Universally Attainable Error-Exponents for Rate-Constrained Denoising of Noisy Sources

Tsachy Weissman
Information Theory Research Group
HP Laboratories Palo Alto

HPL-2002-214
August 1st , 2002*

E-mail: tsachyw@exch.hpl.hp.com

achievable exponents, competitive minimax, compound source, denoising, error exponents, Marton's exponent, Neyman-Pearson, noisy source coding, universal schemes

Consider the problem of rate-constrained reconstruction of a finite-alphabet discrete memoryless signal X^n =(X_1,..., X_n), based on a noise-corrupted observation sequence Z^n, which is the finite-alphabet output of a Discrete Memoryless Channel (DMC) whose input is X^n. Suppose that there is some uncertainty in the source distribution, in the channel characteristics, or in both. Equivalently, suppose that the distribution of the pairs (X_i, Z_i), rather than completely being known, is only known to belong to a set ,. Suppose further that the relevant performance criterion is the probability of excess distortion, i.e., letting \hat X^n (Z^n) denote the reconstruction, we are interested in the behavior of $P_?$ρ (X^n), \hat X^n(Z^n)) > $d_?$), where ρ is a (normalized) block distortion induced by a single-letter distortion measure and $P_?$ denotes the probability measure corresponding to the case where (X*i*, Z*i*) ~?, ? ∈ ,. Since typically this probability will either not decay at all or do so at an exponential rate, it is the rate of this decay which we focus on. More concretely, for a given rate R ≥ 0 and a family of distortion levels {$d_?$} $_{? \in}$ subscript theta, we are interested in families of exponential levels {$I_?$} $_{?\in}$subscript theta which are *achievable* in the sense that for large *n* there exist rate-R schemes satisfying - 1/n log $P_?$ (ρ(X^n), \hat X^n(Z^n)) > $d_?$) ≥ $I_?$ for all ? ∈ ,. Our main result is a complete “single-letter” characterization of achievable levels {$I_?$} $_{?\in}$subscript theta per any given triple (, *R*, {$d_?$}$_{? \in}$ subscript,). Equipped with this result, we later turn to addressing the question of the “right” choice of {$I_?$}$_{?\in}$subscript ,. Relying on methodology recently put forth by Feder and Merhav in the context of the composite hypothesis testing problem, we propose a *competitive minimax* approach for the choice of these levels and apply our main result for characterizing the associated key quantities. Subsequently, we apply the main result to characterize optimal performance in a Neyman-Pearson-like setting, where there are two possible noise-corrupted signals. In this problem, the goal of the observer of the noisy signal, rather than having to determine which of the two it is (as in the hypothesis testing problem), is to reproduce the underlying clean signal with as high a fidelity as possible (e.g., lowest number of symbol errors when distortion measure is Hamming), under the assumption that one source is active, while operating at a limited information rate *R* and subject to a constraint on the fidelity of reconstruction when the other source is active. Finally, we apply our result to characterize a sufficient condition for the source class,, , to be universally encodable in the sense of the existence of schemes attaining the optimal distribution-dependent exponent, simultaneously for all sources in the class. This condition was shown in an earlier work to suffice for universality in expectation.

* Internal Accession Date Only

Approved for External Publication

Universally Attainable Error-Exponents for Rate-Constrained Denoising of Noisy Sources

Tsachy Weissman*

July 26, 2002

Abstract

Consider the problem of rate-constrained reconstruction of a finite-alphabet discrete memoryless signal $X^n = (X_1, \ldots, X_n)$, based on a noise-corrupted observation sequence Z^n, which is the finite-alphabet output of a Discrete Memoryless Channel (DMC) whose input is X^n. Suppose that there is some uncertainty in the source distribution, in the channel characteristics, or in both. Equivalently, suppose that the distribution of the pairs (X_i, Z_i), rather than completely being known, is only known to belong to a set Θ. Suppose further that the relevant performance criterion is the probability of excess distortion, i.e., letting $\hat{X}^n(Z^n)$ denote the reconstruction, we are interested in the behavior of $P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right)$, where ρ is a (normalized) block distortion induced by a single-letter distortion measure and P_θ denotes the probability measure corresponding to the case where $(X_i, Z_i) \sim \theta$, $\theta \in \Theta$. Since typically this probability will either not decay at all or do so at an exponential rate, it is the rate of this decay which we focus on. More concretely, for a given rate $R \geq 0$ and a family of distortion levels $\{d_\theta\}_{\theta\in\Theta}$, we are interested in families of exponential levels $\{I_\theta\}_{\theta\in\Theta}$ which are *achievable* in the sense that for large n there exist rate-R schemes satisfying $-\frac{1}{n}\log P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \geq I_\theta$ for all $\theta \in \Theta$. Our main result is a complete "single-letter" characterization of achievable levels $\{I_\theta\}_{\theta\in\Theta}$ per any given triple $(\Theta, R, \{d_\theta\}_{\theta\in\Theta})$. Equipped with this result, we later turn to addressing the question of the "right" choice of $\{I_\theta\}_{\theta\in\Theta}$. Relying on methodology recently put forth by Feder and Merhav in the context of the composite hypothesis testing problem, we propose a *competitive minimax* approach for the choice of these levels and apply our main result for characterizing the associated key quantities. Subsequently, we apply the main result to characterize optimal performance in a Neyman-Pearson-like setting, where there are two possible noise-corrupted signals. In this problem, the goal of the observer of the noisy signal, rather than having to determine which of the two it is (as in the hypothesis testing problem), is to reproduce the underlying clean signal with as high a fidelity as possible (e.g., lowest number of symbol errors when distortion measure is Hamming), under the assumption that one source is active, while operating at a limited information rate R and subject to a constraint on the fidelity of reconstruction when the other source is active. Finally, we apply our result to characterize a sufficient condition for the source class, Θ, to be universally encodable in the sense of the existence of schemes attaining the optimal distribution-dependent exponent, simultaneously for all sources in the class. This condition was shown in an earlier work to suffice for universality in expectation.

Key words and phrases: Achievable exponents, Competitive minimax, Compound source, Denoising, Error exponents, Marton's exponent, Neyman-Pearson, Noisy source coding, Universal schemes.

1 Introduction

A Limited-Rate Reconstruction of a Noisy Source

Consider a memoryless source $\{X_k\}_{k\geq 1}$, emitting symbols from the finite alphabet $\mathcal{X}$, corrupted by a discrete memoryless channel (DMC), and let $\{Z_k\}_{k\geq 1}$ denote the channel output. The problem of *rate-constrained denoising*,

*T. Weissman is with Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA. E-mail: tsachyw@exch.hpl.hp.com.

also referred to as *noisy source coding*, is that of recovering the source sequence $\{X_k\}_{k\geq 1}$ with as high a fidelity as possible, based on the noisy observation $\{Z_k\}_{k\geq 1}$, while operating at a limited information rate.

More concretely, a *scheme*, or a *rate-constrained denoiser*, or a *block code*, of length n is a mapping $\hat{X}^n : \mathcal{Z}^n \to \hat{\mathcal{X}}^n$, where $\mathcal{Z}, \hat{\mathcal{X}}$ are, respectively, the noisy signal and the reconstruction alphabets. The rate of the scheme is given by $\frac{1}{n}\log_2 |\{\hat{X}^n(z^n) : z^n \in \mathcal{Z}^n\}|$ bits per symbol, $|\cdot|$ denoting cardinality. One natural criterion for evaluating the performance of a scheme is its expected distortion, as measured by the single-letter distortion measure $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$. Letting $\mathcal{S}_n(R)$ denote the class of all block codes of length n and rate $\leq R$, this gives rise to the problem of finding

$$\min_{\hat{X}^n \in \mathcal{S}_n(R)} E\rho(X^n, \hat{X}^n(Z^n)), \tag{1}$$

where $\rho(x^n, \hat{x}^n) = n^{-1}\sum_{i=1}^n \rho(x_i, \hat{x}_i)$ (the left side of the equality being defined by the right). The value of the expression in (1) is well known (cf. [Ber71, Section 3.5], [WZ70, Wit80, EG88, Nat93, EC98, WM02b, DW02a] and references therein) to converge to

$$D(R) = \min E\rho(X, \hat{X}), \tag{2}$$

the minimum taken over all distributions of the triplet $(X, Z, \hat{X}) \in \mathcal{X} \times \mathcal{Z} \times \hat{\mathcal{X}}$ such that $(X, Z) \stackrel{d}{=} (X_1, Z_1)$, $X \to Z \to \hat{X}$ form a Markov chain, and $I(Z; \hat{X}) \leq R$ (the left side denoting the mutual information between Z and $\hat{X}$).

Another natural criterion for assessing the performance of a scheme is the probability of excess distortion. Specifically, given a rate R and a distortion level $d \in [D(R), \rho_{max}]$, where $\rho_{max} = \min_{\hat{x}\in\hat{\mathcal{X}}} \max_{x\in\mathcal{X}} \rho(x, \hat{x})$, the quantity of interest is[1]

$$P_n^{opt}(R, d) \stackrel{\triangle}{=} \min_{\hat{X}^n \in \mathcal{S}_n(R)} \Pr\left(\rho_n(X^n, \hat{X}^n(Z^n)) > d\right). \tag{3}$$

As $P_n^{opt}(R, d)$ will typically converge to zero exponentially rapidly, it is the rate of this exponential convergence, commonly referred to as the "error-exponent", which is of significance in this setting. The error-exponent for this problem was recently characterized in [WM02b]. Specifically, it follows as a special case of [WM02b, Theorem 1] that

$$I(R-0, d) \leq \liminf_{n\to\infty} \frac{1}{n}\log P_n^{opt}(R, d) \leq \limsup_{n\to\infty} \frac{1}{n}\log P_n^{opt}(R, d) \leq I(R+0, d), \tag{4}$$

with the error-exponent function $I(\cdot,\cdot)$ explicitly identified, and $I(R \pm 0, d) \stackrel{\triangle}{=} \lim_{\varepsilon\downarrow 0} I(R \pm \varepsilon, d)$. We shall return to the function $I(\cdot,\cdot)$ and elaborate on its detailed form in the sequel.

B Uncertainty in Source and Channel

Inherent in the problem description of the previous subsection is the complete knowledge of the statistics of the source and the channel. In particular, the optimal schemes attaining the minima in (1) and (3) in general depend, of course, on the distribution of the process $\{(X_k, Z_k)\}_{k\geq 1}$. Suppose now that there is some uncertainty in either the distribution of the clean source sequence, or the DMC corrupting it, or both. Equivalently, suppose that the joint

[1] Note that for values of d outside of $[D(R), \rho_{max}]$, the quantity in (3) is not interesting as it is either bounded away from zero ($d < D(R)$), or is identically zero for $d > \rho_{max}$.

distribution of the i.i.d. pairs (X_k, Z_k), rather than completely being known, is only known to belong to a given set $\Theta \subseteq \mathcal{M}(\mathcal{X} \times \mathcal{Z})$, where $\mathcal{M}(\mathcal{X} \times \mathcal{Z})$ denotes the simplex of all probability distributions on $\mathcal{X} \times \mathcal{Z}$.

Unfortunately, as was recently shown in [DW02a] (and as we shall elaborate on in subsection D below), for a given Θ there will, in general, not exist a (sequence of) scheme(s asymptotically) attaining (1) for all sources in Θ while being independent of the active source $\theta \in \Theta$. Letting E_θ denote expectation under the source corresponding to $\theta \in \Theta$ and $D(\theta, R)$ the associated distortion-rate function defined in (2), a seemingly plausible goal to strive for, under the circumstances, is that of minimizing the worst-case excess expected distortion beyond $D(\theta, R)$ over all sources $\theta \in \Theta$. Mathematically, the problem is that of finding

$$\min_{\hat{X}^n \in \mathcal{S}_n(R)} \max_{\theta \in \Theta} [E_\theta \rho_n(X^n, \hat{X}^n(Z^n)) - D(\theta, R)], \tag{5}$$

and characterizing the minimax-optimal scheme attaining it. This problem was the main theme of [DW02a], where the asymptotic value of the expression in (5), termed "the minimax distortion redundancy", was given a single-letter characterization and an asymptotically optimal sequence of schemes was identified.

Our goal in this work is to characterize the fundamental performance limitations, for a given class of sources Θ, when the performance criterion is the probability of excess distortion, rather than the expected distortion criterion considered in [DW02a].

C Our Setting: Universally Attainable Error Exponents for the Class Θ

As mentioned above, for a general Θ there do not exist schemes which are uniformly optimal in the sense of asymptotically attaining $D(\theta, R)$ for all $\theta \in \Theta$. A fortiori, the error exponent associated with (3) will not, in general, be achievable by a universal scheme for all sources in Θ. We shall introduce, in this context, a notion of achievability. In the ensuing definition, P_θ denotes the probability measure governing the process $\{(X_k, Z_k)\}$ when θ is the active source, and $\{I_\theta\}_{\theta \in \Theta}$, $\{d_\theta\}_{\theta \in \Theta}$ are arbitrary given families of non-negative reals indexed by the elements of Θ, which is a given family of distributions on (X_k, Z_k).

Definition 1 *$\{I_\theta\}_{\theta \in \Theta}$ will be said to be achievable at rate R for distortion levels $\{d_\theta\}_{\theta \in \Theta}$, or simply $(R, \{d_\theta\}_{\theta \in \Theta})$-achievable, if for all $\varepsilon > 0$ and sufficiently large n there exists a block-code $\hat{X}^n(\cdot)$ of rate $\leq R + \varepsilon$ satisfying*

$$-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta \right) \geq I_\theta - \varepsilon \quad \forall \theta \in \Theta. \tag{6}$$

$\{I_\theta\}_{\theta \in \Theta}$ will be said to be maximal at the above given rate and distortion levels if it is achievable and if any other achievable $\{\tilde{I}_\theta\}_{\theta \in \Theta}$ satisfies $\tilde{I}_\theta < I_\theta$ for some $\theta \in \Theta$.

Clearly, among achievable $\{I_\theta\}_{\theta \in \Theta}$, it is the maximal ones which deserve most of the attention, as for any achievable $\{I_\theta\}_{\theta \in \Theta}$ there exists a maximal $\{\tilde{I}_\theta\}_{\theta \in \Theta}$ which is at least as good for all $\theta \in \Theta$. Note that for the case where $\Theta = \{\theta\}$ is a singleton, the infimum of all I_θ which are achievable at rate R for distortion level d_θ is the error exponent function of (4) corresponding to the source θ (evaluated at R, d_θ). Thus, Definition 1 is a generalization of the notion of an achievable error exponent for the general case of a Θ consisting of more than one source. The sense in letting the exponential levels associated with the different sources, as well as the corresponding distortion levels, be

θ-dependent is, loosely speaking, that some sources are harder to recover than others. For one trivial example, it certainly does not make sense to take, for any θ, $I_\theta > I(\theta, R, d)$, the right side denoting the optimal error exponent function appearing in (4) associated with the source θ. Furthermore, as mentioned above and as will be elaborated on in the sequel, unfortunately, taking $\{I_\theta = I(\theta, R, d)\}_{\theta\in\Theta}$ is also overly ambitious in general. The issue of the "right" choice of $\{I_\theta\}_{\theta\in\Theta}$ and $\{d_\theta\}_{\theta\in\Theta}$ for a given family Θ will be dealt with in detail below.

The main result of this work is a characterization of the "achievable region", as defined by Definition 1, and of the achieving schemes. More specifically, in Section 3 we shall exhibit an explicit "single-letter" functional of the source class Θ, the exponential levels $\{I_\theta\}_{\theta\in\Theta}$, the distortion levels $\{d_\theta\}_{\theta\in\Theta}$, and the rate[2] R: $A(\Theta, \{I_\theta\}, \{d_\theta\}, R)$. In terms of this functional, our main result states essentially that $\{I_\theta\}_{\theta\in\Theta}$ is achievable at rate R for distortion levels $\{d_\theta\}_{\theta\in\Theta}$ if and only if $A(\Theta, \{I_\theta\}, \{d_\theta\}, R) \geq 0$. More precisely:

Theorem 1 **(a) Direct part:** *If* $A(\Theta, \{I_\theta\}, \{d_\theta\}, R) \geq 0$ *then* $\{I_\theta\}_{\theta\in\Theta}$ *is achievable at rate* R *for distortion levels* $\{d_\theta\}_{\theta\in\Theta}$.

(b) Converse part: *If* $A(\Theta, \{I_\theta\}, \{d_\theta\}, R) < 0$ *then, for all* $\varepsilon > 0$, $\{I_\theta + \varepsilon\}_{\theta\in\Theta}$ *is* not *achievable at rate* $R - \varepsilon$ *for distortion levels* $\{d_\theta\}_{\theta\in\Theta}$.

As indicated, we shall dedicate Section 3 primarily to the proof of Theorem 1. In doing that, we will be able, in particular, to also characterize in Corollary 2 therein the form of a maximal family (recall Definition 1). Also, the proof of the direct part will be constructive[3] in the sense of demonstrating a sequence of schemes satisfying (6).

Equipped with the necessary and sufficient condition for achievability implied by Theorem 1, we shall turn, in subsequent sections, to exploit its corollaries in several directions.

In Section 4 we shall verify that existing results can be recovered as special cases of Theorem 1. In particular, we shall see how for the noise-free setting we recover Marton's error exponent function [Mar74] and, furthermore, the fact that this error exponent is universally achievable with respect to all sources in the simplex. We shall also verify that for the non-universal setting, when $\Theta = \{\theta\}$ is a singleton, the error exponent of (4) is recovered.

It was mentioned above that while Theorem 1 concerns the achievability of a given set of exponential levels for a given set of distortion levels at a given rate, it does not address the issue of the "right" choices of these sets of levels for a given class of sources Θ. In Section 5 we shall propose a *competitive minimax* approach for the choice of these levels, relying on methodology recently put forth by Feder and Merhav in [FM02] for the composite hypothesis testing problem. Roughly, the initial idea is to optimize the worst-case performance over all sources in the class, where "performance" here is measured by the ratio between the probability of excess distortion of the universal scheme and that of the optimal distribution-dependent one. Unfortunately, as was mentioned in subsection B, universal encodability in our noisy setting is the exception, rather than the rule, even for universality in the expectation sense of [DW02a], a fortiori for the exponential sense under discussion. It would thus, in general, be overly ambitious to compare the performance of the universal scheme with the optimal distribution-dependent exponent. A compromise

[2] Here and throughout, to simplify notation, we shall omit the index set subscript when there is no room for ambiguity. Thus, we shall write, e.g., $\{I_\theta\}, \{d_\theta\}$ for $\{I_\theta\}_{\theta\in\Theta}, \{d_\theta\}_{\theta\in\Theta}$, respectively.

[3] Up to type-covering arguments based on random coding considerations.

is to compare the performance to a discounted version of the optimal distribution-dependent exponent, i.e., the set of exponential levels $\{\xi I(\theta, R, d_\theta)\}$, $\xi > 0$ being the discount factor. In the terminology of Definition 1, $\{\xi I(\theta, R, d_\theta)\}$ being $(R, \{d_\theta\})$-achievable then guarantees the existence of a scheme which is universal in the sense of having a positive exponent whenever the optimal distribution-dependent exponent is positive. Thus, given $(R, \{d_\theta\})$, this motivates choosing $\{\xi^* I(\theta, R, d_\theta)\}$ for the family of exponential levels, ξ^* denoting the maximum ξ for which $\{\xi I(\theta, R, d_\theta)\}$ is $(R, \{d_\theta\})$-achievable. In particular, $\xi^* > 0$ guarantees the existence of a universal scheme attaining a positive exponent for all sources $\theta \in \Theta$ for which $I(\theta, R, d) > 0$. Gratifyingly, Theorem 1 can be applied to obtain a single-letter expression for ξ^* (in terms of $\Theta, R, \{d_\theta\}$). This, as will be argued, can, in turn, also lead to a guideline for a sensible requirement on the choice of distortion levels $\{d_\theta\}$, namely, the associated ξ^* should be positive.

In Section 6 we shall apply Theorem 1 to characterize optimal performance for the following Neyman-Pearson-like setting: Suppose there are two possible noise-corrupted signals. The goal of the observer of the noisy signal, rather than having to determine which of the two it is (as in the hypothesis testing problem), is to reproduce the underlying clean signal with as high a fidelity as possible (e.g., lowest number of symbol errors when distortion measure is Hamming), while operating at a limited information rate R. Clearly, in general, for a given rate R, there is going to be a tradeoff between the probability with which a reconstruction at a given fidelity level can be guaranteed under one source, and the corresponding probability (for a possibly different distortion level) associated with the other source. It is this tradeoff that Theorem 1 will allow us to characterize. Note that the problem we propose here is suited better than the hypothesis testing setting to various signal detection applications, where the goal is to recover an approximate version of the underlying signal rather than to simply determine whether it is one signal or the other. Paraphrasing from the hypothesis testing terminology, we shall consider first the case of simple hypotheses, moving on to the most general case of composite vs. composite. For dealing with the latter, we shall, again, combine ideas which have been recently shown useful for the hypothesis testing framework. More specifically, we shall allow the constraint on the exponent of the error of the "first kind" to be dependent on $\theta_1 \in \Theta_1$, similarly as was motivated in [LM02] for the composite hypothesis testing problem. The motivation here, as well as a plausible choice for the constraint function, derive from considerations similar to those discussed above in the competitive minimax context. Now, subject to a constraint of this type, one can, again, resort to the competitive minimax approach of [FM02] in order to devise the "right" objective function for minimization. Equipped with these choices for a constraint on the composite error exponent of the first kind and the objective function associated with Θ_2 to be minimized, we shall proceed to characterize "single-letterly" the performance of the optimal scheme for this problem. It should be noted that there is an essential difference between the hypothesis testing problem of [LM02] and the one considered here: in the former there existed a scheme complying with the constraint on the error of the first kind which was uniformly optimal for *all* $\theta_2 \in \Theta_2$. Therefore, in that setting, there was no issue of the "right" choice of the objective functional of the composite Θ_2 to be minimized. In our problem there will generally, of course, not exist schemes that are uniformly optimal for all $\theta_2 \in \Theta_2$ under the constraint on the error of the first kind[4], which is why the choice of the optimality criterion for the composite Θ_2 is an issue.

[4] Just as there will not exist, in general, schemes which are universally exponentially optimal when there is no constraint on an error "of the first kind".

One of the contributions of [DW02a] was in identifying necessary and sufficient conditions for a class of noisy sources to be universally encodable in the expectation sense, i.e., in the sense of the existence of a sequence of schemes with expected distortion asymptotically attaining the distortion-rate function for all sources in the class. It was shown that a sufficient condition for universal encodability of a source class is that it be identifiable in the sense that different sources in the class will have different noisy marginal distributions and, furthermore, that there will exist a variational ball (of positive radius) around each noisy marginal not containing any other noisy marginal. Equivalently, the condition is that the class of sources be finite, with different noisy marginal distributions for each of the different sources in the class. In Section 7 we shall apply our results to extend this conclusion and show that it remains true for universality in the error-exponent sense, at least for certain (in some sense, to be argued, most natural) families of distortion levels. More specifically, it will be shown that when Θ satisfies this condition then for any given rate R and $\eta > 0$ sufficiently small, Θ is $(R, \{D(\theta, R) + \eta\}_{\theta\in\Theta})$-universally encodable in the error-exponent sense, i.e., there exist schemes attaining the optimal distribution-dependent exponent (which will be positive for this choice of rate and distortion levels), simultaneously for all sources in the class.

D Divergence from Noise-Free Setting

Universal lossy coding (of noise-free sources) with respect to (w.r.t.) the whole simplex of memoryless sources is known to be feasible in very strong senses. For any given rate there exist schemes that are universal in the sense of:

1. Attaining the distortion-rate function for all sources.
2. Attaining Marton's [Mar74] optimal error-exponent function for all sources. And, furthermore,
3. Being universal with respect to all distortion levels, i.e., not only can the optimal error-exponent from the previous item be universally attained for all sources, it can be attained simultaneously for all distortion levels.

One conceptually straightforward approach for a construction of such a scheme, at rate R, is the following: Partition all source sequences into types. The type covering Lemma [CK81] guarantees the existence of a code-book of size e^{nR} covering *all* source sequences within a type P with distortion essentially $D(P, R)$, where $D(P, R)$ denotes the distortion-rate function of the source P. Taking the union of these code-books gives a scheme of rate essentially R (only polynomial number of types), covering each source sequence x^n with distortion essentially no larger than $D(P_{x^n}, R)$, P_{x^n} denoting the empirical measure induced by x^n. Note that the construction of this scheme does not assume any particular active source, nor is it tuned to any particular distortion level. As for its performance note that if X^n was generated i.i.d. $\sim Q$ then, for any d, the probability that the distortion exceed d is essentially upper bounded by the probability of a type P with $D(P, R) > d$, namely $\approx \exp(-n \min_{P:D(P,R)>d} D(P\|Q))$, which is Marton's optimal lossy source coding exponent for the source Q, at rate R and distortion level d.

In the noisy setting of our present work, the situation is radically different:

1. *Expectation-sense Universality:* Clearly, except for extremely degenerate situations (involving degenerate distortion measures), there do not exist schemes which are universal w.r.t. all noisy sources, i.e., w.r.t. the whole

simplex $\mathcal{M}(\mathcal{X} \times \mathcal{Z})$. This is the first point of divergence between the noise-free and the noisy setting which was one of the central themes of [DW02a]. The bottom line for this setting is that either the class of sources for which universality is sought must be significantly smaller than the whole of $\mathcal{M}(\mathcal{X} \times \mathcal{Z})$, or the goal of universality must be compromised (giving rise to a minimax criterion (5)).

2. *Universality in the Error-Exponent sense:* A further divergence between the noise-free and noisy settings is observed when considering error exponents. While, as discussed above, for the noise-free setting there exist schemes which are universal in both expected and error-exponent sense, for the noisy setting, even the existence of schemes which are universal in the sense of the previous item does *not* guarantee existence of universal schemes in the error-exponent sense.

   To get a feel for why this is true, suppose that $\Theta = \{\theta^{(1)}, \theta^{(2)}\}$ consists of two sources differing in their noisy marginal distributions $\theta_Z^{(1)} \neq \theta_Z^{(2)}$. As is easy to show, there exist schemes which are universal in the expectation sense for this class. Suppose now that one attempts to construct a universal code for error-exponents. In the spirit of the construction outlined above for universally attaining Marton's exponent in the noise-free setting, a seemingly plausible approach is to partition the set of noisy observation sequences according to types, allotting each type approximately nR bits. But now arises the question of how to use the nR bits within a given type P. In the noise-free setting it was very clear: all members of the type can and should be covered with lowest possible distortion, namely, $D(P, R)$, no matter what the active source may be. As we quantify in this work, in the noisy setting, the optimal treatment within a type P is very much dependent on whether the active source is $\theta^{(1)}$ or $\theta^{(2)}$. In general, within a given type, an optimal bit allotment under one source will necessarily be sub-optimal for the other. Thus, the essence of the problem is that, for error-exponent performance, all types must be taken into consideration. Note that for universality in expectation this is not a problem since all one needs to worry about is the two (different) types $\theta_Z^{(1)}, \theta_Z^{(2)}$, attaining optimal performance under both sources by treating noisy sequences of type $\theta_Z^{(1)}$ as if they were emitted from the source $\theta^{(1)}$ and those of type $\theta_Z^{(2)}$ as emitted by $\theta^{(2)}$ (essentially forgetting about the remaining types, which are atypical under both sources and, therefore, inconsequential for an expected performance criterion).

3. *Universality w.r.t. the distortion level:* As was pointed out in [WM02b], even in the non-universal setting of error exponents for noisy source coding, the optimal schemes are dependent on the distortion level. The basic reason is, similarly as was explained in the context of the previous item, that the optimal bit allotment within a given type does not only depend on the source, but also on the distortion level. Indeed, even when there is no rate constraint, an optimal filter maximizing the error-exponent for one distortion level may be strictly sub-optimal for another distortion level.

E Literature Context and Problem Motivation

Shannon's classical theory of lossy source coding [Sha59, Gal68, Ber71] has been subjected in recent decades to extensive research and has been advanced in various directions (cf. [Ber98, Kie93] and the many references therein).

Some of these directions include:

1. *Universal* lossy source coding (cf., e.g., [Ziv72, NGD75, Ziv80, Ris84, LLZ94, CEG96, YK96] for a representative sample).

2. Lossy compression of *noisy* sources (cf. [Ber71, Section 3.5], [WZ70, Wit80, EG88, EC98, WM02b] and references therein).

3. Source coding error exponents (cf., e.g., [Mar74, Bla74, Bla76, Bla87, CK81, KN96, Hum81, Jel68, Wyn74, Mer91, MK01]).

The combination of the first two of the above items, namely, the problem of *universal* coding of *noisy* sources has, prior to this work, been given attention in the recent [DW02a, DW02b] (probabilistic setting) and [WM02a, Section 5] (individual sequence setting). The combination of the second and third of the above items was addressed in [WM02b], where some of the potential obstacles in the way of extending the scope to the universal case were discussed. The present work is the first to address the combination of all three: *error exponents* for *noisy* source coding in the *universal* setting. In fact, even the special instance of our results to the case of the pure denoising problem with no rate constraint (i.e., $R \geq \log |\mathcal{Z}|$) appears to be new.

More than in its generalizing and merging of these three domains, the merit of this work (Theorem 1 in particular) lies in its further characterization, beyond that pertaining to the setting of [DW02a], of the approximation-estimation tradeoff[5] for the noisy source coding problem. As is discussed in [DW02a], in various situations involving statistical modelling there is a need to balance the tradeoff between taking a rich "reference class" of sources Θ (so as to approximate the "true" data-generating mechanism as closely as possible) and the ability to reach optimum performance for the various sources in the class. Quantifying the extent to which a source class, Θ, can be universally encoded, as was done in [DW02a] and as we do in this work, provides one with tools for selecting the most appropriate Θ for a given problem. This is, in some sense, analogous to the application of the MDL principle, which characterizes fundamental limitations on universal noise-free coding (cf. [Ris84, Ris96]) as a model selection rule in various problems (cf. [BRY98] and references therein).

An additional motivation for studying the fundamental limitations on rate-constrained denoising in the universal setting is the relatively recent emergence of the compression-based approach to the denoising problem [Nat93, Ris00, Don02]. In a nutshell, the heuristic underlying this approach is that the noise corrupting the clean signal will constitute that part of the observed noisy signal which is hardest to compress. Thus, by employing a universal lossy compression scheme on the noisy signal, it seems plausible to expect that the part of the noisy signal which is hardest to compress, namely the noise, will be lost. Thus, by appropriately tuning the distortion level according to the noise level, this heuristic seems to suggest using the lossy output (i.e., the output of the lossy compressor whose input is the observed noisy signal) for approximating the underlying noise-free signal. This approach was recently applied to two concrete settings (binary signals corrupted by a BSC and real-valued signals corrupted by additive Gaussian white noise) and its performance rigorously analyzed in [Don02]. It should be emphasized that there is

[5] Also referred to as the "under-fitting versus over-fitting dilemma".

a conceptual difference between the compression-based approach to denoising and our setting: While in the former one's hope is that the rate constraint will *facilitate* the ultimate goal, which is to denoise (i.e., the rate-constraint is only used as a tool for coming up with a denoising scheme, it is not a real requirement), in our setting we assume a veritable constraint on the rate and, subject to such a constraint, our goal is to characterize the fundamental limitations on denoising performance. Nonetheless, ultimately the right benchmarks for assessing the performance of any compression-based denoiser are the fundamental limitations characterized in [DW02a] and herein.

Finally, it should be mentioned that the notion of a universally achievable family of exponential levels, as introduced in Definition 1, is not completely new. An analogous notion of a "universally attainable error exponent" was introduced in [CK81, Definition 5.7] in the context of the compound channel. Unfortunately, the channel coding analogue of our problem, as considered in [CK81], remained open to this day. Indeed, the problem of determining universally attainable error exponents for a given compound channel appears today, as it did when [CK81] came out, to be "very difficult"[6]. It seems somewhat remarkable, in light of this, that the analogous problem considered here of determining universally attainable error exponents for the compound noisy source can be completely solved (Theorem 1).

F Remaining Content

The remainder of this work is organized as follows: In Section 2 we present our notation conventions and recall some standard facts for later use. Section 3 will be dedicated to the proof of Theorem 1 and some of its implications (e.g., the characterization of a maximal family in Corollary 2). In Section 4 we shall make a sanity check and verify that known results, pertaining to settings contained as special cases of the current problem, are recoverable as corollaries of Theorem 1. Sections 5, 6 and 7 deal, respectively, with the competitive minimax setting, the Neyman-Pearson setting, and universality in the error-exponent sense, as elaborated on in subsection C above. Finally, Section 8 summarizes the paper and discusses a direction for related future work.

2 Notation, Conventions, and Preliminaries

For arbitrary finite sets $\mathcal{A}$ and $\mathcal{B}$, we let $\mathcal{M}(\mathcal{A})$ denote the set of all probability measures on $\mathcal{A}$, $\mathcal{M}_+(\mathcal{A})$ the set of probability measures on $\mathcal{A}$ which assign positive mass to all $a \in \mathcal{A}$, and $\mathcal{C}(\mathcal{A} \to \mathcal{B})$ the set of all stochastic matrices (or "channels" or "conditional distributions") from $\mathcal{A}$ to $\mathcal{B}$. The variational distance between two elements of $\mathcal{M}(\mathcal{A})$, P and P', is defined by $d_v(P, P') = \max_{s \subset \mathcal{A}} [P(s) - P'(s)]$.

For $P \in \mathcal{M}(\mathcal{A})$ we let $H(P)$ denote the entropy of a random variable distributed according to P. For $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ we will write $P \times W$ for the distribution of the pair $(A, B) \in \mathcal{A} \times \mathcal{B}$ when A is generated according to P and then B is taken as the output of the channel W whose input is A. Furthermore, if $Q = P \times W$ we will refer to W as the "channel from $\mathcal{A}$ to $\mathcal{B}$ induced by Q". We will also let, in this case, $H(W|P)$ denote the entropy of B given A, i.e.,

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

[6] A quote taken from [CK81, Pg. 172]. As is well known, even the case where the channel is known has yet to be completely solved.

Similarly, we will let $I(P;W)$ denote 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. Logarithms and exponents throughout are assumed to be of base-2, as is assumed to be the case in definitions of mutual information, entropy, etc.

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.

For any $a^n \in \mathcal{A}^n$ we let $P_{a^n} \in \mathcal{M}(\mathcal{A})$ denote the associated empirical measure. For $P \in \mathcal{M}(\mathcal{A})$ we let $T_P^n = \{a^n \in \mathcal{A}^n : P_{a^n} = P\}$ denote the type class of P and omit the superscript n from T_P^n when there is no room for ambiguity. For $n \in \mathbb{N}$ we let $\mathcal{M}_n(\mathcal{A})$, or simply $\mathcal{M}_n$ when the alphabet is clear from the context, denote the set of all $P \in \mathcal{M}(\mathcal{A})$ for which $T_P^n \neq \emptyset$. We shall further let, for $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ and $a^n \in \mathcal{A}^n$, $T_W^n(a^n)$, or simply $T_W(a^n)$, denote the set of all $b^n \in \mathcal{B}^n$ having conditional type W given a^n (cf. [CK81, Definition 2.4]). For $a^n \in \mathcal{A}^n$, $b^n \in \mathcal{B}^n$ we let $P_{a^n,b^n} \in \mathcal{M}(\mathcal{A} \times \mathcal{B})$ denote the associated joint empirical measure.

Following [CK81], for any $P \in \mathcal{M}(\mathcal{A})$ we let $T^n_{[P]_\delta}$ denote the set of all sequences $a^n \in \mathcal{A}^n$ that are P-typical with constant δ (cf. [CK81, Definition 2.8]). We further let, for any stochastic matrix $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ and $a^n \in \mathcal{A}^n$, $T^n_{[W]_\delta}(a^n)$ denote the set of all $b^n \in \mathcal{B}^n$ that are W-typical under the condition $a^n \in \mathcal{A}^n$ with constant δ (cf. [CK81, Definition 2.9]). An immediate consequence of the definitions of δ-typical sequences is (cf. [CK81, Lemma 2.10]):

$$\text{If } a^n \in T^n_{[P]_\delta} \text{ and } b^n \in T^n_{[W]_{\delta'}}(a^n) \text{ then } (a^n, b^n) \in T^n_{[P\times W]_{\delta+\delta'}}. \tag{7}$$

We shall adopt throughout the "delta-convention" used in [CK81]. Namely, we assume a fixed sequence of positive reals $\{\delta_n\}_{n\geq1}$ satisfying

$$\delta_n \to 0, \quad \sqrt{n}\delta_n \to \infty \qquad \text{as} \quad n \to \infty \tag{8}$$

and, for any n, $P \in \mathcal{M}(\mathcal{A})$, $W \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ and $a^n \in \mathcal{A}^n$, we write $T^n_{[P]}$ (or simply $T_{[P]}$) for $T^n_{[P]_{\delta_n}}$ and $T^n_{[W]}(a^n)$ (or simply $T_{[W]}(a^n)$) for $T^n_{[W]_{\delta_n}}(a^n)$.

When dealing with expectations of functions or with functionals of random variables, we shall sometimes subscript the distributions of the associated random variables. Thus, for example, for any $f : \mathcal{A} \to \mathbb{R}$ and $P \in \mathcal{M}(\mathcal{A})$ we shall write $E_P f(A)$ for the expectation of $f(A)$ when A is distributed according to P. Similarly, we shall write, for example, $H_P(f(A))$ to denote the entropy of $f(A)$ when A is distributed according to P and, for $Q \in \mathcal{M}(\mathcal{A} \times \mathcal{B})$, $H_Q(B|A)$ will denote the conditional entropy of B given A when $(A, B) \sim Q$.

In what follows, we assume that $\mathcal{X}$, $\mathcal{Z}$, $\hat{\mathcal{X}}$ are finite alphabets. As previously described, $\mathcal{X}$ is where the components of the clean process take values, $\mathcal{Z}$ is where the noisy observations of the clean process take their values, and $\hat{\mathcal{X}}$ is the reconstruction alphabet. If $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$ and $V \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$, we shall sometimes slightly abuse the notation defined above by writing $P \times V$ to denote the distribution on $(X, Z, \hat{X}) \in \mathcal{X} \times \mathcal{Z} \times \hat{\mathcal{X}}$ where (X, Z) are generated according to P and then $\hat{X}$ is taken as the output of the channel V whose input is Z (so that in this case $X \to Z \to \hat{X}$ form a Markov chain). Also, for $V \in \mathcal{C}(\hat{\mathcal{X}} \times \mathcal{Z} \to \mathcal{X})$, $U \in \mathcal{C}(\mathcal{Z} \to \mathcal{X})$ and $Q \in \mathcal{M}(\hat{\mathcal{X}} \times \mathcal{Z})$, we shall sometimes slightly abuse the notation by writing $D(V\|U|Q)$ for $D(V\|\tilde{U}|Q)$, where $\tilde{U} \in \mathcal{C}(\mathcal{Y} \times \mathcal{Z} \to \mathcal{X})$ is the channel which coincides with U (i.e., the output of the channel $\tilde{U}$ is independent of the $\mathcal{Y}$-valued component of the input). For $\theta \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$ we shall write θ_X, θ_Z, $\theta_{X|Z}$ to denote the induced Z-marginal, X-marginal, and conditional distribution of X given Z, respectively. For cases where $\mathcal{X} = \mathcal{Z}$ we denote the noise-free channel (where $Z = X$ with probability one) by $\delta_{Z|X}$. For $P \in \mathcal{M}(\mathcal{X})$, $R(P, D)$ and $D(P, R)$ will denote, respectively, the standard (noise-free) rate distortion and distortion rate functions associated with the source P (when the reconstruction alphabet is $\hat{\mathcal{X}}$ and w.r.t. the same distortion measure ρ assumed throughout). Similarly, for $\theta \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$, $D(\theta, R)$ will denote the distortion rate function for the noisy source θ, as explained in subsection 1.B.

Throughout we shall adopt the convention that capital letters represent random variables, while the corresponding lower case letters represent specific sample values. We shall use the notation min, max to generally denote a (not necessarily attainable) infimum, supremum, respectively. We define $\log 0 = -\infty$ (to accommodate cases involving zero probabilities). Finally, the infimum and the supremum over the empty set are defined by ∞ and $-\infty$, respectively.

3 Characterization of the Achievable Region

In what follows we assume a given family of sources $\Theta \subseteq \mathcal{M}(\mathcal{X} \times \mathcal{Z})$. We define the function $F : \mathcal{C}(\mathcal{Z} \to \mathcal{X}) \times \mathcal{M}(\mathcal{Z} \times \hat{\mathcal{X}}) \times [0, \infty) \to [0, \infty]$ by

$$F(U, Q, d) = \min_{\left\{ \begin{array}{c} V \in \mathcal{C}(\mathcal{Z} \times \hat{\mathcal{X}} \to \mathcal{X}) : \\ E_{Q \times V} \rho(X, \hat{X}) > d \end{array} \right\}} D\left(V \| U | Q\right). \tag{9}$$

For a set of channels indexed by $\mathcal{M}(\mathcal{Z})$, $\{W_P\}_{P \in \mathcal{M}(\mathcal{Z})} \subseteq \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$, and two sets of positive reals indexed by Θ, $\{I_\theta\}_{\theta \in \Theta}$ and $\{d_\theta\}_{\theta \in \Theta}$, we define

$$G\left(\Theta, \{W_P\}, \{I_\theta\}, \{d_\theta\}\right) = \min_{P \in \mathcal{M}(\mathcal{Z}), \theta \in \Theta} \left[D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) - I_\theta\right] \tag{10}$$

and

$$A(\Theta, \{I_\theta\}, \{d_\theta\}, R) = \max_{\{W_P\} : \max_P I(P; W_P) \leq R} G\left(\Theta, \{W_P\}, \{I_\theta\}, \{d_\theta\}\right). \tag{11}$$

This section is dedicated to the proof of Theorem 1, where the functional $A(\Theta, \{I_\theta\}, \{d_\theta\}, R)$ appears.

A An Explicit Restatement of Theorem 1

In this subsection we recapitulate Theorem 1 in a somewhat more explicit form for convenience in its proof and its later application.

Observe that $A(\Theta, \{I_\theta\}, \{d_\theta\}, R) \geq 0$ if and only if there exists a set of channels[7] indexed by $\mathcal{M}(\mathcal{Z})$, $\{W_P\}_{P \in \mathcal{M}(\mathcal{Z})} \subseteq \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$, satisfying both

$$\min_{P \in \mathcal{M}(\mathcal{Z}), \theta \in \Theta} \left[D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) - I_\theta\right] \geq 0 \tag{12}$$

and

$$\max_{P \in \mathcal{M}(\mathcal{Z})} I(P; W_P) \leq R. \tag{13}$$

[7] In the sequel we shall omit the index set, writing $\{W_P\}$ for $\{W_P\}_{P \in \mathcal{M}(\mathcal{Z})}$.

In other words, being even more explicit, $A(\Theta, \{I_\theta\}, \{d_\theta\}, R) \geq 0$ if and only if the following holds:

Hypothesis 1 *For every $P \in \mathcal{M}(\mathcal{Z})$ there exists $W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$ satisfying both*

$$D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W, d_\theta) \geq I_\theta \quad \forall \theta \in \Theta \tag{14}$$

and

$$I(P; W) \leq R. \tag{15}$$

Thus, the direct part of Theorem 1 asserts that if Hypothesis 1 holds (equivalently, if (12) and (13) hold for some $\{W_P\}$) then $\{I_\theta\}_{\theta\in\Theta}$ is achievable at rate R for distortion levels $\{d_\theta\}_{\theta\in\Theta}$. The converse part asserts that if Hypothesis 1 does not hold then, for all $\varepsilon > 0$, $\{I_\theta + \varepsilon\}_{\theta\in\Theta}$ is *not* achievable at rate $R - \varepsilon$ for distortion levels $\{d_\theta\}_{\theta\in\Theta}$.

Another advantage for the above formulation is in that it conveys the form of a maximal family. Specifically, we have the following corollary to Theorem 1:

Corollary 2 *If $\{I_\theta\}_{\theta\in\Theta}$ is maximal for $(R, \{d_\theta\}_{\theta\in\Theta})$ then there exists $\{W_P\}$ with $\max_P I(P; W_P) \leq R$ such that*

$$I_\theta = \min_{P\in\mathcal{M}(\mathcal{Z})} \left[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta)\right] \quad \forall \theta \in \Theta. \tag{16}$$

Proof: If $\{I_\theta\}_{\theta\in\Theta}$ is maximal for $(R, \{d_\theta\}_{\theta\in\Theta})$ then it is, a fortiori, achievable. Thus, by the direct part of Theorem 1, there exists $\{W_P\}$ with $\max_P I(P; W_P) \leq R$ and

$$I_\theta \leq \min_{P\in\mathcal{M}(\mathcal{Z})} \left[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta)\right] \quad \forall \theta \in \Theta. \tag{17}$$

In other words, denoting $\tilde{I}_\theta \overset{\triangle}{=} \min_P \left[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta)\right]$, $I_\theta \leq \tilde{I}_\theta$ for all $\theta \in \Theta$. Since trivially, by its definition, $\{\tilde{I}_\theta\}$ satisfies Hypothesis (1), it follows from Theorem 1 that it is $(R, \{d_\theta\}_{\theta\in\Theta})$-achievable. Now, if there existed $\theta \in \Theta$ for which $I_\theta < \tilde{I}_\theta$, that would have contradicted the maximality of $\{I_\theta\}$. Thus, $I_\theta = \tilde{I}_\theta$ for all $\theta \in \Theta$. □

B Main Idea behind Proof of Theorem 1

In this subsection we give an informal outline of the main proof idea.

The significance of the functional $F(U, Q, d)$ of equation (9) is in that for all $z^n \in \mathcal{Z}^n$, $\hat{x}^n \in \hat{\mathcal{X}}^n$, $\theta \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$ and $d \geq 0$,

$$P_\theta\left(\rho(X^n, \hat{x}^n) > d | Z^n = z^n\right) \approx \exp\left(-nF(\theta_{X|Z}, P_{z^n,\hat{x}^n}, d)\right). \tag{18}$$

To get a feeling for why this is so note that, conditioned on $Z^n = z^n$ and for any $\hat{x}^n \in \hat{\mathcal{X}}^n$, $\{\rho(X_i, \hat{x}_i)\}$ are distributed, under P_θ, like an arbitrarily varying source with $(\hat{x}_i, z_i)$'s serving as states and with distribution[8] $\mathcal{L}_\theta(\rho(X_i, \hat{x}_i)|Z_i = z_i)$. Thus, the exponential price for a fluctuation in the empirical measure of $\{\rho(X_i, \hat{x}_i)\}$ is a divergence between the measure to which it fluctuates and the true law, $\theta_{X|Z}$, averaged over the frequency of the occurrence of the various "states", namely, according to $P_{z^n,\hat{x}^n}$. This is what $F(\theta_{X|Z}, P_{z^n,\hat{x}^n}, d)$, as defined in (9), is doing, taking the minimum of this weighted divergence over all measures under which the distortion exceeds d.

[8] We let $\mathcal{L}_\theta(\rho(X_i, \hat{x}_i)|Z_i = z_i)$ denote the law of $\rho(X_i, \hat{x}_i)$ conditioned on $Z_i = z_i$, when the active source is θ.

Converse idea: Take any scheme $\hat{X}^n(\cdot)$ restricted to rate R. For any type of Z-sequences, $T_P \subseteq \mathcal{Z}^n$, look at the empirical measures induced by pairs $(z^n, \hat{X}^n(z^n))$ for the various $z^n \in T_P$. Since the number of possible empirical measures is polynomial in n it follows that there exists at least one joint type $T_Q \subseteq (\mathcal{Z} \times \hat{\mathcal{X}})^n$ (where Q is of the form $P \times W$ for some channel from $\mathcal{Z}$ into $\hat{\mathcal{X}}$) such that $(z^n, \hat{X}^n(z^n)) \in T_Q$ for an exponentially non-negligible portion of sequences in T_P. Since each sequence $\hat{x}^n \in T_{Q_{\hat{X}}}$ can "cover" no more than $\approx e^{nH_Q(Z|\hat{X})}$, it follows that the number of distinct code-words must be approximately $\geq |T_P|/e^{nH_Q(Z|\hat{X})} \approx e^{nI(Q)}$. Thus, the restricted rate of $\hat{X}^n(\cdot)$ implies essentially that $I(Q) = I(P;W) \leq R$. From this it follows that, for each type P, there exists an exponentially non-negligible fraction of sequences $z^n \in T_P$ for which $P_\theta\left(\rho(X^n, \hat{X}^n(z^n)) > d_\theta | Z^n = z^n\right) \overset{\sim}{>} \exp\left(-nF(\theta_{X|Z}, P \times W, d_\theta)\right)$, and, consequently,

$$P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \overset{\sim}{>} \exp\left(-n[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W, d_\theta)]\right) \quad \forall \theta \in \Theta, \tag{19}$$

where W is some channel with $I(P;W) \leq R$. Now, if Hypothesis 1 does not hold this implies the existence of a type P such that, whenever $I(P;W) \leq R$,

$$\exists \theta \in \Theta : D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W, d_\theta) < I_\theta.$$

Combined with (19) this implies the existence of $\theta \in \Theta$ for which $P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \overset{\sim}{>} \exp(-nI_\theta)$, essentially completing the proof of the converse part.

Direct idea: Suppose Hypothesis 1 holds. This implies the existence (cf. subsection A) of $\{W_P\}_{P \in \mathcal{M}(\mathcal{Z})}$, $W_P \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$, satisfying both (12) and (13). Now, by type-covering-like arguments one can show that for each $P \in \mathcal{M}_n(\mathcal{Z})$ there exists a "code-book", i.e., a subset of $\mathcal{Z}^n$ of size $\leq e^{nI(P;W_P)} \leq e^{nR}$ (the second inequality owing to (13)) such that for each $z^n \in T_P$ there exists $\hat{x}^n$ for which $(z^n, \hat{x}^n) \in T_{[P \times W_P]}$. Letting $\hat{X}^n(\cdot)$ be the scheme corresponding to the union of these code-books over the different types, it is clear that the rate of this scheme is $\overset{\sim}{<} e^{nR}$ (as there are only a polynomial number of types) and that for each $P \in \mathcal{M}_n(\mathcal{Z})$ and $z^n \in T_P$, $(z^n, \hat{X}^n(z^n)) \in T_{[P \times W_P]}$. Thus, by (18), it follows that for all $\theta \in \Theta$, $P \in \mathcal{M}_n(\mathcal{Z})$ and $z^n \in T_P$, $P_\theta\left(\rho(X^n, \hat{x}^n) > d_\theta | Z^n = z^n\right) \approx \exp\left(-nF(\theta_{X|Z}, P \times W_P, d_\theta)\right)$. This, in turn, essentially concludes the proof of the direct as it implies that, for all $\theta \in \Theta$, $P_\theta\left(\rho(X^n, \hat{x}^n) > d_\theta\right) \approx \exp\left(-n \min_P[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta)]\right) \overset{\sim}{<} e^{-nI_\theta}$, where the (approximate) inequality follows since, by (12), $\min_P[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta)] \geq I_\theta$ for all $\theta \in \Theta$.

The next subsection contains the rigorous version of the proof sketched above.

C Proof of Theorem 1

Recall first that the following was established in [WM02b, Section III-B, equations (11)-(22)]: For all $z^n \in \mathcal{Z}^n$, $\hat{x}^n \in \hat{\mathcal{X}}^n$, $\theta \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$, $d \geq 0$,

$$(n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \exp\left(-nF_n(\theta_{X|Z}, P_{z^n,\hat{x}^n}, d)\right) \leq P_\theta\left(\rho(X^n, \hat{x}^n) > d | Z^n = z^n\right) \leq (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \exp\left(-nF_n(\theta_{X|Z}, P_{z^n,\hat{x}^n}, d)\right), \tag{20}$$

where, for $W \in \mathcal{C}(\mathcal{Z} \to \mathcal{X})$ and $Q \in \mathcal{M}_n(\mathcal{Z} \times \hat{\mathcal{X}})$, $F_n(\cdot,\cdot,\cdot)$ is given by

$$F_n(W, Q, d) = \min_{\{V \in \mathcal{C}(\mathcal{Z} \times \hat{\mathcal{X}} \to \mathcal{X}) : E_{Q \times V}\rho(X,\hat{X}) > d\} \cap \mathcal{C}_n(Q)} D(V\|W|Q), \tag{21}$$

$\mathcal{C}_n(Q)$ denoting the subset of channels $V \in \mathcal{C}(\mathcal{Z} \times \hat{\mathcal{X}} \to \mathcal{X})$ for which $Q \times V \in \mathcal{M}_n(\mathcal{Z} \times \hat{\mathcal{X}} \times \mathcal{X})$.

Proof of Converse Part of Theorem 1: We must prove that if Hypothesis 1 does not hold then for all $\varepsilon > 0$ and sufficiently large n we have the following: For any scheme $\hat{X}^n(\cdot)$ of rate $\leq R - \varepsilon$ there exists a source $\theta \in \Theta$ such that

$$-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta \right) \leq I_\theta + \varepsilon. \tag{22}$$

To this end, suppose that Hypothesis 1 does not hold, i.e., there exists $P \in \mathcal{M}(\mathcal{Z})$ for which there does not exist a channel W satisfying both (14) and (15). In other words, there exists $P \in \mathcal{M}(\mathcal{Z})$ such that for all W satisfying $I(P; W) \leq R$,

$$\exists \theta \in \Theta : D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W, d_\theta) < I_\theta. \tag{23}$$

The continuity properties of the mutual information functional $I(\cdot;\cdot)$ (cf. [CK81]) imply that for any $\varepsilon > 0$ and sufficiently large n, by letting $P^{(n)} \in \mathcal{M}_n(\mathcal{Z})$ be the closest member of $\mathcal{M}_n(\mathcal{Z})$ to P (say, under variational norm) then $\forall W : I(P^{(n)}; W) \leq R - \varepsilon$ equation (23) holds (with $P^{(n)}$ substituted for P). Furthermore, the continuity properties of $D(\cdot\|\cdot)$ and those of $F(\cdot, \cdot, d_\theta)$ (cf. [WM02b, Appendix C]) imply that (when n is sufficiently large) for any θ: $D(P^{(n)} \| \theta_Z) + F(\theta_{X|Z}, P^{(n)} \times W, d_\theta) \leq D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W, d_\theta) + \varepsilon$. Thus we have for any $\varepsilon > 0$ and sufficiently large n the existence of $P^{(n)} \in \mathcal{M}_n(\mathcal{Z})$ such that

$$\forall W : I(P^{(n)}; W) \leq R - \varepsilon \quad \exists \theta \in \Theta : D(P^{(n)} \| \theta_Z) + F(\theta_{X|Z}, P^{(n)} \times W, d_\theta) \leq I_\theta + \varepsilon. \tag{24}$$

Fix now $\varepsilon > 0$, n, $P^{(n)} \in \mathcal{M}_n(\mathcal{Z})$ satisfying (24), and an arbitrary scheme $\hat{X}^n(\cdot)$ of rate $\leq R - 2\varepsilon$. For any z^n the empirical distribution of $\left(z^n, \hat{X}^n(z^n)\right)$ is, by definition, a member of $\mathcal{M}_n(\mathcal{Z} \times \hat{\mathcal{X}})$. Consequently, there exists $Q^{(n)} \in \mathcal{M}_n(\mathcal{Z} \times \hat{\mathcal{X}})$ such that

$$Q_Z^{(n)} = P^{(n)} \tag{25}$$

and for which $S(Q^{(n)}) \triangleq \left\{ z^n \in T_{P^{(n)}} : \left(z^n, \hat{X}^n(z^n)\right) \in T_{Q^{(n)}} \right\} \subseteq T_{P^{(n)}}$ satisfies

$$\left| S(Q^{(n)}) \right| \geq \frac{|T_{P^{(n)}}|}{|\mathcal{M}_n(\mathcal{Z} \times \hat{\mathcal{X}})|} \geq (n+1)^{-(|\mathcal{Z}| + |\mathcal{Z}||\hat{\mathcal{X}}|)} e^{nH(P^{(n)})} \geq e^{n(H(P^{(n)}) - \varepsilon)} \tag{26}$$

(assuming n sufficiently large so that $e^{-n\varepsilon} \leq (n+1)^{-(|\mathcal{Z}| + |\mathcal{Z}||\hat{\mathcal{X}}|)}$). On the other hand, it is clear that for every $\hat{x}^n$, $|\{z^n \in S(Q^{(n)}) : \hat{X}^n(z^n) = \hat{x}^n\}| \leq |T_{V^{(n)}}(\hat{x}^n)| \leq \exp(nH(V^{(n)}|P_{\hat{x}^n})) = \exp(nH_{Q^{(n)}}(Z|\hat{X}))$, $V^{(n)} \in \mathcal{C}(\hat{\mathcal{X}} \to \mathcal{Z})$ being the channel induced by $Q^{(n)}$. Consequently,

$$\begin{aligned}
\exp(n(R - 2\varepsilon)) &\geq |\{\hat{X}^n(z^n) : z^n \in \mathcal{Z}^n\}| \\
&\geq |\{\hat{X}^n(z^n) : z^n \in S(Q^{(n)})\}| \\
&\geq \frac{|S(Q^{(n)})|}{\exp(nH_{Q^{(n)}}(Z|\hat{X}))} \\
&\geq \frac{\exp(n(H(P^{(n)}) - \varepsilon))}{\exp(nH_{Q^{(n)}}(Z|\hat{X}))} = \exp(n(I(Q^{(n)}) - \varepsilon)).
\end{aligned} \tag{27}$$

Thus we have $I(P^{(n)}; V^{(n)}) = I(Q^{(n)}) \leq R - \varepsilon$ implying, by (24), the existence of $\theta \in \Theta$ such that

$$D(P^{(n)} \| \theta_Z) + F(\theta_{X|Z}, Q^{(n)}, d_\theta) \leq I_\theta + \varepsilon. \tag{28}$$

Thus, for this θ,

$$\begin{aligned}
& P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \\
\geq \;& (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{z^n \in T_{P^{(n)}}} \exp\left(-n\left[D(P^{(n)}\|\theta_Z) + H(P^{(n)}) + F_n(\theta_{X|Z}, P_{z^n,\hat{X}^n(z^n)}, d_\theta)\right]\right) \\
\geq \;& (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{z^n \in S(Q^{(n)})} \exp\left(-n\left[D(P^{(n)}\|\theta_Z) + H(P^{(n)}) + F_n(\theta_{X|Z}, P_{z^n,\hat{X}^n(z^n)}, d_\theta)\right]\right) \\
= \;& (n+1)^{-|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} |S(Q^{(n)})| \exp\left(-n\left[D(P^{(n)}\|\theta_Z) + H(P^{(n)}) + F_n(\theta_{X|Z}, Q^{(n)}, d_\theta)\right]\right) \\
\geq \;& \exp\left(-n\left[D(P^{(n)}\|\theta_Z) + F_n(\theta_{X|Z}, Q^{(n)}, d_\theta) + 2\varepsilon\right]\right) \qquad (29) \\
\geq \;& \exp\left(-n\left[D(P^{(n)}\|\theta_Z) + F(\theta_{X|Z}, Q^{(n)}, d_\theta) + 3\varepsilon\right]\right) \\
\geq \;& \exp\left(-n\left[I_\theta + 4\varepsilon\right]\right), \qquad (30)
\end{aligned}$$

where the first inequality follows from the left inequality in (20), the inequality in (29) follows from (26), the inequality before last follows by the fact established in [WM02b, Appendix D] that[9] $|F(W,Q,d) - F_n(W,Q,d)| \to 0$ uniformly in (W,Q,d) (and n is assumed sufficiently large so that $|F - F_n| \leq \varepsilon$). The last inequality follows from (28). Thus, for the arbitrary $\varepsilon > 0$ we have taken an arbitrary scheme of rate $\leq R - 2\varepsilon$ and established the existence of a θ for which, by (30), $-\frac{1}{n}\log P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \leq I_\theta + 4\varepsilon$. This concludes the proof of the converse part.

For the proof of the direct part, we shall employ the following "type-covering" assertion.

Proposition 3 *There exists a sequence of positive reals $\{\varepsilon_n\}_{n\geq 1}$ with $\varepsilon_n \to 0$ as $n \to \infty$, depending only on $|\mathcal{Z}|$ and $|\hat{\mathcal{X}}|$, such that for every n, distribution $P \in \mathcal{M}_n(\mathcal{Z})$, and stochastic matrix $W \in \mathcal{C}(\mathcal{Z} \to \hat{\mathcal{X}})$ there exists a mapping $f_P^n : T_P^n \to \hat{\mathcal{X}}^n$ satisfying both*

$$f_P^n(z^n) \in T_{[W]}^n(z^n) \quad \forall z^n \in T_P^n \qquad (31)$$

and

$$|\{f_P^n(z^n) : z^n \in T_P^n\}| \leq e^{n[I(P;W)+\varepsilon_n]}. \qquad (32)$$

A proof of a slightly stronger version of Proposition 3 can be found in [DW02a] (cf. Proposition 1 and its proof therein, which is based on a random coding argument).

Proof of Direct Part of Theorem 1: We must show that if Hypothesis 1 holds then for all $\varepsilon > 0$ and sufficiently large n (dependent on ε) there exists a block-code $\hat{X}^n(\cdot)$ of rate $\leq R + \varepsilon$ satisfying

$$-\frac{1}{n}\log P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \geq I_\theta - \varepsilon \quad \forall \theta \in \Theta. \qquad (33)$$

To this end, let $\varepsilon > 0$ be fixed. Assume Hypothesis 1 holds and for $P \in \mathcal{M}(\mathcal{Z})$ let W_P denote a channel satisfying both (14) and (15), namely

$$D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) \geq I_\theta \quad \forall \theta \in \Theta \qquad (34)$$

and

$$I(P; W_P) \leq R \qquad (35)$$

[9] More precisely, in [WM02b, Appendix D] it was shown that $|F(W,Q,d) - F_n(W,Q,d)| \to 0$ uniformly in Q for fixed (W,d). The argument used, however, can actually be shown to imply uniformity in the triple (W,Q,d).

(in other words, $\{W_P\}$ is the set of channels satisfying (12) and (13)). For a given n we construct a block-code $\hat{X}^n(\cdot)$ as follows: For each $P \in \mathcal{M}_n(\mathcal{Z})$ let f_P^n be a mapping satisfying the assertion of Proposition 3 for the channel W_P. The block-code is constructed via

$$\hat{X}^n(z^n) = f^n_{P_{z^n}} \quad \forall z^n \in \mathcal{Z}^n, \tag{36}$$

so that

$$\begin{aligned} |\{\hat{X}^n(z^n) : z^n \in \mathcal{Z}^n\}| &= \sum_{P \in \mathcal{M}_n(\mathcal{Z})} |\{\hat{X}^n(z^n) : z^n \in T_P\}| && (37) \\ &\leq \sum_{P \in \mathcal{M}_n(\mathcal{Z})} e^{n(I(P;W_P)+\varepsilon_n)} \\ &\leq |\mathcal{M}_n(\mathcal{Z})| e^{n(R+\varepsilon_n)} \leq e^{n\left(R+\varepsilon_n+\frac{|\mathcal{Z}|}{n}\log(n+1)\right)}, && (38) \end{aligned}$$

the first inequality owing to Proposition 3 and the second to (35).

For the performance of this scheme we have, for all $\theta \in \Theta$,

$$\begin{aligned} & P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \\ \leq\ & (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{P \in \mathcal{M}_n(\mathcal{Z})} \sum_{z^n \in T_P} \exp\left(-n\left[D(P\|\theta_Z) + H(P) + F_n(\theta_{X|Z}, P_{z^n,\hat{X}^n(z^n)}, d_\theta)\right]\right) \\ \leq\ & (n+1)^{|\mathcal{X}||\mathcal{Z}||\hat{\mathcal{X}}|} \sum_{P \in \mathcal{M}_n(\mathcal{Z})} \exp\left(-n\left[D(P\|\theta_Z) + \min_{z^n \in T_P} F_n(\theta_{X|Z}, P_{z^n,\hat{X}^n(z^n)}, d_\theta)\right]\right) \\ \leq\ & \sum_{P \in \mathcal{M}_n(\mathcal{Z})} \exp\left(-n\left[D(P\|\theta_Z) + F_n(\theta_{X|Z}, P \times W_P, d_\theta) - \varepsilon - \varepsilon\right]\right) && (39) \\ \leq\ & \exp\left(-n\left[I_\theta - 4\varepsilon\right]\right), && (40) \end{aligned}$$

where the justification for the last chain of inequalities is as follows: The last inequality follows from (34) (which holds for all $P \in \mathcal{M}_n(\mathcal{Z})$) and the fact discussed in the proof of the converse that $|F_n - F| \to 0$ uniformly in the first two arguments (and n is assumed sufficiently large so that $|F_n - F| \leq \varepsilon$ and $|\mathcal{M}_n(\mathcal{Z})| \leq e^{n\varepsilon}$). For the inequality before last we have used the fact that, by construction of the scheme $\hat{X}^n(\cdot)$ (via the mappings of Proposition 3), $\hat{X}^n(z^n) \in T_{[W_P]}(z^n)\ \forall z^n \in T_P$ and $P \in \mathcal{M}_n(\mathcal{Z})$, implying (by equation (7)) that $(z^n, \hat{X}^n(z^n)) \in T_{[P\times W_p]_{2\delta_n}}\ \forall z^n \in T_P$ (δ_n pertaining to the δ-convention introduced in Section 2). Consequently,

$$\min_{z^n \in T_P} F_n(\theta_{X|Z}, P_{z^n,\hat{X}^n(z^n)}, d_\theta) \geq \min_{Q \in \mathcal{M}_n(\mathcal{Z}\times\hat{\mathcal{X}}) : T_Q \subseteq T_{[P\times W_p]_{2\delta_n}}} F_n(\theta_{X|Z}, Q, d_\theta),$$

the right side being further lower bounded by $F_n(\theta_{X|Z}, P \times W_P, d_\theta) - \varepsilon$ for all sufficiently large n (dependent on ε, yet not on P nor on θ), a consequence of the uniform equicontinuity of the family of functions F_n, which is a straightforward consequence of the definition of F_n (the proof of this fact is similar to, e.g., that of [CK81, Lemma 5.5]). This completes the justification of the inequality in (39). The inequality before (39) is self-evident and the first inequality in the above chain follows from (20). To sum up, for the arbitrary $\varepsilon > 0$ and n sufficiently large we have constructed a block-code of rate $\leq R + \varepsilon$ (recall (38), assuming n is sufficiently large so that $\varepsilon_n + \frac{|\mathcal{Z}|}{n}\log(n+1) \leq \varepsilon$) with $\frac{1}{n}\log P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right) \geq I_\theta - 4\varepsilon$ for all $\theta \in \Theta$. □

Remark1: The uniform nature (i.e., the fact that n and ε do not depend on $\theta \in \Theta$) of the achievability notion, as introduced in Definition 1, may, at first glance, seem somewhat harsh. A seemingly less harsh alternative is to require, instead of (6), that

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log P_\theta\left(\rho(X^n,\hat{X}^n(Z^n))>d_\theta\right)\right]\geq I_\theta \quad \forall\theta\in\Theta. \tag{41}$$

In our finite-alphabet framework, however, this is inconsequential. The proof of the converse part is readily seen to carry over to the achievability notion corresponding to (41).

Remark2: It is easy to see how the above proof (and formulation of the theorem) can be extended to accommodate source-dependent distortion measures. Essentially, the proof carries over by simply subscripting θ to ρ everywhere. There would only remain some technical issues such as showing that the convergence of $F_n(W,Q,d)$ to $F(W,Q,d)$ is uniform not only in (W,Q,d) but also in the distortion measure on which these functions implicitly depend. Such issues would be straightforwardly accounted for under appropriate regularity assumptions on the family of distortion measures $\{\rho_\theta\}_{\theta\in\Theta}$.

D A Note on Terminology

Let $\{W_P\}_{P\in\mathcal{M}(\mathcal{Z})}$ be a set of channels in $\mathcal{C}(\mathcal{Z}\to\hat{\mathcal{X}})$, indexed by the elements of $\mathcal{M}(\mathcal{Z})$. Given this set, we define a sequence of schemes as was done in the above proof of the direct part. Specifically, for each n we let $\hat{X}^n(\cdot)$ be the block-code defined as follows: For each $P\in\mathcal{M}_n(\mathcal{Z})$, letting f_P^n be a mapping satisfying the assertion of Proposition 3 for the channel W_P, we put $\hat{X}^n(z^n)=f^n_{P_{z^n}}(z^n)$ $\forall z^n\in\mathcal{Z}^n$. Note that, similarly as shown in the proof above (cf. (38)), the rate of $\hat{X}^n(\cdot)$ is $\leq \max_P I(P;W_P)+\varepsilon_n+\frac{|\mathcal{Z}|}{n}\log(n+1)\leq \max_P I(P;W_P)+\varepsilon$ for any $\varepsilon>0$ and sufficiently large n. In addition, it satisfies $(z^n,\hat{X}^n(z^n))\in T_{[P\times W_p]_{2\delta_n}}$ for all $z^n\in T_P$ and $P\in\mathcal{M}_n(\mathcal{Z})$. Henceforth we shall refer to the sequence $\{\hat{X}^n(\cdot)\}$ constructed in this way as the "sequence of schemes induced by $\{W_P\}_{P\in\mathcal{M}(\mathcal{Z})}$" or, simply, the "scheme induced by $\{W_P\}$". Note that the direct part of Theorem 1 was established by considering the scheme induced by a $\{W_P\}$ satisfying (12) and (13).

4 Known Results as Corollaries

In this section we verify that results pertaining to settings which are special cases of the setting considered in this work are recoverable from Theorem 1.

A Noise-free Source Coding: Marton's Exponent

Consider the noise-free setting of source coding, which is formally obtained as a special case of our setting whence $\mathcal{X}=\mathcal{Z}$, and one considers only sources θ with $P_\theta(X=Z)=1$, or, equivalently, with $\theta_{Z|X}=\delta_{Z|X}$. It is easy to check from the definition of the function F in equation (9) that

$$F(\delta_{Z|X},Q,d)=\begin{cases} 0 & \text{if } E_Q\rho(Z,\hat{X})>d \\ \infty & \text{otherwise.}\end{cases} \tag{42}$$

So, for the single source $\Theta = \{\theta\}$, Hypothesis 1 becomes

$$\forall P \ \exists W : \left\{ \begin{array}{c} D(P\|\theta_Z) + F(\delta_{Z|X}, P \times W, d_\theta) \geq I_\theta \\ \text{and } I(P;W) \leq R \end{array} \right. \tag{43}$$

which, by (42), is equivalent to

$$\{P : E_{P\times W}\rho(X, \hat{X}) > d_\theta \ \forall W \text{ for which } I(P;W) \leq R\} \subseteq \{P : D(P\|\theta_Z) \geq I_\theta\}. \tag{44}$$

Since the set on the left side of (44) is nothing but $\{P : D(P,R) > d_\theta\}$, we finally get that Hypothesis 1 for the single source in the noise-free setting is

$$\min_{\{P:D(P,R)>d_\theta\}} D(P\|\theta_Z) \geq I_\theta. \tag{45}$$

Thus, Theorem 1 tells us that the exponent $\min_{\{P:D(P,R)>d_\theta\}}$ is achievable at rate R for distortion level d_θ, yet Theorem 1 implies that any $I_\theta < \min_{\{P:D(P,R)>d_\theta\}}$ is not achievable, which is essentially [Mar74, Theorem 1].

For the universal case, let now $\Theta = \mathcal{M}(\mathcal{X})$ and, for each $P \in \mathcal{M}(\mathcal{X})$, let W_P denote the channel achieving $D(P,R)$ (assume it exists, otherwise take ε-achiever and the following argumentation will carry over). From (42) and the definition of $D(P,R)$ it follows that

$$\forall P : \quad F(\delta_{Z|X}, P \times W_P, d) = \left\{ \begin{array}{cc} 0 & \text{if } D(P,R) > d \\ \infty & \text{otherwise} \end{array} \right. \tag{46}$$

so, trivially, this choice of $\{W_P\}$ satisfies for any $\{d_\theta\}_{\theta\in\Theta}$

$$\forall P, \theta \in \Theta \quad D(P\|\theta) + F(\delta_{Z|X}, P \times W_P, d_\theta) \geq \min_{P':D(P',R)>d_\theta} D(P'\|\theta), \quad \text{and} \quad I(P;W_P) \leq R, \tag{47}$$

implying that Hypothesis 1 holds for $\Theta = \mathcal{M}(\mathcal{X})$, any $\{d_\theta\}_{\theta\in\Theta}$, $R \geq 0$, and $\{I_\theta = \min_{P':D(P',R)>d_\theta} D(P'\|\theta)\}_{\theta\in\mathcal{M}(\mathcal{X})}$. On the other hand, according to [Mar74] (and as recovered above), the best attainable exponent even for a single known source θ at distortion level d_θ and rate R is $\min_{P':D(P',R)>d_\theta} D(P'\|\theta_Z)$. Thus, Theorem 1 implies that, for the case of noise-free source coding, there exists a source code, for any given rate R, which is universally optimal in the sense of achieving Marton's exponent for all sources. Indeed, as discussed in [WM02b, Section 4], the choice of $\{W_P\}$ satisfying (47) induces the following conceptually simple source code which universally achieves the optimal exponent: Take a type-covering code-book for each type of size e^{nR} code-words, so that all words of type P are covered to within distortion $D(P,R)$. The overall rate of this scheme is essentially R, and it is readily verified to (asymptotically) attain Marton's optimal exponent, uniformly for all sources. Note further that the construction of the scheme is independent of the distortion levels, thus it attains the optimal exponent uniformly for all distortion levels, as well as sources (technically this is evident from the fact that the choice of $\{W_P\}$ is independent of the distortion levels).

B Non-Universal Noisy Source Coding

Let

$$I(\theta, R, d) \stackrel{\triangle}{=} \min_P \left[D(P\|\theta_Z) + \max_{W:I(P;W)\leq R} F(\theta_{X|Z}, P \times W, d) \right]. \tag{48}$$

Suppose further that $\Theta = \{\theta\}$ consists of a single distribution in $\mathcal{M}(\mathcal{X} \times \mathcal{Z})$. For this case, Hypothesis 1 is readily verified to assume the form

$$I(\theta, R, d_\theta) \geq I_\theta. \tag{49}$$

Thus, by Theorem 1, we get that the optimal exponent for noisy source coding of the single source θ, at rate R, for distortion level d_θ, is the left side of (49). More precisely, denoting

$$P_n^{opt}(\theta, R, d) \stackrel{\triangle}{=} \min_{\hat{X}^n(\cdot) \in \mathcal{S}_n(R)} P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d \right),$$

we get

$$I(\theta, R-0, d) \leq \liminf_{n\to\infty} \left[-\frac{1}{n} \log P_n^{opt}(\theta, R, d) \right] \tag{50}$$

$$\leq \limsup_{n\to\infty} \left[-\frac{1}{n} \log P_n^{opt}(\theta, R, d) \right] \leq I(\theta, R+0, d), \tag{51}$$

where (50) follows by the direct part of Theorem 1 and (51) from its converse part. This result was first obtained in [WM02b, Section 3] (by taking $\lambda = \infty$ in Theorem 1 therein). Let us note, for future reference, that at all continuity points of $I(\theta, \cdot, d)$ (which, by monotonicity, is all points except, possibly, a countable set of points) equations (50) and (51) imply

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log P_n^{opt}(\theta, R, d) \right] = I(\theta, R, d). \tag{52}$$

Comment: The proof of Theorem 1, which utilizes the method of types and Proposition 3, is readily seen to imply that the convergence in (52) holds uniformly rapidly in (θ, R, d).

5 Competitive Minimax Approach

Let $\{\lambda_\theta\}_{\theta\in\Theta}$ be a family of non-negative reals indexed by Θ and define the associated *logarithmic competitive maximin* of the class of sources Θ at rate R, for distortion levels $\{d_\theta\}$, by

$$L_n(\Theta, R, \{d_\theta\}, \{\lambda_\theta\}) \stackrel{\triangle}{=} \max_{\hat{X}^n(\cdot) \in \mathcal{S}_n(R)} \min_{\theta\in\Theta} \left[-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta \right) - \lambda_\theta \right]. \tag{53}$$

Note that, by Definition 1, if $\{\lambda_\theta + \alpha\}_{\theta\in\Theta}$ is achievable at rate R for distortion levels $\{d_\theta\}_{\theta\in\Theta}$ then, for any $\varepsilon > 0$, $\liminf_{n\to\infty} L_n(\Theta, R+\varepsilon, \{d_\theta\}, \{\lambda_\theta\}) \geq \alpha$. Conversely, if $\{\lambda_\theta + \alpha\}_{\theta\in\Theta}$ is *not* achievable at rate R for distortion levels $\{d_\theta\}_{\theta\in\Theta}$ then, $\limsup_{n\to\infty} L_n(\Theta, R, \{d_\theta\}, \{\lambda_\theta\}) \leq \alpha$. Thus, assuming continuity of $\liminf_{n\to\infty} L_n(\Theta, R, \{d_\theta\}, \{\lambda_\theta\})$ at[10] R,

$$\begin{aligned} L(\Theta, R, \{d_\theta\}, \{\lambda_\theta\}) &\stackrel{\triangle}{=} \lim_{n\to\infty} L_n(\Theta, R, \{d_\theta\}, \{\lambda_\theta\}) \\ &= \sup\{\alpha : \{\lambda_\theta + \alpha\}_{\theta\in\Theta} \text{ is achievable at rate } R \text{ for distortion levels } \{d_\theta\}_{\theta\in\Theta}\} \\ &= \sup\{\alpha : \{\lambda_\theta + \alpha\}_{\theta\in\Theta},\ \{d_\theta\}_{\theta\in\Theta},\ R \text{ satisfy Hypothesis 1 }\} \qquad (54) \\ &= \max_{\{\{W_P\}: \max_P I(P;W_P) \leq R\}} \min_{\{\theta\in\Theta, P\in\mathcal{M}(\mathcal{Z})\}} \left[D(P\|\theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) - \lambda_\theta \right] \qquad (55) \end{aligned}$$

[10] $\liminf_{n\to\infty} L_n(\Theta, R, \{d_\theta\}, \{\lambda_\theta\})$ is clearly monotonously non-decreasing in R and, consequently, can have no more than a countable number of discontinuity points. Henceforth, for lucidity of the presentation, we shall assume such continuity without explicit mention. When such continuity does not prevail, the analysis carries over with lower and upper bounds using $\pm\varepsilon$ "slack" (cf. analogous analysis in [WM02b]).

where equality (54) follows from Theorem 1, and equality (55) by writing out Hypothesis 1 explicitly. Thus, we have arrived at a single-letter characterization of the asymptotic logarithmic competitive maxi-min attainable relative to a reference set of exponential levels $\{\lambda_\theta\}_{\theta\in\Theta}$, for any rate R and set of distortion levels $\{d_\theta\}$. Note that taking $\{\lambda_\theta\} = \{I(\theta, R, d_\theta)\}$ corresponds to minimizing the worst-case difference between the optimal distribution-dependent exponent and that of the universal scheme, which is the error-exponent analogue of the minimax distortion redundancy formulation of [DW02a] (recall (5)). Indeed, up to sign (here we use maximin rather than minimax),

$$\max_{\{\{W_P\}:\max_P I(P;W_P)\le R\}} \min_{\{\theta\in\Theta, P\in\mathcal{M}(\mathcal{Z})\}} \left[D(P\|\theta_Z) + F(\theta_{X|Z}, P\times W_P, d_\theta) - I(\theta, R, d_\theta)\right]$$

is the analogue of our setting to the single-letter expression obtained in [DW02a] for the minimax distortion redundancy (cf. Theorem 3 therein).

It should also be clear that the (sequence of) scheme(s) induced by the $\{W_P\}$ attaining the max[11] in (55) is asymptotically maximin optimal. The optimal scheme is, of course, dependent on the reference set $\{\lambda_\theta\}_{\theta\in\Theta}$ through which the maximin criterion (53) is defined.

A special case of the above derivation worth mentioning is the absolute error exponent for a source class Θ defined by

$$P_n^\Theta(R, d) \stackrel{\triangle}{=} \min_{\hat{X}^n\in\mathcal{S}_n(R)} \max_{\theta\in\Theta} \Pr\left(\rho_n(X^n, \hat{X}^n(Z^n)) > d\right). \tag{56}$$

The inner maximum on the right side is the analogue of the "maximum probability of error of a code over a compound channel $\mathcal{W}$" of [CK81, Definition 5.9], where here the role analogous to that of the compound channel is played by the compound source Θ. Unlike for the case of the compound channel, where the optimal exponent is unknown, here we can characterize the precise exponential behavior of $P_n^\Theta(R, d)$. Specifically, noting that $-\frac{1}{n}\log P_n^\Theta(R, d) = L_n(\Theta, R, \{d_\theta \equiv d\}, \{\lambda_\theta \equiv 0\})$, (55) gives for this case

$$\lim_{n\to\infty} -\frac{1}{n}\log P_n^\Theta(R, d) \quad = \quad \max_{\{W_P\}:\max_P I(P;W_P)\le R} \left[\min_{P\in\mathcal{M}(\mathcal{Z}),\theta\in\Theta} \left(D(P\|\theta_Z) + F(\theta_{X|Z}, P\times W_P, d)\right)\right].$$

A question arising at this point is whether, given a rate R and distortion levels $\{d_\theta\}$, there exists a set of exponential levels $\{\lambda_\theta\}_{\theta\in\Theta}$ which is, in some sense, most natural. One plausible answer to this question seems to lie in adopting an approach recently advocated for the problem of composite hypothesis testing [FM02]. Specifically, define the *ξ-factored competitive minimax* by

$$K_n^\xi(\Theta, R, \{d_\theta\}) \stackrel{\triangle}{=} \min_{\hat{X}^n(\cdot)\in\mathcal{S}_n(R)} \max_{\theta\in\Theta} \frac{P_\theta\left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta\right)}{[P_n^{opt}(\theta, R, d_\theta)]^\xi}. \tag{57}$$

Hence, for example, $\xi = 0$ corresponds to the absolute minimax criterion, while $\xi = 1$ corresponds to the *competitive minimax* criterion, analogous to that from the hypothesis testing framework of [FM02] (the reader is referred to [FM02] for a more elaborate discussion of this framework and its motivation, which naturally extend to the setting of the present work). One of the significant features of $K_n^\xi(\Theta, R, \{d_\theta\})$ is that if it decays exponentially with n at a certain rate, $\gamma(\xi)$ (note that $\gamma(\xi)$ may be negative, in which case it is really exponential *growth*), then an error

[11] Here and throughout, when there does not exist a $\{W_P\}$ attaining the max (so that it is really a sup), "$\{W_P\}$ attaining the max" should be understood as "a sequence of sets $\{W_P^{(n)}\}$ asymptotically attaining the sup".

exponent of at least $\gamma(\xi) + \xi I(\theta, R, d_\theta)$ is achieved for all $\theta \in \Theta$. In particular, if $\gamma(\xi^*) \geq 0$ for some $\xi^* > 0$, then an error exponent of at least $\xi^* I(\theta, R, d_\theta)$ is achieved for all $\theta \in \Theta$, implying, in turn, a positive exponent whenever the optimal distribution-dependent scheme has a positive exponent. This thus motivates looking for the largest ξ with this property. Specifically, we let

$$\xi^* \stackrel{\triangle}{=} \sup \left\{ \xi : \liminf_{n \to \infty} -\frac{1}{n} \log K_n^\xi(\Theta, R, \{d_\theta\}) \geq 0 \right\}. \tag{58}$$

We can now use our results to obtain a closed-form expression for ξ^* as follows:

$$\begin{aligned} -\frac{1}{n} \log K_n^\xi(\Theta, R, \{d_\theta\}) \quad &= \quad \max_{\hat{X}^n(\cdot) \in \mathcal{S}_n(R)} \min_{\theta \in \Theta} \left[-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta \right) + \xi \cdot \frac{1}{n} \log P_n^{opt}(\theta, R, d_\theta) \right] \\ &\stackrel{n \to \infty}{\longrightarrow} \quad L(\Theta, R, \{d_\theta\}, \{\xi \cdot I(\theta, R, d_\theta)\}), \end{aligned} \tag{59}$$

the second line owing to the definition of $L(\Theta, R, \{d_\theta\}, \cdot)$, to equation (52) (including the comment following it), and an assumption that the limit in (59) exists. Hence, by (55),

$$\begin{aligned} &\lim_{n \to \infty} \left[-\frac{1}{n} \log K_n^\xi(\Theta, R, \{d_\theta\}) \right] \\ &= \max_{\{\{W_P\} : \max_P I(P; W_P) \leq R\}} \min_{\{\theta \in \Theta, P \in \mathcal{M}(\mathcal{Z})\}} \left[D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) - \xi \cdot I(\theta, R, d_\theta) \right] \end{aligned} \tag{60}$$

and, consequently,

$$\xi^* = \sup \left\{ \xi : \max_{\{\{W_P\} : \max_P I(P; W_P) \leq R\}} \min_{\{\theta \in \Theta, P \in \mathcal{M}(\mathcal{Z})\}} \left[D(P \| \theta_Z) + F(\theta_{X|Z}, P \times W_P, d_\theta) - \xi \cdot I(\theta, R, d_\theta) \right] \geq 0 \right\}. \tag{61}$$

Note that $\xi^* \leq 1$ with equality if and only if the family Θ is $(R, \{d_\theta\})$-"universally encodable in the exponential sense", i.e., there exists a rate-R scheme attaining exponential levels $I(\theta, R, d_\theta)$ uniformly for all $\theta \in \Theta$.

We have thus arrived at what is, in the sense described above, a "canonical" choice for the family of $(R, \{d_\theta\})$-achievable exponential levels, namely, the family $\{\xi^* \cdot I(\theta, R, d_\theta)\}$ (with ξ^* explicitly identified in (61)). As for a choice of the family of distortion levels $\{d_\theta\}$, one seemingly plausible possibility is to take, for some $\eta > 0$, the family $\{D(\theta, R) + \eta\}$. Note that for a general class Θ, $\{D(\theta, R) + \eta\}$ may not be achievable with a set of exponential levels which is everywhere positive, even for arbitrarily small $\eta > 0$. The reason is that, in general, Θ will not even be universally encodable in the expectation sense (recall discussion in Section 1 and cf. [DW02a] for an extensive treatment) and therefore, a fortiori, positive exponential levels will not be attainable simultaneously for all sources in Θ. Nonetheless, the family Θ being finite and "identifiable" in the sense that the noisy marginal θ_Z uniquely determines $\theta \in \Theta$, was shown in [DW02a] to be a sufficient condition for the universal encodability of Θ in the expectation sense. As we show in Section 7, this condition also suffices to guarantee the $(R, \{D(\theta, R) + \eta\})$-"universal encodability in the exponential sense" of Θ, for sufficiently small $\eta > 0$. In any case, it seems that a sensible choice of $\{d_\theta\}$ should have the associated ξ^* be positive, guaranteeing a positive exponent whenever the distribution-dependent exponent is positive.

6 A Neyman-Pearson Type Setting

Consider the following problem: There are two possible signals (or sources) and the goal of an observer of a noise-corrupted version of the true underlying signal is to reproduce the underlying clean signal with as high a fidelity as possible, while operating at a limited information rate R. Clearly, in general, for a given rate R, there is going to be a tradeoff between the probability with which a reconstruction at a given fidelity level can be guaranteed under one source, and the corresponding probability (for a possibly different distortion level) associated with the other source (similarly to the analogous tradeoff in the hypothesis testing problem, cf. [ZG91, LM02] and references therein). Our goal in this section is to characterize this tradeoff. The specialization of the setting considered here for the case where there is no rate constraint (pure denoising) seems particularly well-suited to applications where estimation has to be performed under uncertainty of signal presence. These problems are referred to in various parts of the literature as *simultaneous* or *joint detection and estimation*, cf. [BI95] and the references therein.

Characterization of the tradeoff for our problem, as we show below, follows by an easy specialization of the generic result of Theorem 1. We start with the simple-vs-simple case in subsection A, moving on to the composite-vs-composite case in subsection B, where we combine ideas from the previous section as well.

A Simple vs. Simple

Consider the hypothesis testing setting where there are two possible noisy sources, but the goal of an observer, rather than simply determining which of the two is active, is to come up with a reconstruction of the underlying noise-free active source, while operating at a limited information rate R. More formally, the problem can be stated as follows: Let $\theta^{(1)}, \theta^{(2)}$ be two noisy sources. Suppose $\lambda, d_1, d_2, R \geq 0$ are given and consider the following problem:

$$\text{Minimize} \quad P_{\theta^{(2)}}\left(\rho(X^n, \hat{X}^n(Z^n)) > d_2\right) \tag{62}$$

over $\hat{X}^n(\cdot) \in \mathcal{S}_n(R)$ subject to

$$P_{\theta^{(1)}}\left(\rho(X^n, \hat{X}^n(Z^n)) > d_1\right) \leq e^{-n\lambda}. \tag{63}$$

Let $G_n(d_1, d_2, \lambda, R)$ denote the value of the minimum in (62). Clearly, $\liminf -\frac{1}{n}\log G_n(d_1, d_2, \lambda, R) \geq \alpha$ if $\{\lambda, \alpha\}$ is achievable at rate R for distortion levels $\{d_1, d_2\}$ (the family of sources here being $\Theta = \{\theta^{(1)}, \theta^{(2)}\}$). On the other hand, if $\{\lambda, \alpha\}$ is *not* achievable at rate R for distortion levels $\{d_1, d_2\}$ then $\limsup -\frac{1}{n}\log G_n(d_1, d_2, \lambda, R) \leq \alpha$. Hence, assuming existence of limit (justified, again, for all values of R except, possibly, a countable set of values)

$$\begin{aligned}
& A(d_1, d_2, \lambda, R) \\
\stackrel{\triangle}{=} \; & \lim_{n\to\infty} -\frac{1}{n}\log G_n(d_1, d_2, \lambda, R) \\
= \; & \sup\{\alpha : \{\lambda, \alpha\} \text{ is achievable at rate } R \text{ for distortion levels } \{d_1, d_2\}\ \} \\
= \; & \sup\{\alpha : \text{ Hypothesis 1 holds for } \{\lambda, \alpha\},\ \{d_1, d_2\} \text{ at rate } R\} \qquad (64) \\
= \; & \max_{\left\{\begin{array}{c} \{W_P\} : \max_{P\in\mathcal{M}(\mathcal{Z})} I(P; W_P) \leq R, \\ \min_{P\in\mathcal{M}(\mathcal{Z})}\left[D(P\|\theta_Z^{(1)}) + F(\theta_{X|Z}^{(1)}, P\times W_P, d_1)\right] \geq \lambda \end{array}\right\}} \min_{P\in\mathcal{M}(\mathcal{Z})}\left[D(P\|\theta_Z^{(2)}) + F(\theta_{X|Z}^{(2)}, P\times W_P, d_2)\right],
\end{aligned}$$

where equality (64) follows from Theorem 1, and the last equality by writing out Hypothesis 1 explicitly. Note that the (sequence of) scheme(s) induced by the $\{W_P\}$ attaining the max in the last line is the (asymptotically) optimal one. Note also the particularly simple characterization obtained for the pure denoising problem with no rate constraint: The optimal denoising exponent achievable under $\theta^{(2)}$ for distortion level d_2, subject to a requirement that the "first-kind" exponent for distortion level d_1 be at least λ is given by

$$\max_{\{W_P\}:\min_{P\in\mathcal{M}(\mathcal{Z})}\left[D(P\|\theta_Z^{(1)})+F(\theta_{X|Z}^{(1)},P\times W_P,d_1)\right]\geq\lambda}\ \min_{P\in\mathcal{M}(\mathcal{Z})}\left[D(P\|\theta_Z^{(2)})+F(\theta_{X|Z}^{(2)},P\times W_P,d_2)\right]. \tag{65}$$

Appendix A contains a more explicit evaluation of $A(d_1,d_2,\lambda,R)$ for a binary setting in which under one source (hypothesis) the observation is pure noise, while under the other, the observation is a noise-free signal.

B Composite vs. Composite

Let Θ_1,Θ_2 be given subsets of $\mathcal{M}(\mathcal{X}\times\mathcal{Z})$. Let further $\{\lambda_{\theta_1}^{(1)}\}_{\theta_1\in\Theta_1}$, $\{\lambda_{\theta_2}^{(2)}\}_{\theta_2\in\Theta_2}$, $\{d_{\theta_1}^{(1)}\}_{\theta_1\in\Theta_1}$, $\{d_{\theta_2}^{(2)}\}_{\theta_2\in\Theta_2}$ be given families of non-negative reals and consider the following analogue of the composite hypothesis testing problem:

$$\text{Minimize}\quad\max_{\theta_2\in\Theta_2}\frac{P_{\theta_2}\left(\rho(X^n,\hat{X}^n(Z^n))>d_{\theta_2}^{(2)}\right)}{e^{-n\lambda_{\theta_2}^{(2)}}} \tag{66}$$

over $\hat{X}^n(\cdot)\in\mathcal{S}_n(R)$ subject to

$$P_{\theta_1}\left(\rho(X^n,\hat{X}^n(Z^n))>d_{\theta_1}^{(1)}\right)\leq e^{-n\lambda_{\theta_1}^{(1)}},\quad\forall\theta_1\in\Theta_1. \tag{67}$$

Note that this is the most general possible formulation, as we are allowing the constraint on the error of the "first kind" in (67) to be distribution-dependent (i.e., dependent on $\theta_1\in\Theta_1$) and we seek to minimize a "competitive" version of the error of the "second kind" (i.e., performance under P_{θ_2} is weighted by $e^{-n\lambda_{\theta_2}^{(2)}}$). As far as the formulation involving the latter goes, its relevance for this setting, along with guidelines for plausible choices of $\{\lambda_{\theta_2}^{(2)}\}$ and $\{d_{\theta_2}^{(2)}\}$, derive from argumentation similar to that in Section 5 (which, in turn, was based on the competitive minimax approach of [FM02]). As for the merit in allowing the constraint on the exponential rate in (67) to be θ_1-dependent, this was a central theme in the recent work [LM02], where such an approach was proposed and motivated for the composite hypothesis testing problem. The motivation for this approach, as well as guidelines for the choice of a plausible $\{\lambda_{\theta_1}^{(1)}\}$, carry over from [LM02] to our setting (and will, therefore, not be elaborated on here). The bottom line is that the constraint on the exponent of the first kind in (67) should be allowed to depend on $\theta_1\in\Theta_1$, for reasons similar to those discussed in the context of the setting of Section 5. It should be noted, however, that there is an essential difference between the hypothesis testing problem of [LM02] and the one considered here: while in the former there existed a scheme complying with the constraint on the error of the first kind which was uniformly optimal for all $\theta_2\in\Theta_2$, in our setting this of course will, in general, no longer be the case, which is why the minimax criterion in (66) arises naturally. We now turn to an explicit (single-letter) characterization of the problem in (66) - (67), and the optimal scheme for it.

Letting $J_n\left(\{d_{\theta_1}^{(1)}\},\{d_{\theta_2}^{(2)}\},\{\lambda_{\theta_1}^{(1)}\},\{\lambda_{\theta_2}^{(2)}\},R\right)$ denote the value of the minimum in (66), it is clear that

$$\liminf-\frac{1}{n}\log J_n\left(\{d_{\theta_1}^{(1)}\},\{d_{\theta_2}^{(2)}\},\{\lambda_{\theta_1}^{(1)}\},\{\lambda_{\theta_2}^{(2)}\},R\right)\geq\alpha$$

if $\{\lambda_{\theta_1}^{(1)}\}\cup\{\lambda_{\theta_2}^{(2)}+\alpha\}$ is achievable at rate R for distortion levels $\{d_{\theta_1}^{(1)}\}\cup\{d_{\theta_2}^{(2)}\}$, the family of sources here being $\Theta=\Theta_1\cup\Theta_2$. Conversely, when these exponential levels are not achievable, $\limsup-\frac{1}{n}\log J_n\left(\{d_{\theta_1}^{(1)}\},\{d_{\theta_2}^{(2)}\},\{\lambda_{\theta_1}^{(1)}\},\{\lambda_{\theta_2}^{(2)}\},R\right)\leq$

α. Hence,

$$
\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log J_n\left(\{d^{(1)}_{\theta_1}\},\{d^{(2)}_{\theta_2}\},\{\lambda^{(1)}_{\theta_1}\},\{\lambda^{(2)}_{\theta_2}\},R\right)\\
= \;& \sup\left\{\alpha:\ \{\lambda^{(1)}_{\theta_1}\}\cup\{\lambda^{(2)}_{\theta_2}+\alpha\} \text{ is achievable at rate } R \text{ for distortion levels } \{d^{(1)}_{\theta_1}\}\cup\{d^{(2)}_{\theta_2}\}\ \right\}\\
= \;& \sup\left\{\alpha:\ \text{Hypothesis 1 holds for } \{\lambda^{(1)}_{\theta_1}\}\cup\{\lambda^{(2)}_{\theta_2}+\alpha\},\ \{d^{(1)}_{\theta_1}\}\cup\{d^{(2)}_{\theta_2}\} \text{ at rate } R\right\} \qquad (68)\\
= \;& \max \min_{\left\{\begin{array}{c} P\in\mathcal{M}(\mathcal{Z}),\\ \theta^{(2)}\in\Theta_2\end{array}\right\}} \left[D(P\|\theta^{(2)}_Z)+F(\theta^{(2)}_{X|Z},P\times W_P,d_2)-\lambda^{(2)}_{\theta^{(2)}}\right], \qquad (69)
\end{aligned}
$$

where (68) follows from Theorem 1 and the maximum in (69) is taken over $\{W_P\}$ satisfying

$$
\max_{P\in\mathcal{M}(\mathcal{Z})} I(P;W_P)\le R, \qquad \min_{\left\{\begin{array}{c} P\in\mathcal{M}(\mathcal{Z}),\\ \theta^{(1)}\in\Theta_1\end{array}\right\}} \left[D(P\|\theta^{(1)}_Z)+F(\theta^{(1)}_{X|Z},P\times W_P,d_1)-\lambda^{(1)}_{\theta^{(1)}}\right]\ge 0.
$$

As in previous derivations, the optimal scheme is that induced by the $\{W_P\}$ attaining the maximum in (69).

7 Finite Θ with Distinct Noisy Marginals is Universally Encodable in the Error-Exponent Sense

Let Θ be a finite set such that for all $\theta,\theta'\in\Theta$, $\theta_Z=\theta'_Z$ implies $\theta=\theta'$. In other words, each distribution in Θ is completely determined by its noisy marginal. The results in [DW02a] can easily be shown to imply[12] that such a Θ is universally encodable in the expectation sense, i.e., for any rate R there exists a sequence of schemes with expected distortions asymptotically attaining $D(\theta,R)$, uniformly for all sources $\theta\in\Theta$. This can be qualitatively understood through the fact that by observing the noisy observation sequence for sufficiently long, the Z-marginal of the source can be acquired arbitrarily precisely and reliably, thereby enabling identification of the active source. In this section we establish the stronger fact that such a Θ is universally encodable in the exponential sense, i.e., that there exists a sequence of schemes attaining the optimal (distribution dependent) exponent, $I(\theta,R,d_\theta)$, for all sources in Θ, when the family of distortion levels $\{d_\theta\}=\{D(\theta,R)+\eta\}$ and $\eta>0$ is sufficiently small. Intuitively, the explanation for why this is true is the following: Since there is only a finite number of distinct noisy marginals, there exists an $r>0$ such that the Kullback-Leibler "ball" of radius r around each noisy marginal does not contain any other of the noisy marginals. Thus, considering the conceptually simple scheme which looks at the empirical type of the noisy observation and operates optimally for the source whose noisy marginal is closest to the observed type under Kullback-Leibler distance, it is not hard to see that an exponent associated with the probability for correct identification of the active source of at least $r>0$ is attainable uniformly for all sources in the class. Thus, when the exponent associated with the optimal distribution-dependent scheme for each source is less than r, the identifiability issue is not the bottleneck and universally exponentially optimal performance is attainable. We make this precise in what follows.

[12] This fact can be straightforwardly established based on first principles, without relying on the results of [DW02a].

For $M = |\Theta|$, without loss of generality, let $\Theta = \{\theta^{(j)}\}_{j=1}^{M}$. Define $i : \mathcal{M}(\mathcal{Z}) \to \{1, \ldots, M\}$ by

$$i(P) = \arg \min_{1 \le i \le M} D(P \| \theta_Z^{(i)}), \tag{70}$$

resolving ties arbitrarily. In words, $i(P)$ gives the index of the P-nearest neighbor in $\{\theta_Z^{(j)}\}_{j=1}^{M}$ under Kullback-Leibler distance . Let further, for $1 \le j \le M$,

$$r_j(\Theta) = \min_{P \in \mathcal{M}(\mathcal{Z}) : i(P) \neq j} D(P \| \theta_Z^{(j)}) \tag{71}$$

be the radius of the largest "divergence - ball" around $\theta_Z^{(j)}$ containing solely points P with $i(P) = j$. Finally, define $r(\Theta) = \min_{1 \le j \le M} r_j(\Theta)$ and note that $r(\Theta) > 0$ since, by our assumption on Θ, the members of $\{\theta_Z^{(j)}\}_{j=1}^{M}$ are unique (and $M < \infty$). Note that $r(\Theta)$ can be thought of as a "packing radius" of the sources $\{\theta_Z^{(j)}\}$ in the $\mathcal{M}(\mathcal{Z})$ simplex under Kullback-Leibler distance. In the following corollary to Theorem 1, $\{d_\theta\}_{\theta \in \Theta}$ is assumed to be any set of non-negative distortion levels and $I(\theta, R, d_\theta)$ is the optimal error exponent of the source θ, as defined in (48).

Corollary 4 *The exponential levels $\{\min\{I(\theta, R, d_\theta), r(\Theta)\}\}_{\theta \in \Theta}$ are achievable at rate R for distortion levels $\{d_\theta\}_{\theta \in \Theta}$. Specifically, for any $\varepsilon > 0$ there exists an n and a block code of length n and rate $\le R + \varepsilon$, $\hat{X}^n(\cdot)$, satisfying*

$$-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > d_\theta \right) \ge \min\{I(\theta, R, d_\theta), r(\Theta)\} - \varepsilon \quad \forall \theta \in \Theta. \tag{72}$$

Proof of Corollary 4 will be shortly given below. Note that, in particular, Corollary 4 implies that a positive exponent is universally achievable for all sources in Θ whenever the optimal source-dependent exponent, $I(\theta, R, d_\theta)$, is positive for all $\theta \in \Theta$. Thus, assuming $I(\theta, R, \cdot)$ is continuous at $D(\theta, R)$, we know that[13] $I(\theta, R, D(\theta, R)) = 0$ and, therefore, $I(\theta, R, D(\theta, R) + \eta) \le r(\Theta)$ for all $\theta \in \Theta$ when $\eta > 0$ is sufficiently small. We thus have the following corollary to Corollary 4.

Corollary 5 *The family Θ is $(R, \{D(\theta, R) + \eta\})$-universally encodable for $\eta > 0$ sufficiently small. More explicitly, for any $\varepsilon > 0$ there exists an n and a block code of length n and rate $\le R + \varepsilon$, $\hat{X}^n(\cdot)$, satisfying*

$$-\frac{1}{n} \log P_\theta \left(\rho(X^n, \hat{X}^n(Z^n)) > D(\theta, R) + \eta \right) \ge I(\theta, R, D(\theta, R) + \eta) - \varepsilon \quad \forall \theta \in \Theta. \tag{73}$$

Proof of Corollary 4: Fix $\varepsilon > 0$ and, for $1 \le j \le M$, let $\{W_P^{(j)}\}$ denote an ε-achiever of $I(\theta^{(j)}, R, d_{\theta^{(i)}})$, i.e.,

$$\min_P \left[D(P \| \theta_Z^{(j)}) + F(\theta_{X|Z}^{(j)}, P \times W_P^{(j)}, d_{\theta^{(j)}}) \right] \ge I(\theta^{(j)}, R, d_{\theta^{(j)}}) - \varepsilon, \quad \text{and} \quad \max_P I(P; W_P^{(j)}) \le R. \tag{74}$$

Note that this is possible by the definition (cf. equation (48)) of $I(\theta, R, d)$. Construct now $\{W_P^*\}$ by letting, for each P, $W_P^* = W_P^{(i(P))}$. Thus, clearly, $\max_P I(P; W_P^*) \le R$ and, for each $1 \le j \le M$,

$$\begin{aligned} & \min_P \left[D(P \| \theta_Z^{(j)}) + F(\theta_{X|Z}^{(j)}, P \times W_P^*, d_\theta) \right] \\ = \quad & \min \left\{ \min_{P : i(P) = j} \left[D(P \| \theta_Z^{(j)}) + F(\theta_{X|Z}^{(j)}, P \times W_P^*, d_\theta) \right], \min_{P : i(P) \neq j} \left[D(P \| \theta_Z^{(j)}) + F(\theta_{X|Z}^{(j)}, P \times W_P^*, d_\theta) \right] \right\} \end{aligned}$$

[13] To see this from an operational consideration, note that it follows from the converse to noisy source coding that $I(\theta, R, d_\theta) = 0$ for $d_\theta < D(\theta, R)$. This fact was also verified technically through the explicit form of $I(\theta, R, d_\theta)$ in [WM02b, Subsection 3.D.2].

$$
\begin{aligned}
&\geq \min\left\{\min_{P:i(P)=j}\left[D(P\|\theta_Z^{(j)})+F(\theta_{X|Z}^{(j)},P\times W_P^{(j)},d_\theta)\right],\min_{P:i(P)\neq j}D(P\|\theta_Z^{(j)})\right\}\\
&\geq \min\left\{\min_{P}\left[D(P\|\theta_Z^{(j)})+F(\theta_{X|Z}^{(j)},P\times W_P^{(j)},d_\theta)\right],r_j(\Theta)\right\}\\
&\geq \min\left\{I(\theta^{(j)},R,d_{\theta^{(j)}}),r_j(\Theta)\right\}-\varepsilon\\
&\geq \min\left\{I(\theta^{(j)},R,d_{\theta^{(j)}}),r(\Theta)\right\}-\varepsilon, \qquad (75)
\end{aligned}
$$

the inequality before last owing to (74). It follows that for arbitrary $\varepsilon>0$, $\{\min\{I(\theta,R,d_\theta),r(\Theta)\}-\varepsilon\}_{\theta\in\Theta}$, R, $\{d_\theta\}_{\theta\in\Theta}$ jointly satisfy Hypothesis 1 which, by the direct part of Theorem 1, implies the existence of a block code of rate arbitrarily close to R satisfying (72). □

Note that in this short proof, we have directly applied Theorem 1, avoiding the need for an explicit construction of the "two-part", plug-in, scheme described above[14].

8 Summary, Conclusions, and a related Future Direction

In this work we have characterized the "achievable region" for error exponents in universal noisy source coding. Our principle result, Theorem 1, gives a "single-letter" necessary and sufficient condition for a family of exponential levels to be achievable at a certain rate, for certain distortion levels, for a given family of sources. It was later illustrated how this principle result can be applied to characterize optimal performance in various noisy source coding settings.

It should be pointed out that the noisy source coding settings considered in the sections following Section 3 are not the only conceivable ones. Many other variations are possible. The treatment of those canonical settings, however, should be enough to convince the reader that any other setting involving exponents for noisy source coding when there is more than one possible source can be similarly dealt with by a suitable application of Theorem 1.

Finally, we remark that an extension of the competitive minimax approach of [FM02] to the question of universally attainable error exponents need not be restricted to the noisy source coding setting considered in this work. For example, it can be applied also to the setting of the compound channel [CK81, Definition 5.9]. Indeed, letting $\mathcal{W}$ denote a family of channels, the traditional quantity of interest (cf. also [LT98]) is the *maximum probability of error* of a channel code represented by the encoder-decoder pair (f,ϕ), defined as $e(\mathcal{W},f,\phi)=\sup_{W\in\mathcal{W}}e(W,f,\phi)$. It would seem that a plausible alternative is to look at the competitive minimax analogue: $\sup_{W\in\mathcal{W}}\frac{e(W,f,\phi)}{e^*(W)}$, $e^*(W)$ denoting the optimal error exponent of the channel W (at the specific rate, assumed given and suppressed in the notation). This would be a less pessimistic formulation, incorporating the approach that the better the active channel is, the better the target performance should be. Variants of the form $\sup_{W\in\mathcal{W}}\frac{e(W,f,\phi)}{[e^*(W)]^\xi}$ may also be possible, with guidelines for the optimal choice of ξ following reasoning similar to that in [FM02] and in the present work. Note that at rate regions and channels for which the channel-dependent optimal error exponents are not entirely known, the $e^*(W)$ in the denominator can be replaced by available lower and upper bounds (according to the point of view). It would be interesting to explore this setting and its effect on the optimal (w.r.t. this new criterion) channel code.

[14] Although this result is probably most naturally understood by considering the conceptually simple plug-in approach described earlier.

Appendix

A Pure Noise vs. Pure Signal

We dedicate this appendix to an illustration of a simple setting to which our results apply and where a semi-explicit formula for the achievable exponents can be obtained. We look at the special instance of the simple "hypothesis testing" setting of Subsection 6.A for the case of "Pure Noise vs. Pure Signal". The framework we propose seems particularly relevant for this case as the issue here is to determine whether the observation contains a signal or not, and to come up with a reconstruction in case a signal is present.

Let $\theta^{(1)}$ denote the all-zero source corrupted by a BSC(δ) ("pure noise"), $\theta^{(2)}$ denote the noise-free Bernoulli(π) source ("pure signal"), and ρ denote Hamming loss. So that

$$\theta^{(1)}_Z = \text{Bernoulli}(\delta), \theta^{(2)}_Z = \text{Bernoulli}(\pi), \quad \theta^{(1)}_{X|Z}(0|i) = 1, \quad \theta^{(2)}_{X|Z}(i|i) = 1, \quad i = 0, 1.$$

For $Q \in \mathcal{M}_+(\mathcal{Z} \times \hat{\mathcal{X}})$ and $i = 0, 1$, it is easy to see that

$$D(V \| \theta^{(i)}_{X|Z} | Q) = \left\{ \begin{array}{ll} 0 & \text{if } V = \theta^{(i)}_{X|Z} \\ \infty & \text{otherwise.} \end{array} \right. \tag{A.1}$$

Hence,

$$F(\theta^{(1)}_{X|Z}, Q, d) = \left\{ \begin{array}{cc} 0 & \text{if } \Pr^Q(\hat{X} = 1) > d \\ \infty & \text{otherwise} \end{array} \right. \tag{A.2}$$

and

$$F(\theta^{(2)}_{X|Z}, Q, d) = \left\{ \begin{array}{cc} 0 & \text{if } \Pr^Q(\hat{X} \neq Z) > d \\ \infty & \text{otherwise.} \end{array} \right. \tag{A.3}$$

Thus, evaluating (64) for this case,

$$\begin{aligned}
& A(d_1, d_2, \lambda, R) \\
= & \max_{\left\{ \begin{array}{c} \{W_P\} : \max_{P \in \mathcal{M}(\mathcal{Z})} I(P; W_P) \leq R, \\ \min_{P \in \mathcal{M}(\mathcal{Z})} \left[D(P \| \theta^{(1)}_Z) + F(\theta^{(1)}_{X|Z}, P \times W_P, d_1) \right] \geq \lambda \end{array} \right\}} \min_{P \in \mathcal{M}(\mathcal{Z})} \left[D(P \| \theta^{(2)}_Z) + F(\theta^{(2)}_{X|Z}, P \times W_P, d_2) \right] \\
= & \max_{\left\{ \begin{array}{c} \{W_P\} : \max_{P \in \mathcal{M}(\mathcal{Z})} I(P; W_P) \leq R, \\ P \times W_P(\hat{X} = 1) \leq d_1 \quad \forall P : D(P \| \theta^{(1)}_Z) < \lambda \end{array} \right\}} \min_{P : P \times W_P(\hat{X} \neq Z) > d_2} D(P \| \theta^{(2)}_Z) \qquad \text{(A.4)} \\
= & \max_{\left\{ \begin{array}{c} \{W_p\}_{0 \leq p \leq 1} : \max_{0 \leq p \leq 1} I(p; W_p) \leq R, \\ (1-p) W_p(1|0) + p W_p(1|1) \leq d_1 \quad \forall p : D(p \| \delta) < \lambda \end{array} \right\}} \min_{p : (1-p) W_p(1|0) + p(1 - W_p(1|1)) > d_2} D(p \| \pi), \qquad \text{(A.5)}
\end{aligned}$$

where $D(p \| \pi) = p \log(p/\pi) + (1-p) \log((1-p)/(1-\pi))$, $I(p; W_p)$ denotes mutual information between a Bernoulli(p) random variable and its output from the channel W_p. Note that each W_p here is characterized by the two parameters $0 \leq W_p(1|0) \leq 1$ and $0 \leq W_p(1|1) \leq 1$, so that for specific numerical values of R, d_1, d_2 the variational problem in (A.5) is easily solved numerically. Moreover, from the optimizing $\{W_p\}$ it is easy to construct the optimal (sequence of) schemes.

Acknowledgement

An inspiring conversation with Y. Steinberg is gratefully acknowledged.

References

[Ber71] T. Berger. *Rate-Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall, Englewood Cliffs, N.J., 1971.

[Ber98] T. Berger. Lossy source coding. *IEEE Trans. Inform. Theory*, 44(6):2693–2723, October 1998.

[BI95] B. Baygün and A. O. Hero III. Optimal simultaneous detection and estimation under a false alarm constraint. *IEEE Trans. Inform. Theory*, 41(3):688–703, May 1995.

[Bla74] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inform. Theory*, IT-20:405–417, 1974.

[Bla76] R. E. Blahut. Information bounds of the Fano-Kullback type. *IEEE Trans. Inform. Theory*, IT-22:410–421, 1976.

[Bla87] R. E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, 1987.

[BRY98] A. Barron, J. Rissanen, and B. Yu. The Minimum Description Length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, October 1998.

[CEG96] P. A. Chou, M. Effros, and R. M. Gray. A vector quantization approach to universal noiseless coding and quantization. *IEEE Trans. Inform. Theory*, IT-42:1109–1138, July 1996.

[CK81] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[Don02] D. Donoho. The Kolmogorov sampler. January 2002. Submitted (available at: `http://www-stat.stanford.edu/ donoho/`).

[DW02a] A. Dembo and T. Weissman. The minimax distortion redundancy in noisy source coding. *IEEE Trans. Inform. Theory*, May 2002. Submitted (available at: `http://www-stat.stanford.edu/ amir/`).

[DW02b] A. Dembo and T. Weissman. The minimax distortion redundancy in noisy source coding. *Int. Symposium on Information Theory*, July 2002.

[EC98] E. Erkip and T. M. Cover. The efficiency of investment information. *IEEE Trans. Inform. Theory*, 44(3):1026–1040, May 1998.

[EG88] Y. Ephraim and R. M. Gray. A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization. *IEEE Trans. Inform. Theory*, 34(4):826–834, July 1988.

[FM02] M. Feder and N. Merhav. Universal composite hypothesis testing: A competitive minimax approach. *IEEE Trans. Inform. Theory*, 48(6):1504–1518, June 2002.

[Gal68] R. G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[Hum81] P. A. Humblet. Generalization of Huffman coding to minimize the probability of buffer overflow. *IEEE Trans. Inform. Theory*, IT-27(2):230–237, March 1981.

[Jel68] F. Jelinek. Buffer overflow in variable length coding of fixed rate sources. *IEEE Trans. Inform. Theory*, IT-14(3):490–501, May 1968.

[Kie93] J. C. Kieffer. A survey of the theory of source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, 39:1473–1490, 1993.

[KN96] A. Kanlis and P. Narayan. Error exponents for succesive refinement by partitioning. *IEEE Trans. Inform. Theory*, 42(1):275–282, January 1996.

[LLZ94] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40:1728–1740, November 1994.

[LM02] E. Levitan and N. Merhav. A competitive Neyman - Pearson approach to universal hypothesis testing with applications. *IEEE Trans. Inform. Theory*, August 2002. To appear (available at: `http://tiger.technion.ac.il/users/merhav/`).

[LT98] A. Lapidoth and E. Telatar. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, IT-44:2148–2177, October 1998.

[Mar74] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-20:197–199, March 1974.

[Mer91] N. Merhav. Universal coding with minimum probability of codeword length overflow. *IEEE Trans. Inform. Theory*, 37:556–563, May 1991.

[MK01] N. Merhav and I. Kontoyiannis. Source coding exponents for zero-delay coding with finite memory. *Submitted to: IEEE Transactions on Information Theory*, October 2001. (available at: `http://tiger.technion.ac.il/users/merhav/`).

[Nat93] B. Natarajan. Filtering random noise via data compression. *Data Compression Conference, DCC '93*, pages 60–69, 1993.

[NGD75] D. L. Neuhoff, R. M. Gray, and L. D. Davisson. Fixed rate universal block source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-21:511–523, September 1975.

[Ris84] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, July 1984.

[Ris96] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42(1):40–47, January 1996.

[Ris00] J. Rissanen. MDL denoising. *IEEE Trans. Inform. Theory*, IT-46:2537–2543, November 2000.

[Sha59] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Conv. Rec.*, 7:142–163, 1959.

[Wit80] H. S. Witsenhausen. Indirect rate distortion problems. *IEEE Trans. Inform. Theory*, IT-26(5):518–521, September 1980.

[WM02a] T. Weissman and N. Merhav. Limited-delay lossy coding and filtering of individual sequences. *IEEE Trans. Inform. Theory*, 48(3):721–733, March 2002.

[WM02b] T. Weissman and N. Merhav. Tradeoffs between the excess-code-length exponent and the excess-distortion exponent in lossy source coding. *IEEE Trans. Inform. Theory*, IT-48(2):396–415, February 2002.

[Wyn74] A. D. Wyner. On the probability of buffer overflow under an arbitrary bounded input-output distribution. *SIAM J. Appl. Math.*, 27:544–570, 1974.

[WZ70] J. K. Wolf and J. Ziv. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Trans. Inform. Theory*, IT-16(4):406–411, July 1970.

[YK96] E. H. Yang and J. Kieffer. Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 42:239–245, 1996.

[ZG91] O. Zeitouni and M. Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inform. Theory*, 37:285–290, March 1991.

[Ziv72] J. Ziv. Coding of sources with unknown statistics - part II: Distortion relative to a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-18:389–394, May 1972.

[Ziv80] J. Ziv. Distortion-rate theory for individual sequences. *IEEE Trans. Inform. Theory*, IT-26(2):137–143, March 1980.