

# Source Coding with Decoder Side Information

**Presenter:** Shivaprasad Kotagiri.

**Date:** 27 September 2006.

**Reading:** Elements of Information theory by Cover & Thomas [1, Secs 14.8 and 14.9] and Wyner-ziv paper [2]

## 1 Introduction

In this report, we study source coding problems with decoder side information (SI). Source coding problems with decoder side information are a special case of distributed source coding problems. We outline the relation between various distributed source coding problems in Table 1 with the help of Figure 1. For reference, we also include the standard single-source lossless and lossy compression problems in Table 1. We discuss source coding problems with decoder SI in which an encoder maps a random vector  $x_1^n$  from a random source  $x$  to  $w \in \mathcal{W} = \{1, 2, \dots, M\}$ , and a decoder, provided with  $w$  and the side information correlated with  $x^n$ , computes an estimate of  $x_1^n$  subject to a fidelity criterion<sup>1</sup>. In general, the correlated information that is available at the decoder is called as side information.

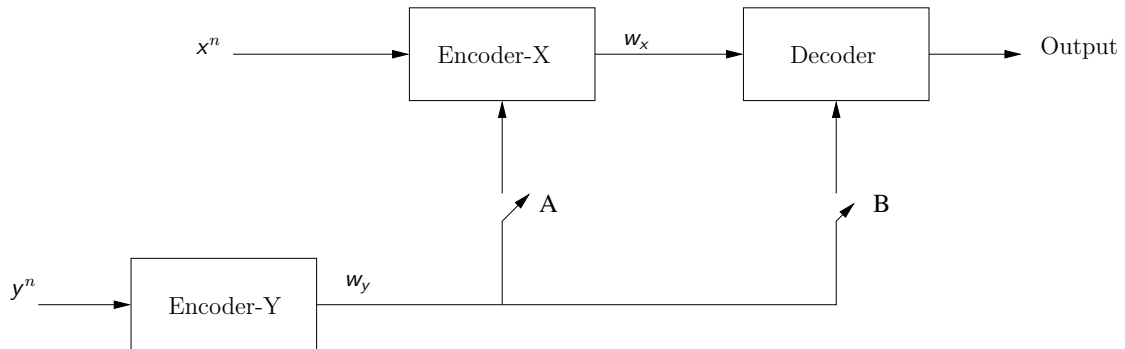


Figure 1: Block diagram of distributed source coding.

The basic idea of source coding with decoder SI is to encode the conditional uncertainty in  $x^n$  given the side information. In this report, we study the rate at which a source can be compressed with the side information at the decoder to achieve some specified distortion  $D_x$  under a given distortion measure  $d(\cdot, \cdot)$ . Let us denote the rate of codes used in Encoder-X and Encoder-Y as  $R_x$  and  $R_y$ , respectively. In this paper, we are interested in two source coding problems when switch B is closed in Figure 1. In these problems,  $y^n$  is correlated with  $x^n$ , the side information available to the decoder

---

<sup>1</sup>In this report, we use the following notation:  $x$ -random variable,  $x$ -sample value of rv  $x$ ,  $x^n = (x_1, x_2, \dots, x_n)$ - $n$  length random vector,  $x^n$ -sample vector of  $x^n$ ,  $\mathcal{X}$ - the alphabet set from which  $x$  takes values, In general, we omit the subscripts of probability distribution functions (pdfs), i.e.,  $p(x) = p_x(x)$

Type of coding	$R_y$	(A,B)	$(D_x, D_y)$	Dec. O/p	Reference
Lossless source coding (of $x^n$ )	N/A	(0,0)	(0, N/A)	$\hat{x}$	[1]
Lossy source coding (SC) (of $x^n$ )	N/A	(0,0)	$(\geq 0, N/A)$	$\hat{x}$	[1]
Slepian-Wolf coding	$\leq H(y)$	(0,1)	(0,0)	$(\hat{x}^n, \hat{y}^n)$	[1, 3]
Lossy multi-terminal source coding	$\leq H(y)$	(0,1)	$(\geq 0, \geq 0)$	$(\hat{x}^n, \hat{y}^n)$	[4]
Conditional Rate-Distortion	$\geq H(y)$	(1,1)	$(\infty, \infty)$	$\hat{x}^n$	[5]
Lossless SC with Lossy SI	$\leq H(y)$	(0,1)	$(0, \infty)$	$\hat{x}^n$	[1, 6]
Lossy SC with Lossless SI	$\geq H(y)$	(0,1)	$(\geq 0, \infty)$	$\hat{x}^n$	[1, 2]

Table 1: Comparison of several problems obtained via different switch configurations in Figure 1. Column 2 indicates whether lossy or lossless side information is available at the decoder. Column 3 indicates which switches are closed. If switch variable is 0, then it is open. Otherwise, it is closed. Column 4 indicates distortion constraints on the decoder estimates. Column 5 indicates whether the decoder is concerned with estimating only one source or both. Appropriate references are mentioned in Column 7.

is knowledge about  $y^n$  in the form of  $w_y$ . The decoder is not concerned with estimating  $y^n$ , but the knowledge about  $y^n$  allows it to reduce the rate  $R_x$ . If  $R_y \leq H(y)$ , the side information is incomplete knowledge of  $y^n$ ; this is the case in the lossless side information problem [1, Sec.14.8]. If  $R_y \geq H(y)$ , the side information is equivalent to a direct observation of  $y^n$ ; this is the case in lossy side information problem [1, Sec.14.9]. In this report, the elements of the random vector pair  $(x^n, y^n)$  are independent and identically distributed with the distribution function  $p(x, y)$ . We discuss the following two source coding techniques with decoder SI here.

- Lossless <sup>2</sup> source coding with lossy SI: In this case, lossless recovery of  $x^n$  is considered. Lossy version of  $y^n$  is available to the decoder in the form of SI. This problem was introduced by Wyner [6].
- Lossy <sup>3</sup> source coding with lossless SI (Wyner-Ziv coding): In this case, lossy recovery of  $x^n$  is considered and lossless version or direct observation of  $y^n$  is provided at the decoder (i.e.,  $R_y \geq H(y)$ ). This problem was introduced by Wyner and Ziv [2]. This problem builds upon the theoretical work of Slepian and Wolf [3] in lossless multi-terminal source coding. This problem for the case in which the sources are jointly Gaussian is studied in [7].

Methods for implementing Slepian-Wolf encoders and side-information encoders are discussed in [8, 9, 10, 11].

## 2 Lossless Source Coding with Lossy Side Information:

**Definition 1** A code  $(f_1, f_2, g)$  consists of encoding functions

$$f_1 : \mathcal{X}^n \rightarrow \mathcal{M}_1 = \{1, 2, \dots, M_1\},$$

$$f_2 : \mathcal{Y}^n \rightarrow \mathcal{M}_2 = \{1, 2, \dots, M_2\},$$

<sup>2</sup>“Lossless” means that  $\Pr[x^n \neq \hat{x}^n] \leq \epsilon$ , where  $\epsilon > 0$  is a very small number

<sup>3</sup>“Lossy” means that  $E[d(x^n, \hat{x}^n)] \leq D_x + \epsilon$ , where  $d(\cdot, \cdot)$  is non-negative bounded distortion measure, and  $D_x$  is distortion constraint.

and a decoding function

$$g : \mathcal{M}_1 \times \mathcal{M}_2 \rightarrow \mathcal{X}^n.$$

The above code is denoted as  $(n, M_1, M_2)$ .

**Definition 2** A rate pair  $(R_x, R_y)$  is said to be achievable if there exists a sequence of codes  $(n, M_1, M_2)$  such that  $P_e^n := \Pr[x^n \neq \hat{x}^n] \rightarrow 0$  as  $n \rightarrow \infty$ .

**Theorem 1** Given a pair of sources  $(x, y)$  jointly distributed according to  $p(x, y)$ , source  $x$  and source  $y$  are encoded at rate  $R_x$  and  $R_y$ , respectively. A rate pair  $(R_x, R_y)$  is achievable if and only if there exists an auxiliary random variable  $u$  such that

(i)  $R_x \geq H(x|u),$

(ii)  $R_y \geq I(y; u),$

(iii)  $x \leftrightarrow y \leftrightarrow u$  (Markov chain) and  $\sum_u p(u, x, y) = p(x, y).$

A converse proof of the above theorem is given in [1] and is not discussed in this report. Most of the steps in the converse use standard information inequalities except the subtle Markov relationship  $x_i \leftrightarrow (f(y^n), y^{i-1}) \leftrightarrow x^{i-1}$  for  $i \in \{1, 2, \dots, n\}$ . At the end, a time-sharing random variable is used to get single letter expressions. The encoding and the decoding methods used in the achievability proof are given below.

## 2.1 Encoding and Decoding

- **Encoder-X:** Randomly assign bin index  $i$  from the set  $\{1, 2, \dots, 2^{nR_x}\}$  to all typical  $x^n$  sequences<sup>4</sup> with uniform probability where  $2^{nR_x} \leq 2^{nH(x)}$ . After bin assignment to all typical sequences, the decoder is informed of bin assignment; this assignment is called a code. If the realized sequence  $x^n$  is not typical, the encoder declares an error. By the strong AEP, the probability of this error can be made arbitrarily small for sufficiently large  $n$ . Otherwise, the encoder transmits the index  $w_1$  of the bin that contains the realized sequence  $x^n$ . Since  $2^{nR_x} \leq 2^{nH(x)}$ , there are approximately  $2^{n(R_x - H(x))}$  sequences in each bin. So, the bin index is not sufficient to decode the realized sequence. The actual sequence can be resolved with the help of the side information at the decoder.
- **Encoder-Y:** Generate  $2^{nR_y}$   $u^n$  codewords i.i.d. according to  $p(u)$ . Label all these sequences with  $j \in \{1, 2, \dots, 2^{nR_y}\}$ . Codebook is revealed to the decoder. Encoder-Y sends  $w_2$  to the decoder if the realized sequence  $y^n$  is jointly typical with  $u^n(w_2)$ . If there is no  $u^n$  sequence jointly typical with the realized sequence, the encoder declares an error. This error can be made small for sufficiently large  $n$  if  $R_y \geq I(u; y)$  using the same arguments used in classical rate-distortion theory. If more than one codeword is jointly typical with the realized sequence  $y^n$ , sends the codeword with the smallest index.
- **Decoder:** Upon receiving  $(w_1, w_2)$ , the decoder estimates the sequence  $x^n$ .  $w_1$  gives the bin index of the realized sequence  $x^n$ . Since there are more than one sequence in each bin, the side-information codeword  $u^n(w_2)$  is used to determine the sequence from the bin  $w_1$ . The decoder looks for a sequence  $x^n$  in bin  $w_1$  that is jointly typical with the codeword  $u^n(w_2)$ . According to the Markov lemma, the realized sequence  $x^n$  in bin  $w_1$  is typical with the codeword  $u^n(w_2)$  with high probability for

---

<sup>4</sup>In this report, “typical” means “strongly typical”.

sufficiently large  $n$  because  $(\mathbf{x}^n, \mathbf{y}^n)$  is a jointly typical pair and  $(\mathbf{y}^n, \mathbf{u}^n)$  is also a jointly typical pair. Another possibility of error here is that  $\tilde{\mathbf{x}}^n$  in bin  $\mathcal{w}_1$ , which is not equal to the realized sequence  $\mathbf{x}^n$ , is typical with  $\mathbf{u}^n(\mathcal{w}_2)$ . The probability of this error can be made arbitrarily small for sufficiently large  $n$  if  $H(\mathbf{x}) - R_x \leq I(\mathbf{x}; \mathbf{u})$ .

## 2.2 Discussion

If  $R_y \geq H(\mathbf{y})$ , then essentially the direct realization of  $\mathbf{y}^n$  is available at the decoder. Then, the problem boils down to special case of the Slepian-Wolf problem. But, in this problem, the decoder is only concerned with the decoding of  $\mathbf{x}^n$ . In this case,  $R_x \geq H(\mathbf{x}|\mathbf{y})$  according to Slepian-Wolf results. Let us investigate what auxiliary random variable in our rate region provides the above Slepian-Wolf result. If  $\mathbf{u} = \mathbf{y}$ , then the constraint on  $R_x$  becomes  $H(\mathbf{x}|\mathbf{y})$  and the constraint on  $R_y$  becomes  $H(\mathbf{y})$ .

From the above, we know that the constraint on  $R_x$  is  $H(\mathbf{x}|\mathbf{y})$  ( $H(\mathbf{x}|\mathbf{u})$ ) when the direct (indirect) knowledge about  $\mathbf{y}^n$  available at the decoder. These two quantities are compared as follows

$$H(\mathbf{x}|\mathbf{u}) \stackrel{(a)}{\geq} H(\mathbf{x}|\mathbf{u}, \mathbf{y}) \stackrel{(b)}{=} H(\mathbf{x}|\mathbf{y}),$$

where (a) follows from the fact that conditioning reduces entropy, and (b) follows from the markov chain  $\mathbf{x} \rightarrow \mathbf{y} \rightarrow \mathbf{u}$ .

According to the support lemma,  $|\mathcal{U}| \leq |\mathcal{Y}| + 1$  is sufficient to compute all points in the region given by Theorem 1.

## 3 Lossy Source Coding with Side Information:

In this section, the direct observation of  $\mathbf{y}^n$  is available at the decoder in the form of SI.

**Definition 3** A code  $(f, g)$  consists of an encoding function

$$f_e : \mathcal{X}^n \rightarrow \mathcal{M}_1 = \{1, 2, \dots, M\},$$

and a decoding function

$$g : \mathcal{Y}^n \times \mathcal{M}_1 \rightarrow \hat{\mathcal{X}}^n.$$

The distortion associated with the above code is  $\Delta^{(n)} = Ed(\mathbf{x}^n, \hat{\mathbf{x}}^n)$ , where  $d(\mathbf{x}^n, \hat{\mathbf{x}}^n) = \frac{1}{n} \sum_{j=1}^n d(x_j, \hat{x}_j)$  and  $d(\cdot, \cdot)$  is non-negative, bounded distortion measure. Such a code is denoted as  $(n, M, \Delta^{(n)})$ .

**Definition 4** A rate pair  $(R, D)$  is said to be achievable if, for arbitrary  $\epsilon > 0$ , there exists a code  $(n, M, \Delta^{(n)})$  with  $M \leq 2^{n(R+\epsilon)}$  and  $\Delta^{(n)} \leq D + \epsilon$ . We define  $\mathcal{R}$  as the set of achievable  $(R, D)$  pairs and define

$$R^*(D) = \min_{(R, D) \in \mathcal{R}} R.$$

**Definition 5** For a given source  $\mathbf{x}$ , side information source  $\mathbf{y}$ , reconstruction alphabet  $\mathcal{X}$ , distortion measure  $d(\cdot, \cdot)$ , and distortion constraint  $D > 0$ , we define  $\mathcal{P}(D)$  as the collection of random variables  $(\mathbf{y}, \mathbf{x}, \mathbf{u}, \hat{\mathbf{x}})$  satisfying the following conditions

1.  $p(x, y, u) = p(x, y)p(u|x)$ .
2. There exists a function  $f : \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  such that  $Ed(x, \hat{x}) \leq D$ , where  $\hat{x} = f(y, u)$ .

For  $D > 0$ , let us define

$$R(D) := \inf_{(\mathbf{y}, \mathbf{x}, \mathbf{u}, \hat{\mathbf{x}}) \in \mathcal{P}(D)} [I(\mathbf{x}; \mathbf{u}) - I(\mathbf{y}; \mathbf{u})]$$

**Theorem 2** For  $D \geq 0$ ,  $R^*(D) = R(D)$ .

### 3.1 Remarks

- Since  $R(D)$  and  $R^*(D)$  are continuous at  $D = 0$ , it is sufficient to prove the above theorem for only  $D > 0$ .
- Since  $y \leftrightarrow x \leftrightarrow u$ , we can write  $R(D)$  as

$$R(D) = \inf_{(y, x, u, \hat{x}) \in \mathcal{P}(D)} I(x; u|y).$$

- Let us investigate what happens to  $R(D)$  as  $D \rightarrow 0$ . When  $D = 0$ , this problem is a special case of Slepian-Wolf problem. Under this condition, the Slepian-Wolf result shows that  $R \geq H(x|y)$ . Using the Markov chain  $y \leftrightarrow x \leftrightarrow u$  and Fano's inequality, it can be shown that  $R(0) = H(x|y)$ .
- $R^*(D) = R(D)$  is a continuous convex function of  $D$ . According to the support lemma,  $|\mathcal{U}| \leq |\mathcal{X}| + 1$  is sufficient to compute  $R(D)$ . Since the image of the function  $[I(x; u) - I(y; u)]$  over  $\mathcal{P}(\Delta)$  is compact, the infimum in the definition of  $R(D)$  can be replaced with the minimum.
- It can be shown that  $R^*(D) \geq R_{x|y}(D)$  for  $D \geq 0$ , where  $R_{x|y}(D)$  is the rate distortion function obtained when the side information is available at both encoder and decoder.

A converse proof for the above theorem is given in [2, 1] and is not discussed here. The encoding and decoding procedures used in the achievability proof are given below.

### 3.2 Encoding and Decoding

Fix  $n$  and  $(y, x, u, \hat{x}) \in \mathcal{P}(D)$ .

- **Encoder:** Generate  $2^{nR_0}$   $u^n$  typical codewords i.i.d. according to distribution  $p(u)$ . Label all these sequences with  $j \in \{1, 2, \dots, 2^{nR_0}\}$ . Randomly assign bin indices  $i \in \{1, 2, \dots, 2^{nR}\}$  to all codewords with uniform probability. These codewords and bin assignments are called codebook. These codebooks are revealed to the decoder. The encoder looks for a codeword  $u^n$  that is jointly typical with the realized typical source sequence  $x^n$ . If there is no such codeword, the encoder declares an error. The probability of this error can be made arbitrarily small for sufficiently large  $n$  if  $R_0 > I(u; x)$ . If one or more codewords are jointly typical with  $x^n$ , choose the codeword with the smallest index. Once such a codeword is found, the encoder transmits the bin index  $w$  of such a codeword to the decoder.
- **Decoder:** Decoder, upon receiving  $w$  and  $y^n$ , looks for a codeword  $u^n$  in bin  $w$  that is jointly typical with  $y^n$ . If an unique codeword in bin  $w$  is jointly typical with  $y^n$ , then the decoder computes the estimate  $\hat{x}^n = \{f(u_1, y_1), \dots, f(u_n, y_n)\}$ . Otherwise, the decoder declares an error. There are two possible sources of error with this decoding procedure. One possibility for error is that  $(u^n, y^n)$  is not jointly typical where  $u^n$  is typical with the realized sequence  $x^n$ . Since the random variable  $y, x$ , and  $u$  form a Markov chain, the probability of this error event can be made arbitrarily small for sufficiently large  $n$  according to Markov lemma. Another possibility for error is that  $\tilde{u}^n$ , which is not equal to the actual codeword  $u^n$  that is typical with the realized sequence  $x^n$ , is typical with  $y^n$ . The probability of this error event can be made arbitrarily small for sufficiently large  $n$  using strong typicality concepts if  $R_0 - R \leq I(y; u)$ . To avoid the encoding error, we need that  $R_0 \geq I(x; u)$ . These two conditions yield  $R \geq I(x; u) - I(y; u)$ .

- **Distortion Constraint:** Using typicality concepts, it can be proved that  $(x^n, \hat{x}^n)$  are jointly typical. Then the distortion between  $x^n$  and  $\hat{x}^n$  can be written as follows

$$\begin{aligned}
d(x^n, \hat{x}^n) &= \frac{1}{n} \sum_{j=1}^n d(x_j, \hat{x}_j) \\
&= \frac{1}{n} \sum_{(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}} N(x, \hat{x} | x^n, \hat{x}^n) d(x, \hat{x}) \\
&\stackrel{(a)}{=} \sum_{(x, \hat{x})} \left[ p(x, \hat{x}) + \frac{\epsilon}{|\mathcal{X}| |\hat{\mathcal{X}}|} \right] d(x, \hat{x}) \\
&\leq E d(x, \hat{x}) + \epsilon d_{max}, \\
&\stackrel{(c)}{\leq} D + \epsilon_0
\end{aligned}$$

where, (a) follows from the definition of strong typicality,  $d_{max}$  is the maximum distortion over  $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$ , and (c) follows from the fact that  $(y, x, u, \hat{x}) \in \mathcal{P}(D)$  and  $\epsilon_0 = d_{max} \epsilon$ .

### 3.3 The Gaussian Source for MSE Distortion

In this subsection, lossy source coding when the source and the side information are Gaussian sources is studied without details. The source  $x$  and the side information source  $y$  are related as  $y = x + z$ , where  $x$  and  $z$  are zero mean Gaussian random variables with variances  $\sigma_x^2$  and  $\sigma_z^2$ , respectively, and  $x$  and  $z$  are independent random variables. In this case, the distortion measure is squared error. Wyner [7] derived the rate distortion function for this problem. In this case, the rate distortion function is

$$R^*(D) = \begin{cases} \frac{1}{2} \log \frac{\lambda_{x|y}}{D} & 0 \leq D \leq \lambda_{x|y} \\ 0 & \lambda_{x|y} < D \end{cases}$$

where  $\lambda_{x|y} = E[x^2] - \frac{E[xy]^2}{E[y^2]}$  is the conditional variance of  $x$  given  $y$ . Interestingly, in this case,  $R^*(D) = R_{x|y}(D)$ , where  $R_{x|y}(D)$  is the rate distortion function when the SI is known at both the encoder and the decoder. In the Gaussian case, the rate-distortion function can be achieved via nested lattice quantizers [8, 12].

## References

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.
- [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. 22, no. 1, pp. 1–10, Jan. 1976.
- [3] D. S. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, July 1973.
- [4] R. Zamir and T. Berger, "Multiterminal Source Coding With High Resolution," *IEEE Trans. Inform. Theory*, vol. 45, pp. 106–117, Jan. 1999.
- [5] R. M. Gray, "Conditional Rate-Distortion Theory," *Technical Report, Stanford Electronics Laboratories*, no. 6502, 1972.

- [6] A. D. Wyner, "A Theorem on the Entropy of Certain Binary Sequences and Applications-II," *IEEE Trans. Inform. Theory*, vol. 19, pp. 772–777, November 1973.
- [7] —, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder-II: General Sources," *Probl. Contr. and Information Theory*, vol. 38, no. 1, pp. 60–80, 1978.
- [8] R. J. Barron, "Systematic Hybrid Analog/Digital Signal Coding," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.
- [9] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic Lossy Source/Channel Coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.
- [10] R. Zamir, "The Rate Loss in Wyner-Ziv Problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2073–2084, Nov. 1996.
- [11] A. D. Wyner, "Recent Results in the Shannon Theory," *IEEE Trans. Inform. Theory*, vol. 20, pp. 2–10, Jan. 1974.
- [12] R. Zamir and S. Shamai, "Nested Linear/Lattice Codes for Wyner-Ziv Encoding," in *Proc. IEEE Information Theory Workshop (ITW)*, 1998, pp. 92–93.