Source Coding With Distortion Side Information

Emin Martinian, Gregory W. Wornell, *Fellow, IEEE*, and Ram Zamir, *Senior Member, IEEE*

***Abstract*—The impact of side information about the distortion measure in problems of quantization is analyzed. It is shown that such "distortion side information" is not only useful in general, but that in many cases knowing it at only the encoder is as good as knowing it at both encoder and decoder, and knowing it at only the decoder is useless. Moreover, it is shown that the strategy of exploiting distortion side information at the encoder by describing it for the decoder is inefficient. Thus, distortion side information is a natural complement to side information about the source signal, as studied by Wyner and Ziv, which if available only at the decoder is often as good as knowing it at both encoder and decoder. When both types of side information are present, conditions are established under which encoder-only distortion side information and decoder-only signal side information are sufficient in the high-resolution limit, and the rate penalty for deviating from this configuration is characterized.**

***Index Terms*—Data compression, distributed source coding, quantization, sensor networks, Wyner–Ziv coding.**

I. Introduction

In settings ranging from sensor networks and communication networks, to distributed control and biological systems, different parts of the system of interest typically have limited, noisy, or incomplete information but must somehow cooperate to achieve some overall functionality.

In such scenarios, it is important to understand a variety of issues. These include: 1) the penalties incurred due to the lack of full, globally shared information; 2) the best ways to encode and combine available information from different sources; and 3) where different kinds of information are most useful in the system. A simple example of such a distributed source coding scenario was introduced by Wyner and Ziv [1], and is illustrated in Fig. 1(a). An encoder observes a signal x^n to be represented digitally for a subsequent decoder having some additional signal side information w^n, which is correlated with x^n. An analysis of the fundamental performance limits for this problem [1]–[5] reveals both that such side information is useful only if available at the decoder, and that in many cases a properly designed system can realize essentially the full benefit of this side information (i.e., as if it were known to both encoder and decoder) even if it is available only at the decoder.

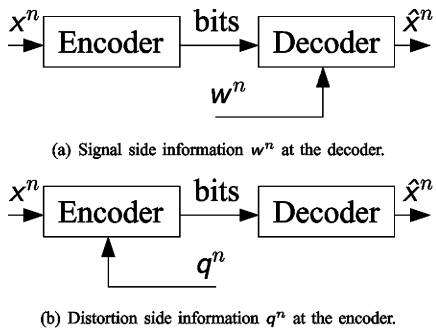


Fig. 1. Compressing a source x^n with side information into a quantized representation $\hat{x}^n$.

In this paper, we introduce and analyze a different scenario, illustrated in Fig. 1(b). As before, the encoder quantizes its observations into a collection of bits, which the decoder uses to reconstruct the observations to some level of fidelity. But now the encoder has some distortion side information q^n describing the relative importance of different components of the observed signal, which enters into our model as a parameter of the distortion measure.

We develop the fundamental performance limits for this problem. Our analysis reveals, in some broad scenarios of interest, both that such side information is useful only if available at the encoder, and that a properly designed system can realize essentially the full benefit of such side information (i.e., as if it were known to both encoder and decoder) even if it is available only at the encoder.[1] As such, distortion side information plays a complementary role to that of signal side information as developed by Wyner and Ziv.

Finally, we show that these kinds of source coding results continue to hold even when both distortion side information q^n and signal side information w^n are jointly considered, under appropriate conditions. Specifically, we establish that a system where only the encoder knows q^n and only the decoder knows w^n can be asymptotically as good as a system with both types of side information known at both the encoder and the decoder. Moreover, we bound the performance gap in the nonasymptotic regime, and also derive the penalty for deviating from the asymptotically sufficient side information configuration.

[1]Clearly, such results can be true only if the side information bears a reasonable statistical relationship to the source, and only if it parameterizes the distortion measure in reasonable ways. Our treatment will make such requirements precise.

Manuscript received December 27, 2004, revised January 16, 2007 and May 20, 2008; current version published September 17, 2008. This work was supported in part by the National Science Foundation under Grant CCF-0515109, the United States–Israel Binational Science Foundation, Hewlett-Packard through the MIT/HP Alliance, and the Draper Laboratory University IR&D Program. The material in this paper was presented in part at the Data Compression Conference, Snowbird, UT, March 2004, and at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

E. Martinian is with the Signals, Information, and Algorithms Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: emin@alum.mit.edu).

G. W. Wornell is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gww@mit.edu).

R. Zamir is with the Department of Electrical Engineering–Systems, Tel-Aviv University, Ramat-Aviv 69978, Israel (e-mail: zamir@eng.tau.ac.il).

Communicated by M. Effros, Associate Editor for Source Coding.

Color versions of Figures 3, 6, and 7 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.928983

In terms of background, an analysis of the value and efficient use of distortion side information available at only the encoder or decoder has received relatively little attention in the information theory and compression communities to date. The rate–distortion function with decoder-only side information, relative to side-information-dependent distortion measures (as an extension of the Wyner–Ziv setting [1]), is given in [4]. And a high-resolution approximation for this rate–distortion function for locally quadratic weighted distortion measures is given in [6]. However, we are not aware of an information-theoretic treatment of encoder-only side information with such distortion measures. In fact, the mistaken notion that encoder-only side information is never useful is common folklore. This may be due to a misunderstanding of Berger's result that side information *that does not affect the distortion measure* is never useful when available only at the encoder [3], [7], a result we will generalize and develop further insight into in this paper.

Before proceeding with our development, it is worth stressing that there are a wide range of applications where distortion side information may be available in some parts of a system but not others. As one example, in a sensor network a node may have information about the reliability of the measurements, which can fluctuate due to calibration or processing. As another example, in audio, image, or video compression systems, the encoder may apply signal analysis to determine which parts of the signal are more or less important (i.e., sensitive to distortion) due to context, masking effects, and other perceptual phenomena [8]. While in practice the conventional approach to exploiting such side information in these kinds of examples involves sharing it with decoders via a side channel, the results of this paper imply that this is both an unnecessary and inefficient use of bandwidth.

An outline of the paper is as follows. Section II summarizes some notation for the paper, and Section III introduces the formal problem model of interest. Section IV then develops the rate–distortion tradeoff for source coding with only distortion side information, and in particular, identifies conditions under which such side information is sufficient at the encoder and useless at the decoder. Section V then extends the problem of interest to include both signal and distortion side information, emphasizing the case of continuous sources in the high-resolution regime. For this scenario, we identify and characterize both ineffective and asymptotically effective partial side information combinations. For ineffective combinations, Section VI then quantifies the penalty incurred by misplaced side information. For asymptotically effective combinations, Section VII develops bounds on the vanishing rate gap relative to the case of complete side information at lower resolutions. Finally, Section VIII contains some concluding remarks.

II. Notation

We use $I(\cdot,\cdot)$, $H(\cdot)$, $h(\cdot)$, $D(\cdot\|\cdot)$, and $J(\cdot)$ to denote mutual information, entropy, differential entropy, information divergence, and Fisher information, respectively. All such quantities are expressed in bits, except for Fisher information, which is defined in terms of the natural logarithm. More generally, $\log(\cdot)$ will refer to base-2 logarithms throughout. In addition, $H_{\mathrm{B}}(p) = -p\log p - (1-p)\log(1-p)$ denotes the binary entropy function, and $d_{\mathrm{H}}(x,\hat{x})$ denotes the Hamming distortion measure, which is 0 if $x=\hat{x}$ and 1 otherwise. We also use $E[\cdot]$ to denote expectation, $|\cdot|$ to denote cardinality for set-valued arguments and absolute value for scalar arguments, and $\|\cdot\|$ to denote the Euclidean norm of its argument.

Finally, sequences are denoted using superscripts and sequence elements with subscripts (e.g., $x^n = (x_1, x_2, \ldots, x_n)$), and random variables are distinguished from sample values by the use of sans-serif fonts for the former (e.g., $\mathsf{x}^n = (\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n)$).

III. Problem Model

Our general rate–distortion problem with side information corresponds to the tuple

$$(\mathcal{X}, \hat{\mathcal{X}}, \mathcal{Z}, p_{\mathsf{x}}(x), p_{\mathsf{z}|\mathsf{x}}(z|x), d(x,\hat{x};z)). \tag{1}$$

Specifically, a source sequence x^n consists of the n samples drawn from the alphabet $\mathcal{X}$, and the side information z^n likewise consists of n samples drawn from the alphabet $\mathcal{Z}$. These random variables are drawn according to the memoryless distribution

$$p_{\mathsf{x}^n,\mathsf{z}^n}(x^n, z^n) = \prod_{i=1}^{n} p_{\mathsf{x}}(x_i) p_{\mathsf{z}|\mathsf{x}}(z_i|x_i). \tag{2}$$

A rate R encoder f_n maps the source x^n as well as possible encoder side information $\phi_n^f(\mathsf{z}^n)$ to an index $\mathsf{m} \in \{1, 2, \ldots, 2^{nR}\}$. The corresponding decoder g_n maps the resulting index m as well as possible decoder side information $\phi_n^g(\mathsf{z}^n)$ to a reconstruction $\hat{\mathsf{x}}^n$ of the source that takes values in the alphabet $\hat{\mathcal{X}}$, which we will generally take to be the same as $\mathcal{X}$. Distortion in a reconstruction $\hat{x}^n$ of a source x^n is measured via the additive measure

$$d_n(x^n, \hat{x}^n; z^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i; z_i), \tag{3}$$

where we explicitly denote the dependence of the distortion measure on the side information. As usual, the rate–distortion function at a particular distortion D is the infimum of rates such that there exists a system where the distortion is less than D with probability approaching 1 as $n \to \infty$.

Of particular interest in this paper is the case in which the side information z can be decomposed into two kinds of side information, which we term "signal side information" w and "distortion side information" q, i.e., $\mathsf{z} = (\mathsf{w}, \mathsf{q})$. The former, whose elements take values in an alphabet $\mathcal{W}$, corresponds to information that is statistically related to the source but does not directly affect the distortion measure, while the latter, whose elements take values in an alphabet $\mathcal{Q}$, corresponds to information that does not have a direct statistical relationship to the source but does directly affect the distortion measure. Formally, we capture this decomposition via the following definition.

Definition 1: A decomposition $\mathsf{z} = (\mathsf{w}, \mathsf{q})$ of side information z into signal side information w and distortion side information q for a rate–distortion problem with source x and additive distortion measure(3) is *admissible* if the following Markov chains are satisfied:

$$\mathsf{q} \leftrightarrow \mathsf{w} \leftrightarrow \mathsf{x} \tag{4a}$$

and

$$d_n(\mathsf{x}^n, \hat{\mathsf{x}}^n; \mathsf{z}^n) \leftrightarrow (\mathsf{x}^n, \hat{\mathsf{x}}^n, \mathsf{q}^n) \leftrightarrow \mathsf{w}^n. \tag{4b}$$

Several remarks are worthwhile before proceeding with our development. First, note that (4a) is equivalent to the condition

$$p_{z|x}(z|x) = p_{w|x}(w|x)p_{q|w}(q|w) \tag{5}$$

for all $x \in \mathcal{X}$ and $z = (w, q) \in \mathcal{W} \times \mathcal{Q}$. Moreover, when (4b) holds, we can (and will), with slight abuse of notation, use $d(x, \hat{x}; q)$ in place of $d(x, \hat{x}; z)$. We restrict our attention to deterministic such functions d.

Second, Definition 1 allows much flexibility in decomposing some side information of interest into signal and distortion components. Indeed, such decompositions always exist—one can always simply let $q = w = z$. Nevertheless, we will see that *any* such decomposition effectively decomposes the side information into a component that is of value at the encoder, and a component that is of value at the decoder.

Third, when separating phenomena that have physically different origins, such decompositions arise quite naturally. Moreover, in such cases, the resulting signal side information w and distortion side information q are often statistically independent, in which case additional results can be obtained on the relative value of different side information availability configurations. Hence, in our treatment we will often impose this further restriction on the side information requirements of Definition 1, which corresponds to a situation in which q and x are independent not just conditioned on w as per (4a), but unconditionally as well:

$$p_{x,q}(x, q) = \sum_{w \in \mathcal{W}} p_{q|w}(q, w)p_{w|x}(w|x)p_x(x) = p_q(q)p_x(x).$$

Moreover, in this case q is also independent of (w, x), i.e.,

$$p_{q|w,x}(q|w, x) = p_{q|w}(q|w) = p_q(q).$$

However, it should be emphasized that admissible decompositions satisfying this further restriction are not always possible, and, later in the paper, for such cases we assess the penalties incurred by the lack of a suitable decomposition.

It is also worth emphasizing that a subclass of side information scenarios with q and w independent corresponds to the case in which signal side information is altogether absent ($w = \varnothing$), so q and x are independent in this case too. This case will also be of special interest in parts of the paper.[2]

Finally, to obtain many of our results, we further restrict the form of the distortion measure $d(x, \hat{x}; q)$, a simple example of which is the modulated quadratic distortion $d(x, \hat{x}; q) = q(x - \hat{x})^2$ for $x, \hat{x} \in \mathbb{R}$. In general, each of our theorems will make clear any particular restrictions on the form of the distortion measure that apply.

In the remainder of the paper, we consider the 16 possible scenarios depicted in Fig. 2, corresponding to where each of q^n and w^n is available. In our notation for the associated rate–distortion functions, the subscripts DEC, ENC, BOTH, and NONE indicate that the associated form of side information is available at the decoder, the encoder, both terminals, or neither terminal, respectively. For example, $R_{[\mathrm{W:DEC}]}(D)$ denotes the Wyner–Ziv rate–distortion function where w^n is available at the decoder [1], corresponding to the scenario depicted in Fig. 1(a). Similarly, when all information is available at both encoder and decoder, $R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$ describes Csiszár and Körner's [4] generalization of Gray's [9] conditional rate–distortion function $R_{[\mathrm{W:BOTH}]}(D)$ to the case where the side information can affect the distortion measure. Finally, $R_{[\mathrm{Q:ENC}]}(D)$ denotes the rate–distortion function corresponding to the scenario of Fig. 1(b).

[2]By contrast, when w exists but is unobserved, q and x can be dependent, as will also arise at times in our development.

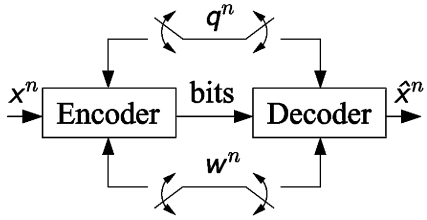


Fig. 2. Scenarios for source coding with distortion side information q^n and signal side information w^n. The four switches may each be open or closed, corresponding to whether the associated side information is available or not at the associated terminal.

We also emphasize that some side information being available at neither terminal (the NONE case) is not in general equivalent to there being no such side information in the problem. Indeed, an unobserved form of side information can affect the problem through the other form of side information if the latter is observed and if the two are statistically dependent. Hence, for example, $R_{[\mathrm{Q:NONE,W:DEC}]}(D)$ in general corresponds to a version of the Wyner–Ziv problem in which the signal side information w^n affects the distortion measure through its correlation with the unobserved q^n, a scenario equivalent to one considered in [6]. However, $R_{[\mathrm{Q:NONE,W:NONE}]}(D)$ is equivalent to $R(D)$, the rate–distortion function without side information (for the corresponding averaged distortion measure), since no side information at all is observed and, hence, any correlation is irrelevant.

As pointed out by Berger [10], all the relevant rate–distortion functions may be derived by considering q as part of x or w (i.e., by considering various combinations of "super sources" and/or "super side information" such as $\tilde{x} = (x, q)$, $\tilde{w} = (w, q)$, etc.) and applying well-known results for source coding, source coding with side information, the conditional rate–distortion theorem, etc. The resulting expressions are a natural starting point for our development. We begin with the simpler set of cases in which there is no signal side information.

IV. Source Coding With Distortion Side Information Alone

It is straightforward to express the rate–distortion tradeoff for quantization when distortion side information is present, but signal side information is not. In particular, we obtain the following.

Proposition 1: The rate–distortion functions for a source x, distortion measure (3), and distortion side information q are

$$R_{[\mathrm{Q:NONE}]}(D) = \inf_{\{p_{\hat{x}|x}: E[d(x,\hat{x};q)] \le D\}} I(\hat{x}; x) \tag{6a}$$

$$R_{[\mathrm{Q:DEC}]}(D) = \inf_{\{p_{u|x}, g: E[d(x,g(u,q);q)] \le D\}} I(u; x|q) \tag{6b}$$

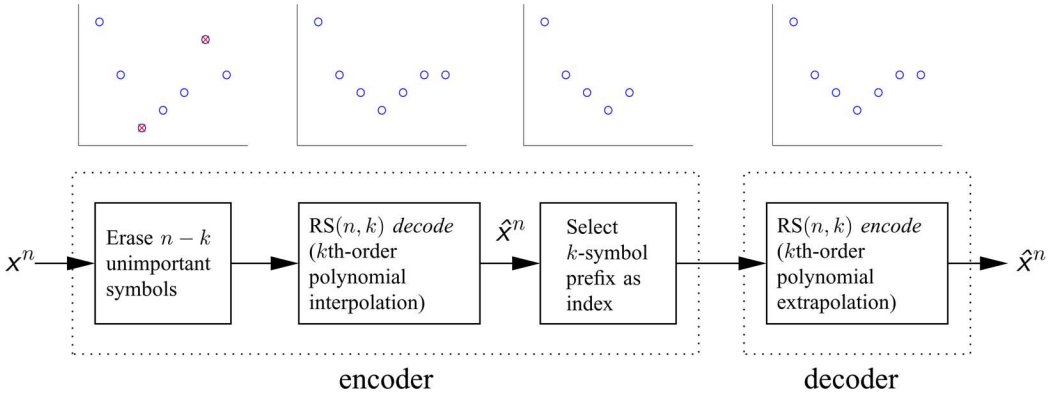


Fig. 3. Source coding with encoder-only distortion side information. In this example, an "erasure-based" distortion is used, whereby the side information indicates which k of n source symbols are important and must be reproduced exactly; the remaining $n-k$ symbols are completely unimportant do not need to be reproduced correctly. The depicted system construction exploits Reed–Solomon coding and its "curve-fitting" interpretation. Depicted is the case $n = 7$ and $k = 5$, with a source alphabet of $|\mathcal{X}| = 8$ possible values.

$$R_{[\mathrm{Q:ENC}]}(D) = \inf_{\{p_{\hat{x}|x,q}: E[d(x,\hat{x};q)] \le D\}} I(\hat{x}; x, q) \tag{6c}$$

$$R_{[\mathrm{Q:BOTH}]}(D) = \inf_{\{p_{\hat{x}|x,q}: E[d(x,\hat{x};q)] \le D\}} I(\hat{x}; x|q). \tag{6d}$$

The rate–distortion functions in (6a)–(6d) follow from standard results (e.g., [1], [3], [4], [7], [9]). To obtain (6c), for example, we apply the classical rate–distortion theorem to the super-source $\tilde{x} = (x, q)$.

In the remainder of this section, we turn our attention to addressing when encoder-only distortion side information is as good as having it at both encoder and decoder, and when decoder-only distortion side information is useless. Before developing our formal results, we first describe some examples of such behavior that provide some preliminary intuition.

A. Illuminating Examples

To develop an appreciation for how having distortion side information available only at the encoder can be as effective as having it at both encoder and decoder, we begin with two motivating examples, corresponding to a discrete and continuous source, respectively. We keep the discussion fairly informal in this section to emphasize the basic insights.

Example 1 (Discrete Source): Consider a source sequence x^n whose n samples are drawn uniformly and independently from the finite alphabet $\mathcal{X}$ with cardinality $|\mathcal{X}| \ge n$. Let q^n correspond to the n binary variables ($\mathcal{Q} = \{0, 1\}$) indicating which source samples are relevant. Specifically, let the distortion measure be of the form $d(x, \hat{x}; q) = 0$ if and only if either $q = 0$ or $x = \hat{x}$. Finally, let the sequence q^n be statistically independent of the source x^n, with q^n drawn uniformly from the $\binom{n}{k}$ subsets with exactly k ones.[3]

If the side information were available at both encoder and decoder, then only the relevant source samples would need to be described to avoid incurring distortion (i.e., achieve $D = 0$), which requires $k \log |\mathcal{X}|$ bits. Thus, this is obviously a lower bound on the number of bits required when side information is available only at the encoder.

At the other extreme, an upper bound is obtained from the scenario in which side information is unavailable or ignored. In this case, representing the source without distortion would require exactly $n \log |\mathcal{X}|$ bits.

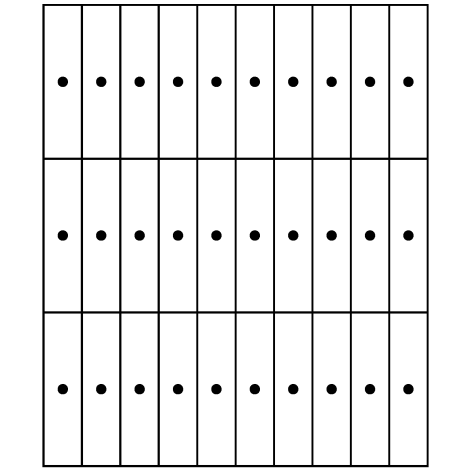
When $H_{\mathrm{B}}(k/n) < (1 - k/n) \log |\mathcal{X}|$, with $H_{\mathrm{B}}(\cdot)$ denoting the binary entropy function as defined in Section II, a better (though still suboptimal) approach when encoder side information is available would be for the encoder to describe for the decoder which samples are relevant and then describe only those samples. Using Stirling's approximation, the former description would require about $nH_{\mathrm{B}}(k/n)$ bits, while the latter description would, again, require $k \log |\mathcal{X}|$ bits. However, it is possible to represent the source without distortion using still fewer bits, as we now describe.

As depicted in Fig. 3, we view the source samples x^n as a codeword of an (n, k) Reed–Solomon code (or more generally any maximal distance separable (MDS) code[4]) with $q_i = 0$ indicating an erasure at sample i. The encoder uses the Reed–Solomon *decoding* algorithm, which coresponds to kth-order polynomial interpolation, to "correct" the erasures and determine the k information symbols, which constitute the source representation. To reconstruct the signal, the decoder uses the Reed–Solomon *encoding* algorithm, which corresponds to kth-order polynomial extrapolation of the k information symbols to produce the reconstruction $\hat{x}^n$. In this way, $\hat{x}_i = x_i$ whenever $q_i = 1$ and the relevant samples are represented without distortion using only $k \log |\mathcal{X}|$ bits. Remarkably, we see this is precisely our lower bound above, corresponding to the best possible rate were the side information also available at the decoder.

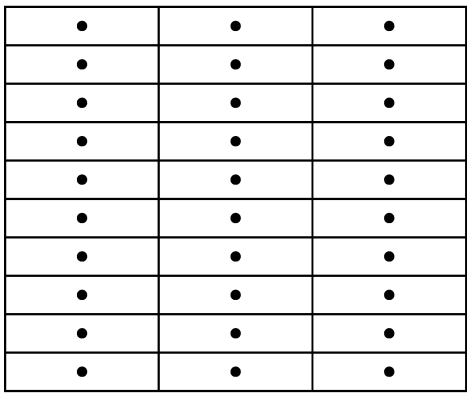
It is also worth remarking that in this example not only is it sufficient to have the side information only at the encoder, but we will show later that having it only at the decoder is useless, i.e., the number of bits required to achieve distortion $D = 0$ is

[3]If the distortion side information were a Bernoulli (k/n) sequence, then there would be about k ones with high probability. We focus on the case with exactly k ones for simplicity.

[4]The desired MDS code always exists since we assumed $|\mathcal{X}| \ge n$. For $|\mathcal{X}| < n$, near-MDS codes exist, which give asymptotically similar performance with an overhead that goes to zero as $n \to \infty$.

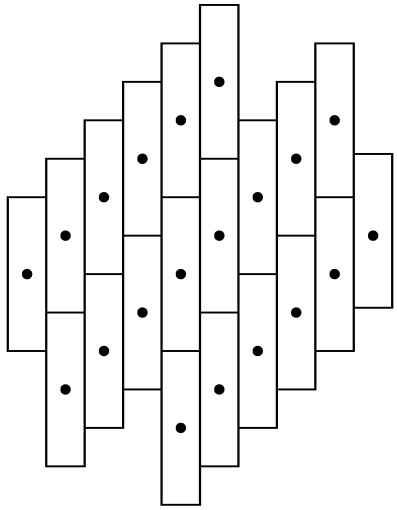


(a) Horizontal coordinate more important.

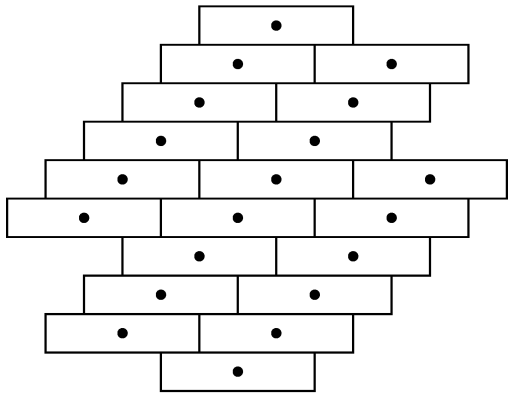


(b) Vertical coordinate more important.

Fig. 4. Quantizers for distortion side information available at encoder and decoder. When the side information q^2 indicates the horizontal (respectively, vertical) coordinate is more important, the encoder and decoder use the upper (respectively, lower) codebook lattice and partition function to increase horizontal (respectively, vertical) accuracy.



(a) Horizontal coordinate more important.



(b) Vertical coordinate more important.

Fig. 5. Quantizers for distortion side information available only at the encoder. A common codebook lattice is used at the decoder, independent of the realized side information, but when the side information indicates that the horizontal (respectively, vertical) coordinate is more important, the encoder uses the upper (respectively, lower) partition to increase horizontal (respectively, vertical) accuracy.

$n \log |\mathcal{X}|$ in this case, the same as if the side information were not available.

An analogous approach can be used for continuous sources. In particular, for such sources the discrete Fourier transform (DFT) plays the role of the Reed–Solomon code. Specifically, to encode the n source samples, we view the k relevant samples as elements of a complex, periodic, Gaussian, sequence with period n, which is band-limited in the sense that only its first k DFT coefficients are nonzero. Using periodic, band-limited, interpolation we can use only the k samples where $q_i = 1$ to find the corresponding nonzero DFT coefficients, which are subsequently quantized. To reconstruct the signal, the decoder reconstructs the temporal signal corresponding to the quantized DFT coefficients.

Rather than developing this analogy further, we instead next develop some additional insights afforded by a rather different approach to continuous sources.

Example 2 (Continuous Source): Consider the quantization of a single pair $(n = 2)$ of samples from a continuous source (i.e., $(x_1, x_2) \in \mathbb{R}^2$) where the side information selects one of two possible distortion measures. In particular, the distortion is of the form

$$d(x^2, \hat{x}^2; q^2) \propto q_1(x_1 - \hat{x}_1)^2 + q_2(x_2 - \hat{x}_2)^2$$

where the corresponding side information pair (q_1, q_2) takes on one of two different values, (a, b) or (b, a) (with $a \neq b$), independently of the realization of the source pair (x_1, x_2). Hence, in one of the distortion measures, the first of the samples is more important (i.e., less distortion tolerant) than the other. In the other measure, it is the second sample that is more important.

If the side information q^n were available to both encoder and decoder, then one could choose a codebook lattice and partition function for each of the two values of the side information. Such a solution is as depicted in Fig. 4.

When the side information is available only at the encoder, then one requires a solution that involves a single, common codebook lattice. However, we can still use two partition functions chosen according to the value of the (binary) side information. For this example, such a solution is as depicted in Fig. 5.

Comparing Fig. 5 with Fig. 4, it is apparent—neglecting edge effects and considering a uniformly distributed source—that having the distortion side information only at the encoder incurs no additional distortion. Later in the paper we will make such statements precise through high-resolution analysis, but our qualitative discussion to this point suffices to reveal the basic

intuition and the fundamental role that fixed-codebook/variable-partition encoders play more generally in the associated systems. Moreover, this encoding strategy generalizes readily to arbitrary block lengths and more efficient partitions, and can be implemented with only linear complexity in the block length, as described in [11].

As a final comment, in this example, too, it turns out that not only is it sufficient to have the side information only at the encoder, but, as we will later establish, having it only at the decoder is useless. In other words, as one might expect intuitively, once the partition (i.e., encoder) has been fixed, there is no way to exploit knowledge of the side information at the decoder to reduce the distortion.

B. Cautionary Counterexamples

At this point, one might begin to get the mistaken impression that encoder-only distortion side information is always sufficient, and that decoder-only distortion side information is never useful. In fact, there must be at least some meaningful structural constraints on the form of the distortion measure for these properties to hold, as the following simple counterexamples make clear.

Claim 1: Consider a source x and side information q each uniformly distributed over the alphabet $\mathcal{X} = \mathcal{Q} = \{0, 1\}$. If the x and q are independent and the distortion measure is of the form

$$d(x, \hat{x}; q) = (x + q - \hat{x})^2 \tag{7}$$

then both

$$R_{[\mathrm{Q:ENC}]}(D) > R_{[\mathrm{Q:BOTH}]}(D) \tag{8}$$

and

$$R_{[\mathrm{Q:DEC}]}(D) < R_{[\mathrm{Q:NONE}]}(D). \tag{9}$$

Moreover, even in the high-resolution limit $(D \to 0)$ the rate gaps corresponding to (8) and (9) are each at least 1/2 bit/sample.

Proof: To establish (8), first note that when the side information is available at the decoder, the problem is equivalent to one without side information, but with effective source x, since due to the form of the distortion measure (7) the side information can be added to the reconstruction to eliminate its effect. However, when the side information is available only at the encoder, then the quantization problem is equivalent to one without side information, but with an effective source $x + q$. Thus, we have for $0 \leq D \leq 1/2$

$$\begin{aligned} &R_{[\mathrm{Q:BOTH}]}(D) \\ &\quad = 1 - H_{\mathrm{B}}(D) < 3/2 - H_{\mathrm{B}}(D/2) \leq R_{[\mathrm{Q:ENC}]}(D) \end{aligned} \tag{10}$$

where the first term on the second line of (10) is the readily computed Shannon lower bound on the rate required to quantize the effective source, whence (8). Finally, to establish (9) it suffices to note that

$$\begin{aligned} &R_{[\mathrm{Q:DEC}]}(D) \\ &\quad = R_{[\mathrm{Q:BOTH}]}(D) < R_{[\mathrm{Q:ENC}]}(D) \leq R_{[\mathrm{Q:NONE}]}(D) \end{aligned}$$

where the first inequality is (8). □

It is also worth emphasizing that if Definition 1 is not satisfied (specifically, if x and q are not independent in this case where there is only distortion side information), encoder-only side information may or may not be sufficient, even when the distortion measure is otherwise reasonably chosen.

We begin with the following proposition.

Proposition 2: If the source x is discrete and the distortion measure $d(x, \hat{x}; q)$ has a unique minimum with respect to $\hat{x}$ at $\hat{x} = x$, then we have

$$R_{[\mathrm{Q:ENC}]}(D_{\min}) - R_{[\mathrm{Q:BOTH}]}(D_{\min}) \leq I(x; q) \tag{11}$$

with $D_{\min} \triangleq E[d(x, x; q)]$. Moreover, for all $D \geq D_{\min}$, we have

$$R_{[\mathrm{Q:NONE}]}(D) - R_{[\mathrm{Q:DEC}]}(D) \leq H(x). \tag{12}$$

Proof: To obtain (11), note that

$$\begin{aligned} R_{[\mathrm{Q:ENC}]}(D) &\leq I(\hat{x}^*; x, q) && (13)\\ &= I(\hat{x}^*; x|q) + I(\hat{x}^*; q) && (14)\\ &= R_{[\mathrm{Q:BOTH}]}(D) + I(\hat{x}^*; q) && (15) \end{aligned}$$

where (13) follows from (6c) with $\hat{x}^*$ corresponding to the distribution that optimizes (6d) instead of (6c), where to obtain (14) we have used the chain rule of mutual information, and where to obtain (15) we have used (6d). In turn, noting that $D = D_{\min}$ implies $\hat{x}^* = x$, we can rewrite (15) in the form of (11), as desired.

To obtain (12), it suffices to note that $R_{[\mathrm{Q:DEC}]}(D) \geq 0$, and that, via (6a) with the particular choice $\hat{x} = x$, which always meets the distortion constraint, we have $R_{[\mathrm{Q:NONE}]}(D) \leq I(x; x) = H(x)$. □

The following claim establishes that the inequalities (11) and (12) can hold with equality, and thus they quantify the maximum insufficiency of encoder-only side information and maximum efficacy of decoder-only side information.

Claim 2: If the source x is discrete, the side information satisfies $q = x$, and the distortion measure is of the form $d(x, \hat{x}; q) = d_{\mathrm{H}}(x, \hat{x})$, where $d_{\mathrm{H}}(\cdot, \cdot)$ is the usual Hamming distortion measure as defined in Section II, then

$$R_{[\mathrm{Q:ENC}]}(0) - R_{[\mathrm{Q:BOTH}]}(0) = H(x) = I(x; q) \tag{16}$$

and

$$R_{[\mathrm{Q:NONE}]}(0) - R_{[\mathrm{Q:DEC}]}(0) = H(x). \tag{17}$$

Proof: To verify (16) and (17), it suffices to note that, for all $D \geq D_{\min}$,

$$R_{[\mathrm{Q:BOTH}]}(D) = R_{[\mathrm{Q:DEC}]}(D) = 0 \tag{18}$$

since $x^n = q^n$ is available at the decoder, and that

$$R_{[\mathrm{Q:ENC}]}(0) = R_{[\mathrm{Q:NONE}]}(0) = H(x), \tag{19}$$

corresponding to lossless encoding of x^n. □

We now turn to developing our main results of the section: comparing the rate–distortion functions in Proposition 1 to identify general conditions under which it is sufficient to have distortion side information at the encoder, i.e., $R_{[\mathrm{Q:ENC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D)$, and under which it is useless to have it at the decoder, i.e., $R_{[\mathrm{Q:DEC}]}(D) = R_{[\mathrm{Q:NONE}]}(D)$. We begin with the former.

C. Sufficiency of Encoder-Only Side Information

First, we have the following equivalent characterization of sufficiency.

Proposition 3: $R_{[\mathrm{Q:ENC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D)$ if and only if $I(\hat{x}; q) = 0$ for some $\hat{x}$ that optimizes (6d).

Proof: It suffices to equate (6c) and (6d), expanding the argument of the former using the chain rule of mutual information as follows:

$$I(\hat{x}; x, q) = I(\hat{x}; x|q) + I(\hat{x}; q),$$

and note that if some $\hat{x}$ minimizes (6d) and satisfies $I(\hat{x}; q) = 0$, it also minimizes (6c). □

Proposition 3 admits a simple interpretation. In particular, since $p_{\hat{x}|q}(\hat{x}|q)$ represents the distribution of the codebook, the condition $I(\hat{x}; q) = 0$ corresponds to the requirement that the codebook distribution be independent of the side information. In the language of Example 2, this says that encoder-only side information can only be sufficient if and only if a common codebook can perform as well as can be achieved by separate codebooks (tuned to each possible value of the side information).

From another perspective, the condition $I(\hat{x}; q) = 0$ tells us that in a system in which encoder-only side information is sufficient and used efficiently, no information about the side information sequence q^n can be inferred from observation of the reconstruction $\hat{x}^n$ (or, equivalently, the quantization index). As such, we can infer that the naive approach to encoder-only side information problems, whereby the side information is conveyed to the decoder via a side channel, can never be efficient in such cases. Indeed, the rate penalty (overhead) associated with such a scheme must be at least $H(q)$, which can be arbitrarily large.

There are two natural scenarios where $I(\hat{x}; q)$ can be zero and hence $R_{[\mathrm{Q:ENC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D)$: the case of arbitrary sources with erasure distortions, and the case of uniform sources with group difference distortions. We consider each separately, in turn, and remark in advance that Example 1 is an instance of both.

1) Arbitrary Sources With Erasure Distortions: The following theorem establishes that for erasure-type distortion measures, encoder-only side information is sufficient.

Theorem 1: For any source x, if the side information q has binary elements, i.e., $\mathcal{Q} = \{0, 1\}$, and the associated distortion measure is of the form

$$d(x, \hat{x}; q) = q\rho(x, \hat{x}) \tag{20}$$

for some function $\rho(\cdot, \cdot)$, then

$$R_{[\mathrm{Q:ENC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D). \tag{21}$$

Before proceeding with a proof, we make some remarks. First, we emphasize that there is no requirement that x and q be independent. Second, not only is encoder side information sufficient in the case of erasure distortions, but the quantizers for optimally exploiting side information at the encoder alone can be especially simple, as Example 1 suggests.

Proof: Let $\hat{x}^*$ correspond to a distribution that optimizes (6d). Define a new random variable $\hat{x}^{**}$ obtained from $\hat{x}^*$ according to

$$p_{\hat{x}^{**}|\hat{x}^*,x,q}(\hat{x}^{**}|\hat{x}^*, x, q) = \begin{cases} \delta_{\hat{x}^{**},\hat{x}^*}, & q = 0 \\ p_{\hat{x}^*|q}(\hat{x}^{**}|q = 0), & q = 1 \end{cases} \tag{22}$$

where $\delta_{\cdot,\cdot}$ is the usual Kronecker delta function. Via this construction, both $\hat{x}^*$ and $\hat{x}^{**}$ have the same expected distortion since they only differ when $q = 0$. Moreover, since $x \leftrightarrow \hat{x}^* \leftrightarrow \hat{x}^{**}$ form a Markov chain (conditioned on q), we have, by the data processing inequality,

$$I(\hat{x}^{**}; x|q) \le I(\hat{x}^*; x|q) \tag{23}$$

so $\hat{x}^{**}$ also optimizes (6d). Finally, since $I(\hat{x}^{**}; q) = 0$, Proposition 3 is satisfied and we obtain the desired result. □

2) Uniform Sources With Group Difference Distortions: Uniform source and group difference distortion measures arise naturally in a variety of applications, such as those where phase is a quantity of particular interest. Example application domains range from magnetic resonance imaging, synthetic aperture radar, and ultrasonic microscopy, to audio, image, and video processing.

The following theorem establishes that for uniform sources and group-difference-type distortion measures, encoder-only side information is also sufficient.

Theorem 2: Suppose the source x of interest is uniformly distributed over a group $\mathcal{X}$ and the distortion measure is of the form

$$d(x, \hat{x}; q) = \rho(x \ominus \hat{x}; q) \tag{24}$$

for some function $\rho(\cdot, \cdot)$ and binary relation $\oplus$,[5] where the distortion side information q is independent of x. Then

$$R_{[\mathrm{Q:ENC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D). \tag{25}$$

Before proceeding with the proof, we remark that this result is rather natural. Indeed, the symmetry in this case ensures that the optimal codebook distribution is uniform independent of the side information, allowing a fixed codebook to perform as well as a variable codebook—provided it is used in conjunction with a suitably chosen variable partition function. The Reed–Solomon-based encoding of Example 1 has precisely this character.

For the most general case of mixed groups with both discrete and continuous components, we provide a direct proof based on symmetry and convexity arguments in Appendix I. However, for the case of either purely discrete or purely continuous

[5]The relation $\ominus$ is, in turn, defined via $a \ominus b \triangleq a \oplus (-b)$, where $-b$ denotes the additive inverse of b in the group.

groups, a simple and intuitive proof is possible via the conditional Shannon lower bound on the right-hand side of (25), which we establish in the sequel.

Proof of Theorem 2: First, consider the case in which $\mathcal{X}$ is a discrete group. With $\hat{x}^*$ denoting the random variable corresponding to the optimizing distribution in (6d), the conditional Shannon lower bound on $R_{[\mathrm{Q:BOTH}]}(D)$ is

$$\begin{aligned} R_{[\mathrm{Q:BOTH}]}(D) &= I(\hat{x}^*; x|q) && (26)\\ &= H(x|q) - H(x|\hat{x}^*, q) \\ &= H(x) - H(\hat{x}^* \ominus x|\hat{x}^*, q) && (27)\\ &\geq H(x) - H(\hat{x}^* \ominus x|q) && (28)\\ &\geq H(x) - H(v_D^*|q) && (29)\\ &= \log|\mathcal{X}| - H(v_D^*|q) && (30) \end{aligned}$$

where (26) follows from (6d), (27) follows from the independence of x and q, (28) follows since conditioning cannot increase entropy, where v_D^* in (29) is a random variable v that maximizes $H(v|q)$ subject to the constraint $E[\rho(v;q)] \leq D$, and where (30) follows from x being uniformly distributed. Note that as a consequence of its maximum entropy property, v_D^* is dependent on q, but independent of x, i.e.,

$$p_{v_D^*,q,x}(v,q,x) = p_{v_D^*|q}(v|q)p_q(q)p_x(x). \tag{31}$$

Next, observe that $\hat{x}^*$ and v_D^* must be related according to $\hat{x}^* = v_D^* \oplus x$, since the inequalities (28) and (29) then hold with equality. To verify that (28) holds with equality, it suffices to note that since v_D^* and x are independent and x is uniform, $\hat{x}^*$ is also uniform. Thus, since the distribution of $\hat{x}^*$ does not depend on v_D^*, the two random variables are independent.

To complete the proof, it suffices to apply Proposition 3, noting that since $\hat{x}^*$ is uniform regardless of the value of q, these two random variables are also independent, whence $I(\hat{x}^*; q) = 0$.

For the case of continuous groups, it suffices, in (30), to replace $|\mathcal{X}|$ with the Lebesgue measure of the group, and $H(v_D^*|q)$ with the differential entropy $h(v_D^*|q)$. □

D. Inefficacy of Decoder-Only Side Information

A fairly general scenario under which decoder-only distortion side information is ineffective is given by the following theorem.

Theorem 3: For any source x, a distortion measure of the scaled form[6]

$$d(x,\hat{x};q) = \gamma(q)\rho(x,\hat{x}) + \tau(q) \tag{32}$$

for some $\gamma(\cdot) \geq 0$, $\tau(\cdot)$, and $\rho(\cdot,\cdot)$, and any side information q that is independent of x, we have

$$R_{[\mathrm{Q:DEC}]}(D) = R_{[\mathrm{Q:NONE}]}(D). \tag{33}$$

Proof: From (6b) we see that if $g(\cdot,\cdot)$ is optimal, then

$$g(u,q) = \arg\min_{\hat{x}} E[d(x,\hat{x};q) \mid u = u, q = q]. \tag{34}$$

[6]Obviously, without loss of generality, we could let $\gamma(\cdot)$ be the identity function by appropriately (re)defining q. Nevertheless, our form will be convenient.

However, substituting (32) into(34) and noting that

$$\begin{aligned} E[d(x,\hat{x};q) \mid q = q, u = u] \\ = \gamma(q)E[\rho(x,\hat{x}) \mid u = u, q = q] + \tau(q) \end{aligned}$$

we obtain

$$\begin{aligned} g(u,q) &= \arg\min_{\hat{x}} E[\rho(x,\hat{x}) \mid u = u, q = q] \\ &= \arg\min_{\hat{x}} E[\rho(x,\hat{x}) \mid u = u] \end{aligned} \tag{35}$$

where to obtain (35) we have used the Markov chain $u \leftrightarrow x \leftrightarrow q$ implicit in (6b), and that x and q are independent. Since the right-hand side of (35) does not depend on q, we conclude that g is only a function of u, in which case we simply replace u and $g(u,q)$ with $\hat{x}$ in (6b). Finally, since $I(u;x) - I(u;x|q) = I(u;q)$ is zero by the same argument used to obtain (35), we see that (6b) specializes to (6a). □

A couple of additional insights are worth developing. In particular, we address the question of companion inefficacy results corresponding to the two special classes of problems considered in Sections IV-C1 and IV-C2.

First, considering the results in Section IV-C1, one might wonder whether for problems with erasure distortion measures that decoder side information would be of no value even without the requirement that the source and side information be independent. In particular, we know from Theorem 3 that this is true when there is such independence. However, without such independence, this is not true, as the following counterexample establishes.

Claim 3: Suppose x is a Bernoulli$(1/2)$ source, t is a Bernoulli(ϵ) process that is independent of x with $0 < \epsilon < 1/2$, and $q = x \oplus t$, where $\oplus$ is the EXCLUSIVE-OR operator corresponding to modulo-2 addition. Then, for the erasure distortion measure $d(x,\hat{x};q) = qd_{\mathrm{H}}(x,\hat{x}) = q(x \oplus \hat{x})$, we have

$$R_{[\mathrm{Q:DEC}]}(0) < R_{[\mathrm{Q:NONE}]}(0). \tag{36}$$

Proof: First, it is straightforward to verify (e.g., via a simple Shannon lower bound) that to achieve $D = 0$ requires rate

$$R_{[\mathrm{Q:NONE}]}(0) = 1 \text{ bit/sample}. \tag{37}$$

On the other hand, via (6b) with the particular choices $g(u,q) = u$ and $u = x$, which satisfies the distortion constraint, we have, in bits per sample

$$R_{[\mathrm{Q:DEC}]}(0) \leq I(x;x|q) = H(x|q) = H(t) = H_{\mathrm{B}}(\epsilon) < 1 \tag{38}$$

where the middle equality follows from the fact that q is also independent of t. Comparing (37) to (38) yields (36). □

Second, considering the results in Section IV-C2, one might wonder whether for problems with group-difference distortion measures and independent side information, that decoder side information would be useless provided the source were uniformly distributed over the group. We know from Claim 1 that this is not true without this constraint on the source. However, even with this additional constraint, the result is still not true, as the following counterexample establishes.

Claim 4: Consider a source x and side information q each uniformly distributed over the alphabet $\mathcal{X} = \mathcal{Q} = \{0,1\}$. If x and q are independent, $\hat{\mathcal{X}} = \mathcal{X}$, and the distortion measure is of the binary group difference form $d(x,\hat{x};q) = x \oplus q \oplus \hat{x}$ (where $\oplus = \ominus$ is the EXCLUSIVE-OR operator), then for any $D < 1/2$

$$R_{[\mathrm{Q:NONE}]}(D) > R_{[\mathrm{Q:DEC}]}(D) = R_{[\mathrm{Q:BOTH}]}(D). \quad (39)$$

Proof: To verify the right equality in (39), note that for the case in which q^n is only at the decoder, we can encode as if the distortion measure were $d'(x,\hat{x}') = x \oplus \hat{x}'$, producing the intermediate reconstruction $(\hat{x}'_1,\ldots,\hat{x}'_n)$. This is the familiar binary-Hamming source coding problem. Then we can produce the final reconstruction via $\hat{x}_i = \hat{x}'_i \oplus q_i$ for $i = 1,\ldots,n$, so the ultimate distortion is $d(x,\hat{x};q) = x \oplus q \oplus \hat{x} = x \oplus \hat{x}'$. Hence, the rate–distortion function is the binary-Hamming one, $R_{[\mathrm{Q:DEC}]}(D) = 1 - H_{\mathrm{B}}(D)$, and is the same as that achievable were q^n also known at the encoder.

Next, to establish the left inequality in (39), note that if q^n is known at neither encoder nor decoder, then $E[d(x,\hat{x};q)] = 1/2$ for any rate since $x \oplus q \oplus \hat{x}$ is uniformly distributed regardless of how $\hat{x}$ is chosen. □

E. Example: A Binary-Hamming Case

More detailed results are possible for the key special case of Theorem 2 in which the source x is binary (i.e., $|\mathcal{X}| = 2$), and where, for convenience, we let the side information be discrete, i.e., $\mathcal{Q} = \{1,2,\ldots,N\}$ for some N. In this case, the distortion measure (24) can be expressed, without loss of generality, in the form

$$d(x,\hat{x};q) = \gamma(q) d_{\mathrm{H}}(x,\hat{x}) + \tau(q) \quad (40)$$

where $\gamma(q) \geq 0$ and $\tau(q)$ are side-information-dependent weights, with $d_{\mathrm{H}}(\cdot,\cdot)$ again denoting the Hamming distortion measure.

Clearly, both Theorems 2 and 3 apply for this case. The associated rate distortion expressions are, when $D \geq E[\tau(q)]$,

$$\begin{aligned} R_{[\mathrm{Q:ENC}]}(D) &= R_{[\mathrm{Q:BOTH}]}(D) \\ &= 1 - E\left[H_{\mathrm{B}}\left(\frac{2^{-\lambda\gamma(q)}}{1+2^{-\lambda\gamma(q)}}\right)\right] \end{aligned} \quad (41a)$$

where λ is chosen to satisfy the distortion constraint

$$E\left[\gamma(q)\cdot\frac{2^{-\lambda\gamma(q)}}{1+2^{-\lambda\gamma(q)}} + \tau(q)\right] = D \quad (41b)$$

and

$$\begin{aligned} R_{[\mathrm{Q:DEC}]}(D) &= R_{[\mathrm{Q:NONE}]}(D) \\ &= 1 - H_{\mathrm{B}}\left(\frac{D - E[\tau(q)]}{E[\gamma(q)]}\right). \end{aligned} \quad (42)$$

The derivations of (41) and (42) are provided in Appendix II.

Two instances of the distortion measure form (40) are worth developing in more detail for additional insight.

1) Noisy Observations: Suppose x is a noisy observation of some underlying source, where the noise is governed by a binary symmetric channel with crossover probability controlled by the side information. Specifically, let the crossover probability of the channel be

$$\epsilon_q = \frac{q-1}{2(N-1)}, \quad (43)$$

which is at most $1/2$. Furthermore, a distortion of 1 is incurred if an error occurs due to either the noise in the observation or the noise in the quantization—but not both—and there is no distortion otherwise, i.e., using (43)

$$\begin{aligned} d(x,\hat{x};q) &= \epsilon_q[1 - d_{\mathrm{H}}(x,\hat{x})] + (1-\epsilon_q) d_{\mathrm{H}}(x,\hat{x}) \\ &= \frac{q-1}{2(N-1)} + \left(1 - \frac{q-1}{N-1}\right) d_{\mathrm{H}}(x,\hat{x}). \end{aligned} \quad (44)$$

Since (44) corresponds to a distortion measure of the form (40) with

$$\tau(q) = \frac{q-1}{2(N-1)} \quad \text{and} \quad \gamma(q) = 1 - \frac{q-1}{N-1}$$

the rate-distortion formulas (41) and (42) apply.[7]

Fig. 6 depicts the associated rate—distortion tradeoffs for the cases $N = 2$ and $N \to \infty$. The solid curves shows the tradeoff achievable when the side information is available at the encoder, while the dashed curves shows the (poorer) tradeoff achievable when it is not.

It should also be emphasized that this is an instance in which the naive encoding method, whereby the encoder losslessly communicates the side information to the decoder and then uses encoding for the case of side information at both encoder and decoder, can require arbitrarily higher rate than the optimal rate-distortion tradeoff. Indeed, to losslessly encode the side information requires an additional rate of $\log N$, which is unbounded in N.

2) Weighted Distortion: In a number of applications, certain samples of a source are inherently more important than others. Such sources are naturally quantized with a weighted distortion measure, an example of which is of the form (40) with $\gamma(q) = \eta^{q/N}$, $\eta > 1$, and $\tau(q) = 0$. In the sequel, we let q be uniformly distributed over $\{0,1,\ldots,N-1\}$.

This weighted Hamming distortion measure can be used to demonstrate that not having (or ignoring) the side information at the encoder can incur an arbitrarily high distortion penalty. To see this, it suffices to restrict attention to the case for which $N = 2$ and observe that as $\eta \to \infty$, a system without encoder side information suffers increasingly more distortion. This is most evident for $R > 1/2$. In this rate region, a simple system with the side information at both encoder and decoder that losslessly encodes the important samples (as revealed by the side information) achieves a distortion that is at most $1/2$. However, a system without encoder side information experiences a distortion

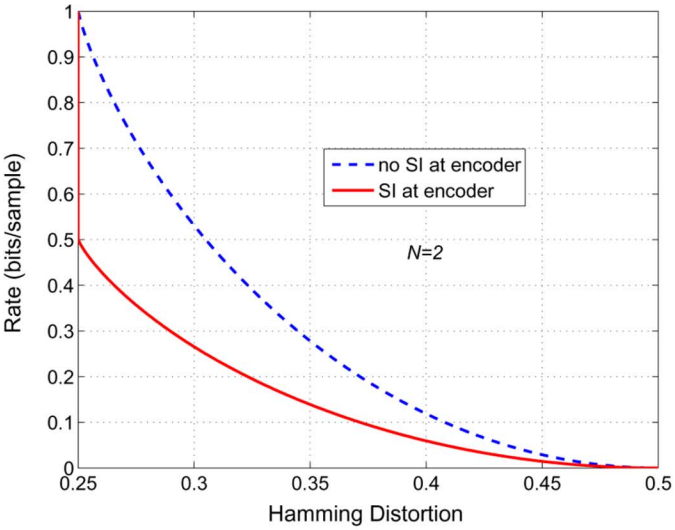
$$D \propto E[\gamma(q)] = (1+\sqrt{\eta})/2,$$

which grows without bound in η. Thus, the extra distortion (and hence rate loss) incurred when q is not available to the encoder can be arbitrarily large.

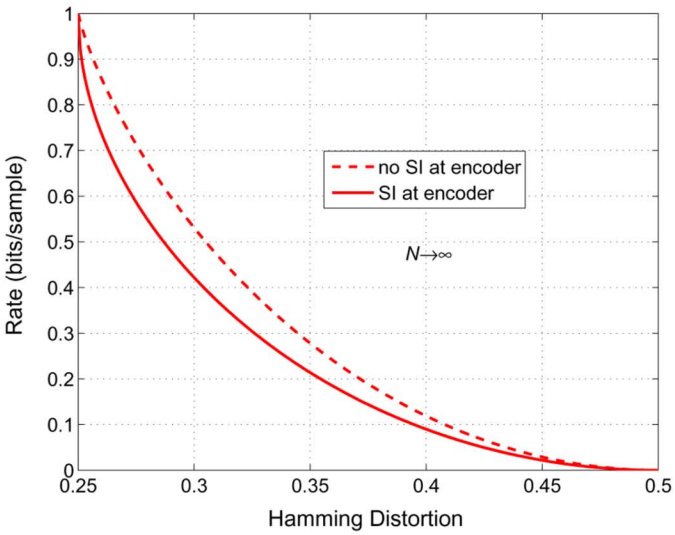
[7]Note that an optimal encoding strategy when the side information is available at both encoder and decoder is to encode the noisy observation directly although with different amounts of quantization depending on the side information [12].

TABLE I
RATE-DISTORTION EQUIVALENCE RESULTS FOR CONTINUOUS SOURCES, AS INDICATED BY ARROWS, ASSOCIATED THEOREMS, AND CONDITIONS(H: HIGH-RESOLUTION; I: q^n AND w^n INDEPENDENT; S: SCALED DIFFERENCE DISTORTION)

	Decoder missing w^n		Decoder has w^n	
Encoder Missing q^n	$R_{[\mathrm{Q:DEC,W:ENC}]}$ $\overset{\text{Th. 4 (I)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:DEC,W:NONE}]}$ $\Updownarrow$ Th. 5 (S+I) $\quad$ $\Updownarrow$ Th. 5 (S+I) $R_{[\mathrm{Q:NONE,W:ENC}]}$ $\overset{\text{Th. 4 (I)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:NONE,W:NONE}]}$		$R_{[\mathrm{Q:DEC,W:BOTH}]}$ $\overset{\text{Th. 5 (S)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:NONE,W:BOTH}]}$ $\Updownarrow$ Th. 6 (H+S+I) $\quad$ $\Updownarrow$ Th. 6 (H+S+I) $R_{[\mathrm{Q:DEC,W:DEC}]}$ $\overset{\text{Th. 5 (S)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:NONE,W:DEC}]}$	
Encoder has q^n	$R_{[\mathrm{Q:ENC,W:ENC}]}$ $\overset{\text{Th. 4}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:ENC,W:NONE}]}$ $\Updownarrow$ Th. 7 (H+I) $\quad$ $\Updownarrow$ Th. 7 (H+I) $R_{[\mathrm{Q:BOTH,W:ENC}]}$ $\overset{\text{Th. 4}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:BOTH,W:NONE}]}$		$R_{[\mathrm{Q:ENC,W:DEC}]}$ $\overset{\text{Th. 6 (H)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:ENC,W:BOTH}]}$ $\Updownarrow$ Th. 7 (H) $\quad$ $\Leftrightarrows$ Lem. 3 (H) $\quad$ $\Updownarrow$ Th. 7 (H) $R_{[\mathrm{Q:BOTH,W:DEC}]}$ $\overset{\text{Th. 6 (H)}}{\Longleftrightarrow}$ $R_{[\mathrm{Q:BOTH,W:BOTH}]}$	



(a) $N = 2$ ($\epsilon = 0$ or $1/2$ with equal probability).



(b) $N \to \infty$ (ϵ uniformly distributed on $[0, 1/2]$).

Fig. 6. Rate-distortion tradeoffs for noisy observations of a binary source. The solid and dashed curves represent the minimum possible Hamming distortion when side information specifying the crossover probability of the observation noise is and is not available at the encoder, respectively.

V. SOURCE CODING WITH DISTORTION AND SIGNAL SIDE INFORMATION

We now turn our attention to the more general scenario in which there is both signal and distortion side information in the problem. Our results take the form of a set of four theorems describing when different types of side information knowledge are equivalent. These results show that the 16 possible information configurations of Fig. 2 can be reduced to the four shown in Table I.

We partition our results into two classes: those establishing where side information is of no value, and those establishing where side information is (asymptotically) of full value. We begin with the former.

A. Inefficacy Theorems

Our first pair of theorems show, respectively, and under appropriate conditions, that w^n known only at the encoder is of no value, and q^n known only at the decoder is of no value. Proofs are provided in Appendices III-A and B, respectively.

Theorem 4: If Definition 1 is satisfied, then

$$R_{[\mathrm{Q:}*,\mathrm{W:ENC}]}(D) = R_{[\mathrm{Q:}*,\mathrm{W:NONE}]}(D) \tag{45}$$

where $* \in \{\mathrm{ENC}, \mathrm{BOTH}\}$ (both $*$'s are identical). Moreover, (45) also holds for $* \in \{\mathrm{DEC,NONE}\}$ provided q and w are independent.

Theorem 5: If Definition 1 is satisfied and the distortion measure is of the scaled form (32), then

$$R_{[\mathrm{Q:DEC,W:}*]}(D) = R_{[\mathrm{Q:NONE,W:}*]}(D) \tag{46}$$

where $* \in \{\mathrm{DEC,BOTH}\}$ (both $*$'s are identical). Moreover, (46) also holds for $* \in \{\mathrm{ENC,NONE}\}$ provided q and w are independent.

These theorems constitute the promised generalization of Berger's result on the inefficacy of signal side information known only at the encoder [7]. Indeed, for the case in which no distortion side information is involved, Theorem 4 specializes to this classical result [3], [7], i.e.,

$$R_{[\mathrm{W:ENC}]}(D) = R_{[\mathrm{W:NONE}]}(D). \tag{47}$$

Moreover, in a similar way, Theorem 5 is the natural generalization of Theorem 3 in Section IV.

The preceding analysis suggests that in pursuing effective side information configurations, we should focus our attention on those in which distortion side information is available at the encoder, and signal side information is available at the decoder. In the sequel, we examine the degree to which having each of these forms of the side information *only* at these respective terminals is sufficient.

B. Admissibility Requirements for Sufficiency Analysis

For our analysis, we restrict our attention to the case of continuous and generally multidimensional sources, and to distortion measures of the difference form

$$d(x,\hat{x};q)=\rho(x-\hat{x};q). \tag{48}$$

Moreover, in contrast to the preceding inefficacy analysis (and to the treatment of Section IV), we focus on an asymptotic analysis. Specifically, we examine conditions under which the relaxed notion of sufficiency in the high-resolution limit is achieved. Such an analysis avoids, for example, some of the edge effects that arise in the formal analysis at finite resolutions of scenarios such as that of Example 2 in Section IV-A.

When there is no signal side information, one would expect to find asymptotic sufficiency of encoder-only distortion side information rather generally. In particular, provided the distortion measure $\rho(v;q)$ has a unique minimum with respect to v at $v=0$, then we would expect $D\to D_{\min}$ with

$$D_{\min}\triangleq E[\rho(0;q)] \tag{49}$$

to imply $\hat{x}^n\to x^n$, and thus $I(\hat{x};q)\to I(x;q)$, which is zero when x and q are independent. Considering Proposition 3, this would then suggest asymptotic sufficiency.

The above intuition turns out to be correct under certain technical conditions, as we establish in this section. Indeed, for the more general side information scenario, we show the asymptotic sufficiency of encoder-only distortion side information and decoder-only signal side information under such conditions.

We begin with the technical conditions we require our source, side information, and distortion measure to satisfy to obtain the anticipated asymptotic behavior. These conditions generalize those in [13].

Definition 2: The collection of a source x, side information pair (q,w), and difference distortion measure (48) is *admissible* if, in addition to the conditions (4) of Definition 1, the following conditions are satisfied.

1) $-\infty<h(x|w=w)<\infty$, for all $w\in\mathcal{W}$.
2) For all $q\in\mathcal{Q}$ and all $D>D_{\min}$, the distribution of the form
$$p_{v|q}(v|q)=ae^{-s\rho(v;q)}, \tag{50}$$
corresponding to the random variable v that maximizes $h(v|q=q)$ subject to
$$E[\rho(v;q)|q=q]\le D, \tag{51}$$
exists and is well behaved, i.e., there exist unique functions $a=a_\rho(D,q)$ and $s=s_\rho(D,q)$ that are continuous functions of their arguments such that
$$\int a_\rho(D,q)e^{-s_\rho(D,q)\rho(v;q)}dv=1 \tag{52a}$$
and
$$\int \rho(v;q)a_\rho(D,q)e^{-s_\rho(D,q)\rho(v;q)}dv=D. \tag{52b}$$
3) The random variable v_D^* that maximizes $h(v|q)$ subject to the constraint $E[\rho(v;q)]\le D$, for $D\ge D_{\min}$, satisfies
$$\lim_{D\to D_{\min}} v_D^*\to 0 \text{ in distribution (conditioned on } w=w),$$
i.e., for all $w\in\mathcal{W}$
$$\lim_{D\to D_{\min}}\int_{-\infty}^{v} p_{v_D^*|w}(v'|w)dv'=\begin{cases}1, & v\ge 0\\ 0, & \text{otherwise.}\end{cases} \tag{53}$$
4) There exists an auxiliary distortion measure $\delta(\cdot)$ such that[8]
   a) the distribution of the form
   $$\mathcal{N}_\delta=ae^{-s\delta(\cdot)}, \tag{54}$$
   which maximizes entropy subject to the constraint $\int\delta(t)\,ae^{-s\delta(t)}dt=\Delta$, exists and is well behaved, i.e., there exist unique functions $a=a_\delta(\Delta)$ and $s=s_\delta(\Delta)$ that are continuous functions of their arguments such that
   $$\int a_\delta(\Delta)e^{-s_\delta(\Delta)\delta(t)}dt=1 \tag{55a}$$
   and
   $$\int \delta(t)a_\delta(\Delta)e^{-s_\delta(\Delta)\delta(t)}dt=\Delta \tag{55b}$$
   for all $\Delta<\infty$;
   b) with
   $$\Delta_D(w)=E[\delta(x+v_D^*;q)\mid w=w] \tag{56a}$$
   $$\Delta_{D_{\min}}(w)=E[\delta(x;q)\mid w=w] \tag{56b}$$
   we have
   $$\lim_{D\to D_{\min}}\Delta_D(w)=\Delta_{D_{\min}}(w) \tag{57}$$
   where the convergence is uniform for $w\in\mathcal{W}$;
   c) there exists an $\epsilon>0$ such that for all $w\in\mathcal{W}$, the maximizing entropy
   $$h_\delta(\Delta)\triangleq h(\mathcal{N}_\delta)=-\log a_\delta(\Delta)+s_\delta(\Delta)\Delta \tag{58}$$
   is a uniformly continuous function on the neighborhood $\{\Delta:|\Delta-\Delta_{D_{\min}}(w)|<\epsilon\}$.

First, some remarks. It should also be noted that the conditions of Definition 2 are not particularly hard to satisfy in practice. For example, any source, side information, and distortion measure where

$$-\infty<h(x|w=w)<\infty,\quad\forall\, w\in\mathcal{W} \tag{59a}$$
$$\exists\epsilon>0 \text{ s.t. } 0<E[\|x\|^\epsilon\mid w=w]<\infty,\quad\forall\, w\in\mathcal{W} \tag{59b}$$
$$d(x,\hat{x};q)=\gamma(q)\|x-\hat{x}\|^{\zeta(q)}+\tau(q) \tag{59c}$$

are admissible in the sense of Definition 2, where $\tau(\cdot),\gamma(\cdot)\ge 0$, and $\zeta(\cdot)\ge 0$ can be freely chosen.[9]

However, it should be emphasized that certain forms of distortion measure are not admissible in the sense of Definition 2, regardless of the nature of the source statistics (unless they are degenerate). An example is the distortion measure (7) in Claim 1.

[8]In practice, $\delta(\cdot)$ can often be taken to be the quadratic measure in problems where the relevant quantities of interest have finite variance.

[9]Note that the auxiliary distortion measure in this class of examples is $\delta(\cdot)=\|\cdot\|^\epsilon$.

It is also worth noting that for our asymptotic analysis to follow, only the local properties of the distortion measure are relevant—specifically, only the behavior of $\rho(x;q)$ in a neighborhood of $x=0$ for each q—in essence only the behavior of the Taylor series expansion of the measure about $x=0$ matters. However, to simplify the exposition, in the sequel we will express our distortion measure requirements in terms of global structure.

Finally, Definition 2 can be readily specialized to the cases of no distortion side information ($q=\varnothing$) or no signal side information ($w=\varnothing$) by eliminating dependences and conditioning on the absent side information throughout the definition.

The random variable v_D^* in Definition 2, which as a consequence of its maximum entropy property satisfies the Markov chain

$$v_D^* \leftrightarrow q \leftrightarrow (w,x), \tag{60}$$

plays a special role in our high-resolution analysis. In particular, it characterizes a fundamental Shannon lower bound of interest, as we now describe.

Lemma 1: When x, w, q satisfy Definition 1 and the distortion measure is of the form (48), then the conditional rate-distortion function

$$R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) = \inf_{\{p_{\hat{x}|x,q,w}:E[\rho(x-\hat{x};q)]\le D\}} I(\hat{x};x|q,w) \tag{61}$$

satisfies

$$R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \ge h(x|w) - h(v_D^*|q) \tag{62}$$

where v_D^* is the random variable defined in condition 3 of Definition 2.

Proof: Using $\hat{x}^*$ to denote the random variable corresponding to the optimizing distribution in (61), we obtain the conditional Shannon lower bound

$$\begin{aligned} R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) &= I(\hat{x}^*;x|q,w) \\ &= h(x|q,w) - h(x|q,w,\hat{x}^*) \\ &= h(x|w) - h(x-\hat{x}^*|q,w,\hat{x}^*) \quad (63)\\ &\ge h(x|w) - h(x-\hat{x}^*|q) \quad (64)\\ &\ge h(x|w) - h(v_D^*|q) \quad (65)\end{aligned}$$

where to obtain (63) we have used the Markov condition (4a), and where in (65) we have used the maximum-entropy property of v_D^*. □

The technical conditions of Definition 2 serve to ensure that a key "continuity of entropy" property is satisfied. This property can be viewed as a natural extension of the corresponding result [13, Theorem 1] to mixtures of entropy maximizing distributions. A proof is provided in Appendix IV.

Lemma 2: If x, w, q, and v_D^* satisfy the conditions of Definition 2, then

$$\lim_{D\to D_{\min}} h(x+v_D^*|w) = h(x|w). \tag{66}$$

The importance of Lemma 2 stems from the fact that it implies the Shannon lower bound above is tight in the high-resolution regime, i.e., an asymptotically optimal test-channel for the case of complete side information takes the form

$$\hat{x} = x + v_D^*. \tag{67}$$

To see this it suffices to note that [cf. (62)]

$$\begin{aligned} R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) &\le I(x+v_D^*;x|q,w) \quad (68)\\ &\le h(x+v_D^*|w) - h(v_D^*|q,w,x) \quad (69)\\ &= h(x+v_D^*|w) - h(v_D^*|q) \quad (70)\\ &\to h(x|w) - h(v_D^*|q) \quad (71)\end{aligned}$$

where to obtain (68) we have used the particular (rather than minimizing) test-channel choice (67) in (61), to obtain (69) we have used that conditioning cannot increase entropy, to obtain (70) we have used the Markov chain (60), and to obtain (71) we have used (66) in Lemma 2.

Finally, two special cases of Lemma 2 will be useful in our development. First, when there is no distortion side information, we obtain the following corollary.

Corollary 1: If x, w, and v_D^* satisfy the conditions of Definition 2 (when specialized to the case $q=\varnothing$), then

$$\lim_{D\to D_{\min}} h(x+v_D^*|w) = h(x|w) \tag{72}$$

where v_D^* is the random variable defined in Condition 3 of the specialization of Definition 2 to the case $q=\varnothing$, and hence is independent of w,x.

Note that here v_D^* maximizes $h(v)$ over all random variables v such that $E[\rho(v)] \le D$.

Second, when there is no signal side information, we obtain the following corollary.

Corollary 2: If x, q, and v_D^* satisfy the conditions of Definition 2 (when specialized to the case $w=\varnothing$), then

$$\lim_{D\to D_{\min}} h(x+v_D^*) = h(x) \tag{73}$$

where v_D^* is the random variable defined in Condition 3 of the specialization of Definition 2 to the case $w=\varnothing$, and hence v_D^*, q are independent of x.

Note, too, that the distribution corresponding to v_D^* is the same as when w *is* present; it is due to (4a) that v_D^* is independent of x when $w=\varnothing$. In fact, it is also straightforward to verify that when w is present, so v_D^* and x are in general correlated, (73) also holds, i.e., that we have continuity of not only conditional entropy as in (66), but of unconditional entropy as well. Specifically, we have the following.

Corollary 3: If x, w, q, and v_D^* satisfy the conditions of Definition 2, then

$$\lim_{D\to D_{\min}} h(x+v_D^*) = h(x). \tag{74}$$

C. Asymptotic Sufficiency Theorems

We begin with the following lemma establishing that, under our technical conditions, having the distortion side information at the encoder and the signal side information at the decoder is sufficient to ensure there is asymptotically no loss relative to the case of complete side information everywhere. A proof is provided in Appendix V-A.

Lemma 3: If Definition 2 is satisfied, then

$$\lim_{D\to D_{\min}} \Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}},\mathrm{W:}\overline{\mathrm{ENC}}]}(D) = 0 \tag{75}$$

where

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}},\mathrm{W:}\overline{\mathrm{ENC}}]}(D) = R_{[\mathrm{Q:ENC,W:DEC}]}(D) - R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \tag{76}$$

is the rate penalty for not knowing q^n at the decoder and w^n at the encoder.

Lemma 3 establishes that there is a natural division of side information between the encoder and decoder (at least asymptotically). Ultimately, this lemma can be viewed as generalizing prior results on the lack of rate loss for the Wyner–Ziv problem in the high-resolution limit [5]. In some ways, Lemma 3 is quite remarkable in its generality. The admissibility conditions (4) require q to be conditionally independent of x given w, and require the distortion to be conditionally independent of w^n given x^n, $\hat{x}^n$, q^n. However, since our model allows for q and w to be statistically dependent, q can be indirectly correlated with x (through w), and w^n can indirectly affect the distortion (through q^n).

It is also worth noting that Lemma 3 is readily specialized to the cases in which there is either no distortion side information q or no signal side information w in the problem. In the former case, the lemma specializes to the results of [5]: if Definition 2 is satisfied (when specialized to the case $q = \emptyset$), then

$$\lim_{D\to D_{\min}} \left[R_{[\mathrm{W:DEC}]}(D) - R_{[\mathrm{W:BOTH}]}(D)\right] = 0. \tag{77}$$

In the latter case, the lemma provides an extension of the results of Section IV-C that encompasses Example 2. In particular, the specialization is as follows.

Corollary 4: If Definition 2 is satisfied (when specialized to the case $w = \emptyset$), then

$$\lim_{D\to D_{\min}} \left[R_{[\mathrm{Q:ENC}]}(D) - R_{[\mathrm{Q:BOTH}]}(D)\right] = 0. \tag{78}$$

Note that satisfying the specialized version of Definition 2 in this case means, among other requirements, that x must be independent of q.

Our second pair of theorems extend Lemma 3 to show that regardless of where signal (respectively, distortion) side information is constrained to be available, having the distortion (respectively, signal) side information at the encoder (respectively, decoder) results in the best possible performance attainable subject to that constraint. Proofs are provided in Appendices V-B and V-C, respectively.

Theorem 6: If Definition 2 is satisfied, then

$$\lim_{D\to D_{\min}} \left[R_{[\mathrm{Q:}*,\mathrm{W:DEC}]}(D) - R_{[\mathrm{Q:}*,\mathrm{W:BOTH}]}(D)\right] = 0 \tag{79}$$

where $* \in \{\mathrm{ENC,BOTH}\}$ (both *'s are identical). Moreover, (79) also holds for $* \in \{\mathrm{DEC,NONE}\}$ provided q and w are independent, and the difference distortion measure is of the scaled form

$$d(x, \hat{x}; q) = \gamma(q)\rho(x - \hat{x}) + \tau(q) \tag{80}$$

for some functions $\gamma(\cdot) \geq 0$, $\tau(\cdot)$, and $\rho(\cdot)$.

Theorem 7: If Definition 2 is satisfied, then

$$\lim_{D\to D_{\min}} \left[R_{[\mathrm{Q:ENC,W:}*]}(D) - R_{[\mathrm{Q:BOTH,W:}*]}(D)\right] = 0 \tag{81}$$

where $* \in \{\mathrm{DEC,BOTH}\}$ (both $*$'s are identical). Moreover, (81) also holds for $* \in \{\mathrm{ENC,NONE}\}$ provided q and w are independent.

In essence, Theorems 6 and 7 collectively establish an approximation result: that, under reasonable conditions, the closer one can get to the ideal of providing q^n to the encoder and w^n to the decoder implied by Lemma 3, the better the system will perform.

VI. Rate Loss for Misplaced Side Information

While the results of Section V-C establish that providing distortion side information to the encoder and signal side information to the decoder is best, in this section we quantify the loss incurred by deviations from this ideal. In particular, our results take the form of yet another pair of theorems, which respectively characterize the rate loss when signal side information is not available at the decoder, and when distortion side information is not available at the encoder. Finally, corollaries of each of these theorems establish how statistical dependencies between the two types of side information influence the associated losses.

Our two theorems are as follows.

Theorem 8: If Definition 2 is satisfied, then the penalty

$$\Delta R_{[\mathrm{Q:}*,\mathrm{W:}\overline{\mathrm{DEC}}]}(D) = R_{[\mathrm{Q:}*,\mathrm{W:ENC}]}(D) - R_{[\mathrm{Q:}*,\mathrm{W:BOTH}]}(D) \tag{82}$$

for not knowing w^n at the decoder satisfies

$$I(x; w|q) \leq \lim_{D\to D_{\min}} \Delta R_{[\mathrm{Q:}*,\mathrm{W:}\overline{\mathrm{DEC}}]}(D) \leq I(x; w) \tag{83}$$

where $* \in \{\mathrm{ENC,BOTH}\}$ (both $*$'s are identical), where the left-hand inequality holds with equality for $* = \mathrm{BOTH}$, and where either left- or right-hand inequalities can hold with equality if $* = \mathrm{ENC}$. Moreover, (83) also holds for $* \in \{\mathrm{DEC,NONE}\}$ provided q and w are independent, and the difference distortion measure is of the scaled form (80); the upper and lower bounds coincide in this case.

Theorem 9: If Definition 2 is satisfied, $\mathcal{X} = \mathbb{R}^k$, and the difference distortion measure is of the particular scaled form

$$d(x, \hat{x}; q) = q\|x - \hat{x}\|^r \tag{84}$$

for some $r > 0$ and with $\min_{\mathcal{Q}} q > 0$, then the penalty

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}},\mathrm{W:}*]}(D) = R_{[\mathrm{Q:DEC,W:}*]}(D) - R_{[\mathrm{Q:BOTH,W:}*]}(D) \tag{85}$$

for not knowing q^n at the encoder satisfies

$$\frac{k}{r}E\left[\log\frac{E[q|w]}{q}\right] \le \lim_{D\to 0}\Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:*]}(D) \le \frac{k}{r}E\left[\log\frac{E[q]}{q}\right] \quad (86)$$

where $* \in \{\mathrm{DEC},\mathrm{BOTH}\}$ (both $*$'s are identical), where the left-hand inequality holds with equality for $* = \mathrm{BOTH}$, and where either left- or right-hand inequalities can hold with equality if $* = \mathrm{DEC}$.[10] Moreover, (86) also holds for $* \in \{\mathrm{ENC},\mathrm{NONE}\}$ provided q and w are independent; the upper and lower bounds coincide in this case.

It should be noted that the cases $* = \mathrm{ENC}$ in Theorem 8 and $* = \mathrm{DEC}$ in Theorem 9 are somewhat special, and their more complicated behavior is a consequence of the potential (unconditional) dependency between x and q.

Proofs of these theorems are provided in Appendix V-B and V-C, respectively. That for Theorem 8 makes use of the following lemma, whose proof is provided first in Appendix VI-A.

Lemma 4: If Definition 2 is satisfied (when specialized to the case $q = \emptyset$), and the distortion measure is of the difference form $d(x,\hat{x}) = \rho(x-\hat{x})$, then

$$\lim_{D\to D_{\min}}\left[R_{[\mathrm{W}:\mathrm{ENC}]}(D) - R_{[\mathrm{W}:\mathrm{BOTH}]}(D)\right] = I(x;w). \quad (87)$$

As the counterpart to Lemma 4, when there is no signal side information, Theorem 9 specializes as follows.

Corollary 5: If Definition 2 is satisfied (when specialized to the case $w = \emptyset$), $\mathcal{X} = \mathbb{R}^k$, and the distortion measure is of the form (84) for some $r > 0$, then

$$\lim_{D\to 0}\left[R_{[\mathrm{Q}:\mathrm{DEC}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH}]}(D)\right] = \frac{k}{r}E\left[\log\frac{E[q]}{q}\right]. \quad (88)$$

To verify Corollary 5, we note that there being no signal side information is equivalent to there being signal side information w^n that is independent of q^n and available at neither encoder nor decoder. Indeed, then

$$\begin{aligned} R_{[\mathrm{Q}:\mathrm{DEC}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH}]}(D) &= R_{[\mathrm{Q}:\mathrm{NONE}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH}]}(D) && (89)\\ &= R_{[\mathrm{Q}:\mathrm{NONE},\mathrm{W}:\mathrm{NONE}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH},\mathrm{W}:\mathrm{NONE}]}(D) && (90)\\ &= R_{[\mathrm{Q}:\mathrm{DEC},\mathrm{W}:\mathrm{NONE}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH},\mathrm{W}:\mathrm{NONE}]}(D) && (91)\end{aligned}$$

where to obtain (89) we have used Theorem 3, and where to obtain (91) we have used the case $* = \mathrm{NONE}$ of Theorem 5. Finally, taking the limit of (91) as $D \to 0$ and applying the case $* = \mathrm{NONE}$ of Theorem 9 yields (88).

In Table II, we evaluate the high-resolution rate penalty (88) for a number of possible distortion side-information distributions. Note that for all of these side information distributions $p_q(q)$ (except the uniform and exponential distributions), the rate penalty can be made arbitrarily large by choosing the appropriate shape parameter to place more probability near $q = 0$ or $q = \infty$. In the former case (lognormal, gamma, or pathological q), the large rate loss occurs because when $q \approx 0$, the informed encoder can transmit almost zero rate while the uninformed encoder must transmit a large rate to achieve high resolution. In the latter case (Pareto or Cauchy q), the large rate loss is caused by the heavy tails of the distribution for q. Specifically, even though q is big only very rarely, it is the rare samples of large q that dominate the moments. Thus, an informed encoder can describe the source extremely accurately during the rare occasions when q is large, while an uninformed encoder must always spend a large rate to obtain a low average distortion.[11]

It is also worth observing that the effects of Theorems 8 and 9 are cumulative, as the following simple corollary indicates.

Corollary 6: If Definition 2 is satisfied, $\mathcal{X} = \mathbb{R}^k$, q and w are independent, and the difference distortion measure is of the particular scaled form (84) for some $r > 0$, then the penalty

$$\Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:\overline{\mathrm{DEC}}]}(D) = R_{[\mathrm{Q}:\mathrm{DEC},\mathrm{W}:\mathrm{ENC}]}(D) - R_{[\mathrm{Q}:\mathrm{BOTH},\mathrm{W}:\mathrm{BOTH}]}(D) \quad (92)$$

for not knowing q^n at the encoder and w^n at the decoder is, asymptotically,

$$\lim_{D\to 0}\Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:\overline{\mathrm{DEC}}]}(D) = I(x;w) + \frac{k}{r}E\left[\log\frac{E[q]}{q}\right]. \quad (93)$$

To verify (93), it suffices to note that

$$\Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:\overline{\mathrm{DEC}}]}(D) = \Delta R_{[\mathrm{Q}:\mathrm{BOTH},\mathrm{W}:\overline{\mathrm{DEC}}]}(D) + \Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:\mathrm{ENC}]}(D) \quad (94)$$

and evaluate the first and second terms of (94) using Theorems 8 and 9, respectively, in the high-resolution limit.

Finally, Theorems 8 and 9 emphasize the case when the side information decomposes naturally into independent signal side information and distortion side information components. When such a decomposition is not possible, it is straightforward to characterize the losses associated with not having the side information everywhere, as we now develop.

Consider a general side information z that both influences the distortion measure via $d(x,\hat{x};z) = \gamma(z)\rho(x-\hat{x}) + \tau(z)$ and is correlated with the source. Then we have the following corollaries of Theorems 8 and 9, respectively.

Corollary 7: If Definition 2 is satisfied with $q = w = z$, and the difference distortion measure is of the scaled form

$$d(x,\hat{x};z) = \gamma(z)\rho(x-\hat{x}) + \tau(z)$$

for some $\gamma(\cdot) \ge 0$, $\tau(\cdot)$, and $\rho(\cdot)$, then the asymptotic penalty for knowing general side information z only at the encoder is bounded via

$$0 \le \lim_{D\to D_{\min}} R_{[\mathrm{Z}:\mathrm{ENC}]}(D) - R_{[\mathrm{Z}:\mathrm{BOTH}]}(D) \le I(x;z) \quad (95)$$

where either the left- or right-hand inequalities can hold with equality.

[10]The $* = \mathrm{DEC}$ case, when $k = 1$ and $r = 2$ specifically, can be deduced, in part, directly from the results of [6].

[11]As an aside, note that while encoder-only distortion side information is asymptotically sufficient for scenarios of Table II, the simple strategy of exploiting encoder-only side information by separately encoding the side information for the decoder is suboptimal. Indeed, for all but one of the distributions in Table II, infinite rate would be required to represent the side information without loss.

TABLE II
ASYMPTOTIC RATE LOSS FOR NOT KNOWING DISTORTION SIDE INFORMATION q AT THE ENCODER WHEN DISTORTION IS MEASURED VIA $d(x,\hat{x};q) = q(x-\hat{x})^2$ AND $\mathcal{X} = \mathbb{R}$

Density $p_q(q)$		**Rate Loss** $\lim_{D\to 0} \Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}}]}(D)$ (nats)
Exponential	$\tau \exp(-q\tau)$	$-\frac{1}{2}\ln\gamma \approx 0.2748$ (γ is Euler's constant)
Uniform	$1_{q\in[0,1]}$	$\frac{1}{2}(1-\ln 2) \approx 0.1534$
Lognormal	$\frac{1}{q\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln q - M)^2}{2\sigma^2}\right]$	$\frac{\sigma^2}{4}$
Pareto	$\frac{a^b}{q^{a+1}}, \quad q \geq b > 0, \quad a > 1$	$\frac{1}{2}\left[\ln\frac{a}{a-1} - \frac{1}{a}\right]$
Gamma	$\frac{b\,(bq)^{a-1}\exp(-bq)}{\Gamma(a)}$	$\frac{1}{2}\left\{\ln a - \frac{d}{dx}[\ln\Gamma(x)]\big\vert_{x=a}\right\} \approx \frac{1}{2a}$
Pathological	$(1-\epsilon)\,\delta(q-\epsilon) + \epsilon\,\delta\left(q-\frac{1}{\epsilon}\right)$	$\frac{1}{2}\ln(1+\epsilon-\epsilon^2) - \frac{1-2\epsilon}{2}\ln\epsilon \approx \frac{1}{2}\ln\frac{1}{\epsilon}$
Positive Cauchy	$\frac{2/\pi}{1+q^2}, \quad q \geq 0$	∞

Corollary 8: If Definition 2 is satisfied with $q = w = z, \mathcal{X} = \mathbb{R}^k$, and the difference distortion measure is of the particular scaled form

$$d(x,\hat{x};z) = z\|x-\hat{x}\|^r$$

for some $r > 0$, the asymptotic penalty for not knowing z at the encoder is bounded via

$$0 \leq \lim_{D\to 0} R_{[\mathrm{Z:DEC}]}(D) - R_{[\mathrm{Z:BOTH}]}(D) \leq \frac{k}{r} E\left[\log\frac{E[z]}{z}\right] \tag{96}$$

where either the left- or right-hand inequalities can hold with equality.

In essence, Corollary 7 establishes that not having general side information at the decoder can only incur a loss if that side information is correlated with the source, while Corollary 8 establishes that not having such side information at the encoder can only incur a loss if that side information influences the distortion measure. However, the corollaries also make clear that there need not be a loss in either case, depending on the specific nature of the distortion measure and side information.

To obtain both Corollaries 7 and 8, it suffices to i) let $q = w = z$ in Theorems 8 and 9, respectively, for the cases $* = \mathrm{ENC}$ and $* = \mathrm{DEC}$, respectively, taking into account the respective footnotes in these theorems; and ii) note that $R_{[\mathrm{Q:ENC,W:BOTH}]}(D) = R_{[\mathrm{Q:BOTH,W:DEC}]}(D) = R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$ when $q = w = z$. Also, the examples for meeting respective left-hand and right-hand inequalities with equality in Theorems 8 and 9 also lead to the corresponding left-hand and right-hand inequalities in Corollaries 7 and 8 being met with equality, since in these examples $q = w$.

VII. SUFFICIENCY RATE GAP AT LOWER RESOLUTIONS

In Section V, we showed that in the quantization of continuous sources, the rate loss due to encoder-only distortion side information and decoder-only signal side information vanishes in the high-resolution limit. In this section, we provide a finer grained analysis of this behavior, upper-bounding how quickly this rate loss decays with increasing resolution. In our treatment, q and w are not constrained to be independent. To simplify our analysis, we restrict our attention to scaled quadratic distortion measures, but briefly suggest how these results can be generalized to other distortion measures. In the sequel, we separately characterize behavior at medium and low resolutions.

A. Medium Resolution

The following theorem bounds the rate penalty incurred by incomplete side information at moderate resolutions; a proof is provided in Appendix VII-A.

Theorem 10: If Definition 1 is satisfied, and the distortion measure is of the particular scaled-difference form (84) with $k = 1$ and $r = 2$, i.e.,

$$d(x,\hat{x};q) = q(x-\hat{x})^2 \tag{97}$$

with $q_{\min} = \min_{\mathcal{Q}} q > 0$, then the rate gap (76) at distortion D is bounded by

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}},\mathrm{W:}\overline{\mathrm{ENC}}]}(D) \leq \frac{\log(e)}{2} J(x|w) \min\left\{\frac{D}{q_{\min}}, \max_{q\in\mathcal{Q}} E\big[\mathrm{var}[x|w] \;\big|\; q = q\big]\right\} \tag{98}$$

where $J(x|w)$ is the Fisher information in estimating a nonrandom parameter τ from $\tau + x$ conditioned on knowing w, i.e.,

$$J(x|w) \triangleq \int p_w(w)dw \int p_{x|w}(x|w)\left[\frac{\partial}{\partial x}\ln p_{x|w}(x|w)\right]^2 dx. \tag{99}$$

A few remarks are worthwhile. First, there is an intuitively satisfying interpretation of the Fisher information factor in our bound. In particular, since $J(x|w)$ is inversely related to the accuracy with which x can be estimated from the additive test channel output (67), it is a measure of the degree to which one can further improve on the additive test channel(67) for achieving a better rate–distortion tradeoff for the encoder-only

distortion side information and decoder-only signal side information configuration. Thus, a large value of $J(x|w)$ means that a large further improvement is possible, so the test channel is far from optimum, which is reflected in a large rate gap. In turn, this implies that the value of the bound is, as a result, limited by the degree to which the Fisher information is a good measure of the accuracy to which estimation is possible; in general it is, itself, of course, only a bound on such accuracy.

Second, we emphasize that the proof of Theorem 10 does not require any extra regularity conditions such as those of Definition 2—whenever the Fisher information of the source is finite, the bound can be applied.

Finally, in principle, similar bounds can be obtained for other distortion measures. A possible approach is suggested at the end of Appendix VII-A. Also, related bounds are discussed in [14, Appendix D].

B. Low Resolution

The Fisher information bound (98) can be quite poor in the low-resolution regime if the source is not smooth. For such scenarios, we develop an alternative bound, which is independent of the distortion level and hence most useful at low resolution. A proof is provided in Appendix VII-B.

Theorem 11: If Definition 1 is satisfied, and the distortion measure is of the particular scaled-difference form (97), then the rate gap (76) at any distortion is at most

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}},\mathrm{W:}\overline{\mathrm{ENC}}]}(D) \leq \frac{1}{2}\log\left(1+\frac{\sigma^2_{\max}}{\sigma^2_{\min}}\right) + D(p_{x|w} \,||\, \mathcal{N}(x|w)), \quad (100)$$

where[12]

$$\sigma^2_{\min} = \min_w \mathrm{var}[x|w=w]$$
$$\sigma^2_{\max} = \max_w \mathrm{var}[x|w=w] \quad (101)$$

and where

$$D(p_{x|w} \,||\, \mathcal{N}(x|w)) = \int D(p_{x|w}(\cdot|w) \,||\, \mathcal{N}(x|w=w))p_w(w)dw \quad (102)$$

with the notation $\mathcal{N}(t)$ defined as follows: for a random variable t corresponding to any distribution p_t (having finite first and second moments), $\mathcal{N}(t)$ denotes a Gaussian distribution with mean $E[t]$ and variance $\mathrm{var}[t]$.

Again, we make some remarks. First, the bound (100) can be readily evaluated in various cases of interest. As an example, in the Wyner–Ziv scenario where x and w are jointly Gaussian, the divergence term on the right-hand side of (100) vanishes, and moreover $\mathrm{var}[x|w=w]$ does not depend on w, so we obtain that the rate loss is at most 1/2 bit/sample, as in [5].

Second, we note that both the bounds (98) and (100) depend on the source distribution, in contrast to, e.g., the bound of [5]. As a result, we conjecture that our bounds are loose. In particular, for a discrete source, the worst case rate loss is at most $H(x|w)$, yet this is not reflected by our bounds since both become infinite in this case. Exploiting techniques from [5], [15], [16] may yield tighter bounds.

Finally, as with our medium-resolution bound, in principle, similar bounds can be developed for other distortion measures; a possible approach is suggested at the end of Appendix VII.

C. Example: A Gaussian-Quadratic Case

To gain some sense for when the asymptotic results take effect, we consider an example involving a Gaussian source and binary-valued distortion side information. Specifically, we consider a zero-mean, unit-variance Gaussian source x, and the quadratic distortion measure $d(x,\hat{x};q) = q(x-\hat{x})^2$, where the distortion side information q is independent of x and $\Pr[q=q_1] = p$, $\Pr[q=q_2] = 1-p$. Without loss of generality, we let $q_1 < q_2$.

The case without side information is equivalent to quantizing a Gaussian random variable with distortion measure $\overline{q}(x-\hat{x})^2$ where

$$\overline{q} = E[q] = pq_1 + (1-p)q_2 \quad (103)$$

and thus the rate-distortion function is

$$R_{[\mathrm{Q:NONE}]}(D) = \begin{cases} 0, & D \geq \overline{q} \\ \frac{1}{2}\log\frac{\overline{q}}{D}, & D < \overline{q} \end{cases} \quad (104)$$

which, of course, is one upper bound on $R_{[\mathrm{Q:ENC}]}(D)$.

To determine $R_{[\mathrm{Q:BOTH,W:NONE}]}(D)$ we must set up a constrained optimization as we did for the binary-Hamming scenario in Appendix II. This optimization results in a "waterfilling" bit allocation, which uses more bits to quantize the source when $q=q_2$ than when $q=q_1$. Specifically, the optimal test channel is a Gaussian distribution where both the mean and the variance depend on q and, thus, $\hat{x}$ has a Gaussian mixture distribution. The resulting rate–distortion tradeoff is

$$R_{[\mathrm{Q:BOTH}]}(D) = \begin{cases} 0, & D \geq \overline{q} \\ \frac{1-p}{2}\log\frac{(1-p)q_2}{(D-pq_1)}, & q_1 \leq D < \overline{q} \\ \frac{p}{2}\log\frac{q_1}{D} + \frac{1-p}{2}\log\frac{q_2}{D}, & D < q_1. \end{cases} \quad (105)$$

Expressions (104) and (105) can be compared in the high-resolution regime via (88). In particular, the asymptotic rate loss (88) for not having the side information at the encoder evaluates to

$$\lim_{D\to 0} \Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}}]}(D) = \frac{1}{2}\log\frac{\overline{q}}{q_1^p q_2^{1-p}}. \quad (106)$$

To assess the rate $R_{[\mathrm{Q:ENC}]}(D)$ achievable by encoder-only side information, we first obtain a good numerical upper bound by evaluating (6c) with the codebook distribution that optimizes (6d).[13] Comparing (6c) to (6d), we see that with this approach the resulting rate loss for not having the side information at the decoder is at most

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}}]}(D) \leq I(\hat{x};q) \quad (107)$$

with this choice of codebook distribution. In turn, this numerical bound can be compared to the medium-resolution analytic

[12]Obviously, the bound is useful only if $\sigma^2_{\min} > 0$.

[13]Actually, we further optimize our bound by allowing time sharing in the code, which corresponds to taking the lower convex envelope of the rate–distortion tradeoff resulting from this choice of test-channel distribution.

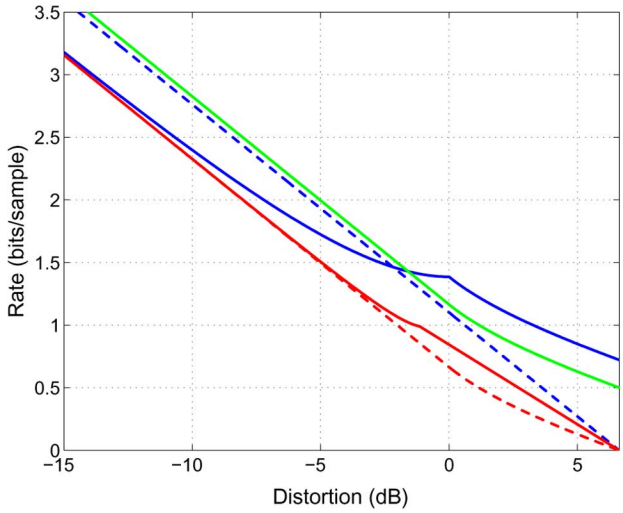


Fig. 7. Rate-distortion tradeoffs for quantizing a zero-mean, unit-variance Gaussian source x with binary side information q, where the distortion in the reconstruction $\hat{x}$ is measured via $d(x, \hat{x}; q) = q(x - \hat{x})^2$. The top and bottom dashed curves correspond to q being at neither encoder nor decoder (104), and q being at both encoder and decoder (105), respectively. From top to bottom on the left-hand side, the solid curves correspond to the analytic bounds (109) and (108), and the numerical bound (107), respectively, on the tradeoffs for q being available only at the encoder. In this example, $\Pr[q = 1] = 3/5$ and $\Pr[q = 10] = 2/5$, i.e., $q_1 = 1$, $q_2 = 10$, and $p = 3/5$.

bound (98) of Theorem 10, which establishes that the rate loss satisfies

$$\Delta R_{[\mathrm{Q}:\overline{\mathrm{DEC}}]}(D) \leq \frac{\log e}{2} \min\left(1, \frac{D}{q_1}\right) \tag{108}$$

in this example. Finally, since x is Gaussian, the low-resolution analytic bound (100) of Theorem 11 is simply

$$\Delta R_{[\mathrm{Q}:\overline{\mathrm{DEC}}]}(D) \leq 1/2 \text{ bit/sample} \tag{109}$$

in our example.

Fig. 7 illustrates the four resulting rate–distortion tradeoffs (104)–(109) for the case in which $q_1 = 1, q_2 = 10$, and $p = 3/5$.

At low rates ($R \ll 1/2$ bit/sample), the rate-distortion functions for the cases of no side information, encoder-only side information, and full side information are close and all converge to the same distortion in the limit since no bits are available for quantization. We also see that our medium-resolution bound (108) is loose at low rates, as is the low-rate bound (109).

At middle resolutions (e.g., in the vicinity of $R = 1/2$ bit/sample, the system with full side information does best because it uses the few available bits to represent only the important source samples (i.e., those for which $q = q_2 = 10$); the associated codebook distribution is a Gaussian mixture. The decoder reconstructs these source samples from the compressed data and reconstructs the less important samples to zero (their mean). In this regime, the system with side information only at the encoder also more accurately quantizes the important source samples. But since the decoder does not know q^n, it does not know which samples of $\hat{x}^n$ to reconstruct to zero, and thus does not perform as well as the system with full side information.

At high rates ($R \gg 1/2$ bit/sample), we see it is sufficient to have side information at the encoder alone. Such systems quantize both more and less important source samples, and the codebook distribution becomes increasingly Gaussian. Note that our numerical bound establishes that beyond even the moderate rate of $R = 1$ bit/sample, the asymptotic sufficiency effects promised by (81) in Theorem 7 can be seen—the benefit of decoder side information is fairly negligible. Note, also, that our medium-resolution analytic bound is still somewhat loose at such rates—it does not start to get tight until closer to $R = 2$ bits/sample. Finally, evaluating (106) for this example, we obtain that the rate loss for not having the side information at the encoder approaches $0.5 \cdot (\log 4.6 - 0.4 \log 10) \approx 0.44$ bits/sample in the high rate limit, as reflected in Fig. 7.

VIII. Concluding Remarks

Through the framework of this paper, we have developed a variety of insights on problems of source coding with side information. Foremost, our analysis indicates that side information that affects the distortion measure can provide significant benefits in compression systems. Perhaps most surprisingly, in a number of cases we also show that having such side information at the encoder alone is just as effective as having it at both encoder and decoder. Furthermore, this "separation theorem" can be combined with the prior result that exploiting signal side information at the decoder is often as effective as exploiting it at both encoder and decoder.

Equally interesting, as an extension of Berger's well-known result on the inefficacy of misplaced side information, we show that not only is encoder-only signal side information generally of no value, but the same is true of decoder-only distortion side information. Our collective results on the impact of having side information in different configurations are summarized in Table I.

More generally, we determine the rate loss for lacking a particular type of side information where it is needed (relative to having full side information available). The resulting theorems show that poorly chosen side information configurations can incur arbitrarily large performance penalties.

Among many possible applications, one area where distortion side information may provide benefits in practice is in designing perceptual coders, which exploit features of the human visual and/or auditory systems to achieve low subjective distortion even when the objective distortion (e.g., mean-square error) is quite large. Recent examples of such systems have shown gains in image coding; see, e.g., [17]. Unfortunately, current systems often communicate the distortion side information (in the form of model parameters or quantizer step sizes) explicitly to the decoder and thus are not as efficient as they could be, as the results of our analysis in this paper reveal. Perhaps more importantly, the abstraction of distortion side information developed in this paper may facilitate future perceptual coder design by allowing quantizer design issues to be considered separately from semantic modeling issues.

Obviously our treatment is a deliberately coarse-scale one. Indeed, we have idealized several aspects of the problem, hiding finer scale detail that will be no doubt ultimately be important to apply this framework. In practice, for example, the distortion side information may not be independent of the source, the source itself may not be memoryless, the distortion measure may not be additive, and there may be a noisy channel through

which the source representation must be conveyed. Thus, there remain several questions and issues for further research.

At least in principle, a number of the required generalizations can be developed. For example, while Corollary 7 indicates that knowing general side information z^n only at the encoder may be suboptimal, the loss is essentially due to lack of signal side information. In particular, even when distortion side information known only at the encoder is correlated with the source, the fixed-codebook/variable-partition code architecture outlined in Section IV and further developed in [11] can still provide significant benefits.

In addition, we believe that information spectrum techniques [18]–[20] can be used to establish both that the familiar source–channel separation theorem holds when a communication channel is also involved, and that the results developed here can be extended to stationary, ergodic scenarios.

Finally, developing efficient practical systems based on the framework, principles, architecture, and techniques of this paper will introduce many interesting issues and constraints, and will obviously be a substantial undertaking in and of itself.

Appendix I
Proof of Theorem 2

Let $p^0_{\hat{x}|x,q}(\hat{x}|x,q)$ be an optimal test-channel distribution. By symmetry, for any $t \in \mathcal{X}$, the shifted distribution

$$p^t_{\hat{x}|x,q}(\hat{x}|x,q) \triangleq p^0_{\hat{x}|x,q}(\hat{x} \oplus t | x \oplus t, q) \tag{110}$$

must also be an optimal test-channel. Since mutual information is convex in the test-channel distribution, yet another optimal test-channel distribution is

$$p^*_{\hat{x}|x,q}(\hat{x}|x,q) \triangleq \int_{\mathcal{X}} p^t_{\hat{x}|x,q}(\hat{x}|x,q) d_{\mathcal{X}}(t) \tag{111}$$

where $d_{\mathcal{X}}(t)$ is the uniform measure. To see that the resulting distribution for $\hat{x}$ given q is uniform for all q (and hence independent of q), it suffices to note that we have, for any $r \in \mathcal{X}$

$$\begin{aligned}
p^*_{\hat{x}|q}(\hat{x}|q) &= \int_{\mathcal{X}} p^*_{\hat{x}|x,q}(\hat{x}|x',q) d_{\mathcal{X}}(x') && (112)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^{t'}_{\hat{x}|x,q}(\hat{x}|x',q) d_{\mathcal{X}}(t') d_{\mathcal{X}}(x') && (113)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^0_{\hat{x}|x,q}(\hat{x} \oplus t' | x' \oplus t', q) d_{\mathcal{X}}(t') d_{\mathcal{X}}(x') && (114)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^0_{\hat{x}|x,q}(\hat{x} \oplus r \oplus t | x' \oplus r \oplus t, q) \\
&\quad \times d_{\mathcal{X}}(r \oplus t) d_{\mathcal{X}}(x') && (115)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^0_{\hat{x}|x,q}(\hat{x} \oplus r \oplus t | x' \oplus r \oplus t, q) \\
&\quad \times d_{\mathcal{X}}(t) d_{\mathcal{X}}(x' \oplus r) && (116)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^0_{\hat{x}|x,q}(\hat{x} \oplus r \oplus t | x \oplus t, q) d_{\mathcal{X}}(t) d_{\mathcal{X}}(x) && (117)\\
&= p^*_{\hat{x}|q}(\hat{x} \oplus r | q) && (118)
\end{aligned}$$

where (112) follows from Bayes' law and the fact that $d_{\mathcal{X}}$ is the uniform measure on $\mathcal{X}$, where (113) follows from (111), and where (114) follows from (110). To obtain (115), we make the change of variable $t' \to r \oplus t$, and then apply the fact that the uniform measure is shift invariant to obtain (116). Similarly, we make the change of variable $x' \oplus r \to x$ to obtain that (117) and (118) follows from (111).

Note that this argument applies regardless of whether the side information is available at the encoder, decoder, both, or neither. □

Appendix II
Derivation of Binary-Hamming Rate–Distortion Functions (41) and (42)

The first equality in (41) follows directly from Theorem 2. To establish the second equality in (41), we note that the optimal rate–distortion code corresponds to simultaneous description of independent random variables [21, Sec. 13.3.3]. Specifically, the source samples for each value of q can be quantized separately using the distribution

$$p_{\hat{x}|x,q}(\hat{x}|x,q) = \begin{cases} 1-\epsilon_q, & \hat{x} = x \\ \epsilon_q, & \hat{x} \neq x \end{cases} \tag{119}$$

for appropriate choices of ϵ_q, $q = 1, 2, \ldots, N$. These ϵ_q correspond to a rate allocation for each value of the side information and are obtained as the values that yield the solution to the constrained optimization problem

$$R_{[\mathrm{Q:BOTH}]}(D) = \min_{\{\epsilon_1,\ldots,\epsilon_N : E[d(x,\hat{x};q)]=D\}} E\left[1 - H_{\mathrm{B}}(\epsilon_q)\right]. \tag{120}$$

Using Lagrange multipliers, we construct the functional

$$\begin{aligned}
&J(\epsilon_1, \ldots, \epsilon_N) \\
&\quad = \sum_{q=1}^{N} p_q(q) \left[1 + \epsilon_q \log \epsilon_q + (1-\epsilon_q)\log(1-\epsilon_q)\right] \\
&\qquad + \lambda \sum_{q=1}^{N} p_q(q) \left[\gamma(q)\epsilon_q + \tau(q)\right]
\end{aligned}$$

whose minimum is easily shown to be attained by the choice

$$\epsilon_q = \frac{2^{-\lambda\gamma(q)}}{1 + 2^{-\lambda\gamma(q)}} \tag{121}$$

where λ is chosen to meet the distortion constraint, whence (41).

Next, the first equality in (42) follows immediately from Theorem 3. To establish the second equality in (42), it suffices to note that the problem is equivalent to conventional quantization a symmetric binary source, but with the side-information-averaged distortion measure

$$d'(x,\hat{x}) = E[\gamma(q)] d_{\mathrm{H}}(x,\hat{x}) + E[\tau(q)]. \tag{122}$$

Thus, the relevant distortion–rate function is simply an affine transformation of the familiar distortion–rate function for the canonical binary-Hamming case, whence the right-hand side of (42).

$$\lim_{D\to D_{\min}}\big[h(x+v_D^*|w)-h(x|w)\big] \le \limsup_{D\to D_{\min}}\int\Big[D\big(p_{x|w}(\cdot|w)\,\big\|\,\mathcal{N}_\delta(x|w=w)\big)-D\big(p_{x+v_D^*|w}(\cdot|w)\,\big\|\,\mathcal{N}_\delta(x+v_D^*|w=w)\big)\Big]p_w(w)dw$$

$$+\lim_{D\to D_{\min}}\int\Big[h\big(\mathcal{N}_\delta(x+v_D^*|w=w)\big)-h\big(\mathcal{N}_\delta(x|w=w)\big)\Big]p_w(w)dw \tag{138}$$

$$\le \lim_{D\to D_{\min}}\int\big[h_\delta(\Delta_D(w))-h_\delta(\Delta_{D_{\min}}(w))\big]p_w(w)dw \tag{139}$$

$$=\int\lim_{D\to D_{\min}}\big[h_\delta(\Delta_D(w))-h_\delta(\Delta_{D_{\min}}(w))\big]p_w(w)dw \tag{140}$$

$$=0, \tag{141}$$

Appendix III
Inefficacy Theorem Proofs

Proof of Theorem 4: When $*\in\{\text{ENC,BOTH}\}$, the encoder can simulate w^n by generating it from (x^n,q^n). When $*\in\{\text{DEC,NONE}\}$

$$p_{w|x,q}(w|x,q)=\frac{p_{w|x}(w|x)p_{q|w}(q|w)}{p_{q|x}(q|x)}=p_{w|x}(w|x), \tag{123}$$

where the second equality follows from the independence of w and q (which, in turn, implies, via the Markov condition (4a), that q and x are also independent). Thus, in this case, the encoder can still simulate w^n correctly. Hence, w^n has no value at the encoder, and (45) follows. □

Proof of Theorem 5: We establish (46) for the four cases separately.

Case: $*=$ DEC*:* We first apply familiar the Wyner–Ziv rate–distortion formula [1] with the super-side-information $\tilde{w}=(w,q)$ to obtain

$$R_{[\text{Q:DEC,W:DEC}]}(D)=\inf_{\{p_{u|x},g:E[d(x,g(u,q,w);q)]\le D}I(u;x|w,q). \tag{124}$$

The optimizing $g(\cdot,\cdot,\cdot)$ in (124) is given by

$$g(u,q,w)=\arg\min_{\hat{x}}E[d(x,\hat{x};q)|q=q,w=w,u=u] \tag{125}$$

$$=\arg\min_{\hat{x}}E[\rho(x,\hat{x})|q=q,w=w,u=u] \tag{126}$$

$$=\arg\min_{\hat{x}}E[\rho(x,\hat{x})|w=w,u=u] \tag{127}$$

where to obtain (126) we have substituted the form of the distortion measure (32), and to obtain (127) we have used the Markov relation $q\leftrightarrow w\leftrightarrow x\leftrightarrow u$, which follows from (4a) and the Markov chain $u\leftrightarrow x\leftrightarrow w,q$ implicit in (124). In turn, (127) implies that the reconstruction does not depend on q^n, and thus the distortion side information is not exploited at the decoder, whence

$$R_{[\text{Q:DEC,W:DEC}]}(D)=R_{[\text{Q:NONE,W:DEC}]}(D). \tag{128}$$

Case: $*=$ BOTH*:* In this scenario, the encoder and decoder agree on a different source coding subsystem for each value of w. The subsystem for a fixed value w corresponds to source coding with distortion side information at the decoder. Specifically, the source has distribution $p_{x|w}(\cdot|w)$, and the distortion side information has distribution $p_{q|w}(\cdot|w)$. The performance of each such subsystem is given by $R^w_{[\text{Q:DEC}]}(D)$, which via Theorem 3 equals the corresponding $R^w_{[\text{Q:NONE}]}(D)$, whence

$$R_{[\text{Q:DEC,W:BOTH}]}(D)=R_{[\text{Q:NONE,W:BOTH}]}(D). \tag{129}$$

Case: $*=$ NONE*:* For this case, w^n is unobserved and does not influence q^n, since w^n and q^n are independent. Thus, it has no impact on the problem, so our scenario is equivalent to one in which there is no signal side information, to which Theorem 3 can be applied, whence

$$R_{[\text{Q:DEC,W:NONE}]}(D)=R_{[\text{Q:NONE,W:NONE}]}(D). \tag{130}$$

Case: $*=$ ENC*:* The independence of q and w implies that

$$R_{[\text{Q:DEC,W:ENC}]}(D)=R_{[\text{Q:DEC,W:NONE}]}(D) \tag{131}$$

since, via (123), an encoder without w^n can generate a simulated w^n from x^n without requiring access to q^n. Moreover, by identical reasoning we obtain

$$R_{[\text{Q:NONE,W:ENC}]}(D)=R_{[\text{Q:NONE,W:NONE}]}(D). \tag{132}$$

Combining (131), (132), and (130) yields

$$R_{[\text{Q:DEC,W:ENC}]}(D)=R_{[\text{Q:NONE,W:ENC}]}(D). \tag{133}$$

□

Appendix IV
Proof of Lemma 2 (Continuity of Entropy)

We begin with some notation. Let $\mathcal{N}_\delta(t)$ be a distribution of the form (54). Specifically, let

$$\mathcal{N}_\delta(t)=ae^{-s\delta(\cdot)} \tag{134}$$

where

$$\int\delta(t)ae^{-s\delta(t)}dt=\int\delta(t)p_t(t)dt=E[\delta(t)]. \tag{135}$$

For such distributions we have [cf. (58)]

$$h(\mathcal{N}_\delta(t))=-\log(a_\delta(t))+\log(e)s_\delta(t)E[\delta(t)] \tag{136}$$

and, hence, the following readily verified identity:

$$D(p_t\,\|\,\mathcal{N}_\delta(t))=h(\mathcal{N}_\delta(t))-h(t), \tag{137}$$

which we use in our proof.

In particular, we have (138)–(141), shown at the top of this page as desired. To obtain (138) we have exploited the identity

(137). In turn, to obtain (139) we have used the definitions (56) and (58), and that

$$\liminf_{D\to D_{\min}} \int D\big(p_{x+v_D^*|w}(\cdot|w) \,\big\|\, \mathcal{N}_\delta(x+v_D^*|w=w)\big)p_w(w)dw$$

$$\geq \int \liminf_{D\to D_{\min}} D\big(p_{x+v_D^*|w}(\cdot|w) \,\big\|\, \mathcal{N}_\delta(x+v_D^*|w=w)\big)p_w(w)dw \tag{142}$$

$$\geq \int D\big(p_{x|w}(\cdot|w) \,\big\|\, \mathcal{N}_\delta(x|w=w)\big)p_w(w)dw \tag{143}$$

where (142) follows from Fatou's lemma [22, p. 78], and where (143) follows from the lower semi-continuity of (conditional) divergence [23], together with the fact that (53) implies

$$\lim_{D\to D_{\min}} p_{x+v_D^*|w}(\cdot|w) = \lim_{D\to D_{\min}} \big[p_{x|w}(\cdot|w) * p_{v_D^*|w}(\cdot|w)\big] \tag{144}$$

$$= p_{x|w}(\cdot|w) \tag{145}$$

where in (144) $*$ denotes the convolution operator.

To obtain (140), we have used Lebesgue's dominated convergence theorem [22, p. 78] to switch the order of limiting and integration, which can be applied since via (57) and Condition 4b of Definition 2, $h_\delta(\Delta_D(w)$ converges uniformly to $h_\delta(\Delta_{D_{\min}}(w))$, whence there exists a $\xi > 0$ such that for all $w \in \mathcal{W}$ and D sufficiently close to $D_{\min}$

$$|h_\delta(\Delta_D(w)) - h_\delta(\Delta_{D_{\min}}(w))|p_w(w) \leq \xi p_w(w) \tag{146}$$

where we note the right-hand side of (146) is nonnegative and integrable. □

APPENDIX V
SUFFICIENCY THEOREM PROOFS

A. Proof of Lemma 3

We first lower bound $R_{[\text{Q:BOTH,W:BOTH}]}(D)$ via (62) of Lemma 1. Next, $R_{[\text{Q:ENC,W:DEC}]}(D)$ can be expressed in the form of the Wyner–Ziv rate–distortion formula [1] applied to the super-source $\tilde{x} = (x, q)$, yielding

$$R_{[\text{Q:ENC,W:DEC}]}(D) = \inf_{\{p_{u|x,q}, g: E[\rho(x-g(u,w);q)]\leq D\}} I(u; q, x|w). \tag{147}$$

In turn, (147) can be upper-bounded via

$$R_{[\text{Q:ENC,W:DEC}]}(D) \leq I(x+v_D^*; q, x|w) \tag{148}$$

$$= h(x+v_D^*|w) - h(v_D^*|q, x, w)$$

$$= h(x+v_D^*|w) - h(v_D^*|q) \tag{149}$$

where to obtain (148) we have made the particular (rather than optimizing) choices $u = x + v_D^*$, with v_D^* as defined via Lemma 1, and $g(u, w) = u$ in (147) that meet the distortion constraint, and where to obtain (149) we have used the Markov property (60).

Finally, using (62) and (149) with (76) we obtain

$$\Delta R_{[\text{Q:}\overline{\text{DEC}}\text{,W:}\overline{\text{ENC}}]}(D) \leq h(x+v_D^*|w) - h(x|w), \tag{150}$$

so to obtain the desired result (75) from (150), it suffices to apply Lemma 2. □

B. Proof of Theorem 6

We establish (79) for the four cases separately.

Case: $* = \text{ENC}$*:* We have

$$\lim_{D\to D_{\min}} R_{[\text{Q:ENC,W:DEC}]}(D) - R_{[\text{Q:ENC,W:BOTH}]}(D)$$

$$\leq \lim_{D\to D_{\min}} \Delta R_{[\text{Q:}\overline{\text{DEC}}\text{,W:}\overline{\text{ENC}}]}(D) \tag{151}$$

$$= 0 \tag{152}$$

where (151) follows from using $R_{[\text{Q:ENC,W:BOTH}]}(D) \geq R_{[\text{Q:BOTH,W:BOTH}]}(D)$ with (76), and where (152) follows from applying Lemma 3.

Case: $* = \text{BOTH}$*:* We have, similarly

$$\lim_{D\to D_{\min}} R_{[\text{Q:BOTH,W:DEC}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D) \tag{153}$$

$$\leq \lim_{D\to D_{\min}} \Delta R_{[\text{Q:}\overline{\text{DEC}}\text{,W:}\overline{\text{ENC}}]}(D) \tag{154}$$

$$= 0 \tag{155}$$

where (154) follows from using $R_{[\text{Q:ENC,W:DEC}]}(D) \geq R_{[\text{Q:BOTH,W:DEC}]}(D)$ with (76), and where (155) follows from applying Lemma 3.

Case: $* = \text{NONE}$*:* Here, q^n is unobserved and does not influence w^n, since q^n and w^n are independent. Thus, it has no impact on the problem, so our scenario is equivalent to one in which there is no distortion side information, to which (77) can be applied, whence

$$\lim_{D\to D_{\min}} R_{[\text{Q:NONE,W:DEC}]}(D) - R_{[\text{Q:NONE,W:BOTH}]}(D) = 0. \tag{156}$$

Note that for this application of asymptotic Wyner–Ziv results, the relevant distortion measure is

$$d'(x, \hat{x}) = E[d(x, \hat{x}; q)|x = x, \hat{x} = \hat{x}, w = w]$$

$$= E[\gamma(q)]\rho(x - \hat{x}) + E[\tau(q)]. \tag{157}$$

Case: $* = \text{DEC}$*:* We have

$$\lim_{D\to D_{\min}} R_{[\text{Q:DEC,W:DEC}]}(D) - R_{[\text{Q:DEC,W:BOTH}]}(D)$$

$$\leq \lim_{D\to D_{\min}} R_{[\text{Q:NONE,W:DEC}]}(D) - R_{[\text{Q:DEC,W:BOTH}]}(D) \tag{158}$$

$$= \lim_{D\to D_{\min}} R_{[\text{Q:NONE,W:DEC}]}(D) - R_{[\text{Q:NONE,W:BOTH}]}(D) \tag{159}$$

$$= 0 \tag{160}$$

where (158) follows from $R_{[\text{Q:NONE,W:DEC}]}(D) \geq R_{[\text{Q:DEC,W:DEC}]}(D)$, where (159) follows from the case $* = \text{NONE}$ of Theorem 5, which applies because the distortion measure is of the scaled form (32), and where (160) follows from (156), which holds because q and w are independent. □

C. Proof of Theorem 7

We establish (81) for the four cases separately.

Case: $* = \text{DEC}$*:* We have

$$\lim_{D\to D_{\min}} R_{[\text{Q:ENC,W:DEC}]}(D) - R_{[\text{Q:BOTH,W:DEC}]}(D)$$

$$\leq \lim_{D\to D_{\min}} \Delta R_{[\text{Q:}\overline{\text{DEC}}\text{,W:}\overline{\text{ENC}}]}(D) = 0 \tag{161}$$

where the inequality follows from (76) and that $R_{[\mathrm{Q:BOTH,W:DEC}]}(D) \geq R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$, and where the final equality follows from applying Lemma 3.

Case: $* = \mathrm{BOTH}$*:* We similarly obtain

$$\lim_{D\to D_{\min}} R_{[\mathrm{Q:ENC,W:BOTH}]}(D) - R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \leq \lim_{D\to D_{\min}} \Delta R_{[\mathrm{Q:}\overline{\mathrm{DEC}}\mathrm{,W:}\overline{\mathrm{ENC}}]}(D) = 0 \tag{162}$$

using that $R_{[\mathrm{Q:ENC,W:DEC}]}(D) \geq R_{[\mathrm{Q:ENC,W:BOTH}]}(D)$ in (76), and applying Lemma 3.

Case: $* = \mathrm{NONE}$*:* Here, w^n is unobserved and does not influence q^n, since w^n and q^n are independent. Thus, it has no impact on the problem, so our scenario is equivalent to one in which there is no signal side information, to which Corollary 4 can be applied, whence

$$\lim_{D\to D_{\min}} R_{[\mathrm{Q:ENC,W:NONE}]}(D) - R_{[\mathrm{Q:BOTH,W:NONE}]}(D) = 0. \tag{163}$$

Case: $* = \mathrm{ENC}$*:* We have

$$\begin{aligned}
&\lim_{D\to D_{\min}} R_{[\mathrm{Q:ENC,W:ENC}]}(D) - R_{[\mathrm{Q:BOTH,W:ENC}]}(D) \\
&\quad \leq \lim_{D\to D_{\min}} R_{[\mathrm{Q:ENC,W:NONE}]}(D) \\
&\qquad - R_{[\mathrm{Q:BOTH,W:ENC}]}(D) && (164)\\
&\quad = \lim_{D\to D_{\min}} R_{[\mathrm{Q:ENC,W:NONE}]}(D) \\
&\qquad - R_{[\mathrm{Q:BOTH,W:NONE}]}(D) && (165)\\
&\quad = 0 && (166)
\end{aligned}$$

where (164) follows from using that $R_{[\mathrm{Q:ENC,W:NONE}]}(D) \geq R_{[\mathrm{Q:ENC,W:ENC}]}(D)$, where (165) follows from applying the case $* = \mathrm{BOTH}$ of Theorem 4, and where (166) follows from (163), which holds because q^n and w^n are independent.

Appendix VI
Asymptotic Rate Loss Theorem Proofs

A. Proof of Lemma 4

First, via (47) we have

$$\begin{aligned}
&R_{[\mathrm{W:ENC}]}(D) - R_{[\mathrm{W:BOTH}]}(D) \\
&\qquad = R_{[\mathrm{W:NONE}]}(D) - R_{[\mathrm{W:BOTH}]}(D). && (167)
\end{aligned}$$

Next, using [13, Theorem 1], we have that

$$\lim_{D\to D_{\min}} R_{[\mathrm{W:NONE}]}(D) = h(x) - h(v_D^*) \tag{168}$$

where v_D^* is a random variable v maximizing $h(v)$ subject to the constraint $E[\rho(v)] \leq D$. Note that as a consequence v_D^* is independent of w, x.

Now for the case with signal side information, we use a different subsystem for each value of the side information w. Moreover, applying [13, Theorem 1] to the subsystem corresponding to w, which is just a system without side information parameterized by w, we have

$$\lim_{D\to D_{\min}} R^w_{[\mathrm{W:NONE}]}(D) = h(x|w=w) - h(\tilde{v}_D^*|w=w) \tag{169}$$

which when averaged over all values of w yields

$$\lim_{D\to D_{\min}} R_{[\mathrm{W:BOTH}]}(D) = h(x|w) - h(\tilde{v}_D^*|w). \tag{170}$$

In this case, $\tilde{v}_D^*$ is a random variable v maximizing $h(v|w=w)$ subject to the constraint $E[\rho(v)] \leq D$. Substituting (168) and (170) into (167) then yields(87) as desired, provided

$$h(\tilde{v}_D^*|w) = h(v_D^*). \tag{171}$$

To verify (171), it suffices to exploit the fact that without loss of generality we can restrict our attention to v that are independent of w in the entropy maximization corresponding to $\tilde{v}_D^*$. Indeed, in this case, the two entropy functions being maximized in the generation of v_D^* and $\tilde{v}_D^*$ are identical, and thus the corresponding entropies are the same.

To confirm that the above restriction does not incur a price in entropy, let $\tilde{v}_D^{*\prime}$ be a random variable independent of w, x such that $p_{\tilde{v}_D^{*\prime}}(\cdot) = p_{\tilde{v}_D^*}(\cdot)$. Then $E[\rho(\tilde{v}_D^{*\prime})] = E[\rho(\tilde{v}_D^*)]$ and

$$\begin{aligned}
h(\tilde{v}_D^{*\prime}|w) &= h(\tilde{v}_D^{*\prime}) && (172)\\
&= h(\tilde{v}_D^*) && (173)\\
&= h\left(\int p_{\tilde{v}_D^*|w}(\cdot|w)p_w(w)dw\right) && (174)\\
&\geq \int h(p_{\tilde{v}_D^*|w}(\cdot|w))p_w(w)dw && (175)\\
&= h(\tilde{v}_D^*|w) && (176)
\end{aligned}$$

where (175) follows form the concavity of differential entropy, and where we have used the notation $h(t)$ and $h(p_t)$ interchangeably. But since $\tilde{v}_D^*$ maximized the conditional entropy, we must have $h(\tilde{v}_D^{*\prime}|w) = h(\tilde{v}_D^*|w)$. □

B. Proof of Theorem 8

First, note that we can apply Theorem 4 to (82) to obtain

$$\Delta R_{[\mathrm{Q:}*\mathrm{,W:}\overline{\mathrm{DEC}}]}(D) = R_{[\mathrm{Q:}*\mathrm{,W:NONE}]}(D) - R_{[\mathrm{Q:}*\mathrm{,W:BOTH}]}(D) \tag{177}$$

for all $* \in \{\mathrm{NONE,ENC,DEC,BOTH}\}$.

We now consider each of the four cases separately.

Case: $* = \mathrm{NONE}$*:* In this case, q^n is unobserved and does not influence w^n, since q^n and w^n are independent. Thus, it has no impact on the problem, so our scenario is equivalent to one in which there is no distortion side information, i.e.,

$$\begin{aligned}
&\Delta R_{[\mathrm{Q:NONE,W:}\overline{\mathrm{DEC}}]}(D) \\
&\quad = R_{[\mathrm{Q:NONE,W:ENC}]}(D) - R_{[\mathrm{Q:NONE,W:BOTH}]}(D) \\
&\quad = R_{[\mathrm{W:ENC}]}(D) - R_{[\mathrm{W:BOTH}]}(D). && (178)
\end{aligned}$$

Noting that the relevant distortion measure is [cf. (157)]

$$\begin{aligned}
d'(x,\hat{x}) &= E[d(x,\hat{x};q)|x=x,\hat{x}=\hat{x},w=w] \\
&= E[\gamma(q)]\rho(x-\hat{x}) + E[\tau(q)],
\end{aligned}$$

and applying Lemma 4 to (178) then yields

$$\lim_{D\to D_{\min}} \Delta R_{[\mathrm{Q:NONE,W:}\overline{\mathrm{DEC}}]}(D) = I(x;w). \tag{179}$$

Case: $* = \text{DEC}$*:* Since the distortion measure (80) is of the scaled form (32) and since w^n and q^n are independent, the cases $* = \text{NONE}, \text{BOTH}$ of Theorem 5 apply, and so we have $R_{[\text{Q:NONE,W:NONE}]}(D) = R_{[\text{Q:DEC,W:NONE}]}(D)$ and $R_{[\text{Q:NONE,W:BOTH}]}(D) = R_{[\text{Q:DEC,W:BOTH}]}(D)$. From (177), we then have

$$\Delta R_{[\text{Q:DEC,W:}\overline{\text{DEC}}]}(D) = \Delta R_{[\text{Q:NONE,W:}\overline{\text{DEC}}]}(D),$$

applying (179) to which gives the desired result.

Case: $* = \text{BOTH}$*:* Since q is known to both encoder and decoder, the systems corresponding, respectively, to $R_{[\text{Q:BOTH,W:ENC}]}(D)$ and $R_{[\text{Q:BOTH,W:BOTH}]}(D)$ in (82) can be realized by using a separate source coding subsystem for each value of q. Consider the two subsystems corresponding to a particular value q, for which the modified source distribution is $p_{x|q}(\cdot|q)$ and the modified signal side information is $p_{w|q}(\cdot|q)$. These subsystems are equivalent to ones without distortion side information but parameterized by q. Thus, using Lemma 4 the associated rate loss is therefore

$$\lim_{D\to D_{\min}} \left[R^q_{[\text{Q:NONE,W:NONE}]}(D) - R^q_{[\text{Q:NONE,W:BOTH}]}(D)\right] = I(x;w|q=q). \quad (180)$$

Averaging over all values of q in (180) then yields

$$\lim_{D\to D_{\min}} \left[R_{[\text{Q:BOTH,W:NONE}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D)\right] = I(x;w|q). \quad (181)$$

Case: $* = \text{ENC}$*:* For the lower bound we have

$$\lim_{D\to D_{\min}} \Delta R_{[\text{Q:ENC,W:}\overline{\text{DEC}}]}(D)$$
$$= \lim_{D\to D_{\min}} \left[R_{[\text{Q:ENC,W:NONE}]}(D) - R_{[\text{Q:ENC,W:BOTH}]}(D)\right] \quad (182)$$
$$\geq \lim_{D\to D_{\min}} \left[R_{[\text{Q:BOTH,W:NONE}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D)\right] \quad (183)$$
$$= I(x;w|q) \quad (184)$$

where (182) follows from (177), where (183) follows from the fact that more side information cannot increase rate to the first term and applying the case $* = \text{BOTH}$ of Theorem 7 to the second term, and where (184) follows from (181).

For the upper bound, we note

$$\lim_{D\to D_{\min}} \Delta R_{[\text{Q:ENC,W:}\overline{\text{DEC}}]}(D)$$
$$= \lim_{D\to D_{\min}} \left[R_{[\text{Q:ENC,W:NONE}]}(D) - R_{[\text{Q:ENC,W:BOTH}]}(D)\right] \quad (185)$$
$$= \lim_{D\to D_{\min}} \left[R_{[\text{Q:ENC,W:NONE}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D)\right] \quad (186)$$

where (185) reproduces (182), where (186) follows from applying the case $* = \text{BOTH}$ of Theorem 7 to the second term.

We lower-bound the second term in (186) via (62) of Lemma 1. The first term in (186) we upper-bound by

$$R_{[\text{Q:ENC,W:NONE}]}(D) \leq I(\hat{x};x,q) \quad (187)$$
$$= h(\hat{x}) - h(\hat{x}|q,x) \quad (188)$$
$$\leq h(x+v^*_D) - h(x+v^*_D|q,x) \quad (189)$$
$$= h(x+v^*_D) - h(v^*_D|q) \quad (190)$$

where (187) follows from (6c) for any $\hat{x}$ satisfying the distortion constraint, where to obtain (189) have made make the particular choice $\hat{x} = x + v^*_D$ in (6c), where v^*_D is the maximum-entropy variable defined via Lemma 1, which satisfies the distortion constraint.

Finally, substituting (190) and (62) into (186) we obtain

$$\lim_{D\to D_{\min}} \Delta R_{[\text{Q:ENC,W:}\overline{\text{DEC}}]}(D)$$
$$\leq \lim_{D\to D_{\min}} [h(x+v^*_D) - h(x|w)] \quad (191)$$
$$= h(x) - h(x|w) + \lim_{D\to D_{\min}} [h(x+v^*_D) - h(x)] \quad (192)$$
$$= I(x;w) \quad (193)$$

where (193) follows from the application of Corollary 3.

It remains only to show that both the upper and lower bounds can be tight. To verify that the lower bound can be tight, consider a scenario in which the source has dimension $k = 2$, i.e., $x = (x_1, x_2)$, the signal and distortion side information are related according to $w = q = x_2 + t$, the distortion measure is of the form $d(x,\hat{x};q) = |q||x_1 - \hat{x}_1|$ with $\hat{x} = (\hat{x}_1, \hat{x}_2)$, and where x_1, x_2, and t are independent, zero-mean, unit-variance Gaussian random variables. This is obviously equivalent to a scenario with source $x' = x_1$, distortion side information $q' = x_2 + t$, distortion measure $d'(x',\hat{x}';q') = |q||x' - \hat{x}'|$ with $\hat{x}' = \hat{x}_1$, and no signal side information. Hence

$$\lim_{D\to D_{\min}} \Delta R_{[\text{Q:ENC,W:}\overline{\text{DEC}}]}(D)$$
$$= \lim_{D\to D_{\min}} [R_{[\text{Q:ENC,W:ENC}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D)] \quad (194)$$
$$= \lim_{D\to D_{\min}} [R_{[\text{Q}':\text{ENC}]}(D) - R_{[\text{Q}':\text{BOTH}]}(D)] \quad (195)$$
$$= 0 \quad (196)$$
$$= h(x|q) - h(x|w,q) \quad (197)$$
$$= I(x;w|q) \quad (198)$$
$$< I(x;w) \quad (199)$$

where to obtain (196) we have used Corollary 4, and where to obtain (197) we have used that $q = w$.

To verify that the upper bound can be tight, consider a scenario in which the source x is Gaussian, the signal and distortion side information are related according to $w = q = x + t$, the distortion measure is of the form $d(x,\hat{x};q) = (x - \hat{x})^2$, and where x and t are independent, zero-mean, Gaussian random variables. This is obviously equivalent to a scenario with no distortion side information, since q does not affect the distortion measure. Hence

$$\lim_{D\to D_{\min}} \Delta R_{[\text{Q:ENC,W:}\overline{\text{DEC}}]}(D)$$
$$= \lim_{D\to D_{\min}} [R_{[\text{Q:ENC,W:ENC}]}(D) - R_{[\text{Q:BOTH,W:BOTH}]}(D)] \quad (200)$$
$$= \lim_{D\to D_{\min}} [R_{[\text{W:ENC}]}(D) - R_{[\text{W:BOTH}]}(D)] \quad (201)$$

$$= \lim_{D\to D_{\min}} [R_{[\mathrm{W:NONE}]}(D) - R_{[\mathrm{W:DEC}]}(D)] \quad (202)$$

$$= \lim_{D\to D_{\min}} \left[\frac{1}{2}\log\frac{\mathrm{var}[x]}{D} - \frac{1}{2}\log\frac{\mathrm{var}[x|w]}{D}\right] \quad (203)$$

$$= \frac{1}{2}\log\frac{\mathrm{var}[x]}{\mathrm{var}[x|w]} \quad (204)$$

$$= \frac{1}{2}\log(2\pi e\mathrm{var}[x]) - \frac{1}{2}\log(2\pi e\mathrm{var}[x|w]) \quad (205)$$

$$= h(x) - h(x|w) \quad (206)$$

$$= I(x;w) \quad (207)$$

$$> I(x;w|q) \quad (208)$$

where (202) follows from (77) and (47), where to obtain (203) we have substituted the particular form of the regular and Wyner–Ziv [2] rate distortion functions for the Gaussian scenario, and where to obtain (206) we have recognized the form of differential entropy for Gaussian random variables. □

C. Proof of Theorem 9

First, note that we can apply Theorem 5 to (85) to obtain

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}},\mathrm{W:}*]}(D) = R_{[\mathrm{Q:NONE,W:}*]}(D) - R_{[\mathrm{Q:BOTH,W:}*]}(D) \quad (209)$$

for all $* \in \{\mathrm{NONE,ENC,DEC,BOTH}\}$.

Moreover, in the sequel, some facts will be convenient. First, the following differential entropy can be computed.

Fact 1 ([24]): Let v be a random variable defined over $\mathbb{R}^k$. Then

$$\max_{\{p_v: E[\|v\|^r]\le\sigma^2\}} h(v) \ge \beta(k,r) + \frac{k}{r}\log\sigma^2 \quad (210)$$

where

$$\beta(k,r) = \frac{k}{r} - \log\left[\frac{r}{kV_k\Gamma(k/r)}\left(\frac{k}{r}\right)^{k/r}\right] \quad (211)$$

with V_k denoting the volume of the unit sphere $S_k = \{x : \|x\| \le 1\}$, and with $\Gamma(\cdot)$ denoting the gamma function.

In turn, Fact 1 is used in establishing the following.

Fact 2 ([13]): For a source coding scenario (without side information) in which $x \in \mathbb{R}^k$ and the distortion measure is of the form $d(x,\hat{x}) = \|x-\hat{x}\|^r$, the rate–distortion tradeoff is asymptotically

$$\lim_{D\to 0}\left[R(D) + \frac{k}{r}\log D\right] = h(x) - \beta(k,r) \quad (212)$$

where $\beta(k,r)$ is as defined in (211).

We now consider the four cases of (86) separately.

Case: $* = \mathrm{NONE}$*:* First, from (209), we have

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}},\mathrm{W:NONE}]}(D) = R_{[\mathrm{Q:NONE,W:NONE}]}(D) - R_{[\mathrm{Q:BOTH,W:NONE}]}(D). \quad (213)$$

Via (6a), we see that $R_{[\mathrm{Q:NONE,W:NONE}]}(D)$ is obtained by conventional source coding with the modified distortion measure

$$d'(x,\hat{x}) = E[d(x,\hat{x};q)|x=x,\hat{x}=\hat{x}] = E[q]\|x-\hat{x}\|^r,$$

which using Fact 2 yields

$$\lim_{D\to 0}\left[R_{[\mathrm{Q:NONE,W:NONE}]}(D) + \frac{k}{r}\log D\right] = h(x) - \beta(k,r) + \frac{k}{r}\log E[q]. \quad (214)$$

Next, via (6d), we see that evaluating the conditional rate distortion function $R_{[\mathrm{Q:BOTH,W:NONE}]}(D)$ involves the familiar waterfilling distortion (and thus rate) allocation (see, e.g., [21, Sec. 13.3.3]). Specifically, for each q, we quantize the corresponding source samples to distortion $D_q = E[\|x-\hat{x}\|^r]$ using rate $R_q(D_q)$, so the overall rate and distortion are, respectively, $E[R_q(D_q)]$ and $E[qD_q]$.

To find the optimal distortion allocation we use Lagrange multipliers, and thus minimize

$$J(D) = E[R_q(D_q)] - \lambda(D - E[qD_q]) \quad (215)$$

with respect to D_q. Since via Fact 2 we have

$$\lim_{D\to 0}\left[R_q(D_q) + \frac{k}{r}\log D_q\right] = h(x) - \beta(k,r) \quad (216)$$

it is straightforward to verify that the optimizing distortion allocation is

$$D_q^* = \frac{D}{q}, \quad (217)$$

whence

$$\lim_{D\to 0}\left[R_{[\mathrm{Q:BOTH,W:NONE}]}(D) + \frac{k}{r}\log D\right] = h(x) - \beta(k,r) + \frac{k}{r}E[\log q]. \quad (218)$$

Subtracting (218) from (214) and using (213) establishes

$$\lim_{D\to 0} R_{[\mathrm{Q:DEC,W:NONE}]}(D) - R_{[\mathrm{Q:BOTH,W:NONE}]}(D) = \frac{k}{r}E\left[\log\frac{E[q]}{q}\right]. \quad (219)$$

Case: $* = \mathrm{ENC}$*:* From Theorem 4, we obtain immediately that

$$R_{[\mathrm{Q:DEC,W:ENC}]}(D) - R_{[\mathrm{Q:BOTH,W:ENC}]}(D) = R_{[\mathrm{Q:DEC,W:NONE}]}(D) - R_{[\mathrm{Q:BOTH,W:NONE}]}(D),$$

which when used with (219) above gives the desired result.

Case: $* = \mathrm{BOTH}$*:* First, via (209) we have

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}},\mathrm{W:BOTH}]}(D) = R_{[\mathrm{Q:NONE,W:BOTH}]}(D) - R_{[\mathrm{Q:BOTH,W:BOTH}]}(D). \quad (220)$$

Next, for the two systems corresponding to the two terms on the right-hand side of (220), the encoder and decoder in both cases agree to use a separate source coding subsystem for each value of w. Thus, the performance loss between each corresponding pair of subsystems for each such value of w is obtained as a conditioned version of the results above for the case $* = \mathrm{NONE}$.

Specifically, via the obvious generalization of the right-hand side of (219), the asymptotic rate loss for the subsystem corresponding to the realized side information w is

$$\frac{k}{r}E\left[\log\frac{E[q|w=w]}{q}\middle|w=w\right], \tag{221}$$

averaging over which with respect to w yields

$$\lim_{D\to 0}[R_{[\mathrm{Q:NONE,W:BOTH}]}(D)-R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)] = \frac{k}{r}E\left[\log\frac{E[q|w]}{q}\right]. \tag{222}$$

Case: $*=\mathrm{DEC}$*:* First, we obtain our lower bound via

$$\lim_{D\to 0}\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}}\mathrm{,W:DEC}]}(D)$$
$$=\lim_{D\to 0}\left[R_{[\mathrm{Q:NONE,W:DEC}]}(D)-R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)\right] \tag{223}$$
$$\geq\lim_{D\to 0}\left[R_{[\mathrm{Q:NONE,W:BOTH}]}(D)-R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)\right] \tag{224}$$
$$=\frac{k}{r}E\left[\log\frac{E[q|w]}{q}\right] \tag{225}$$

where to obtain (223) we have used (209) and the case $*=\mathrm{BOTH}$ in Theorem 6, where to obtain (224) we have used that additional side information cannot increase rate, and where to obtain (225) we have used (222).

We obtain our upper bound starting from (223). In particular, note that $R_{[\mathrm{Q:NONE,W:DEC}]}(D)$ corresponds to Wyner–Ziv source coding with the modified (signal side information dependent) distortion measure $d'(x,\hat{x})=E[q|w=w]\,\|x-\hat{x}\|^r$.

While [6, Theorem 2] can be applied when $k=1$ and $r=2$, for the more general case, we begin by introducing v_D^* as the random variable v that maximizes $h(v)$ and satisfies the constraint $E[\|v\|^r]\leq D/E[q]$. Then we have

$$R_{[\mathrm{Q:NONE,W:DEC}]}(D)$$
$$\leq\inf_{\{p_{u|x},g:E[d(x,g(u,w);q)]\leq D}I(u;x|w) \tag{226}$$
$$\leq\inf_{\{p_{\hat{x}|x}:E[d(x,\hat{x};q)]\leq D}I(\hat{x};x|w) \tag{227}$$
$$=h(x+v_D^*|w)-h(x+v_D^*|x,w) \tag{228}$$
$$=h(x+v_D^*|w)-h(v_D^*) \tag{229}$$
$$\leq h(x+v_D^*|w)-\beta(k,r)-\frac{k}{r}\log\frac{D}{E[q]} \tag{230}$$
$$=[h(x+v_D^*|w)-h(x|w)]+h(x|w)-\beta(k,r)-\frac{k}{r}\log\frac{D}{E[q]} \tag{231}$$

where (226) follows from the obvious Wyner–Ziv style random coding argument based on binning, where (227) follows from the particular choice $g(u,w)=u$ so u can be renamed $\hat{x}$, where (228) follows from the particular test-channel choice $\hat{x}=x+v_D^*$ with v_D^* as defined at the outset, which by construction ensures that the distortion constraint is met, where (229) follows from the independence of v_D^* and w, x, and where to obtain (230) we have used Fact 1. Finally, using Corollary 1 (since q is effectively absent), we have that the entropy-difference term in brackets in (231) vanishes as $D\to D_{\min}$, whence

$$\lim_{D\to 0}\left[R_{[\mathrm{Q:NONE,W:DEC}]}(D)+\frac{k}{r}\log D\right] \leq h(x|w)-\beta(k,r)+\frac{k}{r}\log E[q]. \tag{232}$$

In turn, $R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$ is obtained by using a separate coding subsystem for each value of w. Specifically, for a particular realization w, the rate–distortion function for the corresponding subsystem is equivalent to $R^w_{[\mathrm{Q:BOTH,W:NONE}]}(D)$ for a modified source x' with distribution $p_{x'}(x')=p_{x|w}(x'|w)$ and a modified distortion side information q' with distribution $p_{q'}(q')=p_{q|w}(q'|w)$. Adapting (218) accordingly, we obtain

$$\lim_{D\to 0}\left[R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)+\frac{k}{r}\log D\right] =\int\left[h(x|w=w)-\beta(k,r)+\frac{k}{r}E[\log q|w=w]\right]p_w(w)dw$$
$$=h(x|w)-\beta(k,r)+\frac{k}{r}E[\log q]. \tag{233}$$

Finally, substituting (232) and (233) into(223), we obtain

$$\lim_{D\to 0}\left[R_{[\mathrm{Q:NONE,W:DEC}]}(D)-R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)\right] \leq\frac{k}{r}E\left[\log\frac{E[q]}{q}\right]. \tag{234}$$

It remains only to show that both upper and lower bounds can be tight. To verify that the lower bound can be tight, consider the scenario in which $q=w=x$. In this case, having w^n at the decoder is equivalent to having x^n at the decoder, so

$$R_{[\mathrm{Q:DEC,W:DEC}]}(D)=R_{[\mathrm{Q:BOTH,W:DEC}]}(D)=0$$

and thus

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}}\mathrm{,W:DEC}]}(D)=0 \tag{235}$$
$$=\frac{k}{r}E\left[\log\frac{E[q|w]}{q}\right] \tag{236}$$
$$<\frac{k}{r}E\left[\log\frac{E[q]}{q}\right] \tag{237}$$

where (236) follows from noting that $E[q|w]=q$.

To verify that the upper bound can be tight, consider the scenario in which $q=w$, but this side information is independent of x. Then

$$\Delta R_{[\mathrm{Q:}\overline{\mathrm{ENC}}\mathrm{,W:DEC}]}(D)$$
$$=R_{[\mathrm{Q:DEC,W:DEC}]}(D)-R_{[\mathrm{Q:BOTH,W:DEC}]}(D) \tag{238}$$
$$=R_{[\mathrm{Q:DEC,W:DEC}]}(D)-R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \tag{239}$$
$$=R_{[\mathrm{Q:DEC}]}(D)-R_{[\mathrm{Q:BOTH}]}(D) \tag{240}$$

where to obtain (239) we have used the case $*=\mathrm{BOTH}$ of Theorem 6, and where to obtain (240) we have used that $q=$

w. Finally, taking the limit of (240) as $D \to 0$ and applying Corollary 5 yields

$$\lim_{D\to 0} \Delta R_{[\mathrm{Q}:\overline{\mathrm{ENC}},\mathrm{W}:\mathrm{DEC}]}(D) = \frac{k}{r} E\left[\log \frac{E[q]}{q}\right] \tag{241}$$

$$> \frac{k}{r} E\left[\log \frac{E[q|w]}{q}\right]. \tag{242}$$

□

Appendix VII
Nonasymptotic Rate Gap Theorem Proofs

A. Proof of Theorem 10

We use of the following lemma, which bounds the entropy of the sum of arbitrary random variable and a Gaussian mixture.

Lemma 5: Let x be an arbitrary random variable with variance $\sigma^2 < \infty$. Let s be a zero-mean, unit-variance Gaussian independent of x and let t be a random variable independent of x and s with $0 < t < t_{\max}$. Then

$$h(x + s\sqrt{t}) - h(x) \le \frac{1}{2}\log(1 + t_{\max} J(x)) \tag{243}$$

where

$$J(x) \triangleq \int p_x(x) \left[\frac{\partial}{\partial x} \ln p_x(x)\right]^2 dx \tag{244}$$

with equality if and only if t is a constant and x is Gaussian.

Proof: To obtain (243), we have

$$h(x + s\sqrt{t}) \le h(x + s\sqrt{t_{\max}}) \tag{245}$$

$$= h(x) + \int_0^{t_{\max}} \frac{\partial}{\partial t} h(x + s\sqrt{t}) dt \tag{246}$$

$$= h(x) + \int_0^{t_{\max}} \frac{1}{2}\log(e) J(x + s\sqrt{t}) dt \tag{247}$$

$$\le h(x) + \frac{1}{2}\log(e) \int_0^{t_{\max}} \frac{J(x) J(s\sqrt{t})}{J(x) + J(s\sqrt{t})} dt \tag{248}$$

$$= h(x) + \frac{1}{2}\log(e) \int_0^{t_{\max}} \frac{J(x)}{tJ(x) + 1} dt \tag{249}$$

$$= h(x) + \frac{1}{2}\log\left(1 + t_{\max} J(x)\right) \tag{250}$$

where (245) follows from the concavity of differential entropy, (247) follows from de Bruijn's identity, (248) follows from the convolution inequality for Fisher information [25], [21, p. 497], and where in (249), we have used that the Fisher information for a Gaussian distribution is the reciprocal of its variance.

Finally, the inequality used in (248) is both tight if and only if x is Gaussian, and (245) is tight if and only if t is a constant. □

As an aside, we note that Lemma 5 can be used to bound the (classical) rate–distortion function of an arbitrary source x relative to quadratic distortion. Specifically, using an additive Gaussian noise test-channel $\hat{x} = x + v$ and combining the right-hand side of (243) in Lemma 5 (with $v = s\sqrt{t}$ and $t \equiv D$) to upper-bound $h(x + v)$ yields

$$0 \le R(D) - h(x) + \frac{1}{2}\log(2\pi e D) \le \frac{1}{2}\log\left(1 + D J(x)\right)$$

where the left-hand inequality is the Shannon lower bound [13]. Evidently, the error in the Shannon lower bound is at most $\frac{1}{2}\log(1 + DJ(x))$.

Proof of Theorem 10: To establish (98), we first note that the desired rate gap (76) is bounded via

$$R_{[\mathrm{Q:ENC},\mathrm{W:DEC}]}(D) - R_{[\mathrm{Q:BOTH},\mathrm{W:BOTH}]}(D)$$

$$\le h(x + v_D^*|w) - h(x|w) \tag{251}$$

$$= \int \left[h(x + v_D^*|w = w) - h(x|w = w)\right] p_w(w) dw \tag{252}$$

$$\le \int \frac{1}{2}\log\left(1 + J(x|w = w) \max_{q\in\mathcal{Q}} D_q'\right) p_w(w) dw \tag{253}$$

$$= \int \frac{1}{2}\log\left(1 + J(x|w = w)\frac{D}{q_{\min}}\right) p_w(w) dw \tag{254}$$

$$\le \int \left\{\frac{\log(e)}{2} J(x|w = w)\frac{D}{q_{\min}}\right\} p_w(w) dw \tag{255}$$

$$= \frac{\log(e)}{2} J(x|w)\frac{D}{q_{\min}} \tag{256}$$

where to obtain (251) we have used (150). To obtain (253) we exploit the upper bound from (243) in Lemma 5 with

$$t = D_q' \triangleq E[(v_D^*)^2|q] \tag{257}$$

since conditioned on $q = q$, v_D^* is a zero-mean Gaussian, which maximizes entropy, and so

$$p_{v_D^*|w}(v|w) = \int p_{v_D^*|q}(v|q) p_{q|w}(q|w) dq \tag{258}$$

is a Gaussian mixture. Further maximizing the resulting differential entropy

$$h(v_D^*|q) = \int \frac{1}{2}\log(2\pi e D_q') p_q(q) dq \tag{259}$$

over the choice of D_q' subject to the constraint

$$E[q(v_D^*)^2] = E[qD_q'] \le D \tag{260}$$

yields the waterfilling solution $D_q' = D/q$, which satisfies

$$\max_{q\in\mathcal{Q}} \frac{D}{q} = \frac{D}{q_{\min}}, \tag{261}$$

whence (254). Finally, using the identity

$$\ln(1 + \theta) \le \theta \tag{262}$$

valid for all θ we obtain (255).

A second bound on desired rate gap (76) exploits a slightly different Shannon lower bound. In particular, starting from(64) we can write a Shannon lower bound as

$$R_{[\mathrm{Q:BOTH},\mathrm{W:BOTH}]}(D)$$

$$\ge h(x|w) - \int h(x - \hat{x}^*|q = q) p_q(q) dq \tag{263}$$

$$\ge h(x|w) - \int h(v_D^{**}|q = q) p_q(q) dq, \tag{264}$$

where v_D^{**} is a random variable v that maximizes $h(v|q=q)$ subject to the constraint

$$E[v^2|q=q] = E[(x-\hat{x}^*)^2|q=q] = D_q^* \tag{265}$$

with $\hat{x}^*$ corresponding to the optimizing distribution in (61), and with D_q^* denoting the optimum waterfilling allocation for achieving $R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$. As before, as a consequence of the entropy-maximizing property of v_D^{**}, we have

$$v_D^{**} \leftrightarrow q \leftrightarrow w, x.$$

But then

$$D_q^* = E\big[E[(x-\hat{x}^*)^2|w=w, q=q] \;\big|\; q=q\big] \tag{266}$$

$$\leq E\big[\mathrm{var}[x|w] \;\big|\; q=q\big] \triangleq D_q^+ \tag{267}$$

where, to obtain (266), we have used iterated expectation, and where to obtain (267) we have used that, from minimum mean-square error (MMSE) estimation theory, $\hat{x}^*$ cannot generate greater distortion than the choice $\hat{x} = E[x|w,q] = E[x|w]$.

In turn, we have

$$R_{[\mathrm{Q:ENC,W:DEC}]}(D) - R_{[\mathrm{Q:BOTH,W:BOTH}]}(D)$$

$$\leq h(x+v_D^{**}|w) - h(x|w) \tag{268}$$

$$= \int [h(x+v_D^{**}|w=w) - h(x|w=w)]\, p_w(w)dw \tag{269}$$

$$\leq \int \frac{1}{2}\log\left(1 + J(x|w=w)\max_{q\in\mathcal{Q}} D_q^*\right) p_w(w)dw \tag{270}$$

$$\leq \int \frac{1}{2}\log\left(1 + J(x|w=w)\max_{q\in\mathcal{Q}} D_q^+\right) p_w(w)dw \tag{271}$$

$$\leq \int \left\{\frac{\log(e)}{2} J(x|w=w)\max_{q\in\mathcal{Q}} D_q^+\right\} p_w(w)dw \tag{272}$$

$$= \frac{\log(e)}{2} J(x|w)\max_{q\in\mathcal{Q}} D_q^+. \tag{273}$$

To obtain (268) we have used the counterpart to (149) corresponding to the choice v_D^{**}, which leads to the counterpart of (150) in terms of v_D^{**}. To obtain (270), we have exploited the upper bound from (243) in Lemma 5, again recognizing that conditioned on $q=q$, v_D^{**} is a zero-mean Gaussian, so v_D^{**} is a Gaussian mixture by the counterpart of (258) in terms of v_D^{**}. Finally, to obtain (271) we apply the variance bound (267), and to obtain (272) we again use the identity (262).

Choosing the minimum of the bounds (256) and (273) then yields (98). □

As an aside, one approach for pursuing similar bounds for other distortion would be based on upper-bounding the derivative of $h(x+s\sqrt{t})$ with respect to t. Due to the concavity of differential entropy, such a bound should be obtainable from the derivative at $t=0$. To obtain the desired derivative at $t=0$, one can write

$$h\big(x+s\sqrt{t} \;\big|\; w\big) - h(x|w) = I\big(x+s\sqrt{t}; s\sqrt{t} \;\big|\; w\big) \tag{274}$$

using the Markov chain $s\sqrt{t} \leftrightarrow w \leftrightarrow x$, and note from the results of Prelov and van der Meulen [26] that under certain regularity conditions[14]

$$\frac{\partial}{\partial t}\lim_{t\to 0^+} I\big(x+s\sqrt{t}; s\sqrt{t}\big) = \frac{1}{2}J(x). \tag{275}$$

B. Proof of Theorem 11

The following alternative version of the Shannon lower bound will be useful in our development.

Lemma 6 (Alternative Shannon Lower Bound): Consider a scaled quadratic distortion measure of the form $d(x,\hat{x};q) = q(x-\hat{x})^2$ and let $\hat{x}^*$ denote the random variable corresponding to an optimizing distribution for (61), associated with q^n and w^n both known at both encoder and decoder. If we let $v_D^{*\prime}$ denote a random variable satisfying the Markov chain

$$v_D^{*\prime} \leftrightarrow q \leftrightarrow w, x \tag{276}$$

and having the same distribution as $x-\hat{x}^*$ when conditioned on q, then

$$R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \geq h(x|w) - h(v_D^{*\prime}|q). \tag{277}$$

Proof: Equation (277) is readily verified via

$$R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) = I(\hat{x}^*; x|q,w) \tag{278}$$

$$= h(x|q,w) - h(x|q,w,\hat{x}^*) \tag{279}$$

$$= h(x|w) - h(x-\hat{x}^*|q,w,\hat{x}^*) \tag{280}$$

$$= h(x|w) - h(v_D^{*\prime}|q,w,\hat{x}^*)$$

$$\geq h(x|w) - h(v_D^{*\prime}|q,w)$$

$$= h(x|w) - h(v_D^{*\prime}|q) \tag{281}$$

where (279) follows from the conditional rate–distortion function characterization, (280) follows from the Markov constraint (4), and where (281) follows from the Markov constraint (276). □

The key difference between the bound (277) in Lemma 6 and the corresponding traditional Shannon lower bound (62) is in the choice of the distribution for $v_D^{*\prime}$. The traditional bound uses an entropy-maximizing distribution for v_D^*, which has the advantage of being computable without knowing $\hat{x}^*$. However, v_D^* can have a conditional variance that is unbounded as $D\to\infty$. By contrast, the conditional variance of $v_D^{*\prime}$ remains bounded in this limit, as the following lemma establishes.

Lemma 7: There exists a choice for $v_D^{*\prime}$ in Lemma 6 such that for all w

$$\mathrm{var}[v_D^{*\prime}|w=w] \leq \sigma_{\max}^2 \tag{282}$$

where $\sigma_{\max}^2$ is as defined in (101).

Proof: We establish our proof by contradiction. Suppose every $v_D^{*\prime}$ satisfying the requirements of Lemma 6 violates (282).

[14]Alternatively, (275) can be obtained by rewriting the mutual information in (274) as a relative entropy, and expand the result in a Taylor series [27, Sec. 2.6], assuming certain derivatives of the associated probability distributions exist.

Then from any such $v_D^{*\prime}$ corresponding to some optimizing $\hat{x}^*$ we construct a new random variable $v_D^{*\prime\prime}$ from a corresponding $\hat{x}^{*\prime}$ obtained from $\hat{x}^*$ as follows:

$$p_{\hat{x}^{*\prime}|\hat{x}^*,x,q,w}(\hat{x}^{**}|\hat{x}^*,x,q,w) = \begin{cases} \delta_{\hat{x}^*,\hat{x}^{**}}, & q,w) \in \mathcal{S} \\ \delta_{\hat{x}^*,E[x|q=q,w=w]} & (q,w) \notin \mathcal{S} \end{cases} \quad (283)$$

where $\delta_{.,.}$ is again the Kronecker delta function, and where

$$\mathcal{S} = \left\{(q,w) : E\left[(x-\hat{x}^*)^2|q=q,w=w\right] \le \sigma_{\max}^2\right\} \quad (284)$$

with $\sigma_{\max}^2$ as defined in (101).

First note that $\hat{x}^{*\prime}$ must also be an optimizing distribution for (61). Indeed, by the data processing inequality the associated rate is at least as small

$$I(\hat{x}^{*\prime};x|w,q) \le I(\hat{x}^*;x|w,q), \quad (285)$$

and by optimality of the conditional MMSE estimation, i.e.,

$$E\left[(x-E[x|q,w])^2|q=q,w=w\right] \le E[(x-\hat{x}^*)^2|q=q,w=w] \quad (286)$$

it follows that the distortion induced by $\hat{x}^{*\prime}$ is also at least as small.

Next, we note that the corresponding $v_D^{*\prime\prime}$ satisfies

$$\begin{aligned} &\mathrm{var}[v_D^{*\prime\prime}|w=w] \\ &\quad \le E[(v_D^{*\prime\prime})^2|w=w] && (287)\\ &\quad = \int \left\{\int v^2 p_{x-\hat{x}^{*\prime}|q}(v|q)dv\right\} p_{q|w}(q|w)dq && (288)\\ &\quad = \int E\left[(x-\hat{x}^{*\prime})^2|q=q\right]p_{q|w}(q|w)dq \\ &\quad = \int dq p_{q|w}(q|w) \\ &\qquad \cdot \left\{\int E\left[(x-\hat{x}^{*\prime})^2|q=q,w=w'\right]p_{w|q}(w'|q)dw''\right\} && (289)\\ &\quad \le \sigma_{\max}^2 && (290) \end{aligned}$$

where to obtain (287) we have used that $\mathrm{var}\, t \le E[(t)^2]$ for any random variable t, and to obtain (288) we have used that, for all q and w,

$$p_{v_D^{*\prime\prime}|q,w}(\cdot|q,w) = p_{v_D^{*\prime\prime}|q}(\cdot|q) = p_{x-\hat{x}^{*\prime}|q}(\cdot|q) \quad (291)$$

by the definition of $v_D^{*\prime\prime}$, where to obtain (289) we have used iterated expectation, and where to obtain (290) we have used that, for all q and w,

$$E\left[(x-\hat{x}^{*\prime})^2|q=q,w=w\right] \le \sigma_{\max}^2. \quad (292)$$

To verify (292), note it is trivial for $(q,w) \in \mathcal{S}$. For $(q,w) \notin \mathcal{S}$, we simply note

$$\begin{aligned} &E\left[(x-\hat{x}^{*\prime})^2|q=q,w=w\right] \\ &\quad = \mathrm{var}[x|q=q,w=w] && (293)\\ &\quad \le \mathrm{var}[x|w=w] && (294)\\ &\quad \le \sigma_{\max}^2 && (295) \end{aligned}$$

where (293) follows from the fact that, via (283), $\hat{x}^{*\prime} = E[x|q=q,w=w]$, where (294) follows from the fact that conditioning cannot increase variance (which follows, in turn, from MMSE estimation theory), and where (295) follows from (101).

But (290) establishes that there does exist a distribution satisfying the requirements of Lemma 6 that does not violate (282), thereby establishing the desired contradiction. □

Proof of Theorem 11: First, with $v_D^{*\prime}$ denoting the variable defined in Lemma 6, and following reasoning analogous to that used to obtain (149), we have

$$R_{[\mathrm{Q:ENC,W:DEC}]}(D) \le h(x+v_D^{*\prime}|w) - h(v_D^{*\prime}|q). \quad (296)$$

In turn, we then have

$$\begin{aligned} &R_{[\mathrm{Q:ENC,W:DEC}]}(D) - R_{[\mathrm{Q:BOTH,W:BOTH}]}(D) \\ &\quad \le h(x+v_D^{*\prime}|w) - h(x|w) && (297)\\ &\quad = \int \left[h(\mathcal{N}(x+v_D^{*\prime}|w=w)) - h(\mathcal{N}(x|w=w))\right]p_w(w)dw \\ &\qquad + \int D(p_{x|w}(\cdot|w)||\mathcal{N}(x|w=w))p_w(w)dw \\ &\qquad - \int D(p_{x+v_D^{*\prime}|w}(\cdot|w)||\mathcal{N}(x+v_D^{*\prime}|w=w))p_w(w)dw && (298)\\ &\quad \le \int \left[h(\mathcal{N}(x+v_D^{*\prime}|w=w)) - h(\mathcal{N}(x|w=w))\right]p_w(w)dw \\ &\qquad + \int D(p_{x|w}(\cdot|w)||\mathcal{N}(x|w=w))p_w(w)dw && (299) \end{aligned}$$

where to obtain (297) we have combined (296) with the alternative Shannon lower bound (277), to obtain (298) we have used the following identity [cf. (137)]:

$$D(p_t \,||\, \mathcal{N}(t)) = h(\mathcal{N}(t)) - h(t) \quad (300)$$

valid for any random variable t with finite first and second moments, and where to obtain (299) we have used that divergence is nonnegative.

In addition, we have

$$\begin{aligned} &h(\mathcal{N}(x+v_D^{*\prime}|w=w)) - h(\mathcal{N}(x|w=w)) \\ &\quad = \frac{1}{2}\log\left(1+\frac{\mathrm{var}[v_D^*|w=w]}{\mathrm{var}[x|w=w]}\right) && (301)\\ &\quad \le \frac{1}{2}\log\left(1+\frac{\sigma_{\max}^2}{\mathrm{var}[x|w=w]}\right) && (302)\\ &\quad \le \frac{1}{2}\log\left(1+\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\right) && (303) \end{aligned}$$

where to obtain (301) we have used the identity

$$h(\mathcal{N}(t)) = \frac{1}{2}\log(2\pi e\ \mathrm{var}\ t) \quad (304)$$

together with the fact that x and $v_D^{*\prime}$ are independent given w due to the Markov chain $v_D^{*\prime} \leftrightarrow q \leftrightarrow w \leftrightarrow x$, where to obtain (302) we have used (282) in Lemma 7, and to obtain (303) we have used that by (101)

$$\sigma_{\min}^2 \le \mathrm{var}[x|w=w]$$

for all w.

Finally, using (303) in (299) we obtain (100). □

As a final aside, in principle, analogous bounds for other distortion measures can be developed, and involve replacing, in $D(p_{X|W}||\cdot)$ and $D(p_{X+V}||\cdot)$ above, the Gaussian distribution with the entropy maximizing distribution matched the distortion measure.

REFERENCES

[1] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[2] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: General sources," *Inf., Contr.*, vol. 38, pp. 60–80, 1978.

[3] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1629–1638, Jun. 2002.

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

[5] R. Zamir, "The rate loss in the Wyner–Ziv problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2073–2084, Nov. 1996.

[6] T. Linder, R. Zamir, and K. Zeger, "On source coding with side-information-dependent distortion measures," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2697–2704, Nov. 2000.

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

[8] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[9] R. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 480–489, Jul. 1973.

[10] T. Berger, private communication.

[11] E. Martinian, "Dynamic Information and Constraints in Source and Channel Coding," Ph.D. dissertation, MIT, Cambridge, MA, 2004.

[12] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 406–411, Jul. 1970.

[13] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

[14] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.

[15] H. Feng and M. Effros, "Improved bounds for the rate loss of multiresolution source codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 809–821, Apr. 2003.

[16] L. Lastras and T. Berger, "All sources are nearly successively refinable," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 918–926, Mar. 2001.

[17] M. D. Gaubatz, D. M. Chandler, and S. S. Hemami, "Spatially-selective quantization and coding for wavelet-based image compression," in *Proc. IEEE Int. Conf. Acoustica, Speech, and Signal Processing*, Philadelphia, PA, Mar. 2005, pp. 209–212.

[18] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[19] S. Vembu, S. Verdú, and Y. Steinberg, "The source–channel separation theorem revisited," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.

[20] T. S. Han, "An information-spectrum approach to source coding theorems with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1145–1164, Jul. 1997.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[22] M. Adams and V. Guillemin, *Measure Theory and Probability. Birkäuser*. Cambridge, MA: Birkhäuser, 1996.

[23] I. Csiszár, "On an extremum problem of information theory," *Stud. Sci. Math. Hung.*, pp. 57–70, 1974.

[24] Y. Yamada, S. Tazaki, and R. Gray, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 1, pp. 6–14, Jan. 1980.

[25] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.

[26] V. V. Prelov and E. C. van de Meulen, "An asymptotic expression for the information and capacity of a multidimensional channel with weak input signals," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1728–1735, Sep. 1993.

[27] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.