D_n\}\)) such that for all \(n \geq n_0(\varepsilon, \delta, \eta)\)

\[P_{X^n, Z^n}[r(X^n, X^n(Z^n)) > nd] \geq (n+1)^{-4|V||Z|} \sum_{z^n \in T^n} P_{Z^n}(z^n) \exp\{-nF_n(P_{X^n(Z^n)} \mid \mathcal{Z}, d)\}\]

(23)

\[= (n+1)^{-4|V||Z|} \sum_{\{P : D(P) \geq \lambda - \delta\} \cap \mathcal{M}_n, z^n \in T^n} \sum_{\{P : D(P) \geq \lambda - \delta\} \cap \mathcal{M}_n, z^n \in T^n} \times \exp\{-n[D(P \mid P_Z) + H(P) + F_n(P_{X^n(Z^n)} \mid \mathcal{Z}, d)]\}\]

(24)

\[\geq (n+1)^{-4|V||Z|} \sum_{\{P : D(P) \geq \lambda - \delta\} \cap \mathcal{M}_n, z^n \in T^n} \times \exp\{-n[D(P \mid P_Z) + H(P) + \max_{W \in \mathcal{C}(Z \rightarrow \mathcal{X}) \cap \mathcal{G}_n(P_{X^n})} F_n(P \times W, d)]\}\]

(25)

where we only need to justify the last inequality. To this end, observe first that (24) upper-bounds the first line of (26) as there clearly exists a \(W \in \mathcal{C}(Z \rightarrow \hat{X}) \cap \mathcal{G}_n(P_{Z^n})\) such that \(P_{Z^n \times W} = P_{Z^n, X^n(Z^n)}\). So it remains to argue why (25) upper-bounds the second line of (26). For this purpose, it will suffice to establish the existence of \(\eta_0(\varepsilon, \delta, \eta)\) (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 \(D(P \mid P_Z) < \lambda - \delta\)

\[\sum_{z^n \in T^n} \exp\{-n[D(P \mid P_Z) + H(P) + F_n(P_{X^n(Z^n)} \mid \mathcal{Z}, d)]\} \geq \sum_{z^n \in T^n} \exp\{-n[D(P \mid P_Z) + H(P) + \max_{W \leq R} F_n(P \times W, d)]\} + \sum_{\{W : \varepsilon \leq R\} \cap \mathcal{G}_n(P)} F_n(P \times W, d) + \eta\}

(27)

Inequality (28) is proved in part C of the Appendix. The independence of \(\eta_0(\varepsilon, \delta, \eta)\) on the sequence \(\{D_n\}\) guarantees that (23) can, in fact, be replaced by

\[\min_{D_n \in \mathcal{A}_n(\lambda - \delta, \varepsilon, \rho)} P_{X^n, Z}[r(X^n, X^n(Z^n)) > nd] = G_n(R - \varepsilon, d, \lambda).

Furthermore, since the sums in (28) are over expressions which do not depend on \(z^n \in T^n\), (26) can be further lower-bounded by

\[\geq (n+1)^{-4|V||Z|} \sum_{\{P : D(P) \geq \lambda - \delta\} \cap \mathcal{M}_n, z^n \in T^n} \times \exp\{-n[D(P \mid P_Z) + H(P) + \max_{W \in \mathcal{C}(Z \rightarrow \mathcal{X}) \cap \mathcal{G}_n(P_{X^n})} F_n(P \times W, d)]\}\]

(29)

\[\geq (n+1)^{-4|V||Z|} \sum_{\{P : D(P) \geq \lambda - \delta\} \cap \mathcal{M}_n, z^n \in T^n} \times \exp\{-n[D(P \mid P_Z) + \max_{W \in \mathcal{C}(Z \rightarrow \mathcal{X}) \cap \mathcal{G}_n(P_{X^n})} F_n(P \times W, d)]\}

(30)

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

\[-\frac{1}{n} \log G_n(R - \varepsilon, d, \lambda) \leq \frac{|V||Z||X|}{n} \log(n+1) + \max_{\{W : \varepsilon \leq R\} \cap \mathcal{G}_n(P)} F_n(P \times W, d)\]

(31)
Combining the result of part D of the Appendix (Claim 2) with the definition of $I(R, d, \lambda - \delta)$ (recall (6)) and the arbitrariness of $\eta > 0$ gives
\[
\limsup_{n \to \infty} \left( -\frac{1}{n} \log G_n(R - \varepsilon, d, \lambda) \right) \leq I(R, d, \lambda - \delta) \tag{33}
\]
which, by the arbitrariness of $\varepsilon, \delta > 0$, finally implies (5).

Turning to establish (4), 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 $D(P \| P_{Z^n}) \geq \lambda + \delta$ we take
  \[ \hat{X}^n(z^n) = \arg \max_{\hat{x}^n \in \hat{X}^n} F_n(P_{\hat{X}^n, z^n, d^*}, d). \tag{34} \]
- For $z^n \in T_P$ with $D(P \| P_{Z^n}) < \lambda + \delta$ let
  \[ W^* \in \mathcal{C}(Z \to \hat{X}) \cap \mathcal{C}(P) \]
  achieve
  \[ \max_{W \in \mathcal{C}(Z \to \hat{X}) \cap \mathcal{C}(P)} F_n(P \times W, d). \tag{35} \]

Then, as is well known [6], there exists, for sufficiently large $n$ (dependent on $\varepsilon$ yet not on $P$), a set $B(P) \subseteq \hat{X}^n$ of size
\[ |B(P)| \leq 2^{n(R(P \times W^*) + \varepsilon/2)} \leq 2^{n(R + \varepsilon/2)} \]
and a map $\hat{X}^n(\cdot): T_P \to B(P)$ such that
\[ (z^n, \hat{X}^n(z^n)) \in T_{P \times W}, \quad \text{for all } z^n \in T_P. \]

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 still need to specify the codewords, or, equivalently, the uniquely decodable map $G_n: B_n \to \{0, 1\}^n$. We construct the codewords as follows: use no more than $|Z| \log(n+1)$ bits to convey the type to which $z^n$ belongs. Now, if $z^n \in T_P$ with $D(P \| P_{Z^n}) \geq \lambda + \delta$, 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, once the type to which $z^n$ belongs is known, as $\hat{X}^n(z^n)$ 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 $L_o(Z^n, D_n) > n(R + \varepsilon)$ implies that $Z^n \in T_P$ for some $P$ such that $D(P \| P_{Z^n}) \geq \lambda + \delta$. Consequently, for sufficiently large $n$ (dependent on $\varepsilon$ and $\delta$), we have
\[
P^n_{X, Z} \left\{ L_o(Z^n, D_n) > n(R + \varepsilon) \right\} \leq P^n_{X, Z} \left\{ \bigcup_{P \in \mathcal{M}(Z): D(P \| P_{Z^n}) \geq \lambda + \delta} \{z^n \in T_P\} \right\} < e^{-n\lambda}, \tag{36}
\]
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
\[
P^n_{X, Z} \{ \rho(X^n, \hat{X}^n(Z^n)) > nd \} \tag{37}
\]
\[
\leq (n + 1)^{|Y|/|X|} \left( \sum_{\{P: D(P \| P_{Z^n}) \geq \lambda + \delta\}} \sum_{\{z^n \in T_P\}} F_n \left( P_{X^n(z^n), z^n, d} \right) \right) \tag{38}
\]
\[
+ \sum_{\{P: D(P \| P_{Z^n}) < \lambda + \delta\}} \sum_{\{z^n \in T_P\}} \exp \left\{ \left( -n \left( D(P \| P_{Z^n}) + H(P) + F_n \left( P_{X^n(z^n), z^n, d} \right) \right) \right) \right\} \tag{39}
\]
where the inequality follows by combining (10) with (20) 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
\[
\limsup_{n \to \infty} \left( -\frac{1}{n} \log G_n(R + \varepsilon, d, \lambda) \right) \leq -I(R, d, \lambda + \delta) \tag{40}
\]
which follows when considering the normalized logarithmic limit of the two sides of the last chain of inequalities and invoking Claim 2 of part D in the Appendix. Finally, the arbitrariness of $\varepsilon, \delta > 0$ implies (4).

**Remark:** Note that the optimal scheme constructed in the direct part of the proof is dependent on $d$, since the maximizers in (34) and (35) will, in general, depend on $d$. This is in contrast to Marton’s setting (cf. [17]), 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 $\lambda < \infty$.

1) Marton’s Exponent: Let us verify that for the noise-free case and $\lambda = \infty$, Theorem 1 coincides with Marton’s classical result [17]. To comply with the general formulation, we assume that in this case $Z = X$ 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 $Z$ into $X$.

Taking $\lambda = \infty$ in (6) we have

$$I(R, d, \lambda) = \inf_{P \in M(Z)} \left[ D(P||P_Z) + \sup_{W \in C(Z \rightarrow X)} F(P \times W, d) \right].$$

(41)

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

$$D\left( V||P_{X|Z}Q \right) = D\left( Q \times V||Q \times P_{X|Z} \right)$$

$$= \left\{ \begin{array}{ll}
0, & \text{if } Q \times V = Q \times \delta_{X|Z} \\
\infty, & \text{otherwise}
\end{array} \right.$$

(42)

Consequently,

$$F(Q, d) = \left\{ \begin{array}{ll}
0, & \text{if } E_{Q \times \delta_{X|Z}} \rho(X, \hat{X}) > d \\
\infty, & \text{otherwise}
\end{array} \right.$$

(43)

Therefore,

$$\sup_{W : (P, d) \leq R} F(P \times W, d)$$

$$= \left\{ \begin{array}{ll}
0, & \text{if } \forall W \in C(Z \rightarrow \hat{X}) \text{ with } I(P, W) \leq R, E_{P \times W} \rho(Z, \hat{X}) > d \\
\infty, & \text{otherwise}
\end{array} \right.$$

(44)

Finally, this implies

$$I(R, d, \lambda) = \inf_{P \in M(Z) : D(P||P_Z) \leq \lambda} D(P||P_Z) + \sup_{W \in C(Z \rightarrow \hat{X}) : I(P, W) \leq R} F(P \times W, d)$$

where $F_d(R)$ is precisely the exponent from [17].

2) A Slight Extension: Staying in the noise-free setting, we now generalize the result of [17], 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 $P \in M(\hat{X})$, we have in the noise-free case

$$\sup_{W : (P, d) \leq R} F(P \times W, d) = \infty.$$ 

(45)

Plugging into (6) gives

$$I(R, d, \lambda) = \inf_{P \in M(Z)} \left[ D(P||P_Z) \right]$$

$$+ \sup_{W \in C(Z \rightarrow \hat{X}) : I(P, W) \leq R} F(P \times W, d)$$

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

$$= \inf_{P \in M(Z) : D(P||P_Z) < \lambda, R(P, d) \geq R} D(P||P_Z)$$

(46)

$$= \left\{ \begin{array}{ll}
\infty, & \text{if } \forall P \text{ with } D(P||P_Z) < \lambda, \\
R(P, d) < R, & \text{otherwise}
\end{array} \right.$$

$$= \inf_{P \in M(Z) : D(P||P_Z) < \lambda, R(P, d) \geq R} D(P||P_Z)$$

(47)

where (46) follows from (44) and the rest from the definition of $F_d(R)$. Theorem 1 now gives us the following generalization of the result of [17] (cf. Fig. 1).

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

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

(48)

$$\leq \lim_{n \to \infty} \sup_{R \geq R} \left[ -\frac{1}{n} \log G_n(R, d, \lambda) \right]$$

(49)

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

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

(50)

Corollary 2 tells us the following two things.

- Even when the very strict requirement $\lambda = \infty$ of [17] is considerably relaxed to $\lambda \leq F_d(R)$, the excess-distortion exponent is not improved.
- $P_{X^n, Z^n}(\rho(X^n, \hat{X}^n(Z^n)) > nd)$ may decay at a super-exponential rate, while keeping the excess-code-length 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 I. A more thorough discussion of the dichotomic phenomenon exhibited in Corollary 2 is deferred to Section IV, where the setting of [17] 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 \( P_0(R) \) 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 binary-symmetric channel (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. [8]).

1) Qualitative Investigation of the Function \( I(R, d, \lambda) \): We recall here, for convenience, that

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

(51)

Clearly, we expect \( I(R, d, \lambda) \) to be nonincreasing in \( \lambda \). As this may not be obvious from (51), let us verify that this is indeed the case.

**Claim 1:** \( I(R, d, \lambda) \) is a monotone nonincreasing function of \( \lambda \) for fixed \( R, d \).

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

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

(52)

Since the function \( f_\lambda(\cdot, \infty, d) \) lies above the function \( a(\cdot, R, d) \), it is clear that \( \{f_\lambda(\cdot)\}_{\lambda \in [0, \infty)} \) is a nonincreasing family of functions (with increasing \( \lambda \)). Consequently, the respective minima of these functions is a nonincreasing function of \( \lambda \). But, for each \( \lambda \)

\[
\inf_{P \in \mathcal{M}(Z)} f_\lambda(P) = I(R, d, \lambda),
\]

\( \square \)

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

\[
I_R = \inf_{P \in \mathcal{M}(Z)} a(P, R, d)
\]

(53)

we let \( P_R^* \in \mathcal{M}(Z) \) denote the achiever in the right side of (53), 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 \( P \in \mathcal{M}(Z) \) with \( D(P||P_Z) \) sufficiently small, we also have \( I_\infty < a(P, R, d) \). This means that there is a whole interval \( [0, \hat{\lambda}] \) in which

\[
\inf_{P : D(P||P_Z) < \lambda} a(P, R, d) > I_\infty
\]

for all \( \lambda \in [0, \hat{\lambda}] \), where

\[
\hat{\lambda} = \inf \{D(P||P_Z) : a(P, R, d) \geq I_\infty\}.
\]

Now, since \( \hat{\lambda} \leq \lambda_\infty \) (as \( a(P_Z^*, R, d) \leq a(P_Z^*, \infty, d) = I_\infty \) and therefore clearly, for any \( \lambda > \lambda_\infty \),

\[
\inf_{P : D(P||P_Z) \leq \lambda} a(P, R, d) \leq \inf_{P : D(P||P_Z) \leq \lambda} a(P, R, d) = \inf_{P : D(P||P_Z) \leq D(P_R^*||P_Z)} a(P, R, d) \leq a(P, R, d) \leq I_\infty,
\]

this means that for any \( \lambda \in [0, \hat{\lambda}] \) we have

\[
\inf_{P : D(P||P_Z) \leq \lambda} a(P, R, d) = I_\infty.
\]

Consequently, for all \( \lambda \in [0, \hat{\lambda}] \), we have

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

(54)

By the definition of \( \hat{\lambda} \) it is also clear that for all \( \lambda \geq \hat{\lambda} \), the right branch in (51) “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 \( I_\infty \) until the right branch in (51) “kicks in.” From the point \( \lambda_R \) and on, the function is at a plateau again at level \( I_R \) (see Fig. 2 for a qualitative illustration). Note that the heights of the two plateaus mentioned are, respectively, \( I_\infty \) 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-hand side of (51) does not “kick in” at all except for \( \lambda = 0 \). The reason is that, in this case, we have for any \( \lambda > 0 \)

\[
\inf_{P : D(P||P_Z) \geq \lambda} a(P, R, d) \leq a(P, R, d) < I_\infty \leq \inf_{P : D(P||P_Z) \geq \lambda} a(P, \infty, d)
\]

(55)

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

so that

$$I(R, d, \lambda) = \inf_{P: D(P\|P_Z) < \lambda} \left( P, R, d, \lambda \right).$$

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-hand side of (51) is infinite and we have

$$I(R, d, 0) = \inf_{P, D(P\|P_Z) \geq 0} a(\cdot, \infty, d) = I_\infty.$$

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 $I_\infty$. 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 $a(P_Z, R, d)$ (see Figs. 2 and 3 for a qualitative illustration). Note that the regime of

**Possibility 2** will prevail when $R$ is sufficiently small.

2) The Region Where $I(R, d, \lambda) = 0$: Clearly, if the excess-distortion exponent is positive, we must have

$$E(I(X^n, \hat{X}^n(Z^n))) \leq nd + \epsilon, \quad \text{for all } \epsilon > 0$$

and sufficiently 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 (cf., e.g., [8]). Specifically, it was shown in [8] 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 $\rho^* \cdot Z \times \hat{X} \rightarrow [0, \infty)$, given by

$$\rho^*(z, \hat{x}) = E_{P_{X,Z}}[\rho(X, \hat{X})|Z = z].$$ (54)

Let us now show that our belief was justified.

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

$$I(R, d, \lambda) = 0, \quad \forall R < R^*(P_X, Z, d),$$ (55)

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

$$I(R, d, \lambda) \leq \inf_{P: D(P\|P_Z) < \lambda} \left( D(P\|P_Z) + \sup_{W \in \mathcal{C}(Z \rightarrow Y); \frac{I(PW)}{R(PW) \leq R}} F(P \times W, d) \right).$$ (56)

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

$$\inf_{P: D(P\|P_Z) < \lambda} \left( D(P\|P_Z) + \sup_{W \in \mathcal{C}(Z \rightarrow Y); \frac{I(PW)}{R(PW) \leq R}} F(P \times W, d) \right) = 0.$$ (57)

Taking $P = P_Z$, the left-hand side of (57), is upper-bounded by

$$\sup_{W \in \mathcal{C}(Z \rightarrow Y); \frac{I(P_Z \times W)}{R(P_Z \times W) \leq R}} F(P_Z \times W, d),$$ (58)

It will, therefore, suffice to show that the expression in (58) equals zero. But this will be the case if and only if

$$F(P_Z \times W, d) = 0, \quad \forall W \in \mathcal{C}(Z \rightarrow \hat{X}) \text{ s.t. } I(P_Z; W) < R^*(P_X, Z, d).$$ (59)

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

$$E_{P_Z \times W \times P_X}(\rho(X, \hat{X}) \geq d), \quad \forall W \in \mathcal{C}(Z \rightarrow \hat{X}) \text{ s.t. } I(P_Z; W) < R^*(P_X, Z, d).$$ (60)

But

$$R^*(P_X, Z, d) = \inf_{W \in \mathcal{C}(Z \rightarrow \hat{X})} I(P_Z; W) \quad \text{s.t. } \frac{I(P_Z; W)}{R(P_Z \times W) \times P_X} \leq \frac{\rho^*}{E_{P_Z \times W \times P_X}(\rho(X, \hat{X}) \geq d)} \leq \frac{\rho^*}{d}. $$ (61)

Consequently, if $I(P_Z; W) < R^*(P_X, Z, d)$ then

$$E_{P_Z \times W \times P_X}(\rho(X, \hat{X}) \geq d),$$ (62)

which implies (60). \qed

**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 $(X_k, Z_k, Y_k)$ 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 [21, Sec. 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.

IV. 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 [17], 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 [17], 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 [17].

A. The Error Exponent and the Universally Optimal Scheme

Let \( \mathcal{P}_\Theta \triangleq \{ P_\theta \mid \theta \in \Theta \} \) 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 \( \Theta \) as a subset of (the simplex in) \([0,1]^{[\mathcal{X}]}\). Furthermore, to avoid technical nuances, we assume throughout that \( \Theta \) is a closed (and, hence, compact) subset of \((0,\infty)\) in \([0,1]^{[\mathcal{X}]}\). 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 \( \inf_{\lambda \in \Theta} \int_{\mathcal{X}} \min_{P_{\theta \in \Theta}} P_{\theta}(x) > 0 \), a fact that we will rely on in the sequel. Note that the consideration of a general subset \( \Theta \) 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 tradeoff between the size and structure of the uncertainty set \( \Theta \) and the deterioration of the performance relative to the nonuniversal 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 modeled 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. [20], [19]).

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

\[
epsilon_\Theta(D[\theta]) \triangleq \liminf_{n \to \infty} -\frac{1}{n} \log P_{\theta}^{n} \{ \rho(X^n, \hat{X}^n) > nd \}
\]

(64)

where \( \hat{X}^n = \hat{X}^n(X^n) \) being the mapping from \( \lambda^n \) into \( \mathcal{B}_n \) associated with the scheme \( D_n \) (recall Section III for the complete definition of a coding scheme). Our goal is to find a sequence \( D \) that is independent of \( \theta \) and which maximizes \( \epsilon_\Theta(D[\theta]) \), uniformly for all \( \theta \) if possible, under the requirement that

\[
epsilon_\Theta(D[\theta]) \triangleq \liminf_{n \to \infty} -\frac{1}{n} \log P_{\theta}^{n} \{ L_n(X^n, D_n) > nR \}
\]

(65)

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 (65) can be given as follows. Letting

\[
A_{\lambda}(R, \lambda(\theta)) = \{ D_n \in \mathcal{D}_n : L_n(Z^n, D_n) > nR \} \leq e^{-\lambda n}
\]

one might require that, for all \( \eta \) sufficiently large

\[
D_n \in \bigcap_{\theta \in \Theta} A_{\lambda}(R, \lambda(\theta)),
\]

(66)

Clearly, the nonasymptotic requirement in (66) is slightly stronger than the asymptotic one of (65), though they are similar in spirit. The idea in letting \( \lambda(\theta) \) vary with \( \theta \) is motivated by the fact that certain sources in the parameter space \( \Theta \) 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 \( \bigcap_{\lambda \in \Omega} A_{\lambda}(R, \lambda(\theta)) \) of (66) can, in general, be made considerably larger by considering \( \theta \)-dependent \( \lambda(s) \), rather than some value of \( \lambda \) which would be constant all across the parameter space. This point will be elaborated on in Section IV-B.

A similar approach was recently proven quite fruitful in the context of composite hypothesis testing (cf. [16]). Indeed, there is an intimate relationship between the decision rule proposed in [16] and our coding scheme.

As we show next, for a given \( R > 0 \) there exists a sequence \( D^\mu \) (where the superscript \( \mu \) stands for “universal”) whose elements comply with (66) while maximizing \( \epsilon_\Theta(D[\theta]) \) universally in \( \Theta \). The idea in the construction of \( D^\mu \) 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 \( x^n \in T_\mu \) and \( D(P[\mu]) \leq \lambda(\theta) \), for some \( \theta \in \Theta \), then we cannot afford more than \( nR \) bits, otherwise, under \( P_\theta \), the probability of excess code length would be more than the probability of \( T_\mu \), which would, in turn, be essentially lower-bounded by \( \exp
\{-nD(P[\mu]) \} \leq \exp[-n\lambda(\theta)] \), violating (65). Consequently, if \( D(P[\mu]) \leq \lambda(\theta) \) for some \( \theta \in \Theta \) or, in other words,

\[
u(P) = u(P, \lambda(\cdot)) \triangleq \inf_{\theta \in \Theta} [D(P[\mu]) - \lambda(\theta)] \leq 0
\]

Continuity here is in the standard sense that for each \( \theta \in \Theta \) and \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that \( |\lambda(\theta) - \lambda(\bar{\theta})| \leq \varepsilon \) whenever \( \bar{\theta} \in \Theta \) satisfies \( \| \theta - \bar{\theta} \| \leq \delta \), where \( \infty - \infty \triangleq 0 \). Note, in particular, that when \( \Theta \) is finite, all \( \lambda \)s are continuous.
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 $D^c = \{D^c_n\}_{n \geq 1}$ as follows. For the associated codebook, $B^c_n$, we take the following:

- 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 $B(P) \subseteq \hat{X}^n$ with size $|B(P)| \leq 2n^{(R-\varepsilon)}$ 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 $\rho$-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 $C_n$, $B_n \rightarrow \{0, 1\}^+$ as follows: use no more than $|X|\log(n + 1)$ bits to convey the type of $x^n$. Now, if $x^n \in T_P$ with $u(P) \geq 0$, use $n\log|X|$ 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 $n(k\log(n + 1))/n < \varepsilon$), $\{I_n(X^n, D^c_n) > nR\} \subseteq T_P$ for some $P$ with $u(P) \leq 0$. Consequently, there exists $N(\varepsilon)$, independent of $\theta \in \Theta$, such that for all $n > N(\varepsilon)$ and all $\theta \in \Theta$

\[ P^c_\theta \{ I_n(X^n, D^c_n) > nR \} \leq \frac{1}{n} \sum_{n=0}^{\infty} \exp\{-n\lambda(\theta)\}. \]  

(67)

We now have a sequence of compression schemes $D^c$ for any $\varepsilon > 0$. We construct our scheme, $D^c$, as follows. Let now, for each $m \geq 1$, $e_m = \frac{1}{m} \log(n + 1)$. Since $\delta_n(\varepsilon) \to 0$ for every $\varepsilon > 0$ (recall the second item in the construction of the codebook for the definition of $\delta_n(\varepsilon)$), we can readily construct a (not necessarily strictly) increasing sequence of integers $\{m_n\}_{n \geq 1}$ satisfying

\[ m_n \to \infty \]  

(68)

yet sufficiently slowly such that both

\[ \delta_n(e_{m_n}) \to 0 \]  

(69)

and

\[ m_n \leq n, \quad \forall n \geq 1. \]  

(70)

Note that, in particular, this construction satisfies

\[ e_{m_n} \geq \frac{|X|\log(n + 1)}{n}, \quad \forall n \geq 1. \]  

(71)

By the uniformity of (67) in $\Theta$, we have

\[ P^c_\theta \{ I_n(X^n, D^c_n) > nR \} \leq \exp\{-n\lambda(\theta) - e_{m_n}\} \]  

(72)

for all sufficiently large $n$. Finally, we define $D^c = \{D^c_n\}_{n \geq 1}$. We shall write $D^c(\theta)$ when we want to make the dependence of $D^c$ on $R$ explicit. We begin by assessing the performance of $D^c$. In what follows, we let $\hat{X}_n^c(\cdot)$ denote the mapping from $\hat{X}^n$ into $B_n$, associated with $D^c_n$.

**Theorem 3:**

a) $D^c$ satisfies, for all $n$

\[ D^c_n \in \bigcap_{\theta \in \Theta} \hat{A}^c_\theta(R, \lambda(\theta) - e_{m_n}). \]  

(73)

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

\[ c_d(D^c[*]) \leq \lim\inf_{n \to \infty} \frac{1}{n} \log P^c_\theta \{ \rho(X^n, \hat{X}_n^c) > nd \} \]  

\[ \geq I_n(R, \theta, d = 0, \theta) \]  

(74)

and

\[ \lim\sup_{n \to \infty} \frac{1}{n} \log P^c_\theta \{ \rho(X^n, \hat{X}_n^c) > nd \} \]  

\[ \leq I_n(R, d = 0, \theta) \]  

(75)

where

\[ I_n(R, d = 0, \theta) \]  

(76)

and we define

\[ B(R, d, \lambda(\cdot)) \]  

(77)

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

\[ c_d(D^c[*]) = \lim_{n \to \infty} \frac{1}{n} \log P^c_\theta \{ \rho(X^n, \hat{X}_n^c) > nd \} \]  

\[ = I_n(R, d = 0, \theta) \]  

(78)

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

**Proof of Theorem 3:** Item a) follows from the construction of $D^c$ and, in particular, (72). Turning to the proof of b), we note that, by our construction of $D^c$, we have for any $\xi > 0$ and sufficiently large $n$

\[ P^c_\theta \{ \rho(X^n, \hat{X}_n^c) > nd \} \]  

\[ \leq \frac{1}{n} \sum_{n=0}^{\infty} \exp\{-n\lambda(\theta) - e_{m_n}\} \]  

(79)

Inequality (74) now follows by considering the normalized logarithm of the two ends of the above chain, letting $n \to \infty$, and then $\xi \to 0$.

For (75), we fix $\xi > 0$ and note that, by construction of $D^c$, we have for all sufficiently large $n$

\[ P^c_\theta \{ \rho(X^n, \hat{X}_n^c) > nd \} \]  

\[ \geq \frac{1}{n} \sum_{n=0}^{\infty} \exp\{-n\lambda(\theta) - e_{m_n}\} \]  

(80)
Furthermore, the definition of \( I_\theta(R_i, d, \theta) \), the continuity of \( D(\cdot|P_i) \) and of \( \lambda(\cdot) \) (which together imply the continuity of \( u(\cdot) \)), and the continuity of \( D(P|R) \) in \( P \) (cf. [6, Lemma 2.2]) imply that

\[
\lim_{n\to\infty} \left( \min_{P: u(P) < 0, D(P) > d} D(P|P_0) \right) = I_\theta(R, d, \theta). 
\]  

Thus, considering the normalized logarithm of the two ends of the chain of inequalities (80) and (81) and combining with (82) gives

\[
\limsup_{n\to\infty} -\frac{1}{n} \log P^n_\theta \{ \rho(X^n, \hat{X}^n_\theta) > nd \} \leq I_\theta(R, d + 2\xi, \theta). 
\]  

Finally, we take \( \xi \searrow 0 \) and establish (75). \( \square \)

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

**Theorem 4**: Let \( C(R_i, d, \Theta) \) denote the set of all \( \theta \in \Theta \) for which \( I_\theta(R_i, \cdot, \theta) \) is continuous at \( d \), and \( I_\theta(\cdot, d, \theta) \) is continuous at \( R_i \). For any sequence of schemes \( D \) satisfying

\[
ce_R(D|\theta) > \lambda(\cdot), \quad \forall \theta \in \Theta,
\]  

we have

\[
ce(D|\theta) \leq c_e(D^u|\theta) = I_\theta(R, d, \theta), \quad \forall \theta \in C(R, d, \Theta). 
\]  

Remark: Theorem 4 establishes the universal optimality of the scheme \( D^u \) with respect to the class of all schemes complying with the codeword overflow probability constraint dictated by \( \lambda(\cdot) \). Note that item a) of Theorem 3 implies, in particular, that \( D^u \) satisfies (65) uniformly in \( \Theta \). That is,

\[
\lim_{n\to\infty} \inf_{\theta \in \Theta} \inf_{P} \left[ -\frac{1}{n} \log P^n_\theta \{ L_n(X^n, D^u^n) > nR \} - \lambda(\theta) \right] \geq 0. 
\]  

Proof of Theorem 4: Let \( D = \{ D_n \} \) be any sequence of schemes satisfying (84) and assume, conversely, that there exists a \( \hat{\theta} \in C(\hat{R}, d, \Theta) \) for which

\[
ce(D|\hat{\theta}) > c_e(D^u|\hat{\theta}) = I_\theta(R, d, \hat{\theta}). 
\]  

Before plunging into formalities, let us outline the idea. Inequality (87) essentially implies, for large \( n \), the existence of a type \( T_{R} \) on which \( D \) gives distortion \( \leq nd \) while \( D^u \) gives distortion \( \geq nd \). The latter fact essentially implies that \( R(P, d) > R(D) \).

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 \( \geq nd \) are those for which \( u(P) < 0 \) and \( D(P, d) > d \). Now since \( R(P, d) > R \) and since \( D \) suffers distortion \( \leq nd \) on \( T_{R} \), essentially all \( \hat{x}^n \in T_{R} \) have \( \lambda_n(\hat{x}^n) > nR \). Furthermore, the fact that \( u(P) < 0 \) implies the existence of a \( \hat{\theta} \in \Theta \) for which \( D(P|P_0) \leq \lambda(\hat{\theta}) \).

Consequently, for large \( n \)

\[
\exp\{ -nc_R(D|\hat{\theta}) \} \approx P^n_\theta \{ L_n(X^n, D_n) > nR \} \geq P^n_\theta \{ T_{R} \} \approx \exp\{ -nD(P|P_0) \} \geq \exp\{ -n\lambda(\hat{\theta}) \} 
\]  

which contradicts (84) for \( \hat{\theta} \).

Turning to the formal proof, the fact that \( \hat{\theta} \in C(R, d, \Theta) \) implies that (87) holds for \( \theta = \hat{\theta} \). It thus follows from (87) that we can find an \( \eta = \eta(\hat{\theta}) > 0 \) such that for all sufficiently large\(^7 \)

\[
\frac{1}{n} \log P^n_\theta \{ \rho(X^n, \hat{X}^n_\theta) > nd \} + \eta \leq -\frac{1}{n} \log P^n_\theta \{ \rho(X^n, \hat{X}^n_\theta) > nd \} + e^{-n\eta}. 
\]  

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 \( n \)

\[
P^n_\theta \{ \rho(X^n, \hat{X}^n_\theta) > nd \} \leq \exp\{ -n[I_\theta(R, d + \epsilon, \hat{\theta}) + \eta/2] \} 
\]  

where the first inequality is a rewriting of (87) and the second inequality follows from a reuse of the fact that (87) holds for \( \theta = \hat{\theta} \). The continuity of \( I_\theta(\cdot, d, \theta) \) at \( R \) (recall that \( \hat{\theta} \in C(R, d, \Theta) \)) guarantees the existence of some \( \alpha = \alpha(\hat{\theta}) > 0 \) such that for all sufficiently large \( n \)

\[
P^n_\theta \{ \rho(X^n, \hat{X}^n_\theta) > nd \} \leq \exp\{ -n[I_\theta(R + \alpha, d, \hat{\theta}) + \eta/2] \}. 
\]  

Now, by the definition of \( I_n(R_i|\alpha, d, \hat{\theta}) \) (recall (76)), there exists a \( Q \in M(X) \) and a sequence \( \{ P_n \} \), where \( P_n \in M(X) \cap M_n \) such that \( u(P_n) \leq 0 \). Let \( D_{n,R} + \alpha \geq d \). \( P_n \to Q \), and \( D(Q|P_0) = I_n(R + \alpha, d, \hat{\theta}) \). Hence, by the continuity of \( D(\cdot|P_0) \), we must have for all sufficiently large \( n \)

\[
D(P_n|P_0) \leq I_n(R + \alpha, d, \hat{\theta}) + \eta/4. 
\]  

Also, the fact that \( u(P_n) \leq 0 \) (recall the continuity of \( D_{n,R} \) in \( \theta \), the continuity of \( \lambda(\cdot) \), and the compactness of \( \Theta \)) implies the existence of \( \hat{\theta}_n \in \Theta \) with

\[
D(P_n|P_{\hat{\theta}_n}) \leq \lambda(\hat{\theta}_n) 
\]  

such that \( \theta_n \to \theta \) and

\[
D(Q|P_0) \leq \lambda(\theta). 
\]  

On the other hand, since \( P_n \) satisfies (92), there must exist a \( \hat{\theta} \in C(\hat{R}, d, \Theta) \) with, say, \( |B(P_n)| \geq \frac{1}{4}|T_{R'}| \) such that

\[
\rho(x^n, \hat{X}^n(x^n)) \leq nd, \quad \forall x^n \in B(P_n) 
\]  

since, otherwise, (91) would be violated. By the converse to lossy coding (or type covering), (95) necessarily implies that, for all sufficiently large \( n \) (dependent on \( \alpha \) yet not on the particular sequence \( \{ P_n \} \) chosen),

\[
\left( |\{ x^n \in B(P_n) \} | \geq 2^{n[R(P_n, d) - \alpha/2]} \right. 
\]  

\(^6\)It is tedious but straightforward to show, using the continuity of \( D(\cdot, \cdot) \) in \( M(\lambda) \times [0, \infty) \), that when \( \Theta \) and \( \lambda(\cdot) \) are sufficiently well-behaved, \( C(R, d, \Theta) = \Theta \).

\(^7\)Throughout this proof, the “sufficiently large \( n \)” may be dependent on \( \hat{\theta} \).
which, since \( D(P_n, R + \alpha/2) \geq d \), leads to
\[
\left| \{ x^n \in B(P_n) \} \right| \geq 2^{n(R + \alpha/2)}. \tag{96}
\]
Inequality (96) inevitably implies, for sufficiently large \( n \) (again, independent of the particular sequence \( \{ P_n \} \) chosen), the existence of \( B = B(P_n) \) with, say,
\[
\left| B(P_n) \right| \geq \frac{1}{2} \left| B(P_n) \right| \geq \frac{1}{4} \left| T_{P_n} \right|
\]
such that
\[
L_n(x^n, D_n) \geq n(R + \alpha/4), \quad \forall x^n \in B(P_n) \tag{97}
\]
(this is established easily like in the proof of Theorem 1, cf. (A36) in particular). Consequently, for sufficiently large \( n \)
\[
P_{\lambda_n}^m \{ L_n(x^n, D_n) > nR \} \geq P_{\hat{\theta}_n}^m \{ L_n(x^n, D_n) \geq n(R + \alpha/4) \}
\geq P_{\hat{\theta}_n}^m \{ B(P_n) \}
\geq \frac{1}{4} (n+1)^{-\lambda} \exp \{ -nD(P||\hat{\theta}_n) \}
\geq \frac{1}{4} (n+1)^{-\lambda} \exp \{ -n\lambda(\hat{\theta}_n) \} \tag{98}
\]
where the last inequality follows from (93). Therefore,
\[
\liminf_{n \to \infty} -\frac{1}{n} \log P_{\hat{\theta}_n}^m \{ L_n(x^n, D_n) > nR \}
\leq \liminf_{n \to \infty} \lambda(\hat{\theta}_n) = \lambda(\theta) \tag{99}
\]
where the equality follows from the continuity of \( \lambda(\cdot) \) and the fact that \( \theta \to \hat{\theta} \). 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
\[
\lim_{n \to \infty} -\frac{1}{n} \log \frac{P_{\hat{\theta}_n}^m \{ A_n \}}{P_{\hat{\theta}}^m \{ A_n \}} = 0 \tag{100}
\]
which implies
\[
\liminf_{n \to \infty} -\frac{1}{n} \log P_{\hat{\theta}_n}^m \{ L_n(x^n, D_n) > nR \}
= \liminf_{n \to \infty} -\frac{1}{n} \log P_{\hat{\theta}}^m \{ L_n(x^n, D_n) > nR \}
= e_R(D||\hat{\theta}). \tag{101}
\]
Consequently, we have established
\[
e_R(D||\hat{\theta}) \leq \lambda(\hat{\theta}), \tag{102}
\]
contradicting (84).

B. Discussion

Working with the rate-distortion rather than the distortion-rate function, (76) implies that
\[
I_u(R, d, \theta) = \inf_{\{ P : u(P) \leq \theta, R(P, d) \geq R \}} D(P||P_\theta). \tag{103}
\]
Let further \( F_{d}^\theta(R) \) denote Marton’s exponent [17] corresponding to the source \( P^\theta \)
\[
F_{d}^\theta(R) = \inf_{\{ P : R(P, d) \geq R \}} D(P||P_\theta). \tag{104}
\]
Note first that when \( \Theta \) is a singleton we have
\[
I_u(R, d, \theta) = \inf_{\{ P : D(P||P_\theta) \leq \lambda(\theta), R(P, d) \geq R \}} D(P||P_\theta)
= \begin{cases} 
F_{d}^\theta(R), & \text{if } \lambda(\theta) \geq F_{d}^\theta(R) \\
\infty, & \text{otherwise}
\end{cases} \tag{105}
\]
where the \( \infty \) branch on the right-hand side of (105) 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 \( \Theta \) is a singleton, \( I_u(R, d, \theta) \) coincides with the exponent function derived in Section III-C2.

The next simple observation we make is that when \( \lambda(\theta) \geq F_{d}^\theta(R) \) then the \( P \) achieving the minimum in (104) satisfies \( D(P||P_\theta) \leq \lambda(\theta) \) and, \( a f t o r i o n i s, u(P) \leq 0 \) so that we have
\[
I_u(R, d, \theta) = F_{d}^\theta(R), \quad \forall \theta \in \Theta \text{ s.t. } \lambda(\theta) \geq F_{d}^\theta(R). \tag{106}
\]
Equation (106) can be given both an optimistic and a pessimistic interpretation. On the one hand, as we have seen in Corollary 2, even in the nonuniversal setting, the best achievable exponent for the distortion when subject to an overflow exponent larger than \( F_{d}^\theta(R) \) is \( F_{d}(R) \) itself. Hence, (106) 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 (106) holds for \( \lambda(\cdot) \) given by \( \lambda(\cdot) = \infty \) \( \forall \theta \in \Theta \) in this case, we are back to Marton’s setting [17], where the codeword 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 (106) 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 nonuniversal setting is observed for values of \( \theta \) for which \( \lambda(\theta) < F_{d}^\theta(R) \). While in the source-dependent setting, Corollary 2 gave a rate of \( \infty \), \( I_u(R, d, \theta) \) can well be finite. Technically, this follows from the fact that in the source-dependent setting, where \( \Theta = \{ \theta \} \), the minimizing set in (103), namely,
\[
\{ P \in \mathcal{M}(\mathcal{X}) : D(P||P_\theta) \leq \lambda(\theta) \leq 0, R(P, d) \geq R \}
\]
is empty whenever \( \lambda(\theta) < F_{d}^\theta(R) \) (cf. Section III-C2). In the universal setting, however, even when the set
\[
\{ P \in \mathcal{M}(\mathcal{X}) : u(P) \leq 0, R(P, d) \geq R \}
\]
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 nonuniversal setting, one could allot to each type \( P \) the
\[
\approx nR(P, d) \text{ bits necessary to cover it with distortion no more than } nd, \text{ thereby essentially annihilating the probability for distortion exceeding } nd, \text{ while complying with the requirement on the exponent associated with the codeword length. In the universal setting, on the other hand, the fact that } \lambda(\theta) < F_0^\alpha(R)
\]
does not imply that all types } P \text{ with } R(P, d) > R \text{ can be covered with distortion less than } nd. \text{ It only means, as in the nonuniversal case, that all types } P \text{ with } R(P, d) > R \text{ are such that } D(P' | P_0) > \lambda(\theta). \text{ There might exist some other } \theta' \in \Theta \text{ for which } D(P' | P_0) < \lambda(\theta') \text{ and, if this is the case, complying with the requirement on the codeword exponent forces one to allot no more than } nR \text{ bits to the type } P \text{ and, hence, to suffer distortion greater than } nd \text{ on this type.}

Thus, unlike the nonuniversal setting, } \lambda(\theta) < F_0^\alpha(R) \text{ does not automatically imply } I_0(R, d, \theta) = \infty. \text{ What it does imply, however, is that if all } P_0^\alpha \text{ 's achieving } F_0^\alpha(R) \text{ are such that } u(P_0^\alpha) > 0 \text{ then } I_0(R, d, \theta) \text{ is strictly greater than } F_0^\alpha(R). \text{ The situation in these cases is similar to that in the nonuniversal setting in the dichotomy exhibited between the regime } \lambda(\theta) > F_0^\alpha(R), \text{ where the best achievable exponent is } F_0^\alpha(R), \text{ and the regime } \lambda(\theta) > F_0^\alpha(R), \text{ where a better exponent is achievable.}

Notably, the discussion above tells us that if } \lambda(\cdot) \text{ is such that the set } \{P \in \mathcal{M}(X): u(P) \leq 0, R(P, d) \geq R\} \text{ is empty then, at all } \theta \in \Theta \text{ for which } \lambda(\theta) < F_0^\alpha(R), \text{ the exponential price of universality is infinite. This is because, in this case and for such a } \theta \in \Theta, \text{ the } \theta \text{-dependent scheme complying only with the codeword overflow exponent constraint for the source } P_0 \text{ is infinite, while the universal one is finite. A natural question arising in this context is whether there exists a } \lambda(\cdot) > 0 \text{ such that the } I_0(R, d, \theta) = \infty \text{ regime is attained. For this we would need the set } \{P \in \mathcal{M}(X): u(P) \in \Theta, R(P, d) \geq R\} \text{ to be empty or, in other words, we need } u(P) > 0 \text{ for all } P \text{ with } R(P, d) \geq R. \text{ This implies that } I_0(R, d, \theta) = \infty \text{ if and only if}
\[
\inf_{\Psi(Rd)} u(P) = \inf_{\Psi(Rd)} \inf_{\Theta} [D(P' | P_0) - \lambda(\theta')]
= \min_{\Psi(Rd)} \min_{\Theta} [D(P' | P_0) - \lambda(\theta')] > 0
\]
where
\[
\Psi(R, d) = \{P \in \mathcal{M}(X'); R(P, d) \geq R\}
\]
and the equality holds by our compactness assumption on } \Theta, \text{ the continuity of } \lambda(\cdot), \text{ and the compactness of } \Psi(R, d) \text{ (which follows from the continuity of } I_0(\cdot, \cdot) \text{ in } \mathcal{M}(X) \times [0, \infty)). \text{ The fact that}
\[
\inf_{\Psi(Rd)} u(P) = \min_{\Psi(Rd)} \min_{\Theta} [D(P' | P_0) - \lambda(\theta')]
= \min_{\theta'} [\min_{\Psi(Rd)} D(P' | P_0) - \lambda(\theta')]
\]
implies that a necessary and sufficient condition for
\[
\min_{\Psi(Rd)} D(P' | P_0) > \lambda(\theta'), \quad \forall \theta' \in \Theta,
\]
Finally, (108) (again, by the compactness of } \Psi(R, d) \text{ and } \Theta \text{ implies that a necessary and sufficient condition for the existence of a strictly positive } \lambda(\cdot) \text{ which will satisfy (108) is that}
\[
\liminf_{n \to \infty} \frac{1}{n} \log P(\Omega^n | \theta_1) \geq \lambda(\theta_1), \quad \forall \theta_1 \in \Theta_1
\]
where $P_{o}(\Omega_{o}[\theta])$ is the first kind of error probability associated with the decision rule $\Omega_{o}$, when the active source is $\theta_{1} \in \Theta_{1}$. The optimal decision rule $\Omega_{o}$ for this setting was shown to be one which decides on the first hypothesis when

$$\inf_{\theta_{1} \in \Theta_{1}} [D(F_{o}[P_{b_{2}}] - \lambda(\theta_{1})) \leq 0 \quad (111)$$

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 Section IV-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 [16], assuming that sequences of equal types receive equal treatment (an assumption which is justified by the fact that $P_{o}$ is a sufficient statistic for this problem, cf., e.g., [23, Lemma 1]), the fact that (111) holds implies the existence of $\theta_{1} \in \Theta_{1}$ for which $D(P_{o}[P_{b_{2}}] = \lambda(\theta_{1})$. Consequently, the first hypothesis must be chosen whenever (111) is satisfied since, otherwise, (110) 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 codewords of approximately equal lengths (an assumption which is justified by Theorem 4), the fact that the source sequence belongs to a type $P$ satisfying $\mu(P) \leq 0$ implies the existence of $\theta \in \Theta$ for which $D(P[\theta] \leq \lambda(\theta)$. Hence, such source sequences must be allotted no more than $nR$ bits as, otherwise, the excess-code-length requirement (65) would be violated. Another analogy with [16] is in the motivation for taking a $\theta$-dependent $\lambda(\cdot)$ and in the considerations guiding the search of a sensible $\lambda(\cdot)$. In [16], a judicious choice of $\lambda(\cdot)$ can make the difference between the existence of a decision rule with exponential decay of $P_{o}(\Omega_{o}[\theta])$ for all $\theta_{1} \in \Theta_{1}$ and the nonexistence 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.

1) Special Cases:

1. The Case Where $P_{o} = \mathcal{M}(\mathcal{X})$

As an extreme example, note that when $P_{o} = \mathcal{M}(\mathcal{X})$, we trivially have, for any $\lambda(\cdot)$, $\mu(P) \leq 0$ for all $P \in \mathcal{M}(\mathcal{X})$. Therefore, in this case, $I_{o}(R, d, \theta) = F_{d}^{b}(R)$. 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.

Note that, formally, this case does not conform to our standing assumption that $P_{o} \subseteq \mathcal{M}_{-1}(\mathcal{X})$. We will not be pedantic about that as the discussion of this trivial case is only a qualitative one.

The Case $\Theta = \{\theta_{1}, \theta_{2}\}$

Consider the case where there are only two possible sources, $P_{b_{1}}$ and $P_{b_{2}}$. In this case, we have

$$I_{o}(R, d, \theta_{1}) = \inf_{\{P_{o}|D(P_{o}[P_{b_{2}}] \leq \lambda(\theta_{2})\}} D(P_{o}[P_{b_{2}}]) \quad (112)$$

and

$$F_{d}^{b}(R), \quad \text{if } \lambda(\theta_{2}) \geq F_{d}^{b}(R)$$

$$\infty, \quad \text{if } \left\{ \begin{array}{l} \lambda(\theta_{1}) < F_{d}^{b}(R) \\ \lambda(\theta_{2}) < F_{d}^{b}(R) \end{array} \right\}$$

$$= \left\{ \begin{array}{l} \inf_{\{P_{o}|D(P_{o}[P_{b_{2}}] \leq \lambda(\theta_{2}), R, P_{o} \geq R\}} D(P_{o}[P_{b_{2}}]) \right\},$$

$$\text{if } \left\{ \begin{array}{l} \lambda(\theta_{1}) < F_{d}^{b}(R) \\ \lambda(\theta_{2}) \geq F_{d}^{b}(R) \end{array} \right\}.$$

Note that, by symmetry, $I_{o}(R, d, \theta_{2})$ looks the same with $1 \rightarrow 2$ and $2 \rightarrow 1$. Hence, we see that, as discussed above, we either have $I_{o}(R, d, \theta_{1}) = I_{o}(R, d, \theta_{2}) = \infty$ or both $I_{o}(R, d, \theta_{1})$ and $I_{o}(R, d, \theta_{2})$ finite. These two regimes are reached, respectively, according to whether both $\lambda(\theta_{1}) < F_{d}^{b}(R)$ and $\lambda(\theta_{2}) < F_{d}^{b}(R)$ or not. It is also evident from (113) that if we have both $\lambda(\theta_{1}) < F_{d}^{b}(R)$ and $\lambda(\theta_{2}) \geq F_{d}^{b}(R)$, then

$$F_{d}^{b}(R) < I_{o}(R, d, \theta_{1}) < \infty.$$ 

To see why the inequality $F_{d}^{b}(R) < I_{o}(R, d, \theta_{1})$ is strict note that in this case, where we assume $\lambda(\theta_{1}) < F_{d}^{b}(R)$, if $P^{*}$ achieves $F_{d}^{b}(R)$ then, by the definition of $F_{d}^{b}(R)$, $D(P^{*}[P_{b_{2}}] > \lambda(\theta_{1})$ and, hence, $P^{*}$ cannot belong to the minimizing set in (112). Thus, we see that in the case where $\lambda(\theta_{1}) < F_{d}^{b}(R)$ yet $\lambda(\theta_{2}) \geq F_{d}^{b}(R)$, as in the source-dependent case (cf. Section III-D), a better exponent than $F_{d}^{b}(R)$ is achievable. Unfortunately, unlike the source-dependent case, it is finite.

V. CONCLUSION AND FUTURE DIRECTIONS

The study of error exponents for lossy coding under a constrained probability of codeword 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 IV 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 $P_{b}$, in the noisy setting, as was seen in Section III, this is no longer the case. There would, therefore, no longer be hope of finding a single universal
scheme which would 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, generally, schemes which are optimal for all sources in the reference class do not exist. Thus, in that setting, one is led to consider somewhat less ambitious optimality criteria, such as the minimax criterion (cf. [15], [7]). 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 (100)

Note that by our assumption on \( \Theta, P_\theta \in \mathcal{M}_+(A') \) so that

\[
\alpha(\hat{\theta}) \triangleq \min_{x \in A} P_\theta(x) > 0.
\]

Let \( \varepsilon_n = \| P_\theta - P_\hat{\theta}_n \| \) and recall that, by our choice of \( \{ \hat{\theta}_n \} \),

\[
\varepsilon_n \to 0. \text{ Now, for each } n
\]

\[
P_{\hat{\theta}_n}^{(n)} \{ A_n \} = \sum_{x \in A_n} P_{\hat{\theta}_n}^{(n)} \{ x^n \} = \sum_{x \in A_n} \prod_{i=1}^{n} P_{\hat{\theta}_n}(x_i)
\]

\[
\leq \sum_{x \in A_n} \prod_{i=1}^{n} P_{\hat{\theta}_n}(x_i) + \varepsilon_n
\]

\[
\leq \sum_{x \in A_n} \prod_{i=1}^{n} P_{\hat{\theta}_n}(x_i) \left( 1 + \frac{\varepsilon_n}{a(\hat{\theta})} \right)^n
\]

Consequently,

\[
\limsup_{n \to \infty} \frac{1}{n} \log P_{\hat{\theta}_n}^{(n)} \{ A_n \} \leq \limsup_{n \to \infty} \log \left[ 1 + \frac{\varepsilon_n}{a(\hat{\theta})} \right] = 0.
\]

Analogously to (A5), we can show that

\[
P_{\hat{\theta}_n}^{(n)} \{ A_n \} \geq \left[ 1 - \frac{\varepsilon_n}{a(\hat{\theta})} \right]^n
\]

which, upon taking \( \liminf_{n \to \infty} \frac{1}{n} \log \)’s, completes the proof.

B. Other Proofs

1) Proof That \( I(R - \varepsilon, d, \lambda + \delta) \) and \( I(R + \varepsilon, d, \lambda - \delta) \) are Well-Defined: We need to prove that the limits

\[
\lim_{\varepsilon \to 0^+, \delta \to 0^+} I(R - \varepsilon, d, \lambda + \delta)
\]

exist and are equal. The existence of these limits is guaranteed by the monotonicity of \( I(R, d, \lambda) \) in both \( R \) (increasing) and \( \lambda \) (decreasing). More specifically, this implies that for any \( \eta \) and any \( m_0(\eta) \) (namely, for which \( R_0^{(n)} \leq R_1^{(n)} \) and \( \lambda_0^{(n)} \geq \lambda_1^{(n)} \) for all \( n \geq m_0(\eta) \)) such that

\[
I(R_0^{(n)}), d, \lambda_0^{(n)} \geq I(R_1^{(n)}, d, \lambda_1^{(n)}),
\]

for all \( n \geq m_0(\eta) \), which clearly implies the existence of the limits \( \lim_{n \to \infty} I(R^{(n)}, d, \lambda^{(n)}) \). To establish the equality between the limits it suffices to show that

\[
\lim_{n \to \infty} I(R_1^{(n)}, d, \lambda_1^{(n)}) = \lim_{n \to \infty} I(R_2^{(n)}, d, \lambda_2^{(n)}) \]

(since the arbitrary labeling of the indexes \( i = 1, 2 \) will imply the reverse inequality as well). Inequality (A10), however, follows easily from the monotonicity of \( I(R, d, \lambda) \) as well by observing that for any \( m_0(\eta) \) there exists \( m(\eta) \) (namely, one for which \( R_2^{(n)} \leq R_1^{(n)} \) and \( \lambda_2^{(n)} \geq \lambda_1^{(n)} \) for all \( n \geq m(\eta) \)) such that

\[
I(R_1^{(n)}, d, \lambda_1^{(n)}) \geq I(R_2^{(n)}, d, \lambda_2^{(n)}),
\]

for all \( n \geq m(\eta) \). Thus, the existence of the limit in (A8) is established. The existence of the limit in (A9) is proven similarly.

2) Proof That Equation (9) Holds Outside a Set of Pairs \((R, \lambda)\) of Zero Lebesgue Measure: Since the monotonicity of \( I(R, d, \lambda) \) implies, for all pairs \((R, \lambda)\)

\[
I(R - \varepsilon, d, \lambda + \delta) \leq I(R, d, \lambda) \leq I(R + \varepsilon, d, \lambda - \delta)
\]

it follows from Theorem 1 that it will suffice to show that the set \( S \subseteq \mathbb{R}^2 \) defined by

\[
S = \{(R, \lambda) \in [0, \infty)^2 : I(R - \varepsilon, d, \lambda + \delta) < I(R + \varepsilon, d, \lambda - \delta)\}
\]

has zero Lebesgue measure. To this end we note first that the fact that \( I(R - \varepsilon, d, \lambda + \delta) \) and \( I(R + \varepsilon, d, \lambda - \delta) \) are well-defined, i.e., that the limits which define these respective quantities exist
(cf. part B1) of the Appendix), implies that for each pair \((R, \lambda)\) we have both

\[
I(R - 0, \lambda, \lambda) = \lim_{\epsilon \to 0^+} I(R - \epsilon, \lambda, \lambda + \epsilon) \quad (A12)
\]

and

\[
I(R + 0, \lambda, \lambda) = \lim_{\epsilon \to 0^+} I(R + \epsilon, \lambda, \lambda - \epsilon). \quad (A13)
\]

Consider now the family of univariate monotone functions \(f_c: [0, c] \to [0, \infty)\) defined by

\[
f_c(R) = I(R, d, c - R), \quad 0 \leq R \leq c. \quad (A14)
\]

We note that, by the definition of \(f_c(\cdot)\) and (A12) and (A13), for each \(c \geq 0\) and \(0 \leq R \leq c\) we have

\[
f_c(R + 0) = I(R + 0, d, c - R - 0)
\]

and

\[
f_c(R - 0) = I(R - 0, d, c - R + 0). \quad (A15)
\]

Letting

\[
S_c = \{ R \in [0, c] : f_c(R + 0) > f_c(R - 0) \} \quad (A16)
\]

denote the set of continuity points of \(f_c\), it follows from the definitions of \(S\) and \(S_c\) and from (A15) that

\[
S = \bigcup_{c \geq 0} (R, c - R, R) \quad (A17)
\]

where the union on the right-hand side of (A17) is clearly a disjoint one. Letting \(\mu_1\) and \(\mu_2\) denote the Lebesgue measures on \(\mathbb{R}^1\) and \(\mathbb{R}^2\), respectively, (A17) gives

\[
\mu_2(S) = \int_{c \geq 0} \mu_1(S_c) \, dc. \quad (A18)
\]

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 continuity. In particular, we have \(\mu_2(S_c) = 0\) for all \(c \geq 0\) which annihilates the right-hand side of (A18) and completes the proof. \(\square\)

**C. Proof of Inequality (28)**

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

\[
\sum_{z \in \mathbb{Z}} \exp\left\{-n\left[D(P) + H(P) + F_n(P_{\tilde{X}^{n}(z)}) \right]\right\}
\]

\[
< \sum_{z \in \mathbb{Z}} \exp\left\{-n\left[D(P) + H(P) + \max_{\{W: I(PW) \leq R\}} F_n(P \times W) + \eta\right]\right\} \quad (A19)
\]

or, equivalently,

\[
\sum_{z \in \mathbb{Z}} \exp\left\{-nF_n(P_{\tilde{X}^{n}(z)}) \right\} < |T_P| \cdot \exp\left\{-n\left[\max_{\{W: I(PW) \leq R\}} F_n(P \times W) + \eta\right]\right\}. \quad (A21)
\]

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

\[
\sum_{z \in \mathbb{Z}} \exp\left\{-nF_n(P_{\tilde{X}^{n}(z)}) \right\} = \sum_{z \in \mathbb{Z}} \exp\left\{-nF_n(P_{\tilde{X}^{n}(z)}) \right\} \quad (A23)
\]

\[
+ \sum_{z \in \mathbb{Z}} \exp\left\{-nF_n(P_{\tilde{X}^{n}(z)}) \right\} \quad (A24)
\]

\[
\geq \sum_{z \in \mathbb{Z}} \max_{\{W: I(PW) \leq R\}} F_n(P \times W, d) \quad (A25)
\]

\[
= |T_P| \cdot \tilde{s}(P) \quad (A26)
\]

Combining (A26) with (A22) gives

\[
|\tilde{s}(P)| \geq |T_P| \quad (A27)
\]

or

\[
|\tilde{s}(P)| \geq (1 - e^{-\eta}) \cdot |T_P| \quad (A28)
\]

where the right inequality holds for all sufficiently large \(n\) (dependent on \(\eta\) yet independent of \(P \in \mathcal{M}(Z) \cap \mathcal{M}_n\)). Combining (A28) with the fact that

\[
|\mathcal{C}(Z) \cap \mathcal{C}_n(P)| \leq (n + 1)^{|Z||X|}\quad (A29)
\]

implies the existence of \(W \in \mathcal{C}(Z) \cap \mathcal{C}_n(P)\) with \(I(PW) > R\) such that the set \(s(P) \triangleq \{ z \in T_P : P_{\tilde{X}^{n}(z)} = P \times W \}\) (which is a subset of \(\tilde{s}(P)\)) satisfies

\[
|s(P)| \geq \frac{1}{2} (n + 1)^{-|Z||X|}\cdot |T_P| \quad (A30)
\]

\[
\geq \frac{1}{2} (n + 1)^{-|Z||X|}\cdot |T_P| \quad (A31)
\]

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

\[
|q(P)| = \sum_{z \in s(P)} |\tilde{X}^{n}(z) = \tilde{X}^{n}| \quad (A32)
\]

which, combined with (A29), implies

\[
|q(P)| \geq \frac{1}{2} (n + 1)^{-|Z||X|}\cdot \exp\{n[H(P) - H_{P \times W}(|Z|)]\} \quad (A33)
\]

\[
= \frac{1}{2} (n + 1)^{-|Z||X|}\cdot \exp\{n[H(P) - W]\} \quad (A34)
\]

\[
\geq \frac{1}{2} (n + 1)^{-|Z||X|}\cdot \exp\{nR\}. \quad (A35)
\]

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

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

Consequently,

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where the last (strict) inequality holds for all sufficiently large \( n \) (depending on \( \eta, \epsilon, \delta \)), yet independent of the \( P \in \mathcal{M}(\mathcal{Z}) \), \( P_{X|Z} \), \( P_{X|Z} \), and \( P_{X|Z} \) are defined by (18) and (8), respectively.

**Proof:** We start by recalling that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

Since \( P_{X|Z} \) is a finite convex combination of functions (of the form \( V(x|z) \)), it follows that, in particular, \( P_{X|Z} \) is finite convex combination of functions (of the form \( V(x|z) \)), which are all uniformly continuous in \( X \) when \( P_{X|Z} \) is uniformly continuous in the pair \( (X, Z) \). This, combined with the fact that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

and

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Claim 2:** For any \( P_{X|Z} \) we have

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Proof:** We start by recalling that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

and that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

Since \( P_{X|Z} \) is a finite convex combination of functions (of the form \( V(x|z) \)), it follows that, in particular, \( P_{X|Z} \) is finite convex combination of functions (of the form \( V(x|z) \)), which are all uniformly continuous in \( X \) when \( P_{X|Z} \) is uniformly continuous in the pair \( (X, Z) \). This, combined with the fact that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**D. Technical Claim in Proof of Theorem 1**

In this subsection, we aim to prove the following.

**Claim 2:** For any \( P_{X|Z} \in \mathcal{M}_4(\mathcal{X} \times \mathcal{Z}) \) we have

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Proof:** We start by recalling that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

and

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Claim 2:** For any \( P_{X|Z} \) we have

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Proof:** We start by recalling that

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Claim 2:** For any \( P_{X|Z} \) we have

\[
\frac{\exp\{n(H(P) - \epsilon + 1/n)\}}{\exp\{n(H(P) - \epsilon + 1/n)\}}
\]

where \( F_n(\cdot, \cdot) \) and \( F(\cdot, \cdot) \) are defined by (18) and (8), respectively.

**Acknowledgment**

The authors are grateful to the referees for their insightful comments, which have enhanced the readability of the manuscript.
REFERENCES


