On Source Coding with Side Information Dependent Distortion Measures *

Tamás Linder Ram Zamir Kenneth Zeger

IEEE Trans. on Information Theory
Submitted: April 28, 1998
Revised: September 12, 1999

Abstract

High resolution bounds in lossy coding of a real memoryless source are considered when side information is present. Let X be a "smooth" source and let Y be the side information. First we treat the case when both the encoder and the decoder have access to Y and we establish an asymptotically tight (high-resolution) formula for the conditional rate-distortion function $R_{X|Y}(D)$ for a class of locally quadratic distortion measures which may be *functions of the side information*. We then consider the case when only the decoder has access to the side information (i.e., the "Wyner-Ziv problem"). For side information dependent distortion measures, we give an explicit formula which tightly approximates the Wyner-Ziv rate-distortion function $R^{\mathrm{WZ}}(D)$ for small D under rather general assumptions on the joint distribution of X and Y. These results demonstrate that for side information dependent distortion measures the rate loss $R^{\mathrm{WZ}}(D) - R_{X|Y}(D)$ can be bounded away from zero in the limit of small D. This contrasts the case of distortion measures which do not depend on the side information where the rate loss vanishes as $D \to 0$.

Index Terms: Conditional rate-distortion, Wyner-Ziv problem, source coding, side information, high-resolution theory, general distortion measures, Shannon lower bound.

---

*T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: `linder@mast.queensu.ca`). R. Zamir is with the Department of Electrical Engineering - Systems, Tel-Aviv University, Tel-Aviv, 69978, Israel (email: `zamir@eng.tau.ac.il`). K. Zeger is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla CA 92093-0407 (email: `zeger@ucsd.edu`). The research was supported in part by the National Science Foundation and the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

Consider the source coding scenario depicted in Figure 1 (see Berger [1], Wyner and Ziv [2]). The sequence $\{(X_k, Y_k)\}$ consists of independent and identically distributed copies of a pair of real random variables (X, Y), where X is called the *source* and Y is called the *side information*. Encoding and decoding is done in blocks of length n, and the distortion D between the source block $(X_1, \ldots, X_n)$ and its reproduction $(\hat{X}_1, \ldots, \hat{X}_n)$ is given by $\mathbf{E}\big[n^{-1} \sum_{k=1}^{n} d(X_k, \hat{X}_k)\big]$, where $d(x, \hat{x})$ is a nonnegative single-letter distortion measure.

When switches A and B are closed, both the encoder and the decoder have access to the side information. In this case, let $R_{X|Y}(D)$ denote the minimum rate R such that for any $\epsilon > 0$ and all n large enough there exists an encoder-decoder pair operating at distortion D and rate not exceeding $R + \epsilon$.

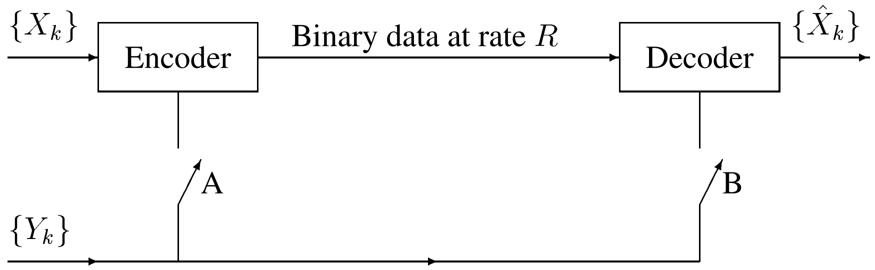

Figure 1: Source coding with side information

Under mild regularity conditions, $R_{X|Y}(D)$, called the *conditional rate-distortion function*, is given by

$$R_{X|Y}(D) = \inf_{\hat{X}} I(X; \hat{X}|Y) \tag{1}$$

where the infimum of the conditional mutual information $I(X; \hat{X}|Y)$ is taken over conditional distributions of $\hat{X}$ given (Y, X) such that $\mathbf{E}[d(X, \hat{X})] \le D$ (see Berger [1], Gray [3], and Wyner [4]).

If switch A is open and switch B is closed, only the decoder knows the side information. In this case, let the minimum rate achievable at distortion D be denoted by $R^{\mathrm{WZ}}(D)$. The quantity $R^{\mathrm{WZ}}(D)$ was determined by Wyner and Ziv [2] for finite alphabets and by Wyner [4] for the general case. Assuming that d satisfies certain mild regularity conditions [4], we have

$$R^{\mathrm{WZ}}(D) = \inf_{Z} I(X; Z|Y) \tag{2}$$

where Z is a random object taking values in an arbitrary measurable space, and where the infimum is taken over all conditional distributions of Z given (X, Y) such that $Y \leftrightarrow X \leftrightarrow Z$ forms a Markov chain (i.e., Y and Z are conditionally independent given X) and there exists a measurable function $f(Y, Z)$ with $\mathbf{E}[d(X, f(Y, Z))] \leq D$. The Wyner-Ziv rate distortion function $R^{\mathrm{WZ}}(D)$ finds applications in coding for communication networks [5] and in systematic data transmission [6].

Although (1) and (2) provide a single-letter characterization[1] of the achievable rates at distortion level D, more explicit expressions are desirable. In this paper we develop explicit formulas which approximate $R_{X|Y}(D)$ and $R^{\mathrm{WZ}}(D)$ with increasing accuracy as $D \to 0$. Moreover, we consider the more general situation in which the average distortion depends on the side information and is given for n-blocks by

$$\mathbf{E}\Big[\frac{1}{n}\sum_{k=1}^{n} d(X_k, Y_k, \hat{X}_k)\Big]$$

for a single-letter distortion measure $d(x, y, \hat{x})$. Intuitively, the distortion between the source X and its reproduction $\hat{X}$ depends also on the current value of a "context" random variable Y.

The need for such a context dependent distortion measure may arise, for example, in video coding where (due to perceptual effects) the visibility threshold at a given pixel location depends on the luminance intensity of a pixel at the same location in a previous frame [7]. In this case, the previous frame can be considered the side information for the coding of the present frame (switches A and B are closed). To motivate the use of a side information dependent distortion measure in the Wyner-Ziv problem, suppose in the presence of acoustical background noise one plays back compressed speech or music. The background noise cannot be eliminated since it depends on the precise location of the listener, but if a microphone system feeds to the decoder a signal Y correlated with the background noise, the perceptual effects (e.g., masking effects in frequency and time) may be modeled by a distortion measure which depends on Y. In this case, one can improve the quality of decoding by making the reconstruction a function of the background noise. Generally, for an encoder-decoder pair using n-blocks, given the received code index $i = f_n(X^n)$ and the side information $Y^n = y^n$, the optimum reconstruction function of the Wyner-Ziv decoder is

$$\hat{x}^n_{\text{opt}}(i, y^n) = \arg\min_{\hat{x}^n} \mathbf{E}\big[d_n(X^n, y^n, \hat{x}^n)\big|\, i, Y^n = y^n\big]$$

where d_n is the single-letter distortion measure generated by d. Therefore, even if the side information Y^n is statistically independent of the source X^n (as is the case in the above example) the

[1]Strictly speaking, (2) is not a single-letter characterization since the alphabet of the auxiliary random object Z is not fixed. However, if X and Y are real-valued, one can prove that Z can be restricted to be a real random variable without changing the defining infimum.

optimum reconstruction may depend on y^n.

In this paper, we will assume that $d(x, y, \hat{x})$ is a sufficiently smooth function such that $d(x, y, \hat{x}) = 0$ if and only if $x = \hat{x}$, and if $|x - \hat{x}|$ is small, then the behavior of $d(x, y, \hat{x})$ is determined by the second order term in its Taylor expansion with respect to $\hat{x}$ around (x, y, x). That is,

$$d(x, y, \hat{x}) = m(x, y)(x - \hat{x})^2 + o(|x - \hat{x}|^2) \tag{3}$$

as $|x - \hat{x}| \to 0$, where

$$m(x, y) = \frac{1}{2} \frac{\partial^2 d(x, y, \hat{x})}{\partial \hat{x}^2}\Big|_{\hat{x}=x}.$$

This definition generalizes the notion of locally quadratic input weighted distortion measures [8] to side information dependent distortion measures. Notice, however, that the dependence on the side information is only through the coefficient of the quadratic term; the optimum reconstruction given x and y is still x, independent of the side information y. Locally quadratic input weighted distortion measures are of particular interest because some important perceptual distortion measures for speech and image coding fall into this category [9, 10].

Let $R_{X|Y}$ and $R^{\mathrm{WZ}}(D)$ denote the obvious extensions of (1) and (2) to a side information dependent distortion measure satisfying (3). As we discuss later, the operational meaning of these quantities does not change with the more general definition of the distortion measure. Theorem 1 in the next section states that for such distortion measures and for "smooth" sources,

$$R_{X|Y}(D) = h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log m(X, Y)\big] + o(1) \tag{4}$$

where $o(1) \to 0$ as $D \to 0$, and where $h(X|Y)$ denotes the conditional differential entropy of the source given the side information. This result generalizes a recently derived asymptotic formula [8] for the rate-distortion function of a smooth source relative to locally quadratic non-difference distortion measures, to conditional rate-distortion functions and to distortion measures which depend on the side information.

In contrast, determining the asymptotics of $R^{\mathrm{WZ}}(D)$ for distortion measures of the form (3) appears to be a more difficult problem and we do not have the complete solution in this case. Assuming that the joint distribution of X and Y satisfies certain conditions (for example, Y is generated by passing X through an additive Gaussian noise channel and then passing the result through an arbitrary memoryless channel), we prove in Theorem 2 that as $D \to 0$,

$$R^{\mathrm{WZ}}(D) = h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log \overline{m}(X)\big] + o(1) \tag{5}$$

where $\overline{m}(X) = \mathbf{E}\big[m(X, Y)|X\big]$. In particular, this formula holds when X and Y are independent. For the general case when the only condition is that $I(X; Y) < \infty$, we prove in Theorem 2 that the right hand side of (5) is an asymptotic upper bound, i.e.,

$$R^{\mathrm{WZ}}(D) \le h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log \overline{m}(X)\big] + o(1) \tag{6}$$

as $D \to 0$. We conjecture that the bound in (6) is also asymptotically tight in this general case (i.e., the reverse inequality also holds) but this is presently unproven.

Expressions (1) and (2) show that $R^{\mathrm{WZ}}(D) \ge R_{X|Y}(D)$ for all D, as expected, since knowledge of the side information can only improve the encoder. Thus there is a nonnegative "rate loss" $R^{\mathrm{WZ}}(D) - R_{X|Y}(D)$ when the side information is known only to the decoder. This rate loss was investigated in [11] and it was established there that for difference distortion measures (i.e., when $d(x,\hat{x}) = \rho(x-\hat{x})$) the loss becomes asymptotically negligible as $D \to 0$. Our conclusion for side information dependent distortion measures is different. At least under the conditions of Theorem 2 on the joint distribution of X and Y, we have by (4), (5), and Jensen's inequality, that

$$\lim_{D\to 0}\Big(R^{\mathrm{WZ}}(D) - R_{X|Y}(D)\Big) = \frac{1}{2}\mathbf{E}\big(\log \mathbf{E}[m(X,Y)|X]\big) - \frac{1}{2}\mathbf{E}\big[\log m(X,Y)\big] \ge 0 \tag{7}$$

where the inequality is strict unless $m(X,Y)$ is a function of X alone with probability one.

Of course, this rate loss would not be surprising if the value of $\hat{x}$ minimizing $d(x,y,\hat{x})$ for given x and y depended on y, the side information that is not available at the encoder. Note, however, that the minimizing $\hat{x}$ in (3) is equal to x for all y and still the rate loss is positive. As we will show in Section 4 through a companding interpretation of the encoding-decoding process, the deeper reason for this rate loss is that the optimum density of the code points of the informed encoder depends on Y, while the uninformed encoder must use a fixed code point density for all values of Y.

We can support the above conclusion by demonstrating that the zero asymptotic rate loss pointed out in [11] is due to $d(x,\hat{x})$ not depending on y rather than to it being a difference distortion measure. Indeed, notice that a locally quadratic non-difference distortion measure $d(x,\hat{x})$ is a special case of a side information dependent distortion measure. In this case, $m(X,Y) = m(X)$, where $m(x) = \frac{1}{2}\frac{\partial^2 d(x,\hat{x})}{\partial \hat{x}^2}\Big|_{\hat{x}=x}$, implying that the right hand sides of (4) and (6) asymptotically coincide. Thus, as a corollary of Theorems 1 and 2, we obtain the new result

$$\lim_{D\to 0}\Big(R^{\mathrm{WZ}}(D) - R_{X|Y}(D)\Big) = 0$$

which states that for locally quadratic side information independent distortion measures the asymptotic rate loss is zero in the Wyner-Ziv problem.

2 The conditional rate-distortion function at high resolution

Let (X, Y) be a pair of real random variables such that $I(X; Y) < \infty$ and X has a density and finite differential entropy $h(X)$. It follows that the conditional differential entropy $h(X|Y)$ is well defined and finite. Suppose that $\mathbf{E}[X^2] < \infty$ and assume the distortion measure $d(x, y, \hat{x})$ satisfies the following conditions:

(a) $d(x, y, \hat{x})$ is three times differentiable with respect to $\hat{x}$ and $\partial^3 d(x, y, \hat{x})/\partial \hat{x}^3$ is uniformly bounded.

(b) $d(x, y, \hat{x}) \geq 0$; and $d(x, y, \hat{x}) = 0$ if and only if $x = \hat{x}$.

(c) $\liminf_{|\hat{x}|\to\infty} d(x, y, \hat{x}) > 0$ for all $x, y \in \mathbb{R}$.

(d) If $m(x, y)$ is defined by

$$m(x, y) = \frac{1}{2} \left. \frac{\partial^2 d(x, y, \hat{x})}{\partial \hat{x}^2} \right|_{\hat{x}=x}$$

then $m(x, y)$ is continuous on an open subset of $\mathbb{R}^2$ whose complement has zero Lebesgue measure. Furthermore,

$$\mathbf{E}|\log m(X, Y)| < \infty \quad \text{and} \quad \mathbf{E}\left[\frac{1}{m(X, Y)}\right] < \infty.$$

Note that conditions (a) and (b) imply $m(x, y) \geq 0$. Note also that the continuity assumption on $m(x, y)$ in (d) does not rule out discrete side information. In fact, for Y having a discrete distribution with a finite number of outcomes $y_1, \ldots, y_m$, only the continuity of $m(x, y_i)$ in x is required for each i, since in this case one can formally redefine $m(x, y)$ to be continuous in both variables.

The conditional rate distortion function $R_{X|Y}(D)$ is defined for $D > 0$ by

$$R_{X|Y}(D) = \inf_{\hat{X}} I(X; \hat{X}|Y) \tag{8}$$

where the infimum of the conditional mutual information $I(X; \hat{X}|Y)$ is taken over all conditional distributions of $\hat{X}$ given (Y, X) such that $\mathbf{E}[d(X, Y, \hat{X})] \leq D$.

Theorem 1 *Let d be a distortion function satisfying conditions* (a)–(d) *and let X be a real source with a density and finite differential entropy $h(X)$ such that $\mathbf{E}[X^2]$ and $I(X; Y)$ are finite. Then as $D \to 0$, the asymptotic behavior of $R_{X|Y}(D)$ is given by*

$$R_{X|Y}(D) = h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log m(X, Y)\big] + o(1).$$

Remark. To attach operational meaning to $R_{X|Y}(D)$, defined in (8), one needs to extend the proof of the coding theorem from the case when d depends only on X and $\hat{X}$ to the case when it also depends on Y. This extension is relatively straightforward under some regularity conditions on the distortion measure. For example, if in addition to conditions (a)–(d), it is also assumed that $d(x, y, \hat{x})$ is bounded, one can check that the corresponding steps in the proof given in [4, Appendix A] carry over to our case. But regardless of the operational meaning of $R_{X|Y}(D)$, Theorem 1 always holds if $d(x, y, \hat{x})$ satisfies conditions (a)–(d).

The proof of the theorem is based on a technique developed in [8]. To prove that

$$\limsup_{D\to 0}\Big[R_{X|Y}(D) + \frac{1}{2}\log(2\pi e D)\Big] \le h(X|Y) + \frac{1}{2}\mathbf{E}\big[\log m(X,Y)\big] \tag{9}$$

we let $\hat{X}$ be the output of the "forward test channel" given by

$$\hat{X} = X + \frac{N_D}{\sqrt{m(X,Y)}}$$

where N_D is a zero-mean Gaussian random variable with variance D which is independent of (X, Y). Then the conditions on d imply that

$$\lim_{D\to 0}\frac{\mathbf{E}[d(X,Y,\hat{X})]}{D} = \lim_{D\to 0}\frac{\mathbf{E}[d(X,Y,\hat{X})]}{\mathbf{E}[m(X,Y)(X-\hat{X})^2]} = 1.$$

On the other hand,

$$I(X;\hat{X}|Y) = h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|Y\Big) - h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|X,Y\Big).$$

By the independence of N_D and (X, Y),

$$h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|X,Y\Big) = \frac{1}{2}\log(2\pi e D) - \frac{1}{2}\mathbf{E}\big[\log m(X,Y)\big].$$

Lemma 1 in Appendix A proves that

$$\limsup_{D\to 0} h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|Y\Big) \le h(X|Y)$$

and (9) follows.

The proof of the converse part, that

$$\liminf_{D\to 0}\Big[R_{X|Y}(D) + \frac{1}{2}\log(2\pi e D)\Big] \ge h(X|Y) + \frac{1}{2}\mathbf{E}\big[\log m(X,Y)\big], \tag{10}$$

is more involved and is deferred to Appendix B.

3 The Wyner-Ziv rate-distortion function at high resolution

For $D > 0$ define

$$R^{\mathrm{WZ}}(D) = \inf_{Z} I(X; Z|Y) \tag{11}$$

where Z is a real random variable, and where the infimum is taken over all conditional distributions of Z given (X, Y) such that $Y \leftrightarrow X \leftrightarrow Z$ forms a Markov chain and for which there exists a measurable function $f(Y, Z)$ with $\mathbf{E}[d(X, Y, f(Y, Z))] \leq D$. This definition is more general than the one originally given in [2] in that the distortion measure is allowed to depend on Y. A coding theorem for discrete alphabets and such distortion measures is proved in [12, Corollary 4.6, Ch. 3]. For general alphabets it can be verified that, with the necessary modifications, the proof of the coding theorem for $R^{\mathrm{WZ}}(D)$ in [4] also works for side information dependent distortion measures (e.g., the convexity of $R^{\mathrm{WZ}}(D)$ in D and the converse coding theorem are straightforward extensions of [4, App. B] and [4, Sec. 4], respectively).

We make the same assumptions on (X, Y) and $d(x, y, \hat{x})$ as in the previous section. The following theorem provides an asymptotically tight expression for $R^{\mathrm{WZ}}(D)$ under rather general conditions on the joint distribution of the source-side information pair (X, Y).

Theorem 2 *Let d be a distortion function satisfying conditions* (a)–(d) *and let X be a real-valued memoryless source with a density such that $\mathbf{E}[X^2] < \infty$. Assume that $I(X; Y) < \infty$ and X has finite differential entropy $h(X)$. Then as $D \to 0$, we have*

$$R^{\mathrm{WZ}}(D) \leq h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log \overline{m}(X)\big] + o(1) \tag{12}$$

where $\overline{m}(X) = \mathbf{E}\big[m(X, Y)|X\big]$. Furthermore, if there exists a sequence of real random variables $\{V_n\}_{n=1}^{\infty}$ such that

(i) $X \leftrightarrow V_n \leftrightarrow Y$ *forms a Markov chain for all n*

(ii) $V_n \to X$ *in probability as $n \to \infty$*

(iii) $I(X; V_n) < \infty$ *for all n*

then, as $D \to 0$, we have

$$R^{\mathrm{WZ}}(D) = h(X|Y) - \frac{1}{2}\log(2\pi e D) + \frac{1}{2}\mathbf{E}\big[\log \overline{m}(X)\big] + o(1). \tag{13}$$

The existence of a sequence $\{V_n\}$ satisfying the assumptions of Theorem 2 can be shown under quite general conditions. In the simplest case when X and Y are statistically independent, V_n can always be chosen as $Q_n(X)$, where Q_n is an (infinite level) uniform quantizer of step size $1/n$. For dependent X and Y, the sequence $\{V_n\}$ exists for example in the following two cases:

(1) Y is generated by passing X through an additive Gaussian noise channel and then passing the result through an arbitrary memoryless channel.

(2) X is generated by passing Y through an arbitrary memoryless channel and then passing the result through an additive Gaussian noise channel.

To see how the sequence $\{V_n\}$ is constructed in these examples, let N be a zero mean Gaussian random variable with variance $\sigma^2 > 0$, and for $n \geq 1$ let N_n and N'_n be independent zero mean Gaussian random variables with variance σ^2/n and $(1-\frac{1}{n})\sigma^2$, respectively, such that $N_n + N'_n = N$. Then in case **(1)** the assumption is that X and N are independent and Y is the output of an arbitrary memoryless channel whose input is $X + N$. In this case we can let (N_n, N'_n) be independent of X and set $V_n = X + N_n$. In case **(2)** the assumption is that Y and N are independent and $X = Y' + N$, where Y' is the output of a memoryless channel whose input is Y. In this case we can choose (N_n, N'_n) to be independent of Y and we set $V_n = Y' + N'_n$. Conditions **(i)-(iii)** are obviously satisfied in both cases.

Note that the variance $\sigma^2 > 0$ of the Gaussian noise in both cases can be arbitrarily small, and therefore this noise can be interpreted as a small Gaussian perturbation of the (otherwise arbitrary) joint distribution of (X, Y). Note also that the assumption that the additive channel is Gaussian is not essential. The examples also work under the more general assumption that the distribution of N is *infinitely divisible* [13], that is, for all $n \geq 1$ there exist n independent and identically distributed random variables $N_1^{(n)}, \ldots, N_n^{(n)}$ such that $N_1^{(n)} + \cdots + N_n^{(n)} = N$.

Proof of Theorem 2 To prove (12), let Z be the output of the test channel given by

$$Z = X + \frac{N_D}{\sqrt{\overline{m}(X)}}$$

where N_D is a Gaussian random variable with zero mean and variance D which is independent of (X, Y). With this choice, Z and Y are conditionally independent given X. Letting $f(y, z) = z$, we have for small D,

$$\mathbf{E}[d(X, Y, Z)] \approx \mathbf{E}[m(X, Y)(X - Z)^2] = D \cdot \mathbf{E}\Big[\frac{m(X, Y)}{\overline{m}(X)}\Big] = D$$

and therefore $R^{\mathrm{WZ}}(D) \leq I(X; Z|Y) + o(1)$ as $D \to 0$. The rest of the proof of (12) follows the corresponding steps in the proof of Theorem 1 with $m(X, Y)$ replaced by $\overline{m}(X)$.

Assume now that there exists $\{V_n\}$ satisfying conditions **(i)-(iii)** of the theorem. Then (12) holds and we only have to prove the reverse inequality to show (13). Assume for simplicity that $d(x, y, \hat{x}) = m(x, y)(x - \hat{x})^2$. The extension of the derivation to the general case is straightforward using the proof technique of [8, Proposition 2]. We will also make the assumption that $m(x, y)$ is uniformly continuous and bounded away from zero on $\mathbb{R}^2$. This assumption on $m(x, y)$ can be relaxed (so that $m(x, y)$ need only satisfy condition (d)) by using the proof technique of the converse part of Theorem 1 in Appendix B.

Since the structure of the two proofs are similar, in the following derivation we will be able to use (without additional justification) some of the bounds developed in proving the converse part of Theorem 1.

Let Z_D and f be such that $Y \leftrightarrow X \leftrightarrow Z_D$ forms a Markov chain, $I(X; Z_D|Y) < \infty$, and

$$\mathbf{E}\big[m(X, Y)(X - f(Y, Z_D))^2\big] \le D. \tag{14}$$

The basic idea of the upcoming derivation is that if n is large enough, then V_n and X are very close with large probability and therefore the distortion $m(X, Y)(X - f(Y, Z_D))^2$ is well approximated by the distortion $m(V_n, Y)(X - f(Y, Z_D))^2$ and thus the latter can be used in place of the former. On the other hand, $m(V_n, Y)(X - f(Y, Z_D))^2$ becomes quadratic in X when conditioned on the event $(V_n = v, Y = y)$. This and the Markov chain condition $X \leftrightarrow V_n \leftrightarrow Y$ allow us to apply the Shannon lower bound [1] "locally". The desired lower bound will follow through Jensen's inequality after averaging these local bounds.

To make this more precise, for all $k, n \ge 1$ define the binary random variable $B_n^{(k)}$ by

$$B_n^{(k)} = \begin{cases} 0 & \text{if } |X - V_n| \le 1/k \\ 1 & \text{otherwise.} \end{cases}$$

Then by condition (**ii**), $\mathbf{P}(B_n^{(k)} = 0) \to 1$ as $n \to \infty$ for all k. Notice that the assumptions on $m(x, y)$ imply that if $|x - x'| \le 1/k$, then

$$\left| \frac{m(x, y)}{m(x', y)} - 1 \right| \le \epsilon_k \tag{15}$$

for some $\epsilon_1, \epsilon_2, \ldots$, such that $\lim_{k\to\infty} \epsilon_k = 0$. It follows that

$$\begin{aligned} &\mathbf{E}\big[m(X, Y)(X - f(Y, Z_D))^2 | V_n = v, B_n^{(k)} = 0\big] \\ &\quad \ge \; (1 - \epsilon_k)\mathbf{E}\big[m(V_n, Y)(X - f(Y, Z_D))^2 | V_n = v, B_n^{(k)} = 0\big]. \end{aligned} \tag{16}$$

Now we have

$$\begin{aligned}
&\mathbf{E}\big[(X - f(Y, Z_D))^2 | Y = y, V_n = v, B_n^{(k)} = 0\big] \\
&\qquad \geq \; \mathbf{E}\Big[(X - \mathbf{E}[X | Z_D, Y, V_n])^2 \big| Y = y, V_n = v, B_n^{(k)} = 0\Big] \\
&\qquad \triangleq \; \mathbf{Var}(X | Z_D, Y = y, V_n = v, B_n^{(k)} = 0).
\end{aligned} \tag{17}$$

The key step in the proof is the observation that the Markov chain conditions $Y \leftrightarrow X \leftrightarrow Z_D$ and $X \leftrightarrow V_n \leftrightarrow Y$ imply that the joint distributions of (Y, X, Z_D) and (X, V_n, Y) can be coupled such that the Markov chain condition $Y \leftrightarrow V_n \leftrightarrow X \leftrightarrow Z_D$ is satisfied. Thus we have

$$\mathbf{Var}(X | Z_D, Y = y, V_n = v, B_n^{(k)} = 0) = \mathbf{Var}(X | Z_D, V_n = v, B_n^{(k)} = 0)$$

(the conditioning on Y can be dropped) and therefore by (17) we obtain

$$\begin{aligned}
&\mathbf{E}\big[m(V_n, Y)(X - f(Y, Z_D))^2 | Y = y, V_n = v, B_n^{(k)} = 0\big] \\
&\qquad = \; m(v, y)\mathbf{E}\big[(X - f(Y, Z_D))^2 | Y = y, V_n = v, B_n^{(k)} = 0\big] \\
&\qquad \geq \; m(v, y)\, \mathbf{Var}(X | Z_D, V_n = v, B_n^{(k)} = 0).
\end{aligned}$$

Combining this with (16) we get

$$\begin{aligned}
D(v) \;&\triangleq\; \mathbf{E}\big[m(X, Y)(X - f(Y, Z_D))^2 | V_n = v, B_n^{(k)} = 0\big] \\
&\geq\; (1 - \epsilon_k)\mathbf{E}[m(V_n, Y) | V_n = v, B_n^{(k)} = 0]\, \mathbf{Var}(X | Z_D, V_n = v, B_n^{(k)} = 0).
\end{aligned} \tag{18}$$

We will now use (18) to lower bound the conditional mutual information $I(X; Z_D | Y)$. Note that the chain rule and the finiteness of $I(X; V_n)$ imply (see B.1)

$$I(X; Z_D | Y) = h(X | Y) - h(X | Y, Z_D, V_n) - I(X; V_n | Y, Z_D). \tag{19}$$

Since $f(Y, Z_D) \to X$ in probability as $D \to 0$, Lemma 2 in Appendix C shows that

$$\lim_{D \to 0} I(X; V_n | Y, Z_D) = 0. \tag{20}$$

Now an argument analogous with (B.4) implies

$$\begin{aligned}
h(X | Y, Z_D, V_n) \;&\leq\; h(X | Y, Z_D, V_n, B_n^{(k)} = 0)\mathbf{P}(B_n^{(k)} = 0) \\
&\qquad + h(X | Y, Z_D, V_n, B_n^{(k)} = 1)\mathbf{P}(B_n^{(k)} = 1) + H(B_n^{(k)})
\end{aligned} \tag{21}$$

where $H(B_n^{(k)})$ is the Shannon entropy of the $B_n^{(k)}$. Using inequality (B.5) and the fact that $\lim_n \mathbf{P}(B_n^{(k)} = 1) = 0$ for all $k \geq 1$, we obtain

$$\limsup_{D \to 0} \Big(h(X | Y, Z_D, V_n, B_n^{(k)} = 1)\mathbf{P}(B_n^{(k)} = 1) + H(B_n^{(k)}) \Big) = \delta_{n,k}, \tag{22}$$

where $\lim_{n\to\infty} \delta_{n,k} = 0$ for all k. On the other hand, since conditioning reduces entropy, using Jensen's inequality and the fact that $h(U) \le \frac{1}{2}\log\bigl(2\pi e\,\mathbf{Var}(U)\bigr)$ for any real random variable U with a density, we obtain

$$\begin{aligned} & h(X|Y, Z_D, V_n = v, B_n^{(k)} = 0) \\ & \quad \le \frac{1}{2}\log\bigl(2\pi e\,\mathbf{Var}(X|Z_D, V_n = v, B_n^{(k)} = 0)\bigr) \\ & \quad \le \frac{1}{2}\log\bigl(2\pi e D(v)\bigr) - \frac{1}{2}\log \mathbf{E}[m(V_n, Y)|V_n = v, B_n^{(k)} = 0] - \frac{1}{2}\log(1-\epsilon_k), \end{aligned}$$

where the second inequality follows by (18). Since $\mathbf{E}[D(V_n)] \le D/\mathbf{P}(B_n^{(k)} = 0)$ (see (B.9), another application of Jensen's inequality gives

$$\begin{aligned} & h(X|Z_D, V_n, B_n^{(k)} = 0) \\ & \quad \le \frac{1}{2}\log\bigl(2\pi e D\bigr) - \frac{1}{2}\mathbf{E}\bigl(\log \mathbf{E}[m(V_n, Y)|V_n, B_n^{(k)} = 0]\bigr) \\ & \qquad - \frac{1}{2}\log \mathbf{P}(B_n^{(k)} = 0) - \frac{1}{2}\log(1-\epsilon_k) \end{aligned} \tag{23}$$

This, (21), and (22) yield

$$\begin{aligned} \limsup_{D\to 0}\left(h(X|Y, Z_D, V_n) - \frac{1}{2}\log\bigl(2\pi e D\bigr)\right) \le {} & \frac{1}{2}\mathbf{E}\bigl(\log \mathbf{E}[m(X, Y)|V_n, B_n^{(k)} = 0]\bigr) \\ & + \hat{\delta}_{n,k} + b_k \end{aligned} \tag{24}$$

where $\lim_n \hat{\delta}_{n,k} = 0$ for all k and $\lim_k b_k = 0$. The assumption that $m(x, y)$ is uniformly continuous and bounded away from zero and the facts that $V_n \to X$ in probability and $\lim_n \mathbf{P}(B_n^{(k)} = 0) = 1$ imply

$$\lim_{n\to\infty} \mathbf{E}\bigl(\log \mathbf{E}[m(V_n, Y)|V_n, B_n^{(k)} = 0]\bigr) = \mathbf{E}\bigl[\log \overline{m}(X)\bigr]$$

where $\overline{m}(X) = \mathbf{E}\bigl[m(X, Y)|X\bigr]$. Thus letting first $n \to \infty$ in (24) and then $k \to \infty$, we get

$$\limsup_{n\to\infty}\limsup_{D\to 0}\left(h(X|Y, Z_D, V_n) - \frac{1}{2}\log\bigl(2\pi e D\bigr)\right) \le \frac{1}{2}\mathbf{E}\bigl[\log \overline{m}(X)\bigr].$$

Combining this with (19) and (20) proves the theorem. $\square$

4 Conclusion

It should be noted that the asymptotic formulas of Theorem 1 and Theorem 2 suggest a companding realization of source coding with side information. Consider first the conditional rate-distortion problem. Let $c(t,y) = \int_0^t m(x,y)^{1/2}\,dx$, where the integral is defined using the convention $\int_a^b = -\int_b^a$ if $a > b$. Assuming $m(x,y) > 0$ for all x and y, $c(\cdot,y)$ has an inverse $c^{-1}(\cdot,y)$ for fixed y. Using $c(\cdot,y)$ and $c^{-1}(\cdot,y)$ as a side information dependent compressor-expander pair, we can construct a coding scheme where first each X_i is passed through the compander $c(\cdot,Y_i)$, then the output is quantized with a lattice vector quantizer, and then the quantizer output is entropy coded conditioned on $\{Y_i\}$. Since the decoder knows $\{Y_i\}$, it can decode the entropy code and then apply the expander $c^{-1}(\cdot,Y_i)$. For large enough lattice quantizer dimensions and small distortion, the rate of this scheme will approach the conditional rate-distortion function.

Similarly, for the Wyner-Ziv problem we can define the compander by $c(t) = \int_0^t \overline{m}(x)^{1/2}\,dx$, where $\overline{m}(x) = \mathbf{E}\big[m(X,Y)|X=x\big]$. Let $c(\cdot)$ and its inverse $c^{-1}(\cdot)$ be the compressor-expander pair, which we combine with lattice quantization and Slepian-Wolf coding (instead of entropy coding conditioned on Y). The performance of this scheme will be arbitrarily close to the upper bound of Theorem 2. Under the conditions for asymptotic tightness in Theorem 2, the coding rate will asymptotically achieve $R^{\mathrm{WZ}}(D)$ for small distortions. A rigorous proof of both of these claims can be given using the techniques developed in [14].

Appendix A

Lemma 1 *Let N_D denote a zero mean Gaussian random variable with variance D which is independent of (X,Y). If $\mathbf{E}[X^2]$, $h(X)$, and $\mathbf{E}[m(X,Y)^{-1}]$, are finite, then*

$$\limsup_{D\to 0} h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|Y\Big) \le h(X|Y). \tag{A.1}$$

Proof. By [8, Appendix B], if $n : \mathbb{R} \to [0,\infty)$ is such that $\mathbf{E}[n(X)^{-1}] < \infty$, then $\limsup_{D\to 0} h\big(X + (N_D/n(X))\big) \le h(X)$. Since $h(X|Y=y)$ is finite a.e. $[P_Y]$, this implies that

$$\limsup_{D\to 0} h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|Y=y\Big) \le h(X|Y=y) \quad \text{a.e. } [P_Y].$$

Assume now that there exists $D_0 > 0$ and a measurable g such that $\mathbf{E}[g(Y)] < \infty$ and for all positive $D \le D_0$,

$$h\Big(X + \frac{N_D}{\sqrt{m(X,Y)}}\Big|Y=y\Big) \le g(y) \quad \text{a.e. } [P_Y].$$

Then Fatou's lemma [15] proves (A.1) since

$$\begin{aligned}\limsup_{D\to 0} h\Big(X+\frac{N_D}{\sqrt{m(X,Y)}}\Big|Y\Big) &= \limsup_{D\to 0}\int h\Big(X+\frac{N_D}{\sqrt{m(X,Y)}}\Big|Y=y\Big)\,P_Y(dy)\\ &\le \int\Big[\limsup_{D\to 0} h\Big(X+\frac{N_D}{\sqrt{m(X,Y)}}\Big|Y=y\Big)\Big]\,P_Y(dy)\\ &= h(X|Y).\end{aligned}$$

To demonstrate the existence of an appropriate g, let $\hat{X}_D = X + m(X,Y)^{-1/2}N_D$ and notice that

$$h(\hat{X}_D|Y=y) \le \frac{1}{2}\log\big(2\pi e\mathbf{E}[\hat{X}_D^2|Y=y]\big) \quad \text{a.e. } [P_Y].$$

By the independence of (X,Y) and N_D,

$$\mathbf{E}[\hat{X}_D^2|Y=y] = \mathbf{E}[X^2|Y=y] + D\mathbf{E}[m(X,Y)^{-1}|Y=y] < \infty \quad \text{a.e. } [P_Y].$$

Define $g_D(y) = \frac{1}{2}\log\big(2\pi e\mathbf{E}[\hat{X}_D^2|Y=y]\big)$. Then $h(\hat{X}_D|Y=y) \le g_{D_0}(y)$ a.e. $[P_Y]$ for $D \le D_0$. On the other hand,

$$\mathbf{E}[g_{D_0}(Y)] \le \frac{1}{2}\log\big(2\pi e\mathbf{E}[\hat{X}_{D_0}^2]\big) < \infty$$

by Jensen's inequality. $\square$

Appendix B

Proof of the converse part of Theorem 1 To keep the proof relatively simple, we will assume that $d(x,y,\hat{x}) = m(x,y)(x-\hat{x})^2$. The extension of the derivation to the general case when $d(x,y,\hat{x}) = m(x,y)(x-\hat{x})^2 + o(|x-\hat{x}|^2)$ is straightforward using the proof technique of [8, Proposition 2].

Let $\{\hat{X}_D : D > 0\}$ be an arbitrary collection of random variables each jointly distributed with (X,Y) such that

$$\mathbf{E}[d(X,Y,\hat{X}_D)] \le D \quad \text{and} \quad I(X;\hat{X}_D|Y) < \infty.$$

Our proof technique will be similar to that of [8, Proposition 2]. The new element in the proof is the introduction of the auxiliary random variable $V_a = Q_a(X)$, where Q_a is an infinite level uniform quantizer of step size $a > 0$. This will substantially ease the technical complications introduced by the conditioning on Y. In fact, using the upcoming derivation, one could give a simpler proof of [8, Proposition 2]. The role of the auxiliary random variable V_a is similar to that played by V_n in the proof of Theorem 2. The basic idea is that for small values of a, for all x such that $Q_a(X) = v$, we have $m(x,y)(x-\hat{x})^2 \approx m(v,y)(x-\hat{x})^2$. On the other hand, $m(v,y)$ is constant in x in each quantization cell. Therefore, $m(V_a,Y)(X-\hat{X}_D)^2$ becomes quadratic in X when conditioned on the event $(V_a = v, Y = y)$, and thus we can apply the Shannon lower bound "locally". The desired lower bound will follow through Jensen's inequality after averaging these local bounds.

For any $a > 0$ let Q_a be the uniform quantizer with codepoints $\{0, \pm a, \pm 2a, \dots\}$. Define the discrete random variable V_a by

$$V_a = Q_a(X).$$

Then by the chain rule and the finiteness of $I(X;V_a)$,

$$\begin{aligned} I(X;\hat{X}_D|Y) &= I(X;V_a|Y) + I(X;\hat{X}_D|V_a,Y) - I(X;V_a|Y,\hat{X}_D) \\ &= h(X|Y) - h(X|Y,\hat{X}_D,V_a) - I(X;V_a|Y,\hat{X}_D). \end{aligned} \tag{B.1}$$

Thus to prove the claim of the theorem it suffices to show that

$$\limsup_{a\to 0}\limsup_{D\to 0}\left[h(X|Y,\hat{X}_D,V_a) - \frac{1}{2}\log(2\pi e D)\right] \le -\frac{1}{2}\mathbf{E}[\log m(X,Y)] \tag{B.2}$$

and that

$$\lim_{D\to 0} I(X;Q_a(X)|Y,\hat{X}_D) = 0. \tag{B.3}$$

It can be easily verified that conditions (c) and (b) together with the fact that $\mathbf{E}[d(X,Y,\hat{X}_D)] \le D$ imply that $\hat{X}_D \to X$ in probability as $D \to 0$. Now letting $V = Q_a(X)$, $Z_D = \hat{X}_D$, and $f(Y,Z_D) = Z_D = \hat{X}_D$ in Lemma 2 in Appendix C, the limit (B.3) holds. The rest of the proof is devoted to establishing (B.2).

Let $A \subset \mathbb{R}^2$ be an open set such that $m(x,y)$ is continuous on A and the complement of A has zero Lebesgue measure (such A exists by condition (d)). If U is defined by

$$U = A \cap \{(x,y) : m(x,y) > 0\},$$

then U is open and $\mathbf{P}((X,Y) \in U) = 1$. By a standard result of measure theory (see, e.g., [16]), there exist compact sets $C_k \subset U$, for $k = 1, 2, \dots$ such that

$$\lim_{k\to\infty} \mathbf{P}((X,Y) \in C_k) = 1.$$

Define the binary random variable B_k by setting $B_k = 0$ if $(X,Y) \in C_k$ and $B_k = 1$ if $(X,Y) \notin C_k$. Using B_k, we can upper bound $h(X|Y,\hat{X}_D,V_a)$ as

$$h(X|Y,\hat{X}_D,V_a) = h(X|Y,\hat{X}_D,V_a,B_k) + I(X;B_k|Y,\hat{X}_D,V_a)$$

$$
\begin{aligned}
&\leq h(X|Y,\hat{X}_D,V_a,B_k=1)\mathbf{P}(B_k=1)+h(X|Y,\hat{X}_D,V_a,B_k=0)\mathbf{P}(B_k=0) \\
&\quad + H(B_k), \qquad \text{(B.4)}
\end{aligned}
$$

where $H(B_k)$ denotes the Shannon entropy of B_k.

If the random variable Z has a density and $\mathbf{E}[Z^2]<\infty$, then $h(Z)\leq \frac{1}{2}\log(2\pi e\mathbf{E}[Z^2])$. Since conditioning reduces differential entropy, this implies

$$
\begin{aligned}
h(X|Y,\hat{X}_D,V_a,B_k=1) &\leq h(X|B_k=1) \\
&\leq \frac{1}{2}\log(2\pi e\mathbf{E}[X^2|B_k=1\,]) \\
&\leq \frac{1}{2}\log\left(2\pi e\frac{\mathbf{E}[X^2]}{\mathbf{P}(B_k=1)}\right). \qquad \text{(B.5)}
\end{aligned}
$$

In the rest of the proof the appropriate upper bound on $h(X|Y,\hat{X}_D,V_a,B_k=0)$ is developed. For each $y\in\mathbb{R}$, let $C_k(y)$ denote the section of C_k at y, defined by $C_k(y)=\{x:(x,y)\in C_k\}$. Then $C_k(y)$ is compact for all y. Let $S_x^{(a)}$ denote the cell of Q_a in which x falls. For $(x,y)\in C_k$ define the function $m_{k,a}(x,y)$ by

$$
m_{k,a}(x,y)=\max_{x'\in S_x^{(a)}\cap C_k(y)} m(x',y). \qquad \text{(B.6)}
$$

Note that $m_{k,a}(x,y)$ is positive for all $(x,y)\in C_k$. Thus we have

$$
\begin{aligned}
h(X|\hat{X}_D,Y,V_a,B_k=0) &= h(X-\hat{X}_D|\hat{X}_D,Y,V_a,B_k=0) \\
&\leq h(X-\hat{X}_D|Y,V_a,B_k=0) \\
&= h\left(\sqrt{m_{k,a}(X,Y)}(X-\hat{X}_D)\big|Y,V_a,B_k=0\right) \\
&\quad -\frac{1}{2}\mathbf{E}[\log(m_{k,a}(X,Y))|B_k=0] \qquad \text{(B.7)} \\
&\leq h\left(\sqrt{m_{k,a}(X,Y)}(X-\hat{X}_D)\big|B_k=0\right) \\
&\quad -\frac{1}{2}\mathbf{E}[\log(m_{k,a}(X,Y))|B_k=0] \\
&\leq \frac{1}{2}\log\Big(2\pi e\mathbf{E}[m_{k,a}(X,Y)(X-\hat{X}_D)^2|B_k=0]\Big) \\
&\quad -\frac{1}{2}\mathbf{E}[\log(m_{k,a}(X,Y))|B_k=0]. \qquad \text{(B.8)}
\end{aligned}
$$

In (B.7) we used the formula $h(\alpha Z)=h(Z)+\log\alpha$, valid for any $\alpha>0$ and random variable Z with a finite differential entropy, combined with the fact that if $(X,Y)\in C_k$, then $m_{k,a}(X,Y)$ is a function of (V_a,Y). In (B.8) we used the same bound as in (B.5).

Since $m(x,y)$ is uniformly continuous and positive on C_k, we have

$$\left|\frac{m_{k,a}(x,y)}{m(x,y)}-1\right| \;\leq\; \frac{\sup_{x',x''\in C_k(y):|x'-x''|\leq a}|m_{k,a}(x',y)-m(x'',y)|}{\inf_{(x,y)\in C_k} m(x,y)} = \epsilon(k,a)$$

where $\lim_{a\to 0}\epsilon(k,a)=0$ for all k. Thus for all $(x,y)\in C_k$,

$$1-\epsilon(k,a)\leq \frac{m_{k,a}(x,y)}{m(x,y)}\leq 1+\epsilon(k,a)$$

which combined with (B.8) and the inequality

$$\mathbf{E}[m(X,Y)(X-\hat{X}_D)^2|B_k=0]\leq \frac{D}{\mathbf{P}(B_k=0)} \tag{B.9}$$

yields

$$\begin{aligned}
h(X|\hat{X}_D,Y,V_a,B_k=0) \;&\leq\; \frac{1}{2}\log\Big(2\pi e\mathbf{E}[m(X,Y)(X-\hat{X}_D)^2|B_k=0]\Big)\\
&\quad -\frac{1}{2}\mathbf{E}[\log(m(X,Y))|B_k=0]+\frac{1}{2}\log\left(\frac{1+\epsilon(k,a)}{1-\epsilon(k,a)}\right)\\
&\leq\; \frac{1}{2}\log\left(2\pi eD\right)-\frac{1}{2}\mathbf{E}[\log(m(X,Y))|B_k=0]\\
&\quad +\frac{1}{2}\log\left(\frac{1+\epsilon(k,a)}{1-\epsilon(k,a)}\right)-\frac{1}{2}\log\mathbf{P}(B_k=0).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\limsup_{D\to 0}\left[h(X|\hat{X}_D,Y,V_a,B_k=0)\mathbf{P}(B=0)-\frac{1}{2}\log(2\pi eD)\right]\\
&\quad\leq\; \limsup_{D\to 0}\left[\left(h(X|\hat{X}_D,Y,V_a,B=0)-\frac{1}{2}\log(2\pi eD)\right)\mathbf{P}(B_k=0)\right]\\
&\quad\leq\; -\frac{1}{2}\mathbf{E}[\log(m(X,Y))|B_k=0]\mathbf{P}(B_k=0)-\frac{1}{2}\mathbf{P}(B_k=0)\log\mathbf{P}(B_k=0)\\
&\qquad +\frac{1}{2}\log\left(\frac{1+\epsilon(k,a)}{1-\epsilon(k,a)}\right)\mathbf{P}(B_k=0).
\end{aligned} \tag{B.10}$$

Combining (B.4),(B.5), and (B.10) we obtain

$$\begin{aligned}
&\limsup_{D\to 0}\left[\,h(X|\hat{X}_D,Y,V_a)-\frac{1}{2}\log(2\pi eD)\,\right]\\
&\qquad\leq\; -\frac{1}{2}\mathbf{E}[\log(m(X,Y))|B_k=0]\mathbf{P}(B_k=0)\\
&\qquad\quad +\frac{1}{2}\mathbf{P}(B_k=1)\log\left(2\pi e\mathbf{E}[X^2]\right)+\frac{3}{2}H(B_k)
\end{aligned}$$

$$+\frac{1}{2}\mathbf{P}(B_k=0)\log\left(\frac{1+\epsilon(k,a)}{1-\epsilon(k,a)}\right). \qquad \text{(B.11)}$$

Since $\epsilon(k,a)\to 0$ as $a\to 0$ for all k and since $\mathbf{P}(B_k=1)\to 1$ as $k\to\infty$, letting $a\to 0$ first and then $k\to\infty$ gives

$$\limsup_{a\to 0}\limsup_{D\to 0}\left[h(X|\hat{X}_D,Y,V_a)-\frac{1}{2}\log(2\pi eD)\right]\le -\frac{1}{2}\mathbf{E}[\log m(X,Y)]$$

which was to be proved. □

Appendix C

Lemma 2 *Assume that $I(X;V)<\infty$ and for any $D>0$, $Y\leftrightarrow V\leftrightarrow X\leftrightarrow Z_D$ forms a Markov chain. Suppose further that there is a measurable function $f(Y,Z_D)$ (which may depend on D) such that $f(Y,Z_D)\to X$ in probability as $D\to 0$. Then*

$$\lim_{D\to 0} I(X;V|Y,Z_D)=0.$$

Proof. Use the chain rule twice to obtain

$$\begin{aligned} I(X;V|Y,Z_D) &= I(X,Z_D;V|Y)-I(Z_D;V|Y)\\ &= I(X;V|Y)+I(Z_D;V|Y,X)-I(Z_D;V|Y)\\ &= I(X;V|Y)-I(Z_D;V|Y) \qquad \text{(C.1)}\end{aligned}$$

where all quantities are finite since $I(X;V)<\infty$, and the third equality holds because $I(Z_D;V|Y,X)=0$ by the Markov chain condition $Y\leftrightarrow\leftrightarrow V\leftrightarrow X\leftrightarrow Z_D$. Since

$$I(Z_D;V|Y)=I(Y,Z_D;V|Y)\ge I(f(Y,Z_D);V|Y)$$

we have

$$\liminf_{D\to 0} I(Z_D;V|Y)\ge \liminf_{D\to 0} I(f(Y,Z_D);V|Y). \qquad \text{(C.2)}$$

Now the the lower semicontinuity of the mutual information [17] and the condition that $f(Y,Z_D)\to X$ in probability imply that

$$\liminf_{D\to 0} I(f(Y,Z_D);V|Y=y)\ge I(X;V|Y=y) \quad \text{a.e. } [P_Y]$$

and therefore by Fatou's lemma [15] we have

$$\liminf_{D\to 0} I(f(Y,Z_D);V|Y)\ge I(X;V|Y).$$

The lemma now follows by (C.1) and (C.2). □

Acknowledgement

The authors would like to thank the reviewers for their comments.

References

[1] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, New Jersey: Prentice–Hall, 1971.

[2] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.

[3] R. M. Gray, “A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, 1973.

[4] 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.

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

[6] S. Shamai (Shitz), S. Verdu, and R. Zamir, “Systematic lossy source/channel coding,” *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 564–579, Mar. 1998.

[7] H. Viswanathan and T. Berger, “Sequential coding of correlated sources.” in *Proc. IEEE Int. Symp. on Information Theory*, p. 272., Ulm, Germany, June, 1997.

[8] T. Linder and R. Zamir, “High-resolution source coding for non-difference distortion measures: the rate distortion function,” *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 533–547, Mar. 1999.

[9] W. R. Gardner and B. D. Rao, “Theoretical analysis of the high-rate vector quantization of LPC parameters,” *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 367–381, Sep. 1995.

[10] J. Li, N. Chaddha, and R. M. Gray, “Asymptotic performance of vector quantizers with the perceptual distortion measure.” preprint, also in *Proc. IEEE Int. Symp. on Information Theory*, p. 55., Ulm, Germany, June, 1997.

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

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

[13] Y. S. Chow and H. Teicher, *Probability Theory, Independence, Interchangeability, Martingales*. New York: Springer Verlag, 1988.

[14] T. Linder, R. Zamir, and K. Zeger, "High-resolution source coding for non-difference distortion measures: multidimensional companding," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 548–561, Mar. 1999.

[15] R. B. Ash, *Real analysis and probability*. New York: Academic Press, 1972.

[16] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 3rd ed., 1987.

[17] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.