Source coding with distortion side information

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

Abstract

We analyze the impact of side information about the distortion measure in problems of quantization. We show that such "distortion side information" is not only useful at the encoder, but that under certain conditions, knowing it only at the encoder is as good as knowing it at both encoder and decoder, and knowing it at only the decoder is useless. 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. Furthermore, when both types of side information are present, we characterize the penalty for deviating from the often sufficient configuration of encoder-only distortion side information and decoder-only signal side information.

Index Terms

Wyner-Ziv coding, distributed source coding, quantization, smart compression, sensor networks

I. Introduction

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

Manuscript received December 27, 2004, revised January 16, 2007. This work was supported in part by the National Science Foundation under Grant No. CCF-0515109, and Hewlett-Packard through the MIT/HP Alliance. This work was presented in part at the Data Compression Conference, Snowbird, UT, Mar. 2004, and at the International Symposium on Information Theory, Chicago, IL, June 2004.

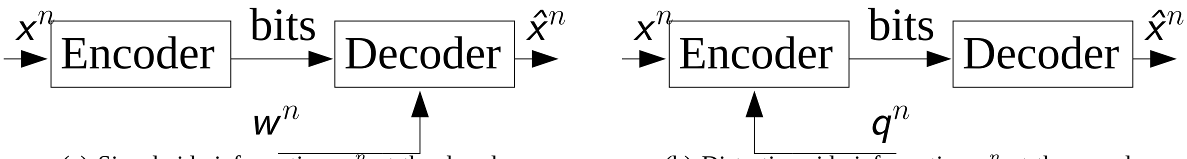
E. Martinian is affiliated with the Signals, Information, and Algorithms Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139. (Email: emin@alum.mit.edu).

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

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

In such scenarios, it is important to understand a variety of issues. These include: 1) the penalties incurred by to the lack of full, globally shared information; 2) the best way to combine available information from different sources; and 3) where different kinds of information is most useful in the system.

A simple example of such a scenario was introduced by Wyner and Ziv [1], and is illustrated in Fig. 1(a). An encoder observes a signal[1] x^n to be conveyed over a digital link to a decoder who also has some additional *signal side information* w^n, which is correlated with x^n. An analysis of the fundamental performance limits for this problem [1], [2] [3], [4], [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.



(a) Signal side information w^n at the decoder.

(b) Distortion side information q^n at the encoder.

Fig. 1. Compressing a source x^n with side information.

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 in a suitable way.

We develop the fundamental rate-distortion trade-off for this problem. Our analysis reveals, under reasonable conditions, both that such side information is useful only if available at the encoder, 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 encoder. As such, distortion side information plays a complementary role to that of signal side information as developed by Wyner and Ziv.

[1] Throughout this paper, sequences are denoted using superscripts and sequence elements with subscripts (e.g., $x^n = (x_1, x_2, \ldots, x_n)$), and random variables and sequences distinguished by the use of sans serif fonts (e.g., $\mathsf{x}^n = (\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n)$).

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 suitable conditions. Specifically, we demonstrate 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. We also derive the penalty for deviating from this often sufficient side information configuration.

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 [7], [3], a point to which we will return in the paper (Theorems 4 and 5 in the sequel) to develop additional insight.

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 can apply signal analysis to determine which parts of the signal are more or less sensitive to distortion due to context, masking effects, and other perceptual phenomena [8]. While the conventional approach to exploiting such side information in practice in these kinds of examples involves sharing it with decoders via a side channel, the results of this paper suggest that this can be an unnecessary and inefficient use of bandwith.

An outline of the paper is as follows. Section II introduces the formal problem model of interest. Section III then develops the rate distortion tradeoffs for source coding with only distortion side information, and in particular identifies conditions under which such side information is sufficient at the encoder. Section IV then extends the problem of interest to include both signal and distortion side information in the case of continuous-sources in the high-resolution regime. For this scenario, several equivalence and loss theorems are developed that quantify the degree to which some side information configurations yield the same and different rate-distortion behaviors. Section V then develops bounds on losses incurred at lower resolution when complete side information is not available. Finally, Section VI contains some concluding remarks. Throughout the paper, most proofs and longer derivations are deferred to appendices.

II. Problem Model

The general rate-distortion problem with side information z 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 likewise consists of n samples drawn from the alphabet $\mathcal{Z}$. These random variables are drawn according to the distribution

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

A rate R encoder $f(\cdot)$ maps the source x^n as well as possible encoder side information to an index $i \in \{1, 2, \ldots, 2^{nR}\}$. The corresponding decoder $g(\cdot)$ maps the resulting index as well as possible decoder side information to a reconstruction $\hat{\mathsf{x}}^n$ of the source, which takes values in the alphabet $\hat{\mathcal{X}}$. Distortion in a reconstruction $\hat{x}^n$ of a source x^n is measured via

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

where we explicitly denote the dependence, in general, of the distortion measure on the side information. As usual, the rate-distortion function is the minimum rate 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^n can be decomposed into two kinds of side information, which we term "signal side information" w^n and "distortion side information" q^n, i.e., $\mathsf{z}^n = (\mathsf{w}^n, \mathsf{q}^n)$. 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 $\mathsf{z}^n = (\mathsf{w}^n, \mathsf{q}^n)$ of side information z^n into signal side information w^n and distortion side information q^n for a rate-distortion problem with source x^n and distortion measure $d(\mathsf{x}, \hat{\mathsf{x}}; \mathsf{z})$ is* admissible *if the following Markov chains are satisfied:*

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

and

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

Several remarks are worthwhile before proceeding with our development.

First, note that (4a) is equivalent to the condition

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

and that when (4b) holds, we can (and will), with slight abuse of notation, use $d(x, \hat{x}; q)$ in place of $d(x, \hat{x}; z)$.

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

Third, when separating phenomena that have physically different origins, such decompositions arise quite naturally. Moreover, in such cases, the resulting signal and distortion side informations 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 — which corresponds to a situation in which q^n and x^n are independent not just conditioned on w^n as per (4a), but unconditionally as well — on the side information requirements of Definition 1. Moreover, in this case q^n is independent of $(\mathsf{x}^n, \mathsf{w}^n)$ as well. However, it should be emphasized that admissible decompositions that satisfy this further restriction are not always possible, and later in the paper we characterize the penalties incurred by the lack of a suitable decomposition.

It is also worth emphasizing that a further subclass of side information scenarios with q^n and w^n independent corresponds to the case in which signal side information is altogether absent ($\mathsf{w}^n = \emptyset$), in which case q^n and x^n are independent. This case will also be of special interest in parts of the paper.

Finally, without any constraints on the structure of the distortion measure $d(x, \hat{x}; q)$ and the nature of its dependency on the side information q, very little can be inferred about the value of such side information at the encoder and/or decoder. Hence, we will typically be interested in special forms of the distortion measure to obtain specific results, a simple example of which would be the modulated quadratic distortion $d(x, \hat{x}; q) = q \cdot (x - \hat{x})^2$ for $x, \hat{x} \in \mathbb{R}$. Each of our key theorems will make clear what particular restrictions on the form of the distortion measure are required.

In the remainder of the paper, we consider the sixteen possible scenarios depicted in Fig. 2, corresponding to where q^n and w^n may each be available at the encoder, decoder, both, or neither. Our notation for the associated rate-distortion functions makes explit where the side-information is available. For example, $R[\text{Q-NONE-W-NONE}](D)$ denotes the rate-distortion function without side information and $R[\text{Q-NONE-W-DEC}](D)$ denotes the Wyner-Ziv rate-distortion function where w^n is available at the decoder [1]. Similarly, when all information is available at both encoder and decoder, $R[\text{Q-BOTH-W-BOTH}](D)$ describes Csiszár and Körner's [4] generalization of Gray's conditional rate-distortion function $R[\text{Q-NONE-W-BOTH}](D)$ [9] to the case where the side information can affect the distortion measure.

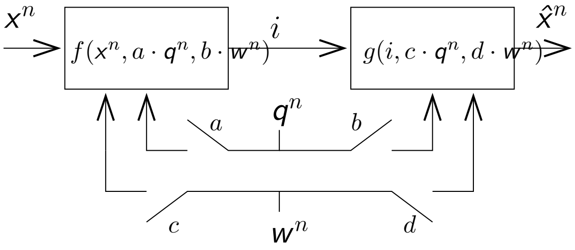

x^n
$f(x^n, a \cdot q^n, b \cdot w^n)$
i
$g(i, c \cdot q^n, d \cdot w^n)$
$\hat{x}^n$
a
q^n
b
c
w^n
d

Fig. 2. Scenarios for source coding with distortion side information q^n and signal side information w^n. The labels a, b, c, and d are 0 (respectively, 1) if the corresponding switch is open (respectively, closed) and the side information is unavailable (respectively, available) to the encoder $f(\cdot)$ or decoder $g(\cdot)$ as shown.

As pointed out by Berger [10], all the rate-distortion functions may be derived by considering q^n as part of x^n or w^n (i.e., by considering the "super-source" $\tilde{x}^n = (x^n, q^n)$ or the "super-side-information" $\tilde{w}^n = (w^n, q^n)$) 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 case in which there is no signal side information.

III. Source Coding with Distortion Side Information Alone

It is straightforward to express the rate-distortion tradeoffs 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 when there is distortion side information q^n but not signal* *side information w^n are:*

$$R[\text{Q-NONE}](D) = \inf_{p_{\hat{x}|x}(\hat{x}|x):E[d(x,\hat{x};q)]\leq D} I(x;\hat{x}) \tag{6a}$$

$$R[\text{Q-DEC}](D) = \inf_{p_{u|x}(u|x),v(\cdot,\cdot):E[d(x,v(u,q);q)]\leq D} I(x;u) - I(u;q) = \inf_{p_{u|x}(u|x),v(\cdot,\cdot):E[d(x,v(u,q);q)]\leq D} I(x;u) \tag{6b}$$

$$R[\text{Q-ENC}](D) = \inf_{p_{\hat{x}|x,q}(\hat{x}|x,q):E[d(x,\hat{x};q)]\leq D} I(x,q;\hat{x}) = \inf_{p_{\hat{x}|x,q}(\hat{x}|x,q):E[d(x,\hat{x};q)]\leq D} I(x;\hat{x}|q) + I(\hat{x};q) \tag{6c}$$

$$R[\text{Q-BOTH}](D) = \inf_{p_{\hat{x}|x,q}(\hat{x}|x,q):E[d(x,\hat{x};q)]\leq D} I(x;\hat{x}|q). \tag{6d}$$

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

In the remainder of this section, we turn our attention to developing conditions under which having the distortion side information only at the encoder is as good as having it at both encoder and decoder. Before developing our formal results, we first describe two simple examples of such behavior in a fairly qualitative manner. These examples both establish that the associated conditions will not be degenerate, and provide preliminary intuition.

A. Motivating 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.

1) Discrete Source: Consider a source x^n whose n samples are drawn uniformly and independently from the finite alphabet $\mathcal{X}$ with cardinality $|\mathcal{X}| \geq n$. Let q^n correspond to the n binary variables 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_i be statistically independent of the source with q_i drawn uniformly from the $\binom{n}{k}$ subsets with exactly k ones.[2]

If the side information were unavailable or ignored, then losslessly communicating the source would require exactly $n \cdot \log|\mathcal{X}|$ bits. When $H_\text{b}(k/n) < (1-k/n)\log|\mathcal{X}|$, a better (though still sub-optimal) approach when encoder side information is available would be for the encoder to first tell the decoder which samples are relevant and then send only those samples. Using Stirling's approximation, this would require about $n \cdot H_\text{b}(k/n)$ bits (where $H_\text{b}(\cdot)$ denotes the binary entropy function) to describe which

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

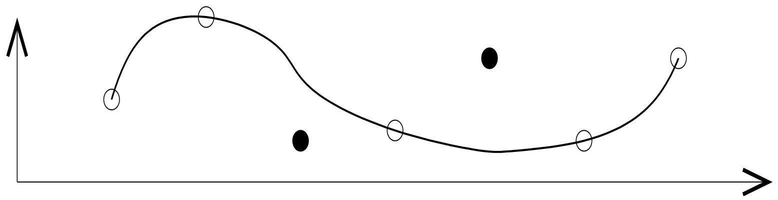


Fig. 3. Source coding with erasure distortion side information model. In this example, only $k = 5$ of the $n = 7$ source samples are relevant (i.e., the unshaded ones). Source encoding can be implemented by exactly fitting a fourth-degree curve to the relevant points, which corresponds to using the Reed-Solomon decoding algorithm, as described in the text. The resulting curve is described by k elements, yielding the optimum achievable compression ratio of k/n.

samples are relevant plus $k \cdot \log |\mathcal{X}|$ bits to describe the relevant source samples. Note that if the side information were also known at the decoder, then the overhead required in telling the decoder which samples are relevant could be avoided and the total rate required would only be $k \cdot \log |\mathcal{X}|$. This overhead can in fact be avoided even without decoder side information.

To see this, 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[3]) with $\mathsf{q}_i = 0$ indicating an erasure at sample i. We use the Reed-Solomon *decoding* algorithm to “correct” the erasures and determine the k corresponding information symbols, which are sent to the receiver. To reconstruct the signal, the receiver *encodes* the k information symbols using the encoder for the (n, k) Reed-Solomon code to produce the reconstruction $\hat{\mathsf{x}}^n$. Only symbols with $\mathsf{q}_i = 0$ could have changed, hence $\hat{\mathsf{x}}_i = \mathsf{x}_i$ whenever $\mathsf{q}_i = 1$ and the relevant samples are losslessly communicated using only $k \cdot \log |\mathcal{X}|$ bits.

As illustrated in Fig. 3, it is worth recalling that Reed-Solomon decoding can be viewed as curve-fitting and Reed-Solomon encoding can be viewed as interpolation. Hence this source coding approach can be interpreted as fitting a curve of degree k to the points of x_i where $\mathsf{q}_i = 1$. The resulting curve can be specified using just k elements. It perfectly reproduces x_i where $\mathsf{q}_i = 1$ and interpolates the remaining points.

Finally, 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,

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

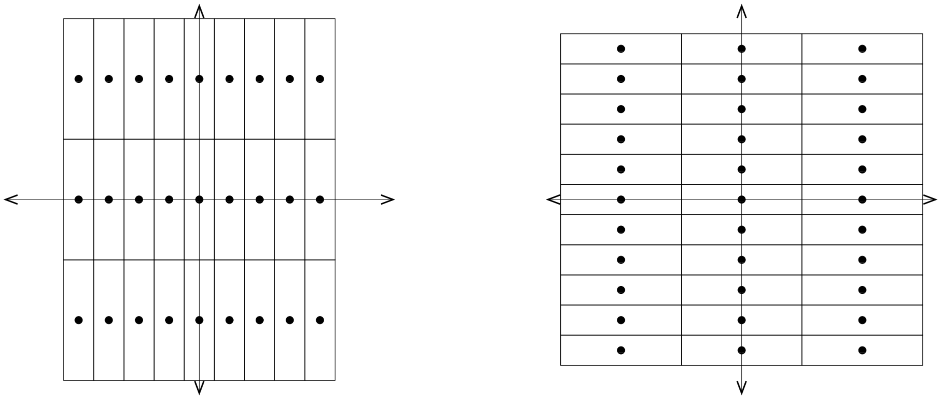


Fig. 4. Quantizers for distortion side information available at encoder and decoder. When the side information q indicates the horizontal error (respectively, vertical error) is more important, the encoder uses the codebook lattice and partition function on the left (respectively, right) to increase horizontal accuracy (respectively, vertical accuracy).

sequence with period n, which is band-limited in the sense that only its first k DFT coefficients are non-zero. 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.

2) Continuous Source: Consider the quantization of a single pair of samples (i.e., $x \in \mathbb{R}^2$) from a continuous source. The distortion side information q is binary, corresponding to two possible additive difference distortion measures. In one measure, the first of the samples is more important than the other. In the other measure, it is the second sample that is more important. As one example, each of the two measures could be weighted quadratic measures.

If the side information q were available to both encoder and decoder, then one could choose a codebook lattice and encoder (i.e., 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 straightforward to see, neglecting edge effects and considering a uniformly distributed source, that having the distortion side information only at the encoder incurs

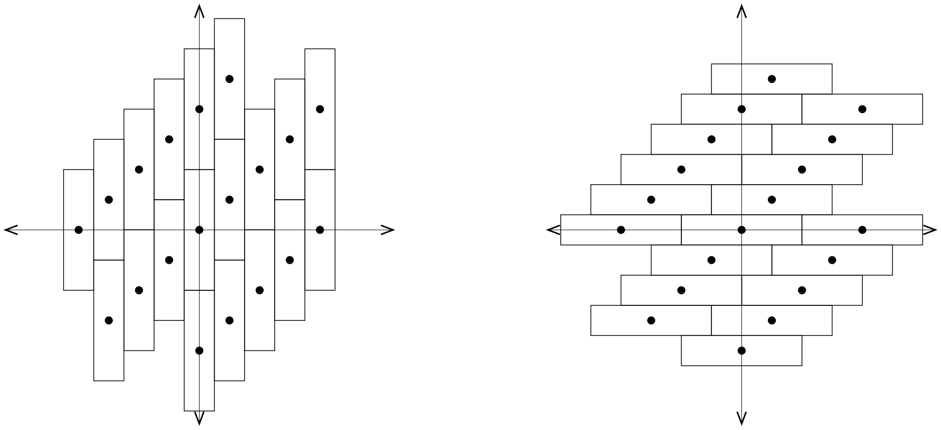


Fig. 5. Quantizers for distortion side information available only at the encoder. A common codebook lattice is used, independent of the realized side information, but when the side information indicates that the horizontal error (respectively, vertical error) is more important, the encoder uses the partition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

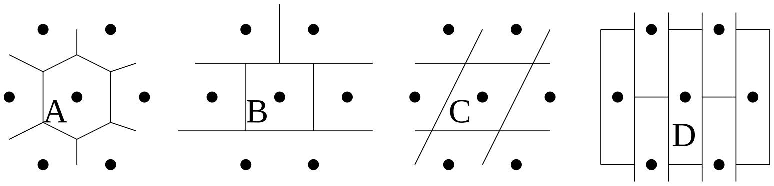


Fig. 6. A fixed-codebook / variable-partition encoder. In this simple example, the codebook is a simple hexagonal lattice in two dimensions, and there are four different partitions, corresponding to two bits of distortion side information.

no additional distortion. Later in the paper we will make such statements more 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 (see, e.g., Fig. 6) play more generally in the associated systems. Moreover, this encoding strategy generalizes readily to arbitrary block lengths, and can be implemented with only linear complexity in the block length, as described in [12].

We know turn developing our main results of the section, characterizing when distortion side information at the encoder is sufficient more generally.

B. Sufficiency of Encoder-Only Side Information

We begin by comparing the rate-distortion functions in Proposition 1. In particular, knowing q^n only at the encoder is as good as knowing it at both encoder and decoder whenever $R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D)$, from which we obtain the following equivalent condition.

Proposition 2 *Knowing q^n only at the encoder is as good as knowing it at both encoder and decoder, i.e.,* $R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D)$*, if and only if* $I(\hat{\mathsf{x}}; \mathsf{q}) = 0$ *for some* $\hat{\mathsf{x}}$ *that optimizes* (6d).

To prove Proposition 2, it suffices to equate (6c) and (6d) and note that if in this case some $\hat{\mathsf{x}}$ optimizes (6d), it also optimizes (6c).

Proposition 2 admits a simple interpretation. In particular, since $p_{\hat{\mathsf{x}}|\mathsf{q}}(\hat{x}|q)$ represents the distribution of the codebook, the condition $I(\hat{\mathsf{x}}; \mathsf{q}) = 0$ corresponds to the requirement that the codebook distribution be independent of the side information. In the language of our example of Section III-A.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).

There are two natural scenarios where $I(\hat{\mathsf{x}}; \mathsf{q})$ can be zero: the case of uniform sources with group difference distortions, and the case of erasure distortions. We consider each separately, in turn.

1) Uniform Sources with Group Difference Distortions: Let the source x be uniformly distributed over a group $\mathcal{X}$ with the binary relation $\oplus$. For convenience, we use the symbol $a \ominus b$ to denote $a \oplus b^{-1}$ (where b^{-1} denotes the additive inverse of b in the group). We define a group difference distortion measure as any distortion measure where

$$d(x, \hat{x}; q) = \rho(\hat{x} \ominus x; q) \tag{7}$$

for some function $\rho(\cdot; \cdot)$. As we will show, the symmetry in this scenario insures that the optimal codebook distribution is uniform. This allows an encoder to design a fixed codebook and vary the quantization partition based on q^n to achieve the same performance as a system where both encoder and decoder know q^n. This uniformity of the codebook, made precise in the following theorem, provides a general information theoretic explanation for the behavior observed in the Reed-Solomon example of Section III-A.1.

Theorem 1 *Consider a source x that is uniformly distributed over a group with a distortion measure of the form* (7)*, where the distortion side information q is independent of x. Then*

$$R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D). \tag{8}$$

For either finite or continuous groups this theorem can be proved by deriving the conditional Shannon Lower Bound (which holds for any source) and showing that this bound is tight for uniform sources. We use this approach below to give some intuition. For more general "mixed groups" with both discrete and continuous components, entropy is not well defined and a more elaborate argument based on symmetry and convexity is provided in Appendix A.

Lemma 1 (Conditional Shannon Lower Bound) *Let the source x be uniformly distributed over a discrete group, $\mathcal{X}$, with a difference distortion measure, $\rho(x \ominus \hat{x}; q)$. Define the conditional maximum entropy variable v^* as the random variable that maximizes $H(\mathsf{v}|\mathsf{q})$ subject to the constraint $E[\rho(\mathsf{v};\mathsf{q})] \leq D$. Then, the rate-distortion function with q^n known at both encoder and decoder (and hence also the rate-distortion function with q^n known only at the encoder) is lower bounded by*

$$R[\text{Q-ENC}](D) \geq R[\text{Q-BOTH}](D) \geq \log |\mathcal{X}| - H(\mathsf{v}^*|\mathsf{q}). \tag{9}$$

For continuous groups, we can replace $|\mathcal{X}|$ and $H(\mathsf{v}^*|\mathsf{q})$ in (9) *(as well as the following proof) with the Lebesgue measure of the group and the differential entropy $h(\mathsf{v}^*|\mathsf{q})$.*

Proof:

$$\begin{aligned} I(\hat{\mathsf{x}};\mathsf{x}|\mathsf{q}) &= H(\mathsf{x}|\mathsf{q}) - H(\mathsf{x}|\mathsf{q},\hat{\mathsf{x}}) && (10)\\ &= \log|\mathcal{X}| - H(\mathsf{x}|\hat{\mathsf{x}},\mathsf{q}) && (11)\\ &= \log|\mathcal{X}| - H(\hat{\mathsf{x}} \ominus \mathsf{x}|\mathsf{x},\mathsf{q}) && (12)\\ &\geq \log|\mathcal{X}| - H(\hat{\mathsf{x}} \ominus \mathsf{x}|\mathsf{q}) && (13)\\ &\geq \log|\mathcal{X}| - H(\mathsf{v}^*|\mathsf{q}), && (14) \end{aligned}$$

where (13) follows since conditioning reduces entropy, and (14) follows from the definition of v^* since $E[\rho(\hat{\mathsf{x}} \ominus \mathsf{x};\mathsf{q})] \leq D$. ■

Proof of Theorem 1: Choosing the test-channel distribution $\hat{\mathsf{x}} = \mathsf{v}^* + \mathsf{x}$ with the pair $(\mathsf{v}^*, \mathsf{q})$ independent of x achieves the bound in (9) with equality and must therefore be optimal. Furthermore, since x is uniform, so is $\hat{\mathsf{x}}$ and therefore $\hat{\mathsf{x}}$ and q are statistically independent. Therefore $I(\hat{\mathsf{x}};\mathsf{q}) = 0$ and thus comparing (6c) to (6d) shows $R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D)$ for finite groups. The same argument holds for continuous groups with entropy replaced by differential entropy and $|\mathcal{X}|$ replaced by Lebesgue measure. ■

Uniform source and group difference distortion measures arise naturally in a variety of applications. One example is phase quantization where applications such as Magnetic Resonance Imaging, Synthetic Aperture Radar, and Ultrasonic Microscopy infer physical phenomena from the phase shifts induced in a probe signal [13], [14], [15]. Alternatively, when magnitude and phase information must both be recorded, there are sometimes advantages to treating these separately, [16], [17], [18], [19]. The key special case when only two phases are recorded corresponds to Hamming distortion. Let us use this special case to illustrate how distortion side information affects quantization.

For a symmetric binary source x with side information q taking values in $\{1, 2, \ldots, N\}$ according to distribution $p_q(q)$, the general side-information dependent Hamming distortion measure of interest takes the form

$$d(x, \hat{x}; q) = \alpha_q + \beta_q \cdot d_H(x, \hat{x}), \tag{15}$$

where $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ and $\{\beta_1, \beta_2, \ldots, \beta_N\}$ are sets of non-negative weights.

For this case, the associated rate distortion expressions are, when $D \geq E[\alpha_q]$,

$$R[\text{Q-NONE}](D) = R[\text{Q-DEC}](D) = 1 - H_\text{b}\left(\frac{D - E[\alpha_q]}{E[\beta_q]}\right) \tag{16a}$$

$$R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D) = 1 - \sum_{i=1}^{N} p_q(i) \cdot H_\text{b}\left(\frac{2^{-\lambda\beta_i}}{1 + 2^{-\lambda\beta_i}}\right), \tag{16b}$$

where λ is chosen to satisfy the distortion constraint

$$\sum_{i=1}^{N} p_q(i) \left[\alpha_i + \beta_i \cdot \frac{2^{-\lambda\beta_i}}{1 + 2^{-\lambda\beta_i}}\right] = D. \tag{16c}$$

The derivations of (16) are provided in Appendix B.

Two special cases of (15) are worth developing in more detail for additional insight.

a) Noisy Observations: One special case of (15) corresponds to quantizing noisy observations. In particular, 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

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

which is always 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:

$$\begin{aligned} d(x, \hat{x}; q) &= \epsilon_q \cdot [1 - d_H(x, \hat{x})] + (1 - \epsilon_q) \cdot d_H(x, \hat{x}) \\ &= \epsilon_q + (1 - 2\epsilon_q) \cdot d_H(x, \hat{x}) \\ &= \frac{q - 1}{2(N - 1)} + \left(1 - \frac{q - 1}{N - 1}\right) \cdot d_H(x, \hat{x}). \end{aligned} \tag{17}$$

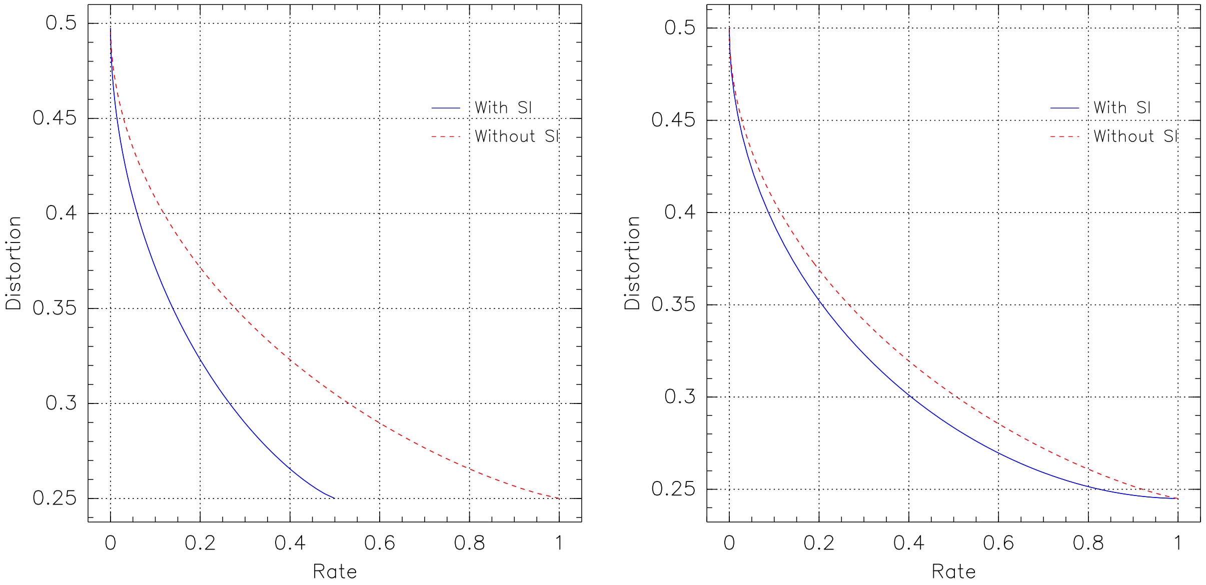


Fig. 7. Rate-Distortion tradeoffs for noisy observations of a binary source. The solid and dashed curves represents the minimum possible Hamming distortion when side information specifying the cross-over probability of the observation noise is and is not available at the encoder, respectively. For the plot on the left the crossover probability for the observation noise is equally likely to be 0 or 1/2, while for the plot on the right it is uniformly distributed over the interval $[0,1/2]$.

Evidently, (17) corresponds to a distortion measure in the form of (15) with

$$\alpha_q = \frac{q-1)}{2(N-1)} \quad \text{and} \quad \beta_q = 1 - \frac{q-1}{N-1},$$

so the rate-distortion formulas of (16) apply. 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 [20].

The rate-distortion tradeoffs for this noisy observations special case are depicted in Fig. 7. The left plot corresponds to $N = 2$, while the right plot corresponds to $N \to \infty$. In each plot, the solid curve shows the tradeoff achievable when the side information is available at the encoder, while the dashed curve shows the (poorer) tradeoff achievable when it is not.

From this special case it is apparent that a naive encoding method whereby the encoder losslessly communicates the side information to the decoder, then uses encoding for the case of side information at both encoder and decoder, can require arbitrarily higher *rate* than the optimal rate-distortion trade-off. Indeed, to losslessly encode the side information requires an additional rate of $\log N$, which is unbounded in N.

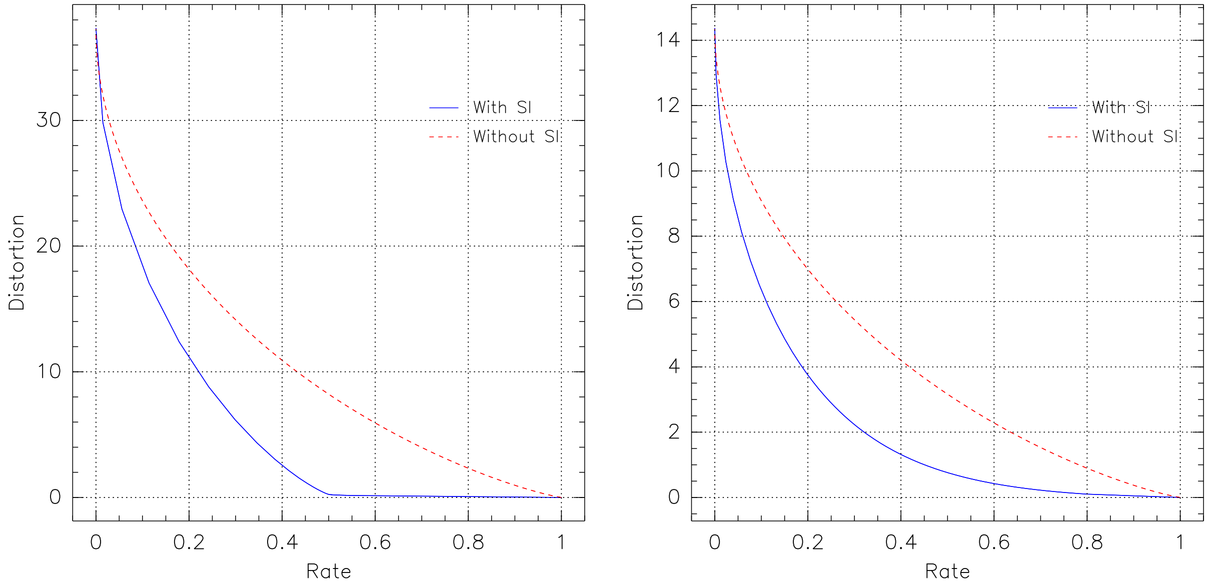


Fig. 8. Rate-Distortion tradeoffs for a binary source, where the Hamming distortion in quantizing each source sample is weighted $\exp(5q)$. The solid and dashed curves represent the minimum possible Hamming distortion when side information specifying the weight is and is not available at the encoder, respectively. In the left plot, q is uniformly distributed over the pair $\{0,1\}$ while in the right plot q is uniformly distributed over the interval $[0,1]$.

b) Weighted Distortion: In a number of applications, certain samples of a source are inherently more important than others — e.g., edges other perceptually important features in an image, or sensor readings in high activity areas. Such a scenario corresponds to quantizing with a weighted distortion measure, which is a different special case of (15). Specifically, we consider a distortion measure of the form (15) where $\beta_i = \exp(\gamma i/N)$, $\alpha_i = 0$, and the side information is uniformly distributed over $\{0, 1, \ldots, N-1\}$.

The rate-distortion tradeoffs for this weighted distortion special case with $\gamma = 5$ are depicted in Fig. 8. The left and right plots correspond to $N = 2$ and $N \to \infty$, respectively. In each plot, the solid curve shows the tradeoff achievable when the side information is available at the encoder, while the dashed curve shows the (poorer) tradeoff achievable when it is not. Note that when the side information is not available, the system is limited to treating all samples equally, while the system with side information will assign more bits to the samples for which the associated weights in the Hamming distortion measure are larger.

This special case of weighted Hamming distortion can also be used to demonstrate that ignoring the side information at the encoder can result in arbitrarily higher *distortion* than the minimum required by optimal schemes. To see this, it suffices to restrict attention to the case $N = 2$ and observe that as

$\gamma \to \infty$, the system not using side information suffers increasingly more distortion. This is most evident for $R > 1/2$. In this rate region, the system with side information losslessly encodes the important samples and the distortion is bounded by $1/2$ while the system without side information has a distortion that scales with $\exp(\gamma/2)$. Thus the extra distortion incurred when q is not available to the encoder can be arbitrarily large.

2) Erasure Distortions: The other natural scenario where it is sufficient for distortion side information to be available only at the encoder is for "erasure distortions" whereby $\mathsf{q} \in \{0, 1\}$ and the distortion measure is of the form

$$d(x, \hat{x}; q) = q \cdot \rho(x, \hat{x}) \tag{18}$$

for some function $\rho(\cdot, \cdot)$ that is itself a valid distortion measure. In particular, we have the following

Theorem 2 *For any source distribution, if the distortion measure is of the form in* (18) *with* $\mathsf{q}^n \in \{0, 1\}^n$, *then*

$$R[\text{Q-ENC}](D) = R[\text{Q-BOTH}](D). \tag{19}$$

Before proceeding with our proof, we remark that as the example of Section III-A.1 suggests, 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 particularly simple.

Proof: Let $\hat{\mathsf{x}}^*$ be a distribution that optimizes (6d). Choose the new random variable $\hat{\mathsf{x}}^{**}$ to be the same as $\hat{\mathsf{x}}^*$ when $\mathsf{q} = 0$ and when $\mathsf{q} = 1$, let $\hat{\mathsf{x}}^{**}$ be independent of x with the same marginal distribution as when $\mathsf{q} = 0$:

$$p_{\hat{\mathsf{x}}^{**}|\mathsf{x},\mathsf{q}}(\hat{x}|x, q) = \begin{cases} p_{\hat{\mathsf{x}}^*|\mathsf{x},\mathsf{q}}(\hat{x}|x, q), & q = 0 \\ p_{\hat{\mathsf{x}}^*|\mathsf{q}}(\hat{x}|q = 0), & q = 1. \end{cases} \tag{20}$$

Both $\hat{\mathsf{x}}^*$ and $\hat{\mathsf{x}}^{**}$ have the same expected distortion since they only differ when $\mathsf{q} = 0$. Furthermore, by the data processing inequality

$$I(\hat{\mathsf{x}}^{**}; \mathsf{x}|\mathsf{q}) \leq I(\hat{\mathsf{x}}^*; \mathsf{x}|\mathsf{q}) \tag{21}$$

so $\hat{\mathsf{x}}^{**}$ also optimizes (6d). Finally, since $I(\hat{\mathsf{x}}^{**}; \mathsf{q}) = 0$, Proposition 2 is satisfied and we obtain the desired result. ■

IV. 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. In contrast to the treatment of Section III, here we will emphasize the case of continuous sources, whose elements take values in $\mathbb{R}^k$ for some integer $k \geq 1$. At the same time, in examining issues of sufficiency of different types of side information, we will consider a looser asymptotic sufficiency in the high resolution limit.

When there is no signal side information, one would expect to find asymptotic sufficiency of encoder-only distortion side information rather generally. Indeed, as $D \to D_{\min}$, where $D_{\min}$ denotes the minimum attainable value, $\hat{\mathsf{x}}^n \to \mathsf{x}^n$. Thus when x^n and q^n are independent we may intuitively expect $\hat{\mathsf{x}}^n \to \mathsf{x}^n$ to imply $I(\hat{\mathsf{x}}; \mathsf{q}) \to I(\mathsf{x}; \mathsf{q}) = 0$. This turns out to be the case under reasonable conditions, as we formally develop in this section.

More generally, when there is both signal and distortion side information, we show the asymptotic sufficiency of encoder-only distortion side information and decoder-only signal side information in the high resolution limit in several natural scenarios of interest.

A. Admissibility Requirements

We begin by defining the class of continuous-source problems of interest. In addition to the side information decomposition implied by Definition 1, our results require a “continuity of entropy” property that essentially states

$$\mathsf{v} \to 0 \text{ in distribution } \Rightarrow h(\mathsf{x} + \mathsf{v} | \mathsf{q}, \mathsf{w}) \to h(\mathsf{x} | \mathsf{q}, \mathsf{w}). \tag{22}$$

The desired continuity follows from [21] provided the source, distortion measure, and side information satisfy some technical conditions related to smoothness. These conditions are not particularly hard to satisfy; for example, any source, side information, and distortion measure where

$$\exists \delta > 0, -\infty < E[\|\mathsf{x}\|^{\delta} \mid \mathsf{w} = w] < \infty \quad \forall w \tag{23a}$$

$$-\infty < h(\mathsf{x} \mid \mathsf{w} = w) < \infty, \quad \forall w \tag{23b}$$

$$d(x, \hat{x}; q) = \alpha(q) + \beta(q) \cdot \|x - \hat{x}\|^{\gamma(q)} \tag{23c}$$

will satisfy the desired technical conditions in [21] provided $\alpha(\cdot)$, $\beta(\cdot)$, and $\gamma(\cdot)$ are non-negative functions.

For more general scenarios we introduce the following definition to summarize the requirements from [21].

Definition 2 *The collection of a source* x*, a side information pair* (q, w)*, and a difference distortion measure* $d(x, \hat{x}; q) = \rho(x - \hat{x}; q)$ *is said to be* admissible *if, in addition to the conditions* (4) *of Definition 1, the following conditions are satisfied:*

1) *the equations*

$$a(D,q)\int \exp[-s(D,q)\rho(x;q)]dx = 1 \tag{24a}$$

$$a(D,q)\int \rho(x;q)\exp[-s(D,q)\rho(x;q)]dx = D \tag{24b}$$

have a unique pair of solutions $(a(D,q), s(D,q))$ *for all* $D > D_{\min}$ *that are continuous functions of their arguments*

2) $-\infty < h(\mathsf{x}|\mathsf{w} = w) < \infty$*, for all* w

3) *For each value of* q*, there exists an auxiliary distortion measure* $\delta(\cdot;q)$ *where the equations*

$$a_\delta(D,q)\int \exp[-s_\delta(D,q)\delta(x;q)]dx = 1 \tag{25a}$$

$$a_\delta(D,q)\int \delta(x;q)\exp[-s_\delta(D,q)\delta(x;q)]dx = D \tag{25b}$$

have a unique pair of solutions $(a_\delta(D,q), s_\delta(D,q))$ *for all* $D > D_{\min}$ *that are continuous functions of their arguments*

4) *The conditional maximum entropy random variable* v^* *that maximizes* $h(\mathsf{v}|\mathsf{q})$ *subject to the constraint* $E[\rho(\mathsf{v};\mathsf{q})] \leq D$ *has the property that*

$$\lim_{D\to D_{\min}} \mathsf{v}^* \to 0 \text{ in distribution } \forall q \tag{26a}$$

$$\lim_{D\to D_{\min}} E[\delta(\mathsf{x}+\mathsf{v}^*,\mathsf{q})|\mathsf{q}=q] = E[\delta(\mathsf{x},\mathsf{q})|\mathsf{q}=q] \quad \forall q. \tag{26b}$$

B. Equivalence Theorems

Our main results for continuous sources are a set of four theorems describing when different types of side information knowledge are equivalent. Our results show that the sixteen possible information configurations of Fig. 2 can be reduced to the four shown in Fig. 9. Our four equivalence theorems presented below are proved in Appendix C.

We begin by establishing that, under suitable conditions, having the distortion side information at the encoder and the signal side information at the decoder is sufficient to ensure there is no loss relative to the case of complete side information everywhere:

Theorem 3 *For a scenario in which Definition 2 is satisfied, and a difference distortion measure of the form* $\rho(x-\hat{x};q)$ *is involved,* q^n *and* w^n *can be divided between the encoder and decoder with no asymptotic penalty, i.e.,*

$$\lim_{D\to D_{\min}} R[\textsc{q-enc-w-dec}](D) - R[\textsc{q-both-w-both}](D) = 0. \tag{27}$$

	Decoder missing w^n			Decoder has w^n		
Encoder Missing q^n	$R[\text{Q-DEC-W-ENC}]$	$\overset{\text{Th. 4 (I)}}{\Longleftrightarrow}$	$R[\text{Q-DEC-W-NONE}]$	$R[\text{Q-DEC-W-BOTH}]$	$\overset{\text{Th. 5 (S)}}{\Longleftrightarrow}$	$R[\text{Q-NONE-W-BOTH}]$
	$\Updownarrow$ Th. 5 (S+I)		$\Updownarrow$ Th. 5 (S+I)	$\Updownarrow$ Th. 6 (H+S+I)		$\Updownarrow$ Th. 6 (H+S+I)
	$R[\text{Q-NONE-W-ENC}]$	$\overset{\text{Th. 4 (I)}}{\Longleftrightarrow}$	$R[\text{Q-NONE-W-NONE}]$	$R[\text{Q-DEC-W-DEC}]$	$\overset{\text{Th. 5 (S)}}{\Longleftrightarrow}$	$R[\text{Q-NONE-W-DEC}]$
Encoder has q^n	$R[\text{Q-ENC-W-ENC}]$	$\overset{\text{Th. 4}}{\Longleftrightarrow}$	$R[\text{Q-ENC-W-NONE}]$	$R[\text{Q-ENC-W-DEC}]$	$\overset{\text{Th. 6 (H)}}{\Longleftrightarrow}$	$R[\text{Q-ENC-W-BOTH}]$
	$\Updownarrow$ Th. 6 (H+I)		$\Updownarrow$ Th. 6 (H+I)	$\Updownarrow$ Th. 6 (H)		$\Updownarrow$ Th. 6 (H)
	$R[\text{Q-BOTH-W-ENC}]$	$\overset{\text{Th. 4}}{\Longleftrightarrow}$	$R[\text{Q-BOTH-W-NONE}]$	$R[\text{Q-BOTH-W-DEC}]$	$\overset{\text{Th. 6 (H)}}{\Longleftrightarrow}$	$R[\text{Q-BOTH-W-BOTH}]$

Fig. 9. Summary of equivalence results for continuous sources. Arrows indicate which theorems demonstrate equality between various rate-distortion functions and list the assumptions required (H = high-resolution, I = q^n and w^n independent, S = scaled difference distortion).

Theorem 3 establishes that there is a natural division of side information between the encoder and decoder (at least asymptotically). Ultimately, this theorem can be viewed as generalizing prior results on the lack of rate loss for the Wyner-Ziv problem in the high-resolution limit [5] [6].

In some ways, Theorem 3 is quite remarkable. The admissibility conditions (4) require q^n to be conditionally independent of x^n given w^n, and require the distortion to be conditionally independent of w^n given q^n, x^n, and $\hat{x}^n$. However, since our model allows for q^n and w^n to be statistically dependent, q^n can be indirectly correlated with x^n (through w^n) and w^n can indirectly affect the distortion (through q^n).

The next pair of theorems show, under appropriate conditions, that w^n known only at the encoder is useless, and q^n known only at the decoder is useless. Hence, deviating from the natural division of Theorem 3 and providing side information in the wrong place makes that side information useless (at least in terms of the rate-distortion function). As such, these theorems generalize Berger's result that signal side information is useless when known only at the encoder [7].

Theorem 4 *For a scenario in which Definition 1 is satisfied, q^n and w^n are independent,*[4]*, and a difference distortion measure of the form $\rho(x - \hat{x}; q)$ is involved, w^n provides no benefit when known only at the encoder, i.e.,*

$$R[\text{Q-*-W-ENC}](D) = R[\text{Q-*-W-NONE}](D), \tag{28}$$

where the wildcard "" may be replaced with an element from* $\{\text{ENC}, \text{DEC}, \text{BOTH}, \text{NONE}\}$ *(both *'s must be*

[4] Independence is only required when $* \in \{\text{DEC}, \text{NONE}\}$; if $* \in \{\text{ENC}, \text{BOTH}\}$, the theorem holds without this independence condition.

replaced with the same element).

Theorem 5 *For a scenario in which Definition 1 is satisfied, q^n and w^n are independent,*[5] *and a scaled distortion measure of the form $d(x, \hat{x}; q) = \rho_0(q)\rho_1(x, \hat{x})$ is involved, q^n provides no benefit when known only at the decoder, i.e.,*

$$R[\text{Q-DEC-W-*}](D) = R[\text{Q-NONE-W-*}](D), \tag{29}$$

where the wildcard "" may be replaced with an element from $\{\text{ENC}, \text{DEC}, \text{BOTH}, \text{NONE}\}$ (both *'s must be replaced with the same element).*

Finally, we can generalize Theorem 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.

Theorem 6 *For a scenario in which Definition 2 is satisfied, q^n and w^n are independent,*[6] *and the difference distortion measure involved is a scaled one of the form*[7] $d(x, \hat{x}; q) = \rho_0(q) \cdot \rho_1(x - \hat{x})$, *$q^n$ (respectively, w^n) is asymptotically only required at the encoder (respectively, at the decoder), i.e.,*

$$\lim_{D \to D_{\min}} R[\text{Q-ENC-W-*}](D) - R[\text{Q-BOTH-W-*}](D) = 0 \tag{30a}$$

$$\lim_{D \to D_{\min}} R[\text{Q-*-W-DEC}](D) - R[\text{Q-*-W-BOTH}](D) = 0, \tag{30b}$$

where the wildcard "" may be replaced with an element from $\{\text{ENC}, \text{DEC}, \text{BOTH}, \text{NONE}\}$ (both *'s must be replaced with the same element).*

In essence, Theorem 6 establishes 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 Theorem 3, the better the system will perform.

[5] Independence is only required when $* \in \{\text{ENC}, \text{NONE}\}$; if $* \in \{\text{DEC}, \text{BOTH}\}$, the theorem holds without this condition.

[6] Independence is only required when $* \in \{\text{ENC}, \text{NONE}\}$ in (30a) or when $* \in \{\text{DEC}, \text{NONE}\}$ in (30b). For $* \in \{\text{DEC}, \text{BOTH}\}$ in (30a) or $* \in \{\text{ENC}, \text{BOTH}\}$ in (30b) the theorem holds without this condition.

[7] The scaled form of the distortion measure is only required when $* \in \{\text{DEC}, \text{NONE}\}$ in (30b). When $* \in \{\text{ENC}, \text{BOTH}\}$, the theorem holds without this restriction.

C. Loss Theorems

While the results of Section IV-B establish that the 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 two 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 side informations influence the associated losses.

Our two theorems are as follows;[8] proofs are proved in Appendix D.

Theorem 7 *For a scenario in which Definition 1 is satisfied, q^n and w^n are independent,*[9] *and the difference distortion measure involved is a scaled one of the form $d(x, \hat{x}; q) = \rho_0(q) \cdot \rho_1(x - \hat{x})$, the penalty for not knowing w^n at the decoder is*

$$\lim_{D \to D_{\min}} R[\text{Q-*-W-\{ENC-OR-NONE\}}](D) - R[\text{Q-*-W-\{DEC-OR-BOTH\}}](D) = I(\mathsf{x}; \mathsf{w}), \tag{31}$$

where the wildcard "" may be replaced with an element from $\{\text{ENC}, \text{DEC}, \text{BOTH}, \text{NONE}\}$ (all *'s must be replaced with the same element).*

Theorem 8 *For a scenario in which Definition 2 is satisfied, q^n and w^n are independent,*[10] *and the difference distortion measure involved is a scaled one of the form $d(x, \hat{x}; q) = q \cdot \|x - \hat{x}\|^r$ for some $r > 0$, the penalty (in nats/sample) for not knowing q^n at the encoder is*

$$\lim_{D \to D_{\min}} R[\text{Q-\{DEC-OR-NONE\}-W-*}](D) - R[\text{Q-\{ENC-OR-BOTH\}-W-*}](D) = \frac{k}{r} E\left[\ln \frac{E[\mathsf{q}]}{\mathsf{q}}\right], \tag{32}$$

where the wildcard "" may be replaced with an element from $\{\text{ENC}, \text{DEC}, \text{BOTH}, \text{NONE}\}$ (both *'s must be replaced with the same element).*

Some remarks are worthwhile. First, Theorem 7 makes clear that the more valuable the signal side information w^n is (i.e., the greater the statistical dependency on the signal as measured by mutual information), the larger the loss incurred by not having it at the decoder. Moreover, there is no loss if and only if the signal side information is useless (i.e., independent of the source).

[8] Note that a special case of Theorem 8 appears in [6] for the case $r = 2$ and the lefthand and righthand $*$'s in (32) being DEC and BOTH, respectively.

[9] Independence is only required when $* \in \{\text{DEC}, \text{NONE}\}$; if $* \in \{\text{ENC}, \text{BOTH}\}$, the theorem holds without this condition.

[10] Independence is only required when $* \in \{\text{ENC}, \text{NONE}\}$; if $* \in \{\text{DEC}, \text{BOTH}\}$, the theorem holds without this condition.

TABLE I

ASYMPTOTIC RATE LOSS (IN NATS) FOR NOT KNOWING DISTORTION SIDE INFORMATION q AT THE ENCODER. DISTORTION IS MEASURED VIA $d(x, \hat{x}; q) = q(x - \hat{x})^2$, AND γ DENOTES EULER'S CONSTANT.

Distribution Name	Density for q	Rate Gap in nats
Exponential	$\tau \exp(-q\tau)$	$-\frac{1}{2}\ln\gamma \approx 0.2748$
Uniform	$1_{q\in[0,1]}$	$\frac{1}{2}(1 - \ln 2) \approx 0.1534$
Lognormal	$\frac{1}{q\sqrt{2\pi Q^2}} \exp\left[-\frac{(\ln q - M)^2}{2Q^2}\right]$	$\frac{Q^2}{4}$
Pareto	$\frac{a^b}{q^{a+1}}, q \geq b > 0, a > 1$	$\frac{1}{2}\left[\ln\frac{a}{a-1} - 1/a\right]$
Gamma	$\frac{b(bq)^{a-1}\exp(-bq)}{\Gamma(a)}$	$\frac{1}{2}\left\{\ln a - \frac{d}{dx}[\ln\Gamma(x)]_{x=a}\right\} \approx \frac{1}{2a}$
Pathological	$(1-\epsilon)\delta(q-\epsilon) + \epsilon\delta(q - 1/\epsilon)$	$\frac{1}{2}\ln(1+\epsilon-\epsilon^2) - \frac{1-2\epsilon}{2}\ln\epsilon \approx \frac{1}{2}\ln\frac{1}{\epsilon}$
Positive Cauchy	$\frac{2/\pi}{1+q^2}, q \geq 0$	∞

For comparison, Theorem 8 makes clear that the more significant the distortion side information (i.e., the greater the range of values this information can take on as measured logarithmically), the larger the loss incurred by not having it at the encoder. Moreover, there is no loss if and only if the distortion side information is a constant with probability 1 (i.e., degenerate).

In Table I, we evaluate the high-resolution rate penalty of Theorem 8 for a number of possible distortion side-information distributions. Note that for all of these side information distributions (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.

Finally, note that all but one of these distributions in Table I would require infinite rate to losslessly communicate the side information. Thus the gains to be had from distortion side information *cannot* be obtained by exactly describing the side information to the decoder.

Theorems 7 and 8 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 influences the distortion measure via $d(x, \hat{x}; z) = \rho_0(z) \cdot \rho_1(x - \hat{x})$ *and* is correlated with the source. Then we have the following corollaries of Theorems 7 and 8, respectively.

Corollary 1 *For a scenario in which Definition 2 is satisfied with $q = w = z$, and the difference distortion measure involved is a scaled one of the form $d(x, \hat{x}; z) = \rho_0(z) \cdot \rho_1(x - \hat{x})$, the penalty for knowing general side information z only at the encoder is*

$$\lim_{D \to D_{\min}} R[\text{Z-ENC}](D) - R[\text{Z-BOTH}](D) = I(x; z). \tag{33}$$

Corollary 2 *For a scenario in which Definition 2 is satisfied, and the difference distortion measure involved is a scaled one of the form $d(x, \hat{x}; z) = z \cdot \|x - \hat{x}\|^r$ for some $r > 0$, the penalty (in nats/sample) for not knowing z at the encoder is*

$$\lim_{D \to D_{\min}} R[\text{Z-DEC}](D) - R[\text{Z-BOTH}](D) = \frac{k}{r} E\left[\ln \frac{E[z]}{z}\right]. \tag{34}$$

In essence, Corollary 1 establishes that not having general side information at the decoder incurs a loss only to the extent that side information is correlated with the source, while Corollary 2 establishes that not having such side information at the encoder incurs a loss only to the extent that side information influences the distortion measure.

To obtain both Corollaries 1 and 2, it suffices to i) let $q = w = z$ in Theorems 7 and 8, respectively, for the cases $R[\text{Q-ENC-W-ENC}](D) - R[\text{Q-ENC-W-BOTH}](D)$ and $R[\text{Q-DEC-W-DEC}](D) - R[\text{Q-BOTH-W-DEC}](D)$, respectively, taking into account the respective footnotes in these theorems; and ii) note that $R[\text{Q-ENC-W-BOTH}](D) = R[\text{Q-BOTH-W-DEC}](D) = R[\text{Q-BOTH-W-BOTH}](D)$ when $q = w = z$.

V. Source Coding with Side Information at Lower Resolutions

While Section IV established the asymptotic sufficiency of encoder-only distortion side information and decoder-only signal side information in the high-resolution limit, they do not tell how quickly this sufficiency is obtained as the resolution is increased. This requires a finer grained analysis, which is the focus of this section. To simplify our analysis, we restrict our attention to scaled quadratic distortion measures, but briefly discuss how these results can be generalized to other distortion measures. As we now develop, our results take the form of two theorems, which characterize behavior at medium and low resolutions, respectively.

A. A Medium Resolution Bound

The following theorem bounds the rate penalties incurred by incomplete side information at medium resolutions; a proof is provided in Appendix E.

Theorem 9 *For a scenario in which Definition 1 is satisfied, and the distortion measure involved is of the form $d(x, \hat{x}; q) = q \cdot (x - \hat{x})^2$ with $q \geq q_{\min} > 0$, the rate gap at distortion D is bounded by*

$$R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq \frac{J(\mathsf{x}|\mathsf{w})}{2} \cdot \min\left[1, \frac{D}{q_{\min}}\right], \tag{35}$$

where $J(\mathsf{x}|\mathsf{w})$ is the Fisher Information in estimating a non-random parameter τ from $\tau + \mathsf{x}$ conditioned on knowing w. Specifically,

$$J(\mathsf{x}|\mathsf{w}) \triangleq \int p_{\mathsf{w}}(w) \left\{ \int p_{\mathsf{x}|\mathsf{w}}(x|w) \left[\frac{\partial}{\partial x} \log p_{\mathsf{x}|\mathsf{w}}(x|w)\right]^2 dx \right\} dw. \tag{36}$$

A few remarks are worthwhile. First, similar bounds can be developed with other distortion measures provided that $D/q_{\min}$ is replaced with a quantity proportional to the variance of the quantization error; see the remark after the proof of Theorem 9 in the Appendix E for details. Also, related bounds are discussed in [22, Appendix D].

Second, Fisher information arises in our bound from a consideration of the underlying additive test-channel distribution $\hat{\mathsf{x}} = \mathsf{x} + \mathsf{v}$. In particular, a clever source decoder could treat each source sample x_i as a parameter to be estimated from the quantized representation $\hat{\mathsf{x}}_i$. If an efficient estimator exists, this procedure could potentially reduce the distortion by the reciprocal of the Fisher Information. But if the distortion can be reduced in this manner without affecting the rate, then the additive test-channel distribution must be sub-optimal and a rate gap must exist.

Exploiting this insight, our bound in Theorem 9 essentially measures the rate gap by assessing how much our additive test-channel distribution could be improved if an efficient estimator existed for x given $\hat{\mathsf{x}}$. This bound will tend to be good when an efficient estimator does exist and poor otherwise. For example, if x is Gaussian with unit-variance conditioned on w, then the Fisher Information term in (35) evaluates to one and the worst-case rate-loss is at most half a bit at maximum distortion. This corresponds to the half-bit bound on the rate-loss for the pure Wyner-Ziv problem derived in [5]. But if x is discontinuous (e.g., if x is uniform), then no efficient estimator exists and the bound in (35) is poor.

We should also emphasize that the proof of Theorem 9 does not require any extra regularity conditions. Hence, if the Fisher Information of the source is finite, it can be immediately applied without the need to check whether the source is admissible according to Definition 2.

B. A Low Resolution Bound

While the Fisher Information bound from (35) can be used at low resolutions, it can be quite poor if the source is not smooth. Therefore, we propose the following alternative bound on the rate penalty, which is independent of the distortion level and hence most useful at low resolution. A proof is provided in Appendix E.

Theorem 10 *For a scenario in which Definition 1 is satisfied, and the distortion measure involved is of the form $d(x, \hat{x}; q) = q \cdot (x - \hat{x})^2$, the rate gap at any distortion is at most*

$$R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq D(p_{\mathsf{x}|\mathsf{w}} \| \mathcal{N}(\mathrm{Var}\,[\mathsf{x}])) + \frac{1}{2}\log\left(1 + \frac{\sigma^2_{\max}}{\sigma^2_{\min}}\right), \tag{37}$$

where $\mathcal{N}(t)$ represents a Gaussian random variable with mean zero and variance t, and where

$$\sigma^2_{\min} = \min_w \mathrm{Var}\,[\mathsf{x}|\mathsf{w} = w] \tag{38a}$$

$$\sigma^2_{\max} = \max_w \mathrm{Var}\,[\mathsf{x}|\mathsf{w} = w]\,. \tag{38b}$$

Again, we make some remarks. First, as with our medium resolution bound, similar bounds can be developed for other distortion measures, which we discuss after the proof of Theorem 10 in Appendix E.

Second, the bound (37) can be readily evaluated in various cases of interest. As one example, consider the familiar Wyner-Ziv scenario where the signal side information is a noisy observation of the source. Specifically, let $\mathsf{w} = \mathsf{x} + \mathsf{v}$ where v is independent of x. In this case, the conditional variance is constant and (37) becomes

$$D(p_{\mathsf{x}|\mathsf{w}} \| \mathcal{N}(\mathrm{Var}\,[\mathsf{x}])) + \frac{1}{2}\log 2 \tag{39}$$

and the rate-loss is at most half a bit plus the deviation from Gaussianity of the source. As another example, if x is Gaussian when conditioned on $\mathsf{w} = w$, then the rate-loss is again seen to be at most half a bit, as in [5].

However, in contrast to the bound of [5], which is independent of the source, both our bounds in (35) and (37) depend on the source distribution. Hence, we conjecture that our bounds are loose. In particular, for a discrete source, the worst case rate loss is at most $H(\mathsf{x}|\mathsf{w})$, but this is not captured by our results since both bounds become infinity. Techniques from [23], [24], [5] may yield tighter bounds.

C. A Finite-Rate Gaussian Example

To gain some sense for when the asymptotic results take effect, we consider a finite-rate Gaussian scenario. Specifically, let the source consist of a sequence of Gaussian random variables with mean zero and variance 1 and consider distortion side information with $\Pr[q = 1] = 0.6$, $\Pr[q = 10] = 0.4$, and distortion measure $d(x, \hat{x}; q) = q \cdot (x - \hat{x})^2$.

The case without side information is equivalent to quantizing a Gaussian random variable with distortion measure $4.6(x - \hat{x})^2$ and thus the rate-distortion function is

$$R[\text{Q-NONE-W-NONE}](D) = \begin{cases} 0, & D \geq 4.6 \\ \frac{1}{2} \ln \frac{4.6}{D}, & D \leq 4.6. \end{cases} \tag{40}$$

To determine $R[\text{Q-BOTH-W-NONE}](D)$ we must set up a constrained optimization as we did for the binary-Hamming scenario in Appendix B. This optimization results in a "water-pouring" bit allocation, which uses more bits to quantize the source when $q = 10$ than when $q = 1$. 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. Going through the details of the constrained optimization yields

$$R[\text{Q-BOTH-W-NONE}](D) = \begin{cases} 0, & 4.6 \leq D \\ \frac{0.4}{2} \ln \frac{4}{(D-0.6)}, & D^* \leq D \leq 4.6 \\ \frac{0.4}{2} \ln \frac{10}{D} + \frac{0.6}{2} \ln \frac{1}{D}, & D \leq D^* \end{cases} \tag{41}$$

for some appropriate threshold D^*. Evaluating (32) for this case indicates that the rate-gap between (40) and (41) goes to $0.5 \cdot (\ln 4.6 - 0.4 \ln 10) \approx 0.3$ nats ≈ 0.43 bits.

Computing $R[\text{Q-ENC-W-NONE}](D)$ analytically seems difficult. Thus, when distortion side information is only available at the encoder we obtain a numerical upper bound on the rate by using the same codebook distribution as when q is known at both encoder and decoder. This yields a rate penalty of $I(\hat{x}; q)$.[11] We can obtain a simple analytic bound from Theorem 9. Specifically, evaluating (35) yields that the rate penalty is at most $(1/2) \cdot \min[1, D]$.

In Fig. 10 we evaluate these rate-distortion trade-offs. We see that at zero rate, the rate-distortion functions for the case of no side information, encoder-only side information, and full side information have the same distortion since no bits are available for quantization. Furthermore, we see that the Fisher

[11] Actually, since the rate distortion function is convex, we take the lower convex envelope of the curve resulting from the optimal test-channel distribution.

Information bound is loose at zero rate. As the rate increases, the system with full distortion side-information does best because it uses the few available bits to represent only the important source samples with $q = 10$. The decoder reconstructs these source samples from the compressed data and reconstructs the less important samples to zero (the mean of x). In this regime, the system with distortion side information at the encoder also more accurately quantizes the important source samples. But since the decoder does not know q, it does not know which samples of $\hat{x}$ to reconstruct to zero. Thus the system with q available at the encoder performs worse than the one with q at both encoder and decoder but better than the system without side information. As the rate increases further, both systems with distortion side information quantize source samples with both $q = 1$ and $q = 10$. Thus the codebook distribution for $\hat{x}$ goes from a Gaussian mixture to become more and more Gaussian and the rate-loss for the system with only encoder side information goes to zero. Finally, we note that even at the modest distortion of -5 dB, the asymptotic effects promised by our theorems have already taken effect.

VI. Concluding Remarks

Our analysis indicates that side information that affects the distortion measure can provide significant benefits in source coding. Perhaps, our most surprising result is that in a number of cases, (e.g., sources uniformly distributed over a group, or in the high-resolution limit) side information at the encoder is just as good as side information known at both encoder and decoder. Furthermore, this “separation theorem” can be composed with the previously known result that having signal side information at the decoder is often as good as having it at both encoder and decoder (e.g., in the high-resolution limit). Our main results regarding when knowing a given type of side information at one place is as good as knowing it at another place are summarized in Fig. 9. Also, we computed the rate-loss for lacking a particular type of side information in a specific place. These penalty theorems show that lacking the proper side information can produce arbitrarily large degradations in performance. Taken together, we believe these results suggest that distortion side information is a useful source coding paradigm.

In practice, one area where distortion side information may provide benefits is in designing perceptual coders which use features of the human visual system (HVS) or human auditory system (HAS) to achieve low subjective distortion even when the objective distortion (e.g., the mean square error) is quite large. Recent examples of such systems have shown gains in image coding [25], [26]. Unfortunately, current systems often communicate the distortion side information (in the form of model parameters or quantizer step sizes) explicitly and thus are not as efficient as they could be. Perhaps more importantly, creating such a perceptual coder often requires the designer to be an expert both in human physiology

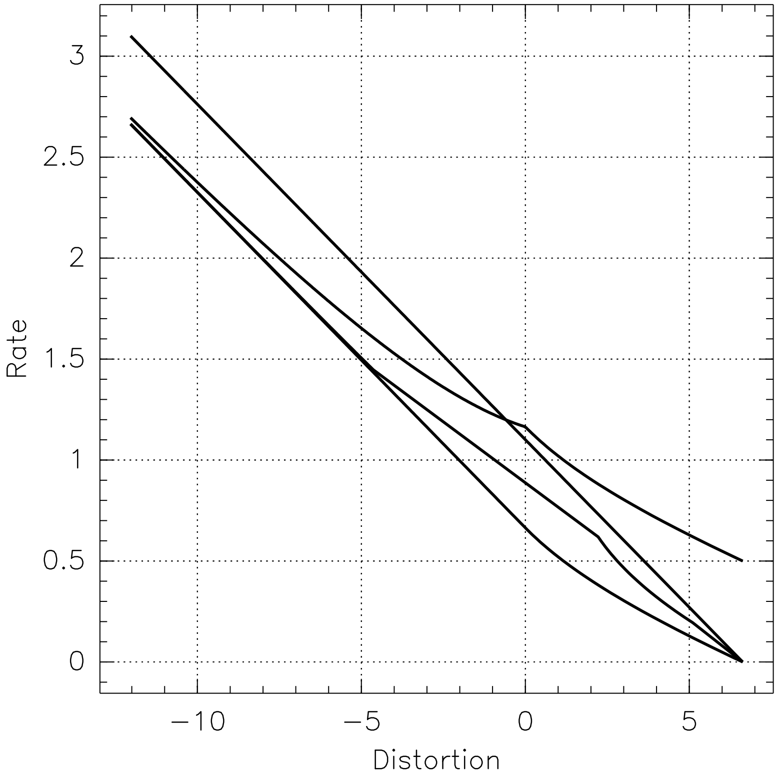


Fig. 10. Rate-distortion curves for quantizing a Gaussian source x with distortion $q(x - \hat{x})^2$ where the side information q is 1 with probability 0.6 or 10 with probability 0.4. From bottom to top on the right the curves correspond to the rate required when both encoder and decoder know q, a numerically computed upper bound to the rate when only the encoder knows q, the rate when neither encoder nor decoder know q, and the Fisher Information upper bound from Theorem 9 for when only the encoder knows q.

as well as quantizer design. Consequently, development becomes expensive and time consuming. Using the abstraction of distortion side information to represent such perceptual effects, however, may help overcome these barriers.

Obviously our model contains many idealizations that may not be exactly accurate for real sources (e.g., the distortion side information may not be independent of the source, the source itself may not be i.i.d., channel coding may be involved, *etc*). On a theoretical level, many of these non-idealities can be addressed. For example, while Corollary 1 indicates that knowing general side information z 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 approach outlined in Section III and developed in more detail in [12] can still provide significant benefits. Finally, we believe that information spectrum techniques can be used to establish that the familiar source–channel separation theorem holds and that the results developed here for i.i.d. models can be generalized to stationary, ergodic scenarios [27], [28], [29].

Appendix

A. Proof of Theorem 1

Assume that $p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q)$ is an optimal test-channel distribution with the conditional $p^*_{\mathsf{\hat{x}}|\mathsf{q}}(\hat{x}|q)$. By symmetry, for any $t \in \mathcal{X}$, the shifted distribution

$$p^t_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) \stackrel{\Delta}{=} p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x} \oplus t|x \oplus t, q) \tag{42}$$

must also be an optimal test-channel. Since mutual information is convex in the test-channel distribution, we obtain an optimal test-channel distribution p^{**} by averaging t over $\mathcal{X}$ via the uniform measure $d_{\mathcal{X}}(t)$:

$$p^{**}_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) \stackrel{\Delta}{=} \int_{\mathcal{X}} p^t_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) d_{\mathcal{X}}(t). \tag{43}$$

To prove that the resulting distribution for $\mathsf{\hat{x}}$ given q is uniform for all q (and hence independent of q), we will show that $p^{**}_{\mathsf{\hat{x}}|\mathsf{q}}(\hat{x}|q) = p^{**}_{\mathsf{\hat{x}}|\mathsf{q}}(\hat{x} \oplus r|q)$ for any $r \in \mathcal{X}$:

$$\begin{aligned}
p^{**}_{\mathsf{\hat{x}}|\mathsf{q}}(\hat{x}|q) &= \int_{\mathcal{X}} p^{**}_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) d_{\mathcal{X}}(x) && (44)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^t_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) d_{\mathcal{X}}(t) d_{\mathcal{X}}(x) && (45)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x} \oplus t|x \oplus t, q) d_{\mathcal{X}}(t) d_{\mathcal{X}}(x) && (46)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x} \oplus r \oplus t|x \oplus r \oplus t, q) d_{\mathcal{X}}(r \oplus t) d_{\mathcal{X}}(x) && (47)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x} \oplus r \oplus t|x \oplus r \oplus t, q) d_{\mathcal{X}}(t) d_{\mathcal{X}}(x \oplus r) && (48)\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p^*_{\mathsf{\hat{x}}|\mathsf{x},\mathsf{q}}(\hat{x} \oplus r \oplus t|x \oplus t, q) d_{\mathcal{X}}(t) d_{\mathcal{X}}(x) && (49)\\
&= p^{**}_{\mathsf{\hat{x}}|\mathsf{q}}(\hat{x} \oplus r|q). && (50)
\end{aligned}$$

Equation (44) follows from Bayes' law and the fact that $d_{\mathcal{X}}$ is the uniform measure on $\mathcal{X}$. The next two lines follow from the definition of p^{**} and p^t respectively. To obtain (47), 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 (48). Similarly, we make the change of variable $x \oplus r \to x$ to obtain (49). The last line follows from the definition in (43).

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

■

B. Derivation of Binary-Hamming Rate-Distortion Functions (16)

Let us first consider when the side information is *not* at the encoder, corresponding to (16a). In this case, the side information is of no value to the decoder, and thus the source coding problem is equivalent to quantizing a symmetric binary source with the distortion measure averaged over the side information, viz.,

$$d(x,\hat{x}) = E[\alpha_{\mathsf{q}} + \beta_{\mathsf{q}} \cdot d_H(x,\hat{x})] = E[\alpha_{\mathsf{q}}] + E[\beta_{\mathsf{q}}] \cdot d_H(x,\hat{x}). \tag{51}$$

Thus, the relevant rate-distortion function is obtained by simply scaling and translating the familiar rate-distortion function for the canonical binary-Hamming case, yielding (16a).

Next, to obtain (16b), the rate-distortion function when the side information *is* at the encoder, we begin by noting that this is the same as that when the side information is available at both encoder and decoder. Hence, we compute $R[\text{Q-ENC}](D)$ and $R[\text{Q-BOTH}](D)$ by considering the latter case and noting that optimal encoding corresponds to simultaneous description of independent random variables [30, Section 13.3.3]. Specifically, the source samples for each value of q can be quantized separately using the distribution

$$p_{\hat{\mathsf{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q) = \begin{cases} 1-p_q, & \hat{x} = x \\ p_q, & \hat{x} = 1-x. \end{cases} \tag{52}$$

The cross-over probabilities p_q correspond to the bit allocations for each value of the side information and are obtained by solving a constrained optimization problem:

$$R[\text{Q-BOTH}](D) = \min_{E[d(\mathsf{x},\hat{\mathsf{x}};\mathsf{q})=D]} \sum_{i=1}^{N} E[1 - H_{\mathrm{b}}(p_{\mathsf{q}})], \tag{53}$$

where $H_{\mathrm{b}}(\cdot)$ is the binary entropy function.

Using Lagrange multipliers, we construct the functional

$$J(D) = \sum_{i=1}^{N} p_{\mathsf{q}}(i) \cdot [1 + p_i \log p_i + (1-p_i)\log(1-p_i)] + \lambda \sum_{i=1}^{N} p_{\mathsf{q}}(i) \cdot [\alpha_i + p_i \beta_i],$$

whose minimum is easily shown to be attained at

$$p_i = \frac{2^{-\lambda\beta_i}}{1 + 2^{-\lambda\beta_i}}, \tag{54}$$

whence (16b) with (16c).

C. *Equivalence Theorem Proofs*

Proof of Theorem 3: To obtain $R[\textsc{q-enc-w-dec}](D)$ we apply the Wyner-Ziv rate-distortion formula in [1] to the "super-source" $\tilde{\mathsf{x}}^n = (\mathsf{x}^n, \mathsf{q}^n)$ yielding

$$R[\textsc{q-enc-w-dec}](D) = \inf_{p_{\hat{\mathsf{x}}|\mathsf{x},\mathsf{q}}(\hat{x}|x,q)} I(\hat{\mathsf{x}}, \mathsf{q}; \mathsf{x}|\mathsf{w}), \tag{55}$$

where the optimization is subject to the constraint that $E[d(\mathsf{x}, v(\hat{\mathsf{x}}, \mathsf{w}); \mathsf{q})] \leq D$ for some reconstruction function $v(\cdot,\cdot)$. To obtain $R[\textsc{q-both-w-both}](D)$ we specialize the well-known conditional rate-distortion function to our notation yielding

$$R[\textsc{q-both-w-both}](D) = \inf_{p_{\hat{\mathsf{x}}|\mathsf{x},\mathsf{q},\mathsf{w}}(\hat{x}|x,q,w)} I(\hat{\mathsf{x}}; \mathsf{x}|\mathsf{w}, \mathsf{q}), \tag{56}$$

where the optimization is subject to the constraint that $E[d(\mathsf{x}, \hat{\mathsf{x}}; \mathsf{q})] \leq D$.

Let us define $\hat{\mathsf{x}}^*$ as the distribution that optimizes (55). Similarly, define $\hat{\mathsf{x}}^*_{\mathsf{w}}$ as the distribution that optimizes (56). Finally, define v given $\mathsf{q} = q$ to be a random variable with a conditional distribution that maximizes $h(\mathsf{v}|\mathsf{q} = q)$ subject to the constraint that

$$E[d(\mathsf{x}, \mathsf{x} + \mathsf{v}; \mathsf{q})|\mathsf{q} = q] \leq E[d(\mathsf{x}, \hat{\mathsf{x}}^*_{\mathsf{w}}; \mathsf{q})|\mathsf{q} = q]. \tag{57}$$

Then we have the following chain of inequalities:

$$\Delta R(D) \overset{\Delta}{=} R[\textsc{q-enc-w-dec}](D) - R[\textsc{q-both-w-both}](D) \tag{58}$$

$$= I(\hat{\mathsf{x}}^*; \mathsf{q}, \mathsf{x}|\mathsf{w}) - [h(\mathsf{x}|\mathsf{q}, \mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}, \hat{\mathsf{x}}^*_{\mathsf{w}})] \tag{59}$$

$$= I(\hat{\mathsf{x}}^*; \mathsf{q}, \mathsf{x}|\mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) + h(\mathsf{x} - \hat{\mathsf{x}}^*_{\mathsf{w}}|\mathsf{q}, \mathsf{w}, \hat{\mathsf{x}}^*_{\mathsf{w}}) \tag{60}$$

$$\leq I(\hat{\mathsf{x}}^*; \mathsf{q}, \mathsf{x}|\mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) + h(\mathsf{x} - \hat{\mathsf{x}}^*_{\mathsf{w}}|\mathsf{q}) \tag{61}$$

$$\leq I(\hat{\mathsf{x}}^*; \mathsf{q}, \mathsf{x}|\mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) + h(\mathsf{v}|\mathsf{q}) \tag{62}$$

$$\leq I(\mathsf{x} + \mathsf{v}; \mathsf{q}, \mathsf{x}|\mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) + h(\mathsf{v}|\mathsf{q}) \tag{63}$$

$$= h(\mathsf{x} + \mathsf{v}|\mathsf{w}) - h(\mathsf{x} + \mathsf{v}|\mathsf{w}, \mathsf{q}, \mathsf{x}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) + h(\mathsf{v}|\mathsf{q}) \tag{64}$$

$$= h(\mathsf{x} + \mathsf{v}|\mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}) \tag{65}$$

$$= h(\mathsf{x} + \mathsf{v}|\mathsf{w}) - h(\mathsf{x}|\mathsf{w}). \tag{66}$$

Eq. (62) follows from the definition of v to be entropy maximizing subject to a distortion constraint. Since v is independent of x and w, the choice $\hat{\mathsf{x}} = \mathsf{x} + \mathsf{v}$ with $v(\hat{x}, w) = \hat{x}$ is an upper bound to (55) and yields (63). We obtain (66) by recalling that according to (4), q and x are independent given w.

Finally, we obtain

$$\lim_{D\to D_{\min}} \Delta R(D) = 0. \tag{67}$$

from (66) using the "continuity of entropy" result from [21, Theorem 1].

Note that although the v in [21, Theorem 1] is an entropy maximizing distribution while our v is a mixture of entropy maximizing distributions, the special form of the density is not required for the continuity of entropy result in [21, Theorem 1]. To illustrate this, we show how to establish the continuity of entropy directly for any distortion measure where $D \to D_{\min} \Rightarrow \mathrm{Var}[\mathsf{v}] \to 0$. One example of such a distortion measure is obtained if we choose $d(x, \hat{x}; q) = q \cdot |x - \hat{x}|^r$ with $r > 0$ and $\Pr[q = 0] = 0$. Denoting $\mathrm{Var}\,[\mathsf{v}|\mathsf{w}]$ as $\sigma^2_{\mathsf{v}|\mathsf{w}}$ and $\mathrm{Var}\,[\mathsf{x}|\mathsf{w}]$ as $\sigma^2_{\mathsf{x}|\mathsf{w}}$ and letting $\mathcal{N}(\alpha)$ represent a Gaussian random variable with variance α yields

$$\limsup_{D\to D_{\min}} h(\mathsf{x}+\mathsf{v}|\mathsf{w}) - h(\mathsf{x}|\mathsf{w}) = \limsup_{\sigma^2\to 0} h(\mathsf{x}+\mathsf{v}|\mathsf{w}) - h(\mathsf{x}|\mathsf{w}) \tag{68}$$

$$\begin{aligned}
&= \limsup_{\sigma^2\to 0} h(\mathsf{x}+\mathsf{v}|\mathsf{w}) \pm h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})|\mathsf{w}) \\
&\quad \pm h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})|\mathsf{w}) - h(\mathsf{x}|\mathsf{w}) && (69) \\
&= \limsup_{\sigma^2\to 0} D(p_{\mathsf{x}|\mathsf{w}}\|\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})) - D(p_{\mathsf{x}+\mathsf{v}|\mathsf{w}}\|\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})) \\
&\quad + h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})|\mathsf{w}) - h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})|\mathsf{w}) && (70) \\
&\leq D(p_{\mathsf{x}|\mathsf{w}}\|\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})) - D(p_{\mathsf{x}|\mathsf{w}}\|\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})) \\
&\quad + \limsup_{\sigma^2\to 0}[h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})|\mathsf{w}) - h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})|\mathsf{w})] && (71) \\
&= \limsup_{\sigma^2\to 0} \int \Big[h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})|\mathsf{w} = w) - h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})|\mathsf{w} = w)\Big]\, dp_{\mathsf{w}}(w) && (72) \\
&= \int \left[\limsup_{\sigma^2\to 0} h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}} + \sigma^2_{\mathsf{v}|\mathsf{w}})|\mathsf{w} = w) - h(\mathcal{N}(\sigma^2_{\mathsf{x}|\mathsf{w}})|\mathsf{w} = w)\right] dp_{\mathsf{w}}(w) && (73) \\
&= 0. && (74)
\end{aligned}$$

We obtain (70) since for any random variable v, the relative entropy from v to a Gaussian takes the special form $D(p_{\mathsf{v}}\|\mathcal{N}(\mathrm{Var}\,[\mathsf{v}])) = h(\mathcal{N}(\mathrm{Var}\,[\mathsf{v}])) - h(\mathsf{v})$ [30, Theorem 9.6.5]. To get (71) we use the fact that relative-entropy (and also conditional relative-entropy) is lower semi-continuous [31]. This could also be shown by applying Fatou's Lemma [32, p.78] to get that if the sequences $p_1(x), p_2(x), \ldots$ and $q_1(x), q_2(x), \ldots$ converge to $p(x)$ and $q(x)$ then

$$\liminf \int p_i(x) \log[p_i(x)/q_i(x)] \geq \int p(x) \log[p(x)/q(x)].$$

Switching the $\limsup$ and integral in (73) is justified by Lebesgue's Dominated Convergence Theorem [32, p.78] since the integrand is bounded for all values of w. In general, this bound is obtained from combining the technical condition requiring $h(\mathsf{x}|\mathsf{w} = w)$ to be finite with the entropy maximizing distribution in (25) and the expected distortion constraint in (26) to bound $h(\mathsf{x} + \mathsf{v}|\mathsf{q} = q)$. For scaled quadratic distortions, $h(\mathsf{x} + \mathsf{v}|\mathsf{q} = q)$ can be bounded above by the entropy of a Gaussian with the appropriate variance. To obtain (74) we first note that $\mathrm{Var}\,[\mathsf{v}] \to 0$ implies $\mathrm{Var}\,[\mathsf{v}|\mathsf{w} = w] \to 0$ except possibly for a set of w having measure zero. This set of measure zero can be ignored because the integrand is finite for all w. Finally, for the set of w where $\mathrm{Var}\,[\mathsf{v}|\mathsf{w} = w] \to 0$, the technical requirement that the entropy maximizing distribution in (25) is continuous shows that the entropy difference (74) goes to zero in the limit. ■

Proof of Theorem 4: When $* \in \{\text{ENC}, \text{BOTH}\}$ in (28), the encoder can simulate w^n by generating it from $(\mathsf{x}^n, \mathsf{q}^n)$. When $* \in \{\text{DEC}, \text{NONE}\}$, the encoder can still simulate w^n correctly provided that w^n and q^n are independent. Thus being provided with w^n provides no advantage given the conditions of the theorem. ■

Proof of Theorem 5: We begin by showing

$$R[\text{Q-DEC-W-DEC}](D) = R[\text{Q-NONE-W-DEC}](D). \tag{75}$$

When side information $(\mathsf{q}^n, \mathsf{w}^n)$ is available only at the decoder, the optimal strategy is Wyner-Ziv encoding [1]. Let us compute the optimal reconstruction function $v(\cdot,\cdot,\cdot)$, which maps an auxiliary random variable u and the side information q and w to a reconstruction of the source:

$$\begin{aligned} v(u,q,w) &= \arg\min_{\hat{x}} E[d(\hat{x},\mathsf{x};\mathsf{q})|\mathsf{q} = q, \mathsf{w} = w, \mathsf{u} = u] && (76)\\ &= \arg\min_{\hat{x}} \rho_0(q) E[\rho_1(\hat{x},\mathsf{x})|\mathsf{q} = q, \mathsf{w} = w, \mathsf{u} = u] && (77)\\ &= \arg\min_{\hat{x}} E[\rho_1(\hat{x},\mathsf{x})|\mathsf{q} = q, \mathsf{w} = w, \mathsf{u} = u] && (78)\\ &= \arg\min_{\hat{x}} E[\rho_1(\hat{x},\mathsf{x})|\mathsf{w} = w, \mathsf{u} = u]. && (79)\end{aligned}$$

We obtain (77) from the assumption that we have a separable distortion measure. To get (79) recall that by assumption q is statistically independent of x given w and also q is statistically independent of u since u is generated at the encoder from x. Thus neither the optimal reconstruction function $v(\cdot,\cdot,\cdot)$ nor the auxiliary random variable u depend on q. This establishes (75).

To show that

$$R[\text{Q-DEC-W-NONE}](D) = R[\text{Q-NONE-W-NONE}](D) \tag{80}$$

we need w^n and q^n to be independent. When this is true, w^n does not affect anything and the problem is equivalent to when $\mathsf{w}^n = 0$ and is available at the decoder. From (75) we see that providing $\mathsf{w}^n = 0$ at the decoder does not help and thus we establish (80). Note that this argument fails when w^n and q^n are not independent since in that case Wyner-Ziv based on q^n could be performed and there would be no w^n at the decoder to enable the argument in (76)–(79).

To show that

$$R[\text{Q-DEC-W-BOTH}](D) = R[\text{Q-NONE-W-BOTH}](D) \tag{81}$$

we note that in this scenario the encoder and decoder can design a different source coding system 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 will have distribution $p_{\mathsf{x}|\mathsf{w}}(x|w^*)$, and the distortion side information will have distribution $p_{\mathsf{q}|\mathsf{w}}(q|w^*)$. Thus the performance of each subsystem is given by $R[\text{Q-DEC-W-NONE}](D)$, which we already showed is the same as $R[\text{Q-NONE-W-NONE}](D)$. This establishes (81).

Finally, to show that

$$R[\text{Q-DEC-W-ENC}](D) = R[\text{Q-NONE-W-ENC}](D) \tag{82}$$

we require the assumption that q^n and w^n are independent. This assumption implies

$$R[\text{Q-DEC-W-ENC}](D) = R[\text{Q-DEC-W-NONE}](D) \tag{83}$$

since an encoder without w^n could always generated a simulated w^n with the correct distribution relative to the other variables. The same argument implies

$$R[\text{Q-NONE-W-ENC}](D) = R[\text{Q-NONE-W-NONE}](D). \tag{84}$$

Combining (83), (84), and (80) yields (82). ■

Proof of Theorem 6: First we establish the four rate-distortion function equalities implied by (30a). Using Theorem 3 we have

$$\lim_{D \to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-DEC}](D) \leq \tag{85}$$

$$\lim_{D \to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \tag{86}$$

$$= 0. \tag{87}$$

Similarly,

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-BOTH}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq \tag{88}$$

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \tag{89}$$

$$= 0. \tag{90}$$

To show that

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) - R[\text{Q-BOTH-W-NONE}](D) = 0 \tag{91}$$

we need q^n and w^n to be independent. When this is true, w^n does not affect anything and the problem is equivalent to when $w^n = 0$ and is available at the decoder and (85)–(87) establishes (91). Without independence this argument fails because we can no longer invoke Theorem 3 since there will be no w^n to make x^n and q^n conditionally independent in (66).

To finish establishing (30a) we again require q^n and w^n to be independent to obtain

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-ENC}](D) - R[\text{Q-BOTH-W-ENC}](D) \leq \tag{92}$$

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) - R[\text{Q-BOTH-W-ENC}](D) = \tag{93}$$

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) - R[\text{Q-BOTH-W-NONE}](D) \tag{94}$$

$$= 0, \tag{95}$$

where (94) follows since the encoder can always simulate w^n from (x^n, q^n) and (95) follows from (91).

Next, we establish the four rate-distortion function equalities implied by (30b). Using Theorem 3 we have

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-ENC-W-BOTH}](D) \leq \tag{96}$$

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \tag{97}$$

$$= 0. \tag{98}$$

Similarly,

$$\lim_{D\to D_{\min}} R[\text{Q-BOTH-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq \tag{99}$$

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \tag{100}$$

$$= 0. \tag{101}$$

To show that

$$\lim_{D \to D_{\min}} R[\text{Q-NONE-W-DEC}](D) - R[\text{Q-NONE-W-BOTH}](D) = 0 \tag{102}$$

we need q^n and w^n to be independent and we need the distortion measure to be of the form $d(\hat{x}, x; q) = \rho_0(q) \cdot \rho_1(x - \hat{x})$. When this is true the two rate-distortion functions in (102) are equivalent to the Wyner-Ziv rate-distortion function and the conditional rate-distortion function for the difference distortion measure $E[\rho_0(\mathsf{q})] \cdot \rho_1(x - \hat{x})$. Thus we can either apply the result from [5] showing these rate-distortion functions are equal in the high-resolution limit or simply specialize Theorem 3 to the case where q^n is a constant.

To complete the proof, we again require the assumptions that q^n and w^n are independent and that the distortion measure is of the form $d(x, \hat{x}; q) = \rho_0(q) \cdot \rho_0(q) \cdot \rho_1(x - \hat{x})$. We have

$$\lim_{D \to D_{\min}} R[\text{Q-DEC-W-DEC}](D) - R[\text{Q-DEC-W-BOTH}](D) \leq \tag{103}$$

$$\lim_{D \to D_{\min}} R[\text{Q-NONE-W-DEC}](D) - R[\text{Q-DEC-W-BOTH}](D) = \tag{104}$$

$$\lim_{D \to D_{\min}} R[\text{Q-NONE-W-DEC}](D) - R[\text{Q-NONE-W-BOTH}](D) \tag{105}$$

$$= 0, \tag{106}$$

where (105) follows from Theorem 5 and (106) follows from (102). ■

D. Loss Theorem Proofs

Proof of Theorem 7: We note that according to Theorem 4 and Theorem 6 we can focus solely on the case

$$R[\text{Q-*-W-NONE}](D) - R[\text{Q-*-W-BOTH}](D). \tag{107}$$

When * = NONE, the rate difference in (107) is the difference between the classical rate-distortion function and the conditional rate-distortion function in the high-resolution limit. Thus the Shannon Lower Bound [21] (and its conditional version) imply that

$$\lim_{D \to D_{\min}} R[\text{Q-NONE-W-NONE}](D) - R[\text{Q-NONE-W-BOTH}](D) = h(\mathsf{x}) - h(\mathsf{x}|\mathsf{w}). \tag{108}$$

Similarly, when * = DEC an identical argument can be combined with Theorem 5.

When * = BOTH, the encoder and decoder can design a separate compression sub-system for each value of q. The rate-loss for each sub-system is then $I(\mathsf{x}; \mathsf{w}|\mathsf{q} = q)$ according to high-resolution Wyner-Ziv theory [5]. Averaging over all values of q yields a total rate-loss of $I(\mathsf{x}; \mathsf{w}|\mathsf{q})$.

Next we consider the case when * = ENC and the rate-loss penalty is

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) - R[\text{Q-ENC-W-BOTH}](D)$$
$$= \lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) - R[\text{Q-BOTH-W-BOTH}](D), \tag{109}$$

where the equality follows from Theorem 6.

Using arguments similar to [21] and the proof of Theorem 3, we can obtain a Shannon Lower Bound for $R[\text{Q-ENC-W-NONE}](D)$, which is of the form

$$R[\text{Q-ENC-W-NONE}](D) \geq h(x) - h(v_D), \tag{110}$$

where v_D is an entropy maximizing random variable subject to the constraint that $E[\rho_0(q)\cdot\rho_1(v_D)] \leq D$. Again using argument similar to the proof of Theorem 3, we have that

$$\lim_{D\to D_{\min}} R[\text{Q-BOTH-W-BOTH}](D) \leq h(x|w) - h(v_D). \tag{111}$$

Combining (110) and (111) shows that the asymptotic difference in (109) is at least $I(x;w)$.

Next, we obtain the Shannon Lower Bound

$$R[\text{Q-BOTH-W-BOTH}](D) \geq h(x|w) - h(v_D) \tag{112}$$

by duplicating the arguments in the proof of Theorem 3 since this lower bound does not require q and w to be independent. Finally, we can obtain the upper bound

$$\lim_{D\to D_{\min}} R[\text{Q-ENC-W-NONE}](D) \leq h(x) - h(v_D) \tag{113}$$

using an additive noise test channel combined with arguments following those in the proof of Theorem 3. Combining (112) and (113) shows that the asymptotic difference in (109) is at most $I(x;w)$. ■

Proof of Theorem 8: To simplify the exposition, we first prove the theorem for the relatively simple case of a one-dimensional source ($k = 1$) with a quadratic distortion ($r = 2$). Then at the end of the proof, we describe how to extend it to general k and r.

We begin with the case where * = NONE. Since Theorems 5 and 6 imply

$$R[\text{Q-NONE-W-NONE}](D) = R[\text{Q-DEC-W-NONE}](D) \tag{114a}$$

and

$$R[\text{Q-ENC-W-NONE}](D) \to R[\text{Q-BOTH-W-NONE}](D)\ , \tag{114b}$$

we focus on showing

$$\lim_{D\to D_{\min}} R[\text{Q-BOTH-W-NONE}](D) - R[\text{Q-NONE-W-NONE}](D) = \frac{1}{2}E\left[\ln\frac{E[q]}{q}\right]. \tag{115}$$

Computing $R[\text{Q-BOTH-W-NONE}](D)$ is equivalent to finding the rate-distortion function for optimally encoding independent random variables and yields the familiar "water-pouring" rate and distortion allocation [30, Section 13.3.3]. For each q, we quantize the corresponding source samples with distortion $D_q = E[(\mathsf{x} - \hat{\mathsf{x}})^2]$ (or $E[\|\mathsf{x}^n - \hat{\mathsf{x}}^n\|^r]$ in the more general case) and rate $R_q(D_q)$. The overall rate and distortion then become $E[R_{\mathsf{q}}(D_{\mathsf{q}})]$ and $E[\mathsf{q} \cdot D_{\mathsf{q}}]$.

Thus to find the rate and distortion allocation we set up a constrained optimization problem using Lagrange multipliers to obtain the functional

$$J(D) = E[R_{\mathsf{q}}(D_{\mathsf{q}})] + \lambda(D - E[\mathsf{q} \cdot D_{\mathsf{q}}]), \tag{116}$$

differentiate with respect to D_q, set equal to zero and solve for each D_q. In the high-resolution limit, various researchers have shown

$$R_{\mathsf{q}}(D_{\mathsf{q}}) \to h(\mathsf{x}) - \frac{1}{2}\log D_q. \tag{117}$$

(e.g., see [21] and references therein). Therefore, it is straightforward to show this process yields the condition $D_q = 1/(2\lambda q)$ with $2\lambda = 1/D$ implying

$$\lim_{D\to 0} R[\text{Q-BOTH-W-NONE}](D) \to h(\mathsf{x}) - \frac{1}{2}\log D + \frac{1}{2}E[\log \mathsf{q}]. \tag{118}$$

To compute $R[\text{Q-NONE-W-NONE}](D)$, we note that since neither encoder nor decoder knows q the optimal strategy is to simply quantize the source according to the distortion $d(\mathsf{q}, \mathsf{x}; \hat{\mathsf{x}}) = E[\mathsf{q}] \cdot (\mathsf{x} - \hat{\mathsf{x}})^2$ to obtain

$$\lim_{D\to 0} R[\text{Q-NONE-W-NONE}](D) \to h(\mathsf{x}) - \frac{1}{2}\log D + \frac{1}{2}\log E[\mathsf{q}]. \tag{119}$$

Comparing (118) with (119) establishes (115).

By applying Theorem 4 we see that the case where * = ENC is the same as * = NONE.

Next we consider the case where * = BOTH in (32). In this case, the encoder and decoder can design a separate compression sub-system for each value of w and the performance for each sub-system is obtained from the case with no signal side information. Specifically, the rate-loss for each sub-system is

$$\frac{1}{2}E\left[\ln \frac{E[\mathsf{q}|\mathsf{w} = w]}{\mathsf{q}}\middle| \mathsf{w} = w\right] \tag{120}$$

according to the previously derived results. Averaging (120) over w then yields the rate-loss in (32).

Finally, we consider the case where * = DEC in (32). Since Theorem 5 implies $R[\text{Q-DEC-W-DEC}](D) = R[\text{Q-NONE-W-DEC}](D)$ and Theorem 3 implies $R[\text{Q-ENC-W-DEC}](D) \to R[\text{Q-BOTH-W-BOTH}](D)$, it suffices to show that

$$\lim_{D\to D_{\min}} R[\text{Q-DEC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) = \frac{1}{2}E\left[\log \frac{E[\mathsf{q}]}{\mathsf{q}}\right]. \tag{121}$$

We can compute $R[\text{Q-BOTH-W-BOTH}](D)$ by considering a separate coding system for each value of w. Specifically, conditioned on $\mathsf{w} = w$, computing the rate-distortion trade-off is equivalent to finding $R[\text{Q-BOTH-W-NONE}](D)$ for a modified source x' with distribution $p_{\mathsf{x}'}(x') = p_{\mathsf{x}|\mathsf{w}}(x'|w)$. Thus we obtain

$$\lim_{D \to D_{\min}} R[\text{Q-BOTH-W-BOTH}](D) \to h(\mathsf{x}|\mathsf{w}) - \frac{1}{2}\log D + \frac{1}{2}\log E[\mathsf{q}]. \tag{122}$$

Applying the standard techniques used throughout the paper, we can compute the Shannon Lower Bound

$$R[\text{Q-DEC-W-DEC}](D) \geq h(\mathsf{x}|\mathsf{w}) - \frac{1}{2}\log(D \cdot E[\mathsf{q}]) \tag{123}$$

and show it is tight in the high-resolution limit. Comparing (122) and (123) establishes the desired result.

This establishes the theorem for $k = 1$ and $r = 2$. For general k and r, the only change is that each component rate-distortion function $R_q(D_q)$ (117) becomes [21, page 2028]

$$R_{\mathsf{q}}(D_{\mathsf{q}}) \to h(\mathsf{x}) - \frac{k}{r}\log D_q - \frac{k}{r} + \log\left[\frac{r}{kV_k\Gamma(k/r)}\left(\frac{k}{r}\right)^{k/r}\right]. \tag{124}$$

and a similar change occurs for all the following rate-distortion expressions. Since we are mainly interested in the difference of rate-distortion functions, most of these extra terms cancel out and the only change is that factors of $1/2$ are replaced with factors of k/r. ■

E. Proofs for Rate Penalties at Lower Resolutions

Before proceeding, we require the following lemma to upper and lower bound the entropy of an arbitrary random variable plus a Gaussian mixture.

Lemma 2 *Let x be an arbitrary random variable with finite variance $\sigma^2 < \infty$. Let v be a zero-mean, unit-variance Gaussian independent of x and let v be a random variable independent of x and v with $0 < v_{\min} \leq \mathsf{v} < v_{\max}$. Then*

$$h(\mathsf{x}) + \frac{1}{2}\log(1 + v_{\min}) \leq h(\mathsf{x} + \mathsf{v}\sqrt{\mathsf{v}}) \leq h(\mathsf{x}) + \frac{1}{2}\log(1 + v_{\max} \cdot J(\mathsf{x})) \tag{125}$$

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

Proof: The concavity of differential entropy yields

$$h(\mathsf{x} + \mathsf{v}\sqrt{v_{\min}}) \leq h(\mathsf{x} + \mathsf{v}\sqrt{\mathsf{v}}) \leq h(\mathsf{x} + \mathsf{v}\sqrt{v_{\max}}). \tag{126}$$

For the lower bound we have

$$h(\mathsf{x}+\mathsf{v}\sqrt{v_{\min}}) = \int_0^{v_{\min}} \frac{\partial}{\partial\tau} h(\mathsf{x}+\mathsf{v}\sqrt{\tau})d\tau + h(\mathsf{x}) \tag{127}$$

$$= \int_0^{v_{\min}} \frac{1}{2} J(\mathsf{x}+\mathsf{v}\sqrt{\tau})d\tau + h(\mathsf{x}) \tag{128}$$

$$\geq \frac{1}{2}\int_0^{v_{\min}} J(\mathsf{v}\sqrt{\sigma^2+\tau})d\tau + h(\mathsf{x}) \tag{129}$$

$$= \frac{1}{2}\int_0^{v_{\min}} \frac{d\tau}{\sigma^2+\tau} + h(\mathsf{x}) \tag{130}$$

$$= \frac{1}{2}\log\left(1+\frac{v_{\min}}{\sigma^2}\right) + h(\mathsf{x}), \tag{131}$$

where (128) follows from de Bruijn's identity [30, Theorem 16.6.2], [33, Theorem 14], (129) follows from the fact that a Gaussian distribution minimizes Fisher Information subject to a variance constraint, and (130) follows since the Fisher Information for a Gaussian is the reciprocal of its variance.

Similarly, for the upper bound we have

$$h(\mathsf{x}+\mathsf{v}\sqrt{v_{\max}}) = \int_0^{v_{\max}} \frac{\partial}{\partial\tau} h(\mathsf{x}+\mathsf{v}\sqrt{\tau})d\tau + h(\mathsf{x}) \tag{132}$$

$$= \int_0^{v_{\max}} \frac{1}{2} J(\mathsf{x}+\mathsf{v}\sqrt{\tau})d\tau + h(\mathsf{x}) \tag{133}$$

$$\leq \frac{1}{2}\int_0^{v_{\max}} \frac{J(\mathsf{x})J(\mathsf{v}\sqrt{\tau})}{J(\mathsf{x})+J(\mathsf{v}\sqrt{\tau})}d\tau + h(\mathsf{x}) \tag{134}$$

$$= \frac{1}{2}\int_0^{v_{\max}} \frac{J(\mathsf{x})d\tau}{\tau J(\mathsf{x})+1} + h(\mathsf{x}) \tag{135}$$

$$= \frac{1}{2}\log\left(1+v_{\max}\cdot J(\mathsf{x})\right) + h(\mathsf{x}), \tag{136}$$

where (133) again follows from de Bruijn's identity, (134) follows from the convolution inequality for Fisher Information [34], [30, p.497], and (135) follows since the Fisher Information for a Gaussian is the reciprocal of its variance.

Combining these upper and lower bounds yields the desired result. Finally, the inequalities used in (129) and (134) are both tight if and only if x is Gaussian. ■

As an aside we note that Lemma 2 can be used to bound the rate-distortion function of an arbitrary unit-variance source x relative to quadratic distortion. Specifically using an additive Gaussian noise test-channel $\hat{\mathsf{x}} = \mathsf{v}+\mathsf{x}$ and combining Lemma 2 to upper bound $h(\mathsf{x}+\mathsf{v})$ with the Shannon Lower Bound [21] yields

$$h(\mathsf{x}) - \frac{1}{2}\log 2\pi e D \leq R(D) \leq h(\mathsf{x}) - \frac{1}{2}\log 2\pi e D + \frac{1}{2}\log[1+DJ(\mathsf{x})]. \tag{137}$$

Evidently, the error in the Shannon Lower Bound is at most $\frac{1}{2}\log[1+DJ(\mathsf{x})]$. Thus, since $J(\mathsf{x}) \geq 1$ with equality only for a Gaussian, the sub-optimality of an additive Gaussian noise test-channel is at least $\frac{1}{2}\log[1+D]$.

Proof of Theorem 9: Starting with the bound for the rate gap from (66), we have

$$R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq h(\mathsf{x}+\mathsf{v}|\mathsf{w}) - h(\mathsf{x}|\mathsf{w}) \tag{138}$$

$$= \int \left[h(\mathsf{x}+\mathsf{v}|\mathsf{w}=w) - h(\mathsf{x}|\mathsf{w}=w)\right] p_{\mathsf{w}}(w)dw \tag{139}$$

$$\leq \int \left\{\frac{1}{2}\log\left(1+\min\left[1,\frac{D}{\mathsf{q}_{\min}}\right]\cdot J(\mathsf{x}|\mathsf{w}=w)\right)\right\} p_{\mathsf{w}}(w)dw \tag{140}$$

$$\leq \int \left\{\frac{J(\mathsf{x}|\mathsf{w}=w)}{2}\cdot\min\left[1,\frac{D}{\mathsf{q}_{\min}}\right]\right\} p_{\mathsf{w}}(w)dw \tag{141}$$

$$= \frac{J(\mathsf{x}|\mathsf{w})}{2}\cdot\min\left[1,\frac{D}{\mathsf{q}_{\min}}\right]. \tag{142}$$

To obtain (140) we note that v is a Gaussian mixture and apply Lemma 2. This follows since, conditioned on $\mathsf{q}=q$, v is a Gaussian with variance $E[d(\mathsf{x},\hat{\mathsf{x}}^*_{\mathsf{w}};q)]$, where $\hat{\mathsf{x}}^*_{\mathsf{w}}$ was defined in the proof of Theorem 3 to be the optimal distribution when both encoder and decoder know the side information. By considering the optimal "water-pouring" distortion allocation for the optimal test-channel distribution $\hat{\mathsf{x}}^*_{\mathsf{w}}$, it can be demonstrated that if the distortion is D, then $E[d(\mathsf{x},\hat{\mathsf{x}}^*_{\mathsf{w}};q)]$ is at most $\min[1,D/q]$ for each q. ■

To develop a similar bound for other distortion measures essentially all we need is an upper bound for the derivative of $h(\mathsf{x}+\sqrt{\tau}\mathsf{v})$ with respect to τ. Since entropy is concave, if we can compute this derivative for $\tau=0$ then it will be an upper bound for the derivative at all τ.

To obtain the desired derivative at $\tau=0$, we can write

$$h(\mathsf{x}+\sqrt{\tau}\mathsf{v}) = I(\mathsf{x}+\sqrt{\tau}\mathsf{v};\sqrt{\tau}\mathsf{v}) - h(\mathsf{x}). \tag{143}$$

The results of Prelov and van der Meulen [35] imply that under certain regularity conditions

$$\frac{\partial}{\partial\tau}\lim_{\tau\to 0^+} I(\mathsf{x}+\sqrt{\tau}\mathsf{v};\sqrt{\tau}\mathsf{v}) = J(\mathsf{x})/2\ , \tag{144}$$

which provides the desired derivative. Similarly if we rewrite the mutual information in (143) as a relative entropy, then a Taylor series expansion of the relative entropy [36, 2.6] can be used to establish (144) provided certain derivatives of the probability distributions exist.

Next, we move to proving Theorem 10. An essential part of our proof is an alternative version of the Shannon Lower Bound, which we develop in the following lemma.

Lemma 3 (Alternative Shannon Lower Bound) *Consider a scaled quadratic distortion measure of the form $d(x, \hat{x}; q) = q \cdot (x - \hat{x})^2$ and let $\hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}$ denote an optimal test-channel distribution when q^n and w^n are known at both encoder and decoder. If we define v to have the same distribution as $\mathsf{x} - \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}$ when conditioned on q and furthermore require v to satisfy the Markov condition $\mathsf{v} \leftrightarrow \mathsf{q} \leftrightarrow \mathsf{w}, \mathsf{x}$, then*

$$R[\text{Q-BOTH-W-BOTH}](D) \geq h(\mathsf{x}|\mathsf{w}) - h(\mathsf{v}|\mathsf{q}). \tag{145}$$

Proof:

$$R[\text{Q-BOTH-W-BOTH}](D) = I(\hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}; \mathsf{x}|\mathsf{q}, \mathsf{w}) \tag{146}$$

$$= h(\mathsf{x}|\mathsf{q}, \mathsf{w}) - h(\mathsf{x}|\mathsf{q}, \mathsf{w}, \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}) \tag{147}$$

$$= h(\mathsf{x}|\mathsf{q}, \mathsf{w}) - h(\mathsf{x} - \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}|\mathsf{q}, \mathsf{w}, \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}) \tag{148}$$

$$= h(\mathsf{x}|\mathsf{q}, \mathsf{w}) - h(\mathsf{v}|\mathsf{q}, \mathsf{w}, \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}) \tag{149}$$

$$\geq h(\mathsf{x}|\mathsf{q}, \mathsf{w}) - h(\mathsf{v}|\mathsf{q}, \mathsf{w}). \tag{150}$$

■

The key difference between Lemma 3 and the traditional Shannon Lower Bound (SLB) is in the choice of the distribution for v. The traditional SLB uses an entropy maximizing distribution for v, which has the advantage of being computable without knowing $\hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}$. The trouble with the entropy maximizing distribution is that it can have an unbounded variance for large distortions. As we show in the following lemma, however, the alternative SLB keeps the variance of v bounded.

Lemma 4 *There exists a choice for v in Lemma 3 such that for all values of w,*

$$\mathrm{Var}[\mathsf{v}|\mathsf{w} = w] \leq \mathrm{Var}[\mathsf{x}|\mathsf{w} = w]. \tag{151}$$

Proof: Imagine that we choose some optimal test-channel distribution $\hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}$ such that the resulting v does not satisfy (151) for some value of w. We will show that it is possible to construct an alternative optimal test-channel distribution $\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}}$, where the resulting v' does satisfy (151) for $\mathsf{w} = w$.

Specifically, if (151) is not satisfied, then it must be that there exists a set $\mathcal{A}$ with

$$\mathrm{Var}[\mathsf{v}|\mathsf{q} = q, \mathsf{w} = w] = \mathrm{Var}[\mathsf{x} - \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}|\mathsf{q} = q, \mathsf{w} = w] > \mathrm{Var}[\mathsf{x}|\mathsf{w} = w], \forall (q, w) \in \mathcal{A}. \tag{152}$$

Define a new random variable $\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}}$ such that $\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}} = \hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}$ for all $(q, w) \notin \mathcal{A}$, but with $\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}} = 0$ for all $(q, w) \in \mathcal{A}$. The distortion is lower for $\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}}$ by construction. Furthermore, the date processing inequality implies that

$$I(\hat{\mathsf{x}}^{*\prime}_{\mathsf{q},\mathsf{w}}; \mathsf{x}|\mathsf{w}, \mathsf{q}) \leq I(\hat{\mathsf{x}}^*_{\mathsf{q},\mathsf{w}}; \mathsf{x}|\mathsf{w}, \mathsf{q}) \tag{153}$$

and so the rate is lower too. Thus if we define $v' = \hat{x}^{*\prime}_{q,w} - x$ analogously to how we defined v, then condition (151) is satisfied with v replaced by v'. ■

Proof of Theorem 10: Using the alternative SLB from Lemma 3 and the test-channel distribution $\hat{x} = x + v$ with v chosen according to Lemma 3 we obtain

$$R[\text{Q-ENC-W-DEC}](D) - R[\text{Q-BOTH-W-BOTH}](D) \leq I(x+v; q, x|w) - [h(x|w,q) - h(v|w,q)] \tag{154}$$

$$= h(x+v|w) - h(x+v|q,x,w) - h(x|w,q) + h(v|w,q) \tag{155}$$

$$= h(x+v|w) - h(v|q,x,w) - h(x|w,q) + h(v|w,q) \tag{156}$$

$$= h(x+v|w) - h(v|q) - h(x|w,q) + h(v|q) \tag{157}$$

$$= h(x+v|w) - h(x|w) \tag{158}$$

$$\begin{aligned} &= D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) - D(p_{x+v|w} \| \mathcal{N}(\sigma^2_{x|w} + \sigma^2_{v|w})) \\ &\quad + h(\mathcal{N}(\sigma^2_{x|w} + \sigma^2_{v|w})|w) - h(\mathcal{N}(\sigma^2_{x|w})|w) \end{aligned} \tag{159}$$

$$\leq D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) + h(\mathcal{N}(\sigma^2_{x|w} + \sigma^2_{v|w})|w) - h(\mathcal{N}(\sigma^2_{x|w})|w) \tag{160}$$

$$\begin{aligned} &= D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) \\ &\quad + \int \left[h(\mathcal{N}(\sigma^2_{x|w} + \sigma^2_{v|w})|w = w) - h(\mathcal{N}(\sigma^2_{x|w})|w = w) \right] p_w(w) dw \end{aligned} \tag{161}$$

$$\leq D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) + \int \left[\frac{1}{2} \log \left(1 + \frac{\sigma^2_{v|w}}{\sigma^2_{x|w}} \right) \right] p_w(w) dw \tag{162}$$

$$\leq D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) + \int \left[\frac{1}{2} \log \left(1 + \frac{\sigma^2_{\max}}{\sigma^2_{x|w}} \right) \right] p_w(w) dw \tag{163}$$

$$\leq D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) + \int \left[\frac{1}{2} \log \left(1 + \frac{\sigma^2_{\max}}{\sigma^2_{\min}} \right) \right] p_w(w) dw \tag{164}$$

$$= D(p_{x|w} \| \mathcal{N}(\sigma^2_{x|w})) + \frac{1}{2} \log \left(1 + \frac{\sigma^2_{\max}}{\sigma^2_{\min}} \right). \tag{165}$$

To obtain (158)–(162) we use the same arguments as in (68)–(74) plus the additional observation that relative entropy is positive and can be dropped in obtaining (160). Next, we apply Lemma 4 to keep the variance of the test-channel noise to be at most $\sigma^2_{\max}$ to get (163). Finally, the assumption that $\sigma^2_{x|w} \geq \sigma^2_{\min}$ yields (164). ■

To develop a similar bound for other distortion measures, we would use an entropy maximizing distribution for the appropriate distortion measure in $D(p_{x|w} \| \cdot)$ and $D(p_{x+v} \| \cdot)$ above.

REFERENCES

[1] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. 22, 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,” *Information and Control*, 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. Inform. Theory*, vol. 48, pp. 1629–1638, June 2002.

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

[5] R. Zamir, “The rate loss in the Wyner-Ziv problem,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 2073–2084, Nov. 1996.

[6] T. Linder, R. Zamir, and K. Zeger, “On source coding with side-information-dependent distortion measures,” *IEEE Trans. Inform. Theory*, vol. 46, 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,” *Proceedings of the IEEE*, vol. 81, 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. Inform. Theory*, vol. 19, pp. 480–489, July 1973.

[10] T. Berger. Private Communication.

[11] R. Zamir, “The half a bit loss of robust source/channel codebooks,” in *Information Theory Workshop*, (Bangalore, India), pp. 123–126, Oct. 2002.

[12] E. Martinian, *Dynamic Information and Constraints in Source and Channel Coding*. PhD thesis, Massachusetts Institute of Technology, 2004.

[13] Z.-P. Liang and P. C. Lauterbur, *Principles of Magnetic Resonance Imaging*. IEEE Press, 1999.

[14] G. Franceschetti, S. Merolla, and M. Tesauro, “Phase quantized sar signal processing: theory and experiments,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 35, pp. 201–204, Jan. 1999.

[15] B. L. Douglas, S. D. Kent, and H. Lee, “Parameter estimation and the importance of phase in acoustic microscopy,” in *Proceedings of the IEEE Ultrasonics Symposium*, vol. 2, pp. 715–718, Oct. 1992.

[16] A. Oppenheim, J. Lim, G. Kopec, and S. Pohlig, “Phase in speech and pictures,” in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 632–637, 1979.

[17] A. Hung and T. Meng, “Multidimensional rotations for robust quantization of image data,” *IEEE Transactions on Image Processing*, vol. 7, pp. 1–12, Jan 1998.

[18] R. Vafin and W. Kleijn, “Entropy-constrained polar quantization: theory and an application to audio coding,” in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1837–1840, 2002.

[19] M. Hayes, J. Lim, and A. Oppenheim, “Phase-only signal reconstruction,” in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 437–440, 1980.

[20] J. K. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Trans. Inform. Theory*, vol. 16, pp. 406–411, July 1970.

[21] T. Linder and R. Zamir, “On the asymptotic tightness of the Shannon lower bound,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 2026–2031, Nov. 1994.

[22] R. Zamir and M. Feder, “Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 141–154, Jan. 1995.

[23] H. Feng and M. Effros, “Improved bounds for the rate loss of multiresolution source codes,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 809–821, Apr. 2003.

[24] L. Lastras and T. Berger, “All sources are nearly successively refinable,” *IEEE Trans. Inform. Theory*, vol. 47, pp. 918–926, Mar. 2001.

[25] M. D. Gaubatz, D. M. Chandler, and S. S. Hemami, “Spatially-selective quantization and coding for wavelet-based image compression,” in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 209–212, 2005.

[26] M. Gaubatz, D. M. Chandler, and S. S. Hemami, “Spatial quantization via local texture masking,” in *Proc. Human Vision and Electronic Imaging*, (San Jose, CA), Jan. 2005.

[27] T. S. Han and S. Verdu, “Approximation theory of output statistics,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.

[28] S. Vembu, S. Verdu, and Y. Steinberg, “The source-channel separation theorem revisited,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 44–54, Jan. 1995.

[29] T. S. Han, “An information-spectrum approach to source coding theorems with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 1145–1164, July 1997.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

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

[32] M. Adams and V. Guillemin, *Measure Theory and Probability*. Birkäuser, 1996.

[33] A. Dembo, T. M. Cover, and J. A. Thomas, “Information theoretic inequalities,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501–1518, Nov. 1991.

[34] N. Blachman, “The convolution inequality for entropy powers,” *IEEE Trans. Inform. Theory*, vol. 11, pp. 267–271, Apr. 1965.

[35] V. V. Prelov and E. C. van der Meulen, “An asymptotic expression for the information and capacity of a multidimensional channel with weak input signals,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 1728–1735, Sep. 1993.

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