Side-information Scalable Source Coding

Chao Tian, *Member, IEEE*, Suhas N. Diggavi, *Member, IEEE*

Abstract

The problem of side-information scalable source coding is considered in this work, where the encoder constructs a progressive description, such that the receiver with high quality side information will be able to truncate the bitstream and reconstruct in the rate distortion sense, while the receiver with low quality side information will have to receive further data in order to decode. We provide inner and outer bounds for general discrete sources. The achievable region is shown to be tight for the case that either of the stages requires a lossless reconstruction. Furthermore we show that the gap between the achievable region and the outer bounds can be bounded by a constant when square error distortion measure is used. Complete characterization is provided for the important quadratic Gaussian source with jointly Gaussian side-informations, where the outer bounds match the achievable region. Partial result is provided for the doubly symmetric binary source with Hamming distortion when the worse side information is a constant, for which one of the outer bound is strictly tighter than the other one.

I. Introduction

Consider the following scenario where a server is to broadcast multimedia data to multiple users with different side informations, however the side informations are not available at the server. A user may have side information so strong that only minimal additional information is required from the server to satisfy a fidelity criterion, or a user may have barely any side information and expect the server to provide virtually everything to satisfy a (possibly different) fidelity criterion.

A naive strategy is to form a single description and broadcast it to all the users, who can decode only after receiving it completely regardless of the quality of their individual side informations. However, for the users with good-quality side information (who will be simply referred to as the good users), most of the information received is redundant, which introduces a delay caused simply by the existence of users with poor-quality side informations (referred to as the bad users) in the network. It is natural to ask whether an opportunistic method exists, i.e., whether it is possible to construct a two-layer description, such that the good users can decode with only the first layer, and the bad users receive both the first and the second layer to reconstruct. Moreover, it is of importance to investigate whether such a coding order introduces any performance loss. We call this coding strategy *side-information scalable* (SI-scalable) source coding, since the scalable coding direction is from the good users to the bad users. In this work, we consider mostly two-layer systems, except the quadratic Gaussian source for which the solution to an even more general problem is given.

This work is related to the successive refinement problem, where a source is to be encoded in a scalable manner to satisfy different distortion requirement at each individual stage. This problem was studied by Koshelev [1], and by Equitz and Cover [2]; Rimoldi [3] later provided a complete characterization of the rate-distortion region.

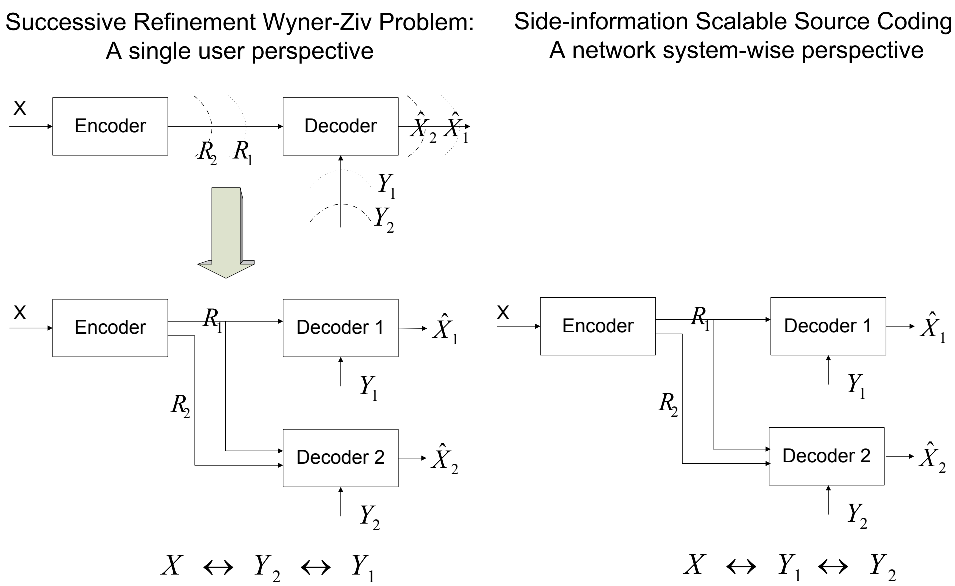

Fig. 1. The SR-WZ system vs. the SI-scalable system.

Another related problem is the rate-distortion for source coding with side information at the decoder [4], for which Wyner and Ziv provided conclusive result (now widely known as the Wyner-Ziv problem). Steinberg and Merhav [5] recently extended the successive refinement problem in the Wyner-Ziv setting (SR-WZ) for the special case of the two stage system, when the second stage side information Y_2 is better than that of the first stage Y_1, in the sense that $X \leftrightarrow Y_2 \leftrightarrow Y_1$ forms a Markov string. The extension to multistage systems with degraded side informations in such a direction was recently completed in [6]. Also relevant is the work by Heegard and Berger [7] (see also [8]), where the problem of source coding when side information may be present at the decoder was considered; the result was extended to the multistage case when the side informations are degraded. This is quite similar to the problem being considered here, however without the scalable coding requirement.

The current work differs from that in [5][6] in terms of mathematical formulation, as well as their possible applications. Roughly speaking, in the SI-scalable problem, the side information Y_2 at the later stage is worse than the side information Y_1 at the early stage, while in the SR-WZ problem, the order is reversed. In more mathematically precise terms, for the SI-scalable problem, the side informations are degraded as $X \leftrightarrow Y_1 \leftrightarrow Y_2$, in contrast to the SR-WZ problem where the reversed order is specified as $X \leftrightarrow Y_2 \leftrightarrow Y_1$. The two problems are also different in terms of their possible applications. The SR-WZ problem is more applicable for a single server-user pair, when the user is receiving side information through another channel, and at the same time receiving the description(s) from the server; for this scenario, two decoders can be extracted to provide a simplified model. On the other hand, the SI-scalable problem is more applicable when multiple users exist in the network, and the server wants to provide a scalable description, such that good users are not jeopardized unnecessarily (see Fig. 1). It is also worth pointing out that Heegard and Berger showed when the scalable coding requirement is removed, the optimal encoding by itself is in fact naturally progressive from the bad user to the good one; as such, the SI-scalable problem is expected to be more difficult than the SR-WZ problem, since the encoding order is reversed from the natural one. Despite the differences, both problems can be though as special cases of the general problem of scalable source coding with side information at the decoders with no structure imposed on the side informations. This general problem appears difficult; in fact, even without the scalable requirement, a complete solution was not found.

The problem is in fact well understood for the lossless case. The key difference from the lossy case is that the quality of the side informations can be naturally determined by the value of $H(X|Y)$. By the seminal work of Slepian and Wolf [9], $H(X|Y)$ is the minimum rate of encoding X losslessly with side information Y at the decoder, thus in a sense a larger $H(X|Y)$ corresponds to weaker side information. If $H(X|Y_1) < H(X|Y_2)$, then the rate $(R_1, R_2) = (H(X|Y_1), H(X|Y_2) - H(X|Y_1))$ is achievable, as noticed by Feder and Shulman [10]. Extending this observation and a coding scheme in [11], Draper [12] proposed a universal incremental Slepian-Wolf coding scheme when the distribution is unknown, which inspired Eckford and Yu [13] to design rateless Slepian-Wolf LDPC code. For the lossless case, there is no loss of optimality by using a scalable coding approach; an immediate question is to ask whether the same is true for the lossy case in terms of rate distortion, which we will show to be not so in general. In this rate-distortion setting, the order of goodness by the value of $H(X|Y)$ is not sufficient because of the presence of the distortion constraints. The Markov condition is therefore introduced as $X \leftrightarrow Y_1 \leftrightarrow Y_2$ for the SI-scalable coding problem. From this point of view, the SI-scalable problem is also applicable in the single user setting, when the source encoder does not know exactly which side information statistics the receiver has within a given set, i.e., a special case of universal rate distortion coding.

In this work, we formulate the problem of side information scalable source coding, and provide two inner bounds and two outer bounds for the rate-distortion region. One of the inner-bounds has the same distortion and rate expressions as one of the outer bound, and they differ only by a Markov string requirement. Though the inner and the outer bounds do not coincide in general, the inner bounds are indeed tight for the case when either the first stage or the second stage requires a lossless reconstruction. Furthermore, a conclusive result is given for the quadratic Gaussian source with (more than two) arbitrary correlated Gaussian side informations by using these bounds.

With this set of inner and outer bounds, the problem of *perfectly scalability* is investigated, defined as when both of the layers can achieve the corresponding Wyner-Ziv bounds; this is similar to the notion of (strictly) successive refinement in the SR-WZ problem [5][6][1]. Necessary and sufficient conditions are derived for general discrete memoryless sources to be perfectly scalable under a mild support condition. By using the tool of rate-loss introduced by Zamir [14], we further show that the gap between the inner bounds and the outer bounds are bounded by a constant when square error distortion measure is used, and thus the inner bounds are "nearly sufficient", in the sense as given in [15].

[1]In the rest of the paper, decoder one, respectively decoder two, will also be referred to as the first stage decoder, respectively second stage decoder, depending on the context.

In addition to the result for the Gaussian source, partial result is provided for the doubly symmetric binary source (DSBS) with Hamming distortion measure when the second stage does not have side information, for which the inner bounds and outer bounds coincide in certain distortion regimes. Furthermore, it is shown one of the outer bound can be strictly better than the other for this source.

This paper is organized as follows. In Section II we define the problem and establish the notation. In Section III, we provide inner and outer bounds to the rate-distortion regio.

and the gap between the inner bound and the outer bounds is investigated for the special case of mean squared distortion measure. The special cases are discussed in VI. We conclude the paper in Section VII.

II. Notation and Preliminaries

Let $\mathcal{X}$ be a finite set and let $\mathcal{X}^n$ be the set of all n-vectors with components in $\mathcal{X}$. Denote an arbitrary member of $\mathcal{X}^n$ as $x^n = (x_1, x_2, ..., x_n)$, or alternatively as $\boldsymbol{x}$. Upper case is used for random variables and vectors. A discrete memoryless source (DMS) $(\mathcal{X}, P_X)$ is an infinite sequence $\{X_i\}_{i=1}^{\infty}$ of independent copies of a random variable X in $\mathcal{X}$ with a generic distribution P_X with $P_X(x^n) = \prod_{i=1}^{n} P_X(x_i)$. Similarly, let $(\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, P_{XY_1Y_2})$ be a discrete memoryless three-source with generic distribution $P_{XY_1Y_2}$; the subscript will be drop when it is clear from the context as $P(X, Y_1, Y_2)$.

Let $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$ be finite reconstruction alphabets. Let $d_j : \mathcal{X} \times \hat{\mathcal{X}}_j \to [0, \infty)$, $j = 1, 2$ be two distortion measures. The single-letter distortion extension of d_j to a vector is defined as

$$d_j(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n} \sum_{i=1}^{n} d_j(x_i, \hat{x}_i), \quad \forall \boldsymbol{x} \in \mathcal{X}^n, \quad \hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_j^n, \qquad j = 1, 2. \tag{1}$$

***Definition** 1:* An (n, M_1, M_2, D_1, D_2) rate distortion (RD) SI-scalable code for source X with side information (Y_1, Y_2) consists of two encoding functions ϕ_i and two decoding functions ψ_i, $i = 1, 2$:

$$\phi_1 : \mathcal{X}^n \to I_{M_1}, \qquad \phi_2 : \mathcal{X}^n \to I_{M_2}, \tag{2}$$

where $I_k = \{1, 2, ..., k\}$ and

$$\psi_1 : I_{M_1} \times \mathcal{Y}_1^n \to \hat{\mathcal{X}}_1^n, \qquad \psi_2 : I_{M_1} \times I_{M_2} \times \mathcal{Y}_2^n \to \hat{\mathcal{X}}_2^n, \tag{3}$$

such that

$$\mathbb{E} d_1(X^n, \psi_1(\phi_1(X^n), Y_1^n)) \leq D_1, \tag{4}$$

$$\mathbb{E} d_2(X^n, \psi_2(\phi_1(X^n), \phi_2(X^n), Y_2^n)) \leq D_2, \tag{5}$$

where $\mathbb{E}$ is the expectation operation.

***Definition** 2:* A rate pair (R_1, R_2) is said to be (D_1, D_2)-achievable for SI-scalable encoding with side information (Y_1, Y_2), if for any $\epsilon > 0$ and sufficiently large n, there exist an $(n, M_1, M_2, D_1 + \epsilon, D_2 + \epsilon)$ RD SI-scalable code, such that $R_1 + \epsilon \geq \frac{1}{n} \log(M_1)$ and $R_2 + \epsilon \geq \frac{1}{n} \log(M_2)$.

Denote the collection of all the (D_1, D_2)-achievable rate pair (R_1, R_2) for SI-scalable encoding as $\mathcal{R}(D_1, D_2)$, and we seek to characterize this region when $X \leftrightarrow Y_1 \leftrightarrow Y_2$ forms a Markov string (see similar but reversed degradedness conditions in [5] and [7]). The Markov condition in effect specifies the *goodness* of the side informations.

The rate-distortion function for degraded side-informations was established in [7] for the non-scalable coding problem. In light of the discussion in Section I, it gives a lower bound on the sum-rate for any RD SI-scalable code. More precisely, in order to achieve distortion D_1 with side information Y_1, and achieve distortion D_2 with side information Y_2, when $X \leftrightarrow Y_1 \leftrightarrow Y_2$, the rate-distortion function is

$$R_{HB}(D_1, D_2) = \min_{p(D_1,D_2)} [I(X; W_2|Y_2) + I(X; W_1|W_2, Y_1)], \tag{6}$$

where $p(D_1, D_2)$ is the set of all random variable $(W_1, W_2) \in \mathcal{W}_1 \times \mathcal{W}_2$ jointly distributed with the generic random variables (X, Y_1, Y_2), such that the following conditions are satisfied[2]: (i) $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string; (ii) $\hat{X}_1 = f_1(W_1, Y_1)$ and $\hat{X}_2 = f_2(W_2, Y_2)$ satisfy the distortion constraints. Notice that the rate distortion function $R(D_1, D_2)$ given above suggests an encoding and decoding order from the bad user to the good user.

Wyner and Ziv [4] showed that under the following quite general assumption that the distortion measure is chosen in the set Γ_d defined as

$$\Gamma_d \triangleq \{d(\cdot, \cdot) : d(x, x) = 0, \text{and } d(x, \hat{x}) > 0 \text{ if } \hat{x} \neq x\}, \tag{7}$$

then the rate distortion function satisfies $R^*_{X|Y}(0) = H(X|Y)$, where $R^*_{X|Y}(D)$ is the well-known Wyner-Ziv rate distortion function with side information Y. If the same assumption is made on the distortion measure $d_1(\cdot, \cdot) \in \Gamma_d$, then it is easy to show using a similar argument as remark (3) in [4] that

$$R_{HB}(0, D_2) = \min_{p(D_2)} [I(X; W_2|Y_2) + H(X|W_2, Y_1)], \tag{8}$$

where $p(D_2)$ is the set of all random variable W_2 such that $W_2 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string, and $\hat{X}_2 = f_2(W_2, Y_2)$ satisfies the distortion constraint.

III. Inner and Outer Bounds

To provide intuition into the the SI-scalable problem, we first examine a simple Gaussian source under the mean squared error (MSE) distortion measure, and describe the coding schemes informally.

Let $X \sim \mathcal{N}(0, \sigma_x^2)$ and $Y_1 = Y = X + N$, where $N \sim \mathcal{N}(0, \sigma_N^2)$ is independent of X; Y_2 is simply a constant, i.e., no side information at the second decoder. $X \leftrightarrow Y_1 \leftrightarrow Y_2$ is indeed a Markov string. To avoid lengthy discussion on degenerate regimes, assume $\sigma_N^2 \approx \sigma_x^2$, and consider only the following extreme cases.

- $D_1 \gg D_2 \gg \sigma_x^2$: It is known binning with a Gaussian codebook, generated single-letter-wisely as $W_1 = X + Z_1$, where Z_1 is a zero-mean Gaussian random variable independent of X such that $D_1 = \mathbb{E}[X - \mathbb{E}(X|Y, W_1)]^2$, is optimal for Wyner-Ziv coding. This coding scheme can still be used for the first stage. In the second stage, by direct enumeration in the list of possible codewords in the particular bin specified in the first stage, the exact

[2] This form is slightly different than the one in [7] where f_1 was defined as $f_1(W_1, W_2, Y)$, but it is straightforwardly to verify that they are equivalent. The cardinality bound is also ignored, which is not essential here.

codeword can be recovered by decoder two, who does not have any side information. Since $D_1 \gg D_2 \gg \sigma_x^2$, W_1 alone is not sufficient to guarantee a distortion D_2, i.e., $D_2 \ll \mathbb{E}[X - \mathbb{E}(X|W_1)]^2$. Thus a successive refinement codebook, say using a Gaussian random variable W_2 such that $D_2 = \mathbb{E}[X - \mathbb{E}(X|W_1, W_2)]^2$, is needed. This leads to the achievable rates:

$$R_1 \geq I(X; W_1|Y), \quad R_1 + R_2 \geq I(X; W_1|Y) + I(W_1; Y) + I(X; W_2|W_1) = I(X; W_1, W_2). \tag{9}$$

- $D_2 \gg D_1 \gg \sigma_x^2$: If we choose $W_1 = X + Z_1$ such that $D_1 = \mathbb{E}[X - \mathbb{E}(X|Y, W_1)]^2$ and use the coding method in the previous case, then since $D_2 \gg D_1$, W_2 is more than sufficient to achieve distortion D_2, i.e., $D_2 \gg \mathbb{E}[X - \mathbb{E}(X|W_1)]^2$. The rate needed for the enumeration is $I(W_1; Y)$, and it is rather wasteful since W_2 is more than we need. To solve this problem, we construct an coarser description using random variable $W_2 = X + Z_1 + Z_2$, such that $D_2 = \mathbb{E}[X - \mathbb{E}(X|W_2)]^2$. The encoding process has three effective layers for the needed two stages: (i) the first layer uses Wyner-Ziv coding with codewords generated by P_{W_2} (ii) the second layer uses successive refinement Wyner-Ziv coding with $P_{W_1|W_2}$ (iii) the third layer enumerates the specific W_2 codeword within the first layer bin. Note that the first two layers form a SR-WZ scheme with identical side information Y at the decoder. For decoding, decoder one decodes the first two layers, while decoder two decodes the first and the third layer. By the Markov string $X \leftrightarrow W_1 \leftrightarrow W_2$, this scheme gives the following rates:

$$\begin{aligned} R_1 &\geq I(X; W_1, W_2|Y) = I(X; W_1|Y) \\ R_1 + R_2 &\geq I(X; W_1|Y) + I(W_2; Y) = I(X; W_2) + I(X; W_1|Y, W_2). \end{aligned} \tag{10}$$

It is seen in the above discussion the specific coding schemes depend on the distortion values, which is not desirable since this usually suggests difficulty in proving the converse. The two coding schemes can be unified into a single one by introducing an auxiliary random variable, as will be shown in the sequel, however, it appears the converse is indeed quite difficult to prove.

In the rest of this section, inner and outer bounds for $\mathcal{R}(D_1, D_2)$ are provided. The coding schemes for the above Gaussian example are naturally generalized to give the inner bounds. It is further shown when either one of the stages requires lossless reconstruction, the inner bounds are in fact tight.

A. Two inner bounds

Define the region $\mathcal{R}_{in}(D_1, D_2)$ to be the set of all rate pairs (R_1, R_2) for which there exist random variables (W_1, W_2, V) in finite alphabets $\mathcal{W}_1, \mathcal{W}_2, \mathcal{V}$ such that the following condition are satisfied.

1) $(W_1, W_2, V) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string.
2) There exist deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \rightarrow \hat{\mathcal{X}}_j$ such that

$$\mathbb{E} d_j(X, f_j(W_j, Y_j)) \leq D_j, \quad j = 1, 2. \tag{11}$$

3) The non-negative rate pairs satisfy:

$$R_1 \geq I(X; V, W_1|Y_1), \quad R_1 + R_2 \geq I(X; V, W_2|Y_2) + I(X; W_1|Y_1, V). \tag{12}$$

4) $W_1 \leftrightarrow (X, V) \leftrightarrow W_2$ is a Markov string.
5) The alphabets $\mathcal{V}$, $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy

$$|\mathcal{V}| \leq |\mathcal{X}| + 3, \quad |\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 1, \quad |\mathcal{W}_2| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 1. \tag{13}$$

The last two conditions can be removed without causing essential difference to the region $\mathcal{R}_{in}(D_1, D_2)$; with them removed, no specific structure is required on the joint distribution of (X, V, W_1, W_2). To see the last two conditions indeed do not cause loss of generality, apply the support lemma [11] as follows. For an arbitrary joint distribution of (X, V, W_1, W_2) satisfying the first three conditions, we first reduce the cardinality of $\mathcal{V}$. To preserve P_X and the two distortions and two mutual information values, $|\mathcal{X}| + 3$ letters are needed. With this reduced alphabet, observe that both the distortion and rate expressions depend only on the marginal of (X, V, W_1) and (X, V, W_2), respectively, hence requiring $W_1 \leftrightarrow (X, V) \leftrightarrow W_2$ being a Markov string does not cause any loss of generality. Next to reduce the cardinality of $\mathcal{W}_1$, it is seen $|\mathcal{X}||\mathcal{V}| - 1$ letters are needed to preserve the joint distribution of (X, V), one more is needed to preserve D_1 and another is needed to preserve $I(X; W_1|Y_1, V)$. Thus $|\mathcal{X}|(|\mathcal{X}| + 3) + 1$ letters suffice. Note that we do not need to preserve the value of D_2 (and the value of the other mutual information term) because of the aforementioned Markov string. A similar argument holds for $|\mathcal{W}_2|$.

The following theorem asserts that $\mathcal{R}_{in}(D_1, D_2)$ is an achievable region.

***Theorem** 1:* For any discrete memoryless stochastically source with side informations under the Markov condition $X \leftrightarrow Y_1 \leftrightarrow Y_2$,

$$\mathcal{R}(D_1, D_2) \supseteq \mathcal{R}_{in}(D_1, D_2).$$

This theorem is proved in Appendix II, and here we outline the coding scheme for this achievable region in an intuitive manner. The encoder first encodes $\boldsymbol{V}$ using a "coarse" binning, such that decoder one is able to decode it with side information $\boldsymbol{Y_1}$. A Wyner-Ziv successive refinement coding (with side information $\boldsymbol{Y_1}$) is then added conditioned on the codeword $\boldsymbol{V}$ also for decoder one using $\boldsymbol{W_1}$. The encoder then enumerates the binning of $\boldsymbol{V}$ up to a level such that $\boldsymbol{V}$ is decodable by decoder two using the weaker side information $\boldsymbol{Y_2}$. By doing so, decoder two is able to reduce the number of possible codewords in the (coarse) bin to a smaller number, which essentially forms a "finer" bin; with the weaker side information $\boldsymbol{Y_2}$, the $\boldsymbol{V}$ codeword is then decoded correctly with high probability. Another Wyner-Ziv successive refinement coding (with side information $\boldsymbol{Y_2}$) is finally added conditioned on the codeword $\boldsymbol{V}$ for decoder two using random variable $\boldsymbol{W_2}$.

As seen in the above argument, in order to reduce the number of possible $\boldsymbol{V}$ codewords from the first stage to the second stage, the key idea is to construct a nested binning structure as illustrated in Fig. 2. Each of the coarser bin contains the same number of finer bins; each finer bin holds certain number of codewords. They are constructed in such a way that given the specific coarser bin index, the first stage decoder can decode in it with the strong side information; at the second stage, additional bitstream is received by the decoder, which further specifies one of the finer bin in the coarser bin, such that the second stage decoder can decode in this finer bin using the weaker side information. If we assign each codeword to a finer bin independently, then its coarser bin index is also independent of that of the other codewords.

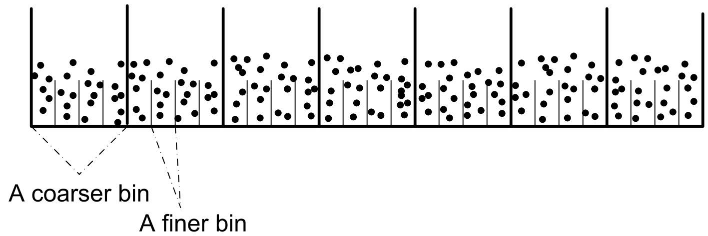

Fig. 2. An illustration of the codewords in the nested binning structure.

Specializing the region $\mathcal{R}_{in}(D_1, D_2)$ gives another inner bound. Let $\hat{\mathcal{R}}_{in}(D_1, D_2)$ be the set of all rate pairs (R_1, R_2) for which there exist random variables (W_1, W_2) in finite alphabets $\mathcal{W}_1, \mathcal{W}_2$ such that the following condition are satisfied.

1) $W_1 \leftrightarrow W_2 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ or $W_2 \leftrightarrow W_1 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string.
2) There exist deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \to \hat{\mathcal{X}}_j$ such that

$$\mathbb{E} d_j(X, f_j(W_j, Y_j)) \leq D_j, \quad j = 1, 2. \tag{14}$$

3) The non-negative rate pairs satisfy:

$$R_1 \geq I(X; W_1|Y_1), \quad R_1 + R_2 \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_2). \tag{15}$$

4) The alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy

$$|\mathcal{W}_1| \leq (|\mathcal{X}| + 3)(|\mathcal{X}|(|\mathcal{X}| + 3) + 1), \quad |\mathcal{W}_1| \leq (|\mathcal{X}| + 3)(|\mathcal{X}|(|\mathcal{X}| + 3) + 1). \tag{16}$$

***Corollary** 1:* For any discrete memoryless stochastically source with side informations under the Markov condition $X \leftrightarrow Y_1 \leftrightarrow Y_2$,

$$\mathcal{R}_{in}(D_1, D_2) \supseteq \hat{\mathcal{R}}_{in}(D_1, D_2).$$

The region $\hat{\mathcal{R}}_{in}(D_1, D_2)$ is particular interesting for the following reasons. Firstly, it can be explicitly matched back to the coding scheme for the simple Gaussian example. Secondly, it will be shown that one of the outer bounds has the same rate and distortion expressions as $\hat{\mathcal{R}}_{in}(D_1, D_2)$, only without the second Markov string requirement. We now prove this corollary.

Proof of Corollary 1

When $W_1 \leftrightarrow W_2 \leftrightarrow X$, let $V = W_1$. Then the rate expressions in Theorem 1 gives

$$R_1 \geq I(X; W_1|Y_1), \quad R_1 + R_2 \geq I(X; V, W_2|Y_2) + I(X; W_1|V, Y_1) = I(X; W_2|Y_2), \tag{17}$$

and therefore $\mathcal{R}_{in}(D_1, D_2) \supseteq \hat{\mathcal{R}}_{in}(D_1, D_2)$ for this case. When $W_2 \leftrightarrow W_1 \leftrightarrow X$, let $V = W_2$. Then the rate expressions in Theorem 1 gives

$$\begin{aligned} R_1 &\geq I(X; V, W_1|Y_1) = I(X; W_1|Y_1) \\ R_1 + R_2 &\geq I(X; V, W_2|Y_2) + I(X; W_1|V, Y_1) = I(X; W_2|Y_2) + I(X; W_1|W_2, Y_1), \end{aligned}$$

and therefore $\mathcal{R}_{in}(D_1, D_2) \supseteq \hat{\mathcal{R}}_{in}(D_1, D_2)$ for this case.

The cardinality bound here is larger than that in Theorem 1 because of the requirement to preserve the Markov conditions. ■

B. Two outer bounds

Define the following two regions, which will be shown to be two outer bounds. An obvious outer bound is given by the intersection of the Wyner-Ziv rate distortion function and the rate-distortion function for the problem considered by Heegard and Berger [7] with degraded side information $X \leftrightarrow Y_1 \leftrightarrow Y_2$

$$\mathcal{R}_{\cap}(D_1, D_2) = \{(R_1, R_2) : R_1 \geq R^*_{X|Y_1}(D_1), \quad R_1 + R_2 \geq R_{HB}(D_1, D_2)\}. \tag{18}$$

A tighter outer bound is now given as follows: define the region $\mathcal{R}_{out}(D_1, D_2)$ to be the set of all rate pairs (R_1, R_2) for which there exist random variables (W_1, W_2) in finite alphabets $\mathcal{W}_1, \mathcal{W}_2$ such that the following conditions are satisfied.

1) $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$.
2) There exist deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \rightarrow \hat{\mathcal{X}}_j$ such that

$$\mathbb{E} d_j(X, f_j(W_j, Y_j)) \leq D_j, \quad j = 1, 2. \tag{19}$$

3) $|\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 2$, $|\mathcal{W}_2| \leq |\mathcal{X}| + 3$.
4) The non-negative rate vectors satisfies:

$$R_1 \geq I(X; W_1|Y_1), \quad R_1 + R_2 \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_2). \tag{20}$$

The main result of this subsection is the following theorem.

***Theorem** 2:* For any discrete memoryless stochastically source with side informations under the Markov condition $X \leftrightarrow Y_1 \leftrightarrow Y_2$,

$$\mathcal{R}_{\cap}(D_1, D_2) \supseteq \mathcal{R}_{out}(D_1, D_2) \supseteq \mathcal{R}(D_1, D_2).$$

The first inclusion of $\mathcal{R}_{\cap}(D_1, D_2) \supseteq \mathcal{R}_{out}(D_1, D_2)$ is obvious, since $\mathcal{R}_{out}(D_1, D_2)$ takes the same form as $R^*_{X|Y_1}(D_1)$ and $R_{HB}(D_1, D_2)$ when the rates R_1 and $R_1 + R_2$ are considered individually. Thus we will focus on the latter inclusion, whose proof is given in Appendix III.

Note that the inner bound $\hat{\mathcal{R}}_{in}(D_1, D_2)$ and $\mathcal{R}_{out}(D_1, D_2)$ have the same rate and distortion expressions and they differ only by a Markov string requirement (ignoring the non-essential cardinality bounds). This is quite similar to the case of distributed lossy source coding problem, for which the Berger-Tung inner bound requires a long Markov string and the Berger-Tung outer bound requires only two short Markov strings [16], but their rate and distortion expressions are the same.

C. Lossless reconstruction at one of the decoders

Since decoder one has better quality side information, it is reasonable for it to require a higher quality reconstruction. Alternatively, from the point of view of universal coding, when the encoder does not know the quality of the side information, it might assume the better quality one exists at the decoder and aim to reconstruct with a higher quality, comparing with the case when the poorer quality side information is available. In the extreme case, decoder one might require a lossless reconstruction. In this subsection, we consider the setting where either decoder one or decoder two requires lossless reconstruction. More precisely, we have the following theorem.

***Theorem** 3:* If $D_1 = 0$ with $d_1(\cdot,\cdot) \in \Gamma_d$, or $D_2 = 0$ with $d_2(\cdot,\cdot) \in \Gamma_d$, then $\mathcal{R}(D_1, D_2) = \mathcal{R}_{in}(D_1, D_2)$.

Remark: Zero distortion under a distortion measure $d \in \Gamma_d$ can be interpreted as *lossless*, however, it is a weaker requirement than the traditional Shannon sense that the probability of error is arbitrarily small. Nevertheless, it is rather straightforward to specialize the coding scheme for these cases, and show that the same conclusion is true for lossless coding in the Shannon sense.

Proof of Theorem 3:

For $D_2 = 0$, let $W_1 = V$ and $W_2 = X$. The achievable rate vector is given by

$$R_1 \geq I(X; W_1|Y_1), \quad R_1 + R_2 \geq H(X|Y_2). \tag{21}$$

It is easily seen that this rate region is tight by the converse of Wyner-Ziv coding for rate R_1, and the converse of Slepian-Wolf coding (or more precisely, Wyner-Ziv rate distortion function with $d_2(\cdot,\cdot) \in \Gamma_d$ as given in [4]) for rate $R_1 + R_2$.

For $D_1 = 0$, let $W_1 = X$ and $V = W_2$. The achievable rate vector is given by

$$R_1 \geq H(X|Y_1), \quad R_1 + R_2 \geq I(X; W_2|Y_2) + H(X|Y_1, W_2). \tag{22}$$

It is seen that this rate region is tight by the converse of Slepian-Wolf coding for rate R_1, and by (8) of Heegard-Berger coding for rate $R_1 + R_2$. ■

The key difference from the general case when both stages are lossy is the elimination of the need to generate one of codebooks using an auxiliary random variables, which simplifies the matter tremendously. For example when $D_2 = 0$, since the first stage encoder guarantees that $\boldsymbol{w}_1$ and $\boldsymbol{x}$ are jointly typical, the second stage only needs to construct a codebook of $\boldsymbol{x}$ by binning the approximately $2^{H(X|W_1)}$ such $\boldsymbol{x}$ vector directly. Subsequently the second stage encoder does not search for a vector $\boldsymbol{x}^*$ to be jointly typical with both $\boldsymbol{w}_1$ and $\boldsymbol{x}$, but instead just sends the bin index of the observed source vector $\boldsymbol{x}$ directly. Alternatively, it can be understood as both the encoder and decoder at the second stage have access to a side information vector $\boldsymbol{w}_1$, and thus a conditional Slepian-Wolf coding with decoder side information Y_2 suffices.

IV. Perfect Scalability

Similarly as the notion of the (strictly) successive refinability defined in [5], we now introduce the notion of perfect scalability for the SI-scalable problem.

***Definition** 3:* A source X is said to be *perfect scalable* for distortion pair (D_1, D_2), with side informations under the Markov string $X \leftrightarrow Y_1 \leftrightarrow Y_2$, if

$$(R^*_{X|Y_1}(D_1), R^*_{X|Y_2}(D_2) - R^*_{X|Y_1}(D_1)) \in \mathcal{R}(D_1, D_2).$$

***Theorem** 4:* A source X with side informations under the Markov string $X \leftrightarrow Y_1 \leftrightarrow Y_2$, for which there exists $y_1 \in \mathcal{Y}_1$ such that $P_{XY_1}(x, y_1) > 0$ for each $x \in \mathcal{X}$, is perfect scalable for distortion pair (D_1, D_2) if and only if there exist random variables (W_1, W_2) and deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \to \hat{\mathcal{X}}_j$ such that the following conditions hold simultaneously:

1) $R^*_{X|Y_j}(D_j) = I(X; W_j|Y_j)$ and $\mathbb{E}d_j(X, f_j(W_1, Y_j)) \leq D_j$, for $j = 1, 2$.
2) $W_1 \leftrightarrow W_2 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ forms a Markov string.
3) The alphabet $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy $|\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 2$, and $|\mathcal{W}_2| \leq |\mathcal{X}|(|\mathcal{X}| + 3)$.

The Markov string condition is the most crucial one, and the substring $W_1 \leftrightarrow W_2 \leftrightarrow X$ is the same as one of the condition for successive refinability without side information [2][3]. The cardinality bounds on the alphabet can be removed without essential difference, but it is included here for easy comparison with the outer bound $\mathcal{R}_{out}(D_1, D_2)$, for which the alphabet size of W_2 is in fact smaller.

Proof of Theorem 4

Without loss of generality, assume $P_X(x) > 0$ for all $x \in \mathcal{X}$. By Theorem 2, if $(R^*_{X|Y_1}(D_1), R^*_{X|Y_2}(D_2) - R^*_{X|Y_1}(D_1)$ is achievable for (D_1, D_2), then there exist random variable W_1, W_2 in finite alphabet, whose sizes is bounded as $|\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}|+3)+2$ and $|\mathcal{W}_2| \leq |\mathcal{X}|+3$, and functions f_1, f_2 such that $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string, $\mathbb{E}d_j(X, f_j(W_j, Y_j)) \leq D_j$ for $j = 1, 2$ and

$$R^*_{X|Y_1}(D_1) \geq I(X; W_1|Y_1), \quad R^*_{X|Y_2}(D_2) \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_1). \tag{23}$$

It follows

$$R^*_{X|Y_2}(D_2) \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_2) \geq I(X; W_2|Y_2) \overset{(a)}{\geq} R^*_{X|Y_2}(D_2), \tag{24}$$

where (a) follows the converse of rate-distortion theorem for Wyner-Ziv coding. Since the leftmost and the rightmost quantities are the same, all the inequalities must be equalities in (24), and it follows $I(X; W_1|Y_1, W_2) = 0$. Similarly we have

$$R^*_{X|Y_1}(D_1) \geq I(X; W_1|Y_1) \geq R^*_{X|Y_1}(D_1), \tag{25}$$

thus (25) also holds with equality.

Notice that if $W_1 \leftrightarrow W_2 \leftrightarrow X$ is a Markov string, then we can use Corollary 1 to claim the sufficiency and complete the proof. However, this Markov condition is not true in general. This is where the support condition is needed.

For convenience, define the set $F(w_2)=\{x\in\mathcal{X}:P(x,w_2)>0\}$. By the Markov string $(W_1,W_2)\leftrightarrow X\leftrightarrow Y_1$, the joint distribution of (w_1,w_2,x,y_1) can be factorized as follows

$$P(w_1,w_2,x,y_1)=P(x,y_1)P(w_2|x)P(w_1|x,w_2). \tag{26}$$

Furthermore, $I(X;W_1|Y_1,W_2)=0$ implies the Markov string $X\leftrightarrow(W_2,Y_1)\leftrightarrow W_1$, and thus the joint distribution of (w_1,w_2,x,y_1) can also be factorized as follows

$$P(w_1,w_2,x,y_1)=P(x,y_1,w_2)p(w_1|y_1,w_2)\stackrel{(a)}{=}P(x,y_1)P(w_2|x)P(w_1|y_1,w_2), \tag{27}$$

where (a) follows by the Markov substring $W_2\leftrightarrow X\leftrightarrow Y_1\leftrightarrow Y_2$. Fix an arbitrary (w_1^*,w_2^*) pair, by the assumption that $P(x,y_1)>0$ for any $x\in\mathcal{X}$, we have

$$P(w_2^*|x)P(w_1^*|x,w_2^*)=P(w_2^*|x)P(w_1^*|y_1,w_2^*) \tag{28}$$

for any $x\in\mathcal{X}$. Thus for any $x\in F(w_2^*)$ such that $P(w_1|x,w_2^*)$ is well defined, we have

$$p(w_1^*|y_1^*,w_2^*)=p(w_1^*|x,w_2^*) \tag{29}$$

and it further implies

$$p(w_1^*|w_2^*)=\frac{\sum_x P(x,w_1^*,w_2^*)}{\sum_x P(x,w_2^*)}=\frac{\sum_x P(x,w_2^*)P(w_1^*|x,w_2^*)}{\sum_x P(x,w_2^*)}=p(w_1^*|x,w_2^*) \tag{30}$$

for any $x\in F(w_2^*)$. This indeed implies $W_1\leftrightarrow W_2\leftrightarrow X$ is a Markov string, which completes the proof. ■

The following two cases will be examined in Section VI: Gaussian source with MSE distortion measure and the doubly symmetric binary source with Hamming distortion measure. It will be shown that for some distortion pairs, both sources are perfectly scalable, while for others this is not possible.

V. A Near Sufficiency Result

By using the tool of rate loss introduced by Zamir [14], which was further developed in [15], [17], [18], it can be shown that when both the source and reconstruction alphabets are reals, and the distortion measure is MSE, the gap between the achievable region and the out bounds are bounded by a constant. To do this, we distinguish the two cases $D_1\ge D_2$ and $D_1\le D_2$. The source X is assumed to have finite variance σ_x^2. The result of this section is summarized in Fig. 3.

A. The case $D_1\ge D_2$

Construct two random variable $W_1'=X+N_1+N_2$ and $W_2'=X+N_2$, where N_1 and N_2 are zero mean independent Gaussian random variables, independent of everything else, with variance σ_1^2 and σ_2^2 such that $\sigma_1^2+\sigma_2^2=D_1$ and $\sigma_2^2=D_2$. Let U be optimal random variable to achieve the Wyner-Ziv rate at distortion D_1 given decoder

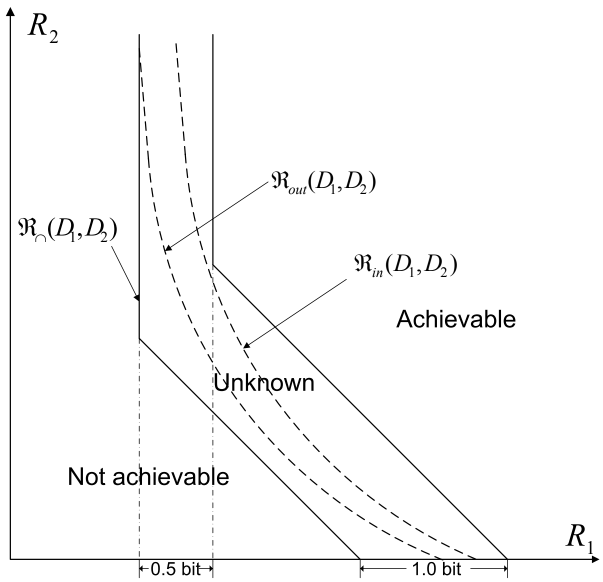

Fig. 3. An illustration of the gap between the inner bound and the outer bounds when MSE is the distortion measure. The two regions $\mathcal{R}_{in}(D_1, D_2)$ and $\mathcal{R}_{out}(D_1, D_2)$ are given in dashed lines, since it is unknown whether they are indeed the same.

side information Y_1. Then

$$
\begin{aligned}
& I(X; X+N_1+N_2|Y_1) - I(X; U|Y_1) \\
\stackrel{(a)}{=} \; & I(X; X+N_1+N_2|UY_1) - I(X; U|Y_1, X+N_1+N_2) \\
\leq \; & I(X; X+N_1+N_2|UY_1) \\
= \; & I(X-\hat{X}_1; X-\hat{X}_1+N_1+N_2|UY_1) \\
\leq \; & I(X-\hat{X}_1, U, Y_1; X-\hat{X}_1+N_1+N_2) \\
= \; & I(X-\hat{X}_1; X-\hat{X}_1+N_1+N_2) + I(U, Y_1; X-\hat{X}_1+N_1+N_2|X-\hat{X}_1) \\
= \; & I(X-\hat{X}_1; X-\hat{X}_1+N_1+N_2) \\
\stackrel{(b)}{\leq} \; & \frac{1}{2}\log_2 \frac{D_1+D_1}{D_1} = 0.5
\end{aligned}
\tag{31}
$$

where (a) is by applying chain rule to $I(X; X+N_1+N_2, U|Y_1)$ in two different ways; (b) is true because $\hat{X}_1$ is the decoding function given (U, Y_1), the distortion between X and $\hat{X}_1$ is bounded by D_1, and $X - \hat{X}_1$ is independent of (N_1, N_2).

Now we turn to bound the gap for the sum rate $R_1 + R_2$. Let W_1 and W_2 be the two random variables to achieve the rate distortion function $R_{HB}(D_1, D_2)$. First notice the following two identities due to the Markov string $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ and (N_1, N_2) are independent of (X, Y_1, Y_2)

$$I(X; W_2|Y_2) + I(X; W_1|W_2Y_1) = I(X; W_1W_2|Y_1) + I(Y_1; W_2|Y_2) \tag{32}$$

$$I(X; X + N_2|Y_2) = I(X; X + N_2|Y_1) + I(Y_1; X + N_2|Y_2). \tag{33}$$

Next we seek to bound the following difference

$$\begin{aligned} & I(X; X + N_2|Y_2) - I(X; W_2|Y_2) - I(X; W_1|W_2Y_1) \\ = \; & \{I(X; X + N_2|Y_1) - I(X; W_1W_2|Y_1)\} + \{I(Y_1; X + N_2|Y_2) - I(Y_1; W_2|Y_2)\}. \end{aligned} \tag{34}$$

To bound the first bracket, notice that

$$\begin{aligned} & I(X; X + N_2|Y_1) - I(X; W_1W_2|Y_1) \\ = \; & I(X; X + N_2|W_1W_2Y_1) - I(X; W_1W_2|Y_1, X + N_2) \\ \leq \; & I(X; X + N_2|W_1W_2Y_1) \\ \overset{(a)}{=} \; & I(X; X + N_2|W_1W_2Y_1Y_2) \\ = \; & I(X - \hat{X}_2; X - \hat{X}_2 + N_2|W_1W_2Y_1Y_2) \\ \leq \; & I(X - \hat{X}_2, W_1, W_2, Y_1, Y_2; X - \hat{X}_2 + N_2) \\ = \; & I(X - \hat{X}_2; X - \hat{X}_2 + N_2) + I(W_1, W_2, Y_1, Y_2; X - \hat{X}_2 + N_2|X - \hat{X}_2) \\ = \; & I(X - \hat{X}_2; X - \hat{X}_2 + N_2) \leq \frac{1}{2}\log\frac{D_2 + D_2}{D_2} = 0.5 \end{aligned} \tag{35}$$

where (a) is due to the Markov string $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$, $\hat{X}_2$ is the decoding function given (W_2, Y_2), and the other inequalities follow similar arguments as in Eqn. (31). To bound the second bracket, we write the following

$$\begin{aligned} I(Y_1; X + N_2|Y_2) - I(Y_1; W_2|Y_2) &= I(Y_1; X + N_2|W_2Y_2) - I(Y_1; W_2|Y_2, X + N_2) \\ &\leq I(Y_1; X + N_2|W_2Y_2) \\ &\leq I(XY_1; X + N_2|W_2Y_2) \\ &= I(X; X + N_2|W_2Y_2) \leq \frac{1}{2}\log\frac{D_2 + D_2}{D_2} = 0.5 \end{aligned} \tag{36}$$

By letting $W_1' = V' = X + N_1 + N_2$ and $W_2' = X + N_2$, it is obvious that the following rates are achievable for distortion (D_1, D_2) from Theorem 1

$$R_1 = I(X; X + N_1 + N_2|Y_1), \quad R_1 + R_2 = I(X; X + N_2|Y_2). \tag{37}$$

Thus we have shown that for $D_1 \geq D_2$, the gap between the outer bound $\mathcal{R}_\cap(D_1, D_2)$ and the inner bound $\mathcal{R}_{in}(D_1, D_2)$ is bounded. More precisely, the gap for R_1 is bounded by 0.5 bit, while the gap for the sum rate is bounded by 1.0 bit.

B. The case $D_1 \leq D_2$

Construct random variable $W_1' = X + N_1$ and $W_2' = X + N_1 + N_2$, where N_1 and N_2 are zero mean independent Gaussian random variables, independent of everything else, with variance σ_1^2 and σ_2^2 such that $\sigma_1^2 = D_1$ and $\sigma_1^2 + \sigma_2^2 = D_2$.

Apparently, the argument for the first stage R_1 still holds with minor changes. To bound the sum-rate gap, notice the following identity

$$\begin{aligned} & I(X; X + N_1 + N_2 | Y_2) + I(X; X + N_1 | Y_1, X + N_1 + N_2) \\ = \quad & I(X; X + N_1 + N_2 | Y_1) + I(Y_1; X + N_1 + N_2 | Y_2) + I(X; X + N_1 | Y_1, X + N_1 + N_2) && (38) \\ = \quad & I(Y_1; X + N_1 + N_2 | Y_2) + I(X; X + N_1 | Y_1). && (39) \end{aligned}$$

Next we seek to upper bound the following quantity

$$\begin{aligned} & I(X; X + N_1 + N_2 | Y_2) + I(X; X + N_1 | Y_1, X + N_1 + N_2) - I(X; W_2 | Y_2) - I(X; W_1 | W_2 Y_1) \\ & = \{I(X; X + N_1 | Y_1) - I(X; W_1 W_2 | Y_1)\} + \{I(Y_1; X + N_1 + N_2 | Y_2) - I(Y_1; W_2 | Y_2)\}. && (40) \end{aligned}$$

For the first bracket, we have

$$\begin{aligned} & I(X; X + N_1 | Y_1) - I(X; W_1 W_2 | Y_1) \\ = \quad & I(X; X + N_1 | W_1 W_2 Y_1) - I(X; W_1 W_2 | Y_1, X + N_1) \\ \leq \quad & I(X; X + N_1 | W_1 W_2 Y_1) \\ = \quad & I(X - \hat{X}_1; X - \hat{X}_1 + N_2 | W_1 W_2 Y_1) \\ \leq \quad & I(X - \hat{X}_1, W_1, W_2, Y_1; X - \hat{X}_1 + N_2) \\ = \quad & I(X - \hat{X}_1; X - \hat{X}_1 + N_1) + I(W_1, W_2, Y_1; X - \hat{X}_1 + N_1 | X - \hat{X}_1) \\ = \quad & I(X - \hat{X}_1; X - \hat{X}_1 + N_1) \leq \frac{1}{2} \log \frac{D_1 + D_1}{D_1} = 0.5, && (41) \end{aligned}$$

where $\hat{X}_1$ is the decoding function given (W_1, Y_1). For the second bracket, following a similar approach as (36), we have

$$\begin{aligned} I(Y_1; X + N_1 + N_2 | Y_2) - I(Y_1; W_2 | Y_2) \quad & \leq \quad I(X; X + N_1 + N_2 | W_2 Y_2) && (42) \\ & \leq \quad I(X - \hat{X}_2, W_2, Y_2; X - \hat{X}_2 + N_1 + N_2) && (43) \\ & = \quad I(X - \hat{X}_2; X - \hat{X}_2 + N_1 + N_2) \leq 0.5 && (44) \end{aligned}$$

By letting $V' = W_2' = X + N_1 + N_2$, it is easily seen that the following rates are achievable for distortion (D_1, D_2)

$$\begin{aligned} R_1 \quad & = \quad I(X; X + N_1 | Y_1) \\ R_1 + R_2 \quad & = \quad I(X; X + N_1 + N_2 | Y_2) + I(X; X + N_1 | Y_1, X + N_1 + N_2). \end{aligned}$$

Thus we conclude that for both cases the gap between the inner bound and the outer bound is bounded. Fig. 3 illustrates the inner bound and outer bounds, as well as the gap in between.

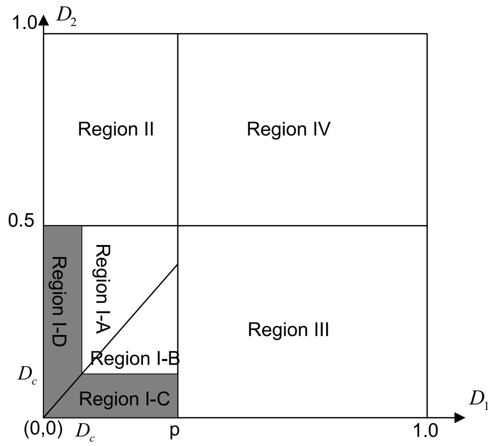

Fig. 4. The partition of the distortion region, where d_c is the critical distortion in [4] below which time sharing is not necessary.

VI. Two Special Sources

A. The Quadratic Gaussian Case

Consider the Gaussian source X and two jointly Gaussian side informations Y_1 and Y_2. The degraded side information assumption, either $X \leftrightarrow Y_1 \leftrightarrow Y_2$ or $X \leftrightarrow Y_2 \leftrightarrow Y_1$, for the quadratic Gaussian case is specially interesting. Since physically degradedness and statistically degradedness do not cause essential difference in terms of the rate-distortion region [5], and furthermore jointly Gaussian source-side information is always statistically degraded, these two degraded cases provide a complete solution to the jointly Gaussian two-decoder case. This can be generalized one more step to a system with more than two decoders, which introduces slight complication because the quality of side informations may not be monotonic along the scalable coding order. We next consider this general case, and the solution for the two stage case can be easily reduced from the general solution.

B. The Doubly Symmetric Binary Source with Hamming Distortion Measure

Consider the following source: X is a memoryless binary source $X \in \{0,1\}$ and $P(X=0)=0.5$. The first stage side information Y can be taken as the output of a binary symmetric channel with input X, and crossover probability $p < 0.5$. The second stage does not have side information. Despite various attempts [7], [19], [20], an explicit calculation of $R_{HB}(D_1, D_2)$ was not found until recently [6].

With this explicit calculation, it can be shown that in the shaded region in Fig. 4, the outer bound $\mathcal{R}_\cap(D_1, D_2)$ is in fact achievable (as well as in Region II, III and IV; however these three regions are degenerate cases, and will be ignored in what follows). Recall the definition of the critical distortion d_c in the Wyner-Ziv problem for the DSBS source in [4]

$$\frac{G(d_c)}{d_c - p} = G'(d_c),$$

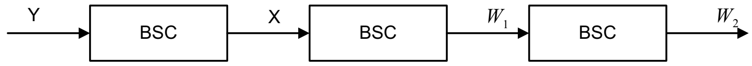

Fig. 5. The forward test channel in Region I-D. The crossover probability for the BSC between X and W_1 is D_1, while the crossover probability η for the BSC between W_1 and W_2 is such that $D_1 * \eta = D_2$.

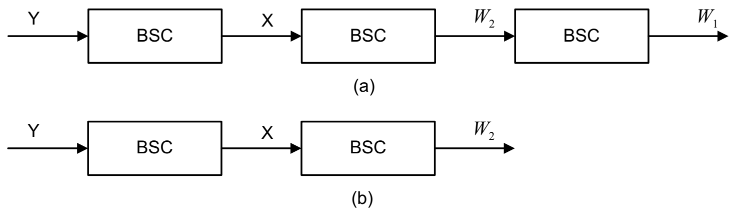

Fig. 6. The forward test channels in Region I-C. The crossover probability for the BSC between X and W_2 is D_2 in both the channels, while the crossover probability η for the BSC between W_2 and W_1 in (a) is such that $D_2 \leq D_1 * \eta = \eta' \leq d_c$. Note for (b), W_1 can be taken as a constant.

where $G(u) = h(p*u) - h(u)$, $h(u)$ is the binary entropy function $h(u) = -u \log u - (1-u)\log(1-u)$, and $u * v$ is the binary convolution for $0 \leq u, v \leq 1$ as $u * v = u(1-v) + v(1-u)$. It was shown in [4] that if $D \leq d_c$, then $R^*_{X|Y}(D) = G(D)$. We will use the following result from [6].

***Theorem** 5:* For distortion pairs (D_1, D_2) such that $0 \leq D_2 \leq 0.5$ and $0 \leq D_1 \leq \min(d_c, D_2)$ (i.e., Region I-D),

$$R_{HB}(D_1, D_2) = 1 - h(D_2 * p) + G(D_1).$$

This result implies that for the shaded region I-D, the forward test channel to achieve this lower bound is in fact a cascade of two BSC channels depicted in Fig. 5. This choice clearly satisfy the condition in Corollary 1 with the rates given by the outer bound $\mathcal{R}_\cap(D_1, D_2)$, which shows that this outer bound is indeed achievable. Note the following inequality

$$R_{HB}(D_1, D_2) = 1 - h(D_2 * p) + h(p * D_1) - h(D_1) \geq 1 - h(D_2) = R(D_2), \tag{45}$$

where the inequality is due to the monotonicity of $G(u)$ in $0 \leq u \leq 0.5$, we conclude that in this regime the source is not perfectly scalable. To see $\mathcal{R}_\cap(D_1, D_2)$ is also achievable in region I-C, recall the result in [4] that the optimal forward test channel to achieve $R^*_{X|Y}(D)$ has the following structure: it is the time-sharing between zero-rate coding and a BSC with crossover probability d_c if $D \geq d_c$, or a single BSC with crossover probability D otherwise. Thus it is straightforward to verify that $\mathcal{R}_\cap(D_1, D_2)$ is achievable by time sharing the following two forward test channels in Fig. 6. From this time-sharing channel, an equivalent forward test channel can be found such that the Markov condition $W_1' \leftrightarrow W_2 \leftrightarrow X$ is satisfied, and furthermore it satisfies the condition given in Theorem 4, and thus in this regime, the source is in fact perfectly scalable.

Unfortunately, we were not able to find the complete characterization for the regime I-A and I-B. Using an

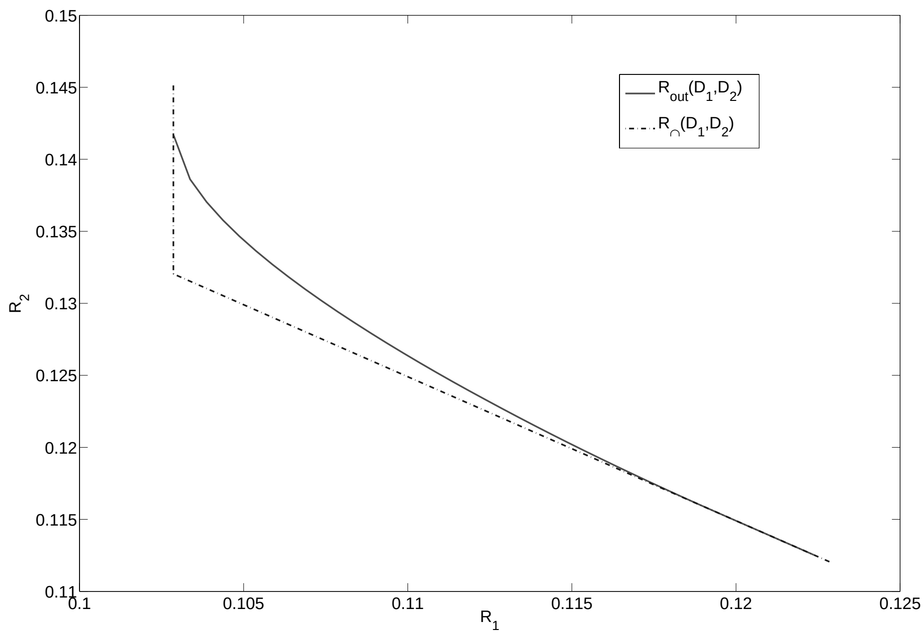

Fig. 7. The rate outer bounds for a particular choice of D_1, D_2 in Region I-B of Figure 4.

approach similar to [6], an explicit outer bound can be derived from $\mathcal{R}_{out}(D_1, D_2)$. It can then be shown numerically that for certain distortion pairs in this regime, $\mathcal{R}_{out}(D_1, D_2)$ is strictly tighter than $\mathcal{R}_{\cap}(D_1, D_2)$. This calculation is relegated to Appendix IV. An example is given in Fig. 7 for the two outer bounds with a non-zero gap in between for a specific distortion pair in Region I-B.

VII. Conclusion

We studied the problem of scalable source coding with reversely degraded side-information and gave an achievable rate region as well as two outer bounds. Furthermore we provided a complete solution the Gaussian source with quadratic distortion measure, which was shown to be generalized SI-scalable without rate loss. Furthermore, for some distortion pairs, it is strictly SI-scalable without rate loss, which suggests such an opportunistic approach does not cause any loss of optimality. For the doubly symmetric binary source with Hamming distortion, we provided partial results which show that it is generalized SI-scalable without rate loss for low distortion pairs, but fails to be so for others. The result illustrates the difference between the lossless and the lossy source coding: though a universal approach does exist with uncertain side informations at the decoder, such uncertainty generally causes loss of performance in the lossy case.

APPENDIX I

NOTATION AND BASIC PROPERTIES OF TYPICAL SEQUENCES

We will follow the definition of typicality in [11], but use a slightly different notation to make the small positive quantity δ explicit (see [5]).

***Definition** 4:* A sequence $\boldsymbol{x} \in \mathcal{X}^n$ is said to be δ-strongly-typical with respect to a distribution $P_X(x)$ on $\mathcal{X}$ if

1) For all $a \in \mathcal{X}$ with $P_X(a) > 0$

$$\left|\frac{1}{n}N(a|\boldsymbol{x}) - P_X(a)\right| < \delta, \tag{46}$$

2) For all $a \in \mathcal{X}$ with $P_X(a) = 0$, $N(a|\boldsymbol{x})$=0,

where $N(a|\boldsymbol{x})$ is the number of occurrences of the symbol a in the sequence $\boldsymbol{x}$. The set of sequences $\boldsymbol{x} \in \mathcal{X}^n$ that is δ-strongly-typical is called the δ-strongly-typical set and denoted as $T^{\delta}_{[X]}$, where the dimension n is dropped.

The following properties are well-known and will be used in the proof:

1) Given a $\boldsymbol{x} \in T^{\delta}_{[X]}$, for a $\boldsymbol{y}$ whose component is drawn i.i.d according to P_Y and any $\delta' > \delta$, we have

$$2^{-n(I(X;Y)+\lambda_1)} \leq P[(\boldsymbol{x},\boldsymbol{y}) \in T^{\delta'}_{[XY]}] \leq 2^{-n(I(X;Y)-\lambda_1)} \tag{47}$$

where λ_1 is a small positive quantity $\lambda_1 \to 0$ as $n \to \infty$ and both $\delta, \delta' \to 0$.

2) Similarly, given $(\boldsymbol{x},\boldsymbol{y}) \in T^{\delta'}_{[XY]}$, for any $\delta'' > \delta'$, let the component of $\boldsymbol{z}$ be drawn i.i.d according to the conditional marginal $P_{Z_i|Y_i}(y_i)$, then

$$2^{-n(I(X;Z|Y)+\lambda_2)} \leq P[(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \in T^{\delta''}_{[XYZ]}] \leq 2^{-n(I(X;Z|Y)-\lambda_2)} \tag{48}$$

where λ_2 is a small positive quantity $\lambda_2 \to 0$ as $n \to \infty$ and both $\delta', \delta'' \to 0$.

3) *Markov Lemma [16]:* If $X \leftrightarrow Y \leftrightarrow Z$ is a Markov string, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are such that their component is drawn independently according to P_{XY}. Then for all $\delta > 0$

$$\lim_{n\to\infty} P[(\boldsymbol{X},\boldsymbol{z}) \in T^{|\mathcal{Y}|\delta}_{[XZ]}\,|(\boldsymbol{Y},\boldsymbol{z}) \in T^{\delta}_{[YZ]}] \to 1. \tag{49}$$

furthermore,

$$\lim_{n\to\infty} P[(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{z}) \in T^{\delta}_{[XYZ]}\,|(\boldsymbol{Y},\boldsymbol{z}) \in T^{\delta}_{[YZ]}] \to 1. \tag{50}$$

APPENDIX II

PROOF OF THEOREM 1

Codebook generation: Let a probability distribution $P_{W_1W_2XY_1Y_2} = P_{XVW_1W_2}P_{Y_1|X}P_{Y_2|Y_1}$, and two reconstruction functions $f_1(Y_1,W_1)$ and $f_2(Y_2,W_2)$ be given. First construct 2^{nR_A} coarser bins and $2^{nR_A+R'_A}$ finer bins, where R_A and R'_A are to be specified later. Generate 2^{R_V} length-n codewords according to $P_V(\cdot)$, denote this set of codewords as $\mathcal{C}_v$; assign each of them into one of the finer bins independently. For each codeword $\boldsymbol{v} \in \mathcal{C}_v$, generate $2^{nR_{W_1}}$ length-n codewords according to $P_{W_1|V}(\boldsymbol{w_1}|\boldsymbol{v}) = \prod_{k=1}^{n} P_{W_1|V}(w_{1,k}|v_k)$, denote this set of codewords as $\mathcal{C}_{W_1}(\boldsymbol{v})$; independently assign each codeword to one of the 2^{nR_B} bins. Again for each $\boldsymbol{V}$ codeword, independently generate $2^{nR_{W_2}}$ length-n codewords according to $P_{W_2|V}(\boldsymbol{w_2}|\boldsymbol{v}) = \prod_{k=1}^{n} P_{W_2|V}(w_{2,k}|v_k)$, denote this set of codewords as $\mathcal{C}_{W_2}(\boldsymbol{v})$; independently assign each codeword to one of the 2^{nR_C} bins. Reveal this codebook to the encoders and decoders.

Encoding: For a given $\boldsymbol{x}$, find in $\mathcal{C}_v$ a codeword $\boldsymbol{v}^*$ such that $(\boldsymbol{x}, \boldsymbol{v}^*) \in T^{2\delta}_{[XV]}$; calculate the coarser bin index $i(\boldsymbol{v}^*)$, and the finer bin index within the coarser bin $j(\boldsymbol{v}^*)$. Then in the $\mathcal{C}_{w_1}(\boldsymbol{v}^*)$ codebook, find a codeword $\boldsymbol{w}_1^*$ such that $(\boldsymbol{w}_1^*, \boldsymbol{v}^*, \boldsymbol{x}^*) \in T^{3\delta}_{[W_1VX]}$, and calculate its corresponding bin index k. In $\mathcal{C}_{w_2}(\boldsymbol{v}^*)$ codebook, find a codeword $\boldsymbol{w}_2^*$ such that $(\boldsymbol{w}_2^*, \boldsymbol{v}^*, \boldsymbol{x}) \in T^{3\delta}_{[W_2VX]}$, and calculate its corresponding bin index l. The first-stage encoder sends i and k, and the second-stage encoder sends j and l. In the above procedure, if there is more than one joint-typical sequence, choose the least; if there is none, choose a default codeword and declare an error.

Decoding: The first stage decoder finds $\hat{\boldsymbol{v}}$ in the coarser bin i, such that $(\hat{\boldsymbol{v}}, \boldsymbol{y_1}) \in T^{3|\mathcal{X}|\delta}_{[VY_1]}$; then in the $\mathcal{C}_{w_1}(\hat{\boldsymbol{v}})$ codebook, find $\hat{\boldsymbol{w_1}}$ such that $(\hat{\boldsymbol{w_1}}, \hat{\boldsymbol{v}}, \boldsymbol{y_1}) \in T^{4|\mathcal{X}|\delta}_{[W_1VY_1]}$. In the second stage, the decoder finds $\hat{\boldsymbol{v}}$ in the finer bin specified by (i, j) such that $(\hat{\boldsymbol{v}}, \boldsymbol{y_2}) \in T^{3|\mathcal{X}|\delta}_{[VY_2]}$; then in the $\mathcal{C}_{w_2}(\hat{\boldsymbol{v}})$ codebook, find $\hat{\boldsymbol{w_2}}$ such that $(\hat{\boldsymbol{w_2}}, \hat{\boldsymbol{v}}, \boldsymbol{y_2}) \in T^{4|\mathcal{X}|\delta}_{[W_2VY_2]}$. In the above procedure, if there is none or there are more than one, an error is declared and the decoding stops. The first decoder reconstructs as $\hat{x}_{1,k} = f_1(\hat{w}_{1,k}, y_{1,k})$ and the second decoder as $\hat{x}_{2,k} = f_2(\hat{w}_{2,k}, y_{2,k})$.

Probability of error: First define the encoding errors:

$$
\begin{aligned}
E_0 &= \{\boldsymbol{X} \notin T^{\delta}_{[X]}\} \cup \{\boldsymbol{Y_1} \notin T^{\delta}_{[Y_1]}\} \cup \{\boldsymbol{Y_2} \notin T^{\delta}_{[Y_2]}\} \\
E_1 &= E_0^c \cap \{\forall \boldsymbol{v} \in \mathcal{C}_v, (\boldsymbol{X}, \boldsymbol{v}) \notin T^{2\delta}_{[XV]}\} \\
E_2 &= E_0^c \cap E_1^c \cap \{\forall \boldsymbol{w_1} \in \mathcal{C}_{w_1}(\boldsymbol{v}^*), (\boldsymbol{w_1}, \boldsymbol{v}^*, \boldsymbol{X}) \notin T^{3\delta}_{[W_1VX]}\} \\
E_3 &= E_0^c \cap E_1^c \cap \{\forall \boldsymbol{w_2} \in \mathcal{C}_{w_2}(\boldsymbol{v}^*), (\boldsymbol{w_2}, \boldsymbol{v}^*, \boldsymbol{X}) \notin T^{3\delta}_{[W_2VX]}\}.
\end{aligned}
$$

Next define the decoding errors:

$$
\begin{aligned}
E_4 &= E_0^c \cap E_1^c \cap \{(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_1}) \notin T^{2\delta}_{[VXY_1]}\} \\
E_5 &= E_0^c \cap E_1^c \cap \{(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_2}) \notin T^{2\delta}_{[VXY_2]}\} \\
E_6 &= E_0^c \cap E_1^c \cap \{\exists \boldsymbol{v}' \neq \boldsymbol{v}^* : i(\boldsymbol{v}') = i(\boldsymbol{v}^*) \text{ and } (\boldsymbol{v}', \boldsymbol{Y_1}) \in T^{3|\mathcal{X}|\delta}_{[VY_1]}\} \\
E_7 &= E_0^c \cap E_1^c \cap \{\exists \boldsymbol{v}' \neq \boldsymbol{v}^* : i(\boldsymbol{v}') = i(\boldsymbol{v}^*) \text{ and } j(\boldsymbol{v}') = j(\boldsymbol{v}^*) \text{ and } (\boldsymbol{v}', \boldsymbol{Y_2}) \in T^{3|\mathcal{X}|\delta}_{[VY_2]}\} \\
E_8 &= E_0^c \cap E_1^c \cap E_2^c \cap E_4^c \cap E_6^c \cap \{(\boldsymbol{w_1^*}, \boldsymbol{v^*}, \boldsymbol{X}, \boldsymbol{Y_1}) \notin T^{3\delta}_{[W_1VXY_1]}\} \\
E_9 &= E_0^c \cap E_1^c \cap E_3^c \cap E_5^c \cap E_7^c \cap \{(\boldsymbol{w_2^*}, \boldsymbol{v^*}, \boldsymbol{X}, \boldsymbol{Y_2}) \notin T^{3\delta}_{[W_2VXY_2]}\} \\
E_{10} &= E_0^c \cap E_1^c \cap E_2^c \cap E_4^c \cap E_6^c \cap \{\exists \boldsymbol{w_1'} \neq \boldsymbol{w_1^*} : l(\boldsymbol{w_1'}) = l(\boldsymbol{w_1^*}) \text{ and } (\boldsymbol{w_1'}, \boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{4|\mathcal{X}|\delta}_{[W_1VY_1]}\} \\
E_{11} &= E_0^c \cap E_1^c \cap E_3^c \cap E_5^c \cap E_7^c \cap \{\exists \boldsymbol{w_2'} \neq \boldsymbol{w_2^*} : l(\boldsymbol{w_2'}) = l(\boldsymbol{w_2^*}) \text{ and } (\boldsymbol{w_2'}, \boldsymbol{v}^*, \boldsymbol{Y_2}) \in T^{4|\mathcal{X}|\delta}_{[W_2VY_2]}\}
\end{aligned}
$$

Apparently, for any ϵ', for $n > n_1(\epsilon', \delta)$, $P(E_0) \leq \epsilon'$. We have also

$$\begin{aligned} P(E_1) &\leq P(\boldsymbol{X} \in T^{\delta}_{[X]}) P(\{\forall\ \boldsymbol{v} \in \mathcal{C}_v,\ (\boldsymbol{X}, \boldsymbol{v}) \notin T^{2\delta}_{[XV]}\} | \boldsymbol{X} \in T^{\delta}_{[X]}) \\ &\leq \sum_{\boldsymbol{x} \in T^{\delta}_{[X]}} P_X(\boldsymbol{x})(1 - 2^{-n(I(X;V)+\lambda)})^{nR_1} \\ &\leq \exp(-2^{-n(I(X;V)+\lambda - R_V)}), \end{aligned} \tag{51}$$

where Property 1) of the typical sequences and $(1-x)^y < e^{-xy}$ are used. Thus $P(E_1) \to 0$, provided that $R_V > I(X;V) + \lambda$.

$P(E_4)$ and $P(E_5)$ both tends to zero due to the Markov lemma; it requires the condition $(\boldsymbol{v}^*, \boldsymbol{X}) \in T^{2\delta}_{[VX]}$ to hold, which is indeed so given E_1 does not happen. Similarly, both $P(E_8)$ and $P(E_9)$ tends to zero for the same reason. Notice that if $(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_1}) \in T^{2\delta}_{[VXY_1]}$, then $(\boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{3|\mathcal{X}|\delta}_{[VY_1]}$, thus $\boldsymbol{v}^*$ can be correctly decoded if there is no other codewords in the same bin satisfying the typicality test.

Conditioned on E_1^c, we have $(\boldsymbol{X}, \boldsymbol{v}) \in T^{2\delta}_{[XV]}$. Thus

$$\begin{aligned} P(E_2) &\leq \sum_{(\boldsymbol{x},\boldsymbol{v}) \in T^{2\delta}_{[XV]}} Pr(\boldsymbol{x}, \boldsymbol{v})(1 - 2^{-n(I(X;W_1|V)+\lambda)})^{nR_2} \\ &\leq \exp(-2^{-n(I(X;W_1|V)+\lambda_2 - R_2)}) \end{aligned} \tag{52}$$

where property 2) of the typical sequences is used. Thus $P(E_2)$ tends to zero provided $R_{W_1} > I(X;W_1|V) + \lambda_1$. Similarly $P(E_3')$ tends to zero provided $R_{W_2} > I(X;W_2|V) + \lambda_2$.

Conditioned on E_1^c, $\boldsymbol{y_1} \in T^{\delta}_{[Y_1]}$, since codeword in $\mathcal{C}_v$ are generated independently according to $P_U(\cdot)$

$$\begin{aligned} P(E_6) &\leq \sum_{\boldsymbol{v} \in \mathcal{C}_v} 2^{-nR_A} 2^{-n(I(Y_1;V)-\lambda_1)} \\ &= 2^{n(R_V - R_A - I(Y_1;V)+\lambda_1)} \end{aligned} \tag{53}$$

where we have used property 2) of the typical sequences and the fact the bin to which $\boldsymbol{v}$ is assigned is independent. Thus $P(E_6) \to 0$ provided that $R_A > R_V - I(Y_1;V) + \lambda_3$. Similarly $P(E_7) \to 0$ provided that $R_A + R_A' > R_V - I(Y_2;V) + \lambda_4$.

Conditioned on E_4^c, $(\boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{2|\mathcal{X}|\delta}_{[VY_1]}$. Thus

$$\begin{aligned} P(E_{10}) &\leq 2^{nR_{W_1}} 2^{-nR_B} 2^{-n(I(Y_1;W_1|V)-\lambda_3)} \\ &= 2^{n(R_{W_1} - R_B - I(Y_1;W_1|V)+\lambda_3)} \end{aligned} \tag{54}$$

where property 3) of the typical sequences is used. Thus $P(E_{10})$ tends to zero provided $R_B > R_{W_1} - I(Y_1;W_1|V) + \lambda_5$. Similarly, $P(E_{11})$ tends to zero provided $R_C > R_{W_2} - I(Y_2;W_2|V) + \lambda_6$. Thus the rates only need to satisfy

$$R_1 = R_A + R_B > I(X;VW_1|Y_1) + \lambda' \tag{55}$$

$$R_1 + R_2 = R_A + R_A' + R_B + R_C > I(X;VW_2|Y_2) + I(X;W_2|VY_1) + \lambda'' \tag{56}$$

where λ' and λ'' are both small positive quantities and vanish as $\delta \to 0$ and $n \to \infty$; then $P_e \leq \sum_{i=0}^{11} P(E_i) \to 0$. It only remains to show that the distortions constraints are satisfied as well. When no error occurs, then $(\hat{\boldsymbol{W}}_{\boldsymbol{1}}, \boldsymbol{X}, \boldsymbol{Y_1}) \in$

$T^{3|\mathcal{V}|\delta}_{[W_1XY]}$ and $(\hat{\boldsymbol{W_2}}, \boldsymbol{X}, \boldsymbol{Y_1}) \in T^{3|\mathcal{V}|\delta}_{[W_2XY]}$. By standard argument using the definition of the typical sequences, it can be shown that

$$d(\boldsymbol{x}, \hat{\boldsymbol{x_1}}) \leq \mathbb{E}d[X, f_1(W_1, Y_1)] + \epsilon' \tag{57}$$

where $\epsilon' = \max(d(x,\hat{x}))(3|\mathcal{V}\times\mathcal{W}_1\times\mathcal{X}\times\mathcal{Y}_1|\delta + P_e)$. Thus the distortion can be made arbitrarily small by choosing sufficiently small δ and sufficiently large n. Similar arguments holds for the second stage decoder. This completes the proof. ■

APPENDIX III
PROOF OF THE THEOREM 2

Assume the existence of (n, M_1, M_2, D_1, D_2) RD SI-scalable code, there exist encoding and decoding functions ϕ_i and ψ_i for $1 = 1, 2$. Denote $\phi_i(X^n)$ as T_i. $\boldsymbol{X}_k^-$ will be used to denote the vector $(X_1, X_2, ..., X_{k-1})$ and $\boldsymbol{X}_k^+$ to denote $(X_{k+1}, X_{k+2}, ..., X_n)$; the subscript k will be dropped when it is clear from the context. The proof follows the same line as the converse proof in [7]. The following chain of inequalities is standard (see page 440 of [21]).

$$\begin{aligned}
nR_1 &\geq H(T_1) \\
&\geq H(T_1|\boldsymbol{Y_1}) \\
&= I(\boldsymbol{X}; T_1|\boldsymbol{Y_1}) \\
&= \sum_{k=1}^{n} I(X_k; T_1|\boldsymbol{Y_1}\boldsymbol{X}_k^-) \\
&= \sum_{k=1}^{n} H(X_k|\boldsymbol{Y_1}\boldsymbol{X}_k^-) - H(X_k|T_1\boldsymbol{Y_1}\boldsymbol{X}_k^-) \\
&= \sum_{k=1}^{n} H(X_k|Y_{1,k}) - H(X_k|T_1\boldsymbol{Y_1}\boldsymbol{X}_k^-) \\
&\geq \sum_{k=1}^{n} I(X_k; T_1\boldsymbol{Y_1^-}\boldsymbol{Y_1^+}|Y_k).
\end{aligned} \tag{58}$$

Next we bound the sum rate as follows

$$\begin{aligned}
n(R_1+R_2) &\geq H(T_1T_2) \\
&\geq H(T_1T_2|\boldsymbol{Y_2}) \\
&= I(\boldsymbol{X}; T_1T_2|\boldsymbol{Y_2}) \\
&= I(\boldsymbol{X}; T_1T_2\boldsymbol{Y_1}|\boldsymbol{Y_2}) - I(\boldsymbol{X}; \boldsymbol{Y_1}|T_1T_2\boldsymbol{Y_2}) \\
&= \sum_{k=1}^{n} [I(X_k; T_1T_2\boldsymbol{Y_1}|\boldsymbol{Y_2}\boldsymbol{X}^-) - I(\boldsymbol{X}; Y_{1,k}|T_1T_2\boldsymbol{Y_2}\boldsymbol{Y_1^-})].
\end{aligned}$$

Since $(X_k, Y_{2,k})$ is independent of $(\boldsymbol{X}^-, \boldsymbol{Y_2^-}, \boldsymbol{Y_2^+})$, we have

$$\begin{aligned}
I(X_k; T_1T_2\boldsymbol{Y_1}|\boldsymbol{Y_2}\boldsymbol{X}^-) &= I(X_k; T_1T_2\boldsymbol{Y_1}\boldsymbol{Y_2^-}\boldsymbol{Y_2^+}\boldsymbol{X}^-|Y_{2,k}) \\
&\geq I(X_k; T_1T_2\boldsymbol{Y_1}\boldsymbol{Y_2^-}\boldsymbol{Y_2^+}|Y_{2,k})
\end{aligned} \tag{59}$$

The Markov condition $Y_{1,k} \leftrightarrow (X_k, Y_{2,k}) \leftrightarrow (\boldsymbol{X}^- \boldsymbol{X}^+ T_1 T_2 \mathbf{Y_1}^- \mathbf{Y_2}^- \mathbf{Y_2}^+)$ gives

$$I(\boldsymbol{X}; Y_{1,k} | T_1 T_2 \mathbf{Y_2} \mathbf{Y_1}^-) = I(X_k; Y_{1,k} | T_1 T_2 \mathbf{Y_2} \mathbf{Y_1}^-). \tag{60}$$

Thus we have

$$\begin{aligned} n(R_1 + R_2) &\geq \sum_{k=1}^{n} [I(X_k; T_1 T_2 \mathbf{Y_1} \mathbf{Y_2}^- \mathbf{Y_2}^+ | Y_{2,k}) - I(X_k; Y_{1,k} | T_1 T_2 \mathbf{Y_2} \mathbf{Y_1}^-)] \\ &= \sum_{k=1}^{n} [I(X_k; T_1 T_2 \mathbf{Y_1}^- \mathbf{Y_2}^- \mathbf{Y_2}^+ | Y_{2,k}) + I(X_k; \mathbf{Y_1}^+ | T_1 T_2 \mathbf{Y_2} \mathbf{Y_1}^- Y_{1,k})]. \end{aligned} \tag{61}$$

The degradedness gives $Y_{2,k} \leftrightarrow Y_{1,k} \leftrightarrow (X_k, T_1 T_2, \mathbf{Y_1}^- \mathbf{Y_2}^- \mathbf{Y_2}^+)$, which implies

$$n(R_1 + R_2) \geq \sum_{k=1}^{n} [I(X_k; T_1 T_2 \mathbf{Y_2}^- \mathbf{Y_2}^+ \mathbf{Y_1}^- | Y_{2,k}) + I(X_k; \mathbf{Y_1}^+ | T_1 T_2 \mathbf{Y_2}^- \mathbf{Y_2}^+ \mathbf{Y_1}^- Y_{1,k})]. \tag{62}$$

Define $W_{1,k} = (T_1 \mathbf{Y_1}^- \mathbf{Y_1}^+)$ and $W_{2,k} = (T_1 T_2 \mathbf{Y_2}^- \mathbf{Y_2}^+ \mathbf{Y_1}^-)$, by which we have

$$nR_1 \geq \sum_{k=1}^{n} I(X_k; W_{1,k} | Y_{1,k}) \tag{63}$$

$$n(R_1 + R_2) \geq \sum_{k=1}^{n} [I(X_k; W_{2,k} | Y_{2,k}) + I(X_k; W_{1,k} | W_{2,k} Y_{1,k})]. \tag{64}$$

Therefore the Markov condition $(W_{1,k}, W_{2,k}) \leftrightarrow X_k \leftrightarrow Y_{1,k} \leftrightarrow Y_{2,k}$ is true. Next introduce the time sharing random variable Q, which is independent of the multisource, and uniformly distributed over I_n. Define $W_j = (W_{j,Q}, Q),\ j = 1, 2$. The existence of function f_j follows by defining

$$f_1(W_1, Y_1) = \psi_{1,Q}(\phi_1(\boldsymbol{X}), \mathbf{Y_1}) \tag{65}$$

$$f_2(W_2, Y_2) = \psi_{2,Q}(\phi_1(\boldsymbol{X}), \phi_2(\boldsymbol{X}), \mathbf{Y_2}) \tag{66}$$

which leads the fulfillment of the distortion constraints. It only remains to show both the bound can be written in single letter form in W_1, W_2, which is straightforward following the approach in (page 435 of) [21]. This completes the proof for $\mathcal{R}_{out}(D_1, D_2) \supseteq \mathcal{R}(D_1, D_2)$. ■

APPENDIX IV
AN EXPLICIT OUTER BOUND FOR THE DSBS

In this appendix, we provide an explicit lower bound for the doubly symmetric binary source considered in Section VI-B. To simply the notations, we reformulate the problem as follows:

$$R_{sum}(D_1, D_2, R_1) = \min_{(R_1, R_2) \in \mathcal{R}(D_1, D_2)} [R_1 + R_2]. \tag{67}$$

We will show the following is true for the DSBS in question:

$$R_{sum}(D_1, D_2, R_1) \geq \min_{(\alpha, \beta, \theta, \theta_1, D_2') \in \mathcal{Q}} [1 - h(D_2' * p) + \theta_1 G(\beta) + (\theta - \theta_1) G(\alpha) + (1 - \theta) G(\gamma)] \tag{68}$$

where

$$\gamma = \frac{D_1 - (\theta - \theta_1)(1 - \alpha) - \theta_1 \beta}{1 - \theta}, \tag{69}$$

and the minimization is within the set

$$\mathcal{Q} = \{(\alpha,\beta,\theta,\theta_1,D_2'): 0 \le D_2' - (\theta-\theta_1)(1-\alpha) - \theta_1\beta \le (1-\theta), \quad 0 \le \theta_1 \le \theta \le 1,$$
$$0 \le \alpha,\beta \le 1, \quad (\theta-\theta_1)\alpha + \theta_1\beta + (1-\theta)p \le D_1, \quad D_2' \le D_2, \quad \theta\, G(\frac{\theta_1\beta + (\theta-\theta_1)\alpha}{\theta}) \le R_1\}.$$

We will need the following lemma from [19] to simplify the calculation.

***Lemma** 1:* For (W_1, W_2) such that $(W_1, W_2) \leftrightarrow X \leftrightarrow Y$ forms a Markov string,

$$I(X;W_2) + I(X;W_1|YW_2) = H(X) - H(Y|W_2) + H(Y|W_1W_2) - H(X|W_1W_2). \tag{70}$$

Proof of (68)

Let (W_1, W_2) define a joint distribution with (X, Y). Furthermore, assume the functions f_1 and f_2 are optimal for these random variables, i.e., there do not exist f_1' (or f_2'), such that $\mathbb{E}d(X, f_1'(W_1, Y)) < \mathbb{E}d(X, f_1(W_1, Y))$ (or $\mathbb{E}d(X, f_2'(W_2)) < \mathbb{E}d(X, f_2(W_2))$), because otherwise we can consider the alternative functions f_1' (or f_2') without loss of optimality.

Similar as in [4], define the following set

$$A = \{(w_1) : f_1(w_1, 0) = f_1(w_1, 1)\}, \tag{71}$$

which makes it complement

$$A^c = \mathcal{W}_1 - A = \{w_1 : f_1(w_1, 0) \neq f_1(w_1, 1)\}. \tag{72}$$

For each $w_2 \in \mathcal{W}_2$, define the following two sets

$$B(w_2) = \{w_1 \in A : f_2(w_2) = f_1(w_1, 0)\}, \quad B^*(w_2) = \{w_1 \in A : f_2(w_2) \neq f_1(w_1, 0)\}.$$

Notice that for each fixed $w_2^* \in \mathcal{W}_2$, we have $\mathcal{W}_1 = B(w_2^*) \cup B^*(w_2^*) \cup \{w_1 : w_1 \in A^c\}$, and the three sets are disjoint. To simplify the notations, write $P\{(W_1W_2) = (w_1w_2)\}$ as $P_{w_1w_2}$, $P\{W_1 = w_1\}$ as P_{w_1}, and $P\{W_2 = w_2\}$ as P_{w_2}. Define the following quantities for each $w_1 \in A$

$$D_{1,w_1} \triangleq \mathbb{E}[d(X, \hat{X}_1)|W_1 = w_1] = P\{X \neq f_1(w_1, 0)|W_1 = w_1\}$$
$$D_{1,w_1w_2} \triangleq \mathbb{E}[d(X, \hat{X}_1)|(W_1, W_2) = (w_1, w_2)] = P\{X \neq f_1(w_1, 0)|(W_1, W_2) = (w_1, w_2)\},$$

and define the following quantity for each $w_2 \in \mathcal{W}_2$,

$$D_{2,w_2} \triangleq \mathbb{E}[d(X, \hat{X}_2)|W_2 = w_2] = P\{X \neq f_2(w_2)|W_2 = w_2\}.$$

By the Markov string $Y \leftrightarrow X \leftrightarrow (W_1, W_2)$, it follows that for each $w_2 \in \mathcal{W}_2$

$$H(X|W_2 = w_2) = h(D_{2,w_2}), \quad H(Y|W_2 = w_2) = h(p * D_{2,w_2}), \tag{73}$$

and for each each (w_1, w_2) such that $w_1 \in A$,

$$H[X|(W_1, W_2) = (w_1, w_2)] = h(D_{1,w_1w_2}), \quad H[Y|(W_1, W_2) = (w_1, w_2)] = h(p * D_{1,w_1w_2}). \tag{74}$$

We will also need the following quantities

$$\theta \triangleq P\{(W_1,W_2):W_1\in A\},\quad \theta_1 \triangleq P\{(W_1,W_2):W_1\in B(W_2)\}. \tag{75}$$

Apparently, we have

$$\begin{aligned} H(X)-H(Y|W_2) &= 1-\sum_{w_2\in\mathcal{W}_2} P_{w_2}H(Y|W_2=w_2) \\ &= 1-\sum_{w_2\in\mathcal{W}_2} P_{w_2}h(p*D_{2,w_2}) \\ &\geq 1-h(p*D_2') \end{aligned} \tag{76}$$

where we have used the concavity of function $h(p*u)$ in the last step and

$$D_2' \triangleq \sum_{w_2\in\mathcal{W}_2} P_{w_2}D_{2,w_2}.$$

Furthermore we have

$$\begin{aligned} & H(Y|W_1W_2)-H(X|W_1W_2) \\ = & \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A} P_{w_1,w_2}[H(Y|(W_1,W_2)=(w_1,w_2))-H(X|(W_1,W_2)=(w_1,w_2))] \\ & +\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A^c} P_{w_1,w_2}[H(Y|(W_1,W_2)=(w_1,w_2))-H(X|(W_1,W_2)=(w_1,w_2))] \end{aligned}$$

The first term can be bounded as follows

$$\begin{aligned} & \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A} P_{w_1,w_2}[H(Y|(W_1,W_2)=(w_1,w_2))-H(X|(W_1,W_2)=(w_1,w_2))] \\ = & \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B(w_2)} P_{w_1,w_2}[h(p*D_{1,w_1w_2})-h(D_{1,w_1w_2})]+\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B^*(w_2)} P_{w_1,w_2}[h(p*D_{1,w_1w_2})-h(D_{1,w_1w_2})] \\ \geq & \theta_1 G(\beta)+(\theta-\theta_1)G(\alpha), \end{aligned} \tag{77}$$

where

$$\alpha \triangleq \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B^*(w_2)} \frac{P_{w_1w_2}}{\theta-\theta_1}D_{1,w_1w_2},\quad \beta \triangleq \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B(w_2)} \frac{P_{w_1w_2}}{\theta_1}D_{1,w_1w_2}, \tag{78}$$

and the convexity of function $G(u)$ is used in the last step. Next, notice the identity that for each $w_2\in\mathcal{W}_2$

$$\begin{aligned} P_{w_2}D_{2,w_2} &= P\{X\neq f_2(w_2),W_2=w_2\} \\ &= \sum_{w_1\in B(w_2)} P\{X\neq f_1(w_1,0),W_1=w_1,W_2=w_2\} \\ &\quad +\sum_{w_1\in B^*(w_2)} P\{X=f_1(w_1,0),W_1=w_1,W_2=w_2\} \\ &\quad +\sum_{w_1\in A^c} P\{X\neq f_2(w_2),W_1=w_1,W_2=w_2\} \\ &= \sum_{w_1\in B(w_2)} P_{w_1w_2}D_{1,w_1w_2}+\sum_{w_1\in B^*(w_2)} P_{w_1w_2}(1-D_{1,w_1w_2}) \\ &\quad +\sum_{w_1\in A^c} P_{w_1w_2}P\{X\neq f_2(w_2)|W_1=w_1,W_2=w_2\}. \end{aligned} \tag{79}$$

It follows that

$$
\begin{aligned}
& \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A^c} P_{w_1,w_2}[H(Y|(W_1,W_2)=(w_1,w_2))-H(X|(W_1,W_2)=(w_1,w_2))] \\
= & \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A^c} P_{w_1,w_2}G[P\{X\neq f_2(w_2)|(W_1,W_2)=(w_1,w_2)\}] \\
\geq & (1-\theta)G(\gamma), \qquad (80)
\end{aligned}
$$

where again the convexity of function $G(u)$ is used, and because of the identity (79), we have

$$
\begin{aligned}
\gamma &= \sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A^c}\frac{P_{w_1w_2}}{1-\theta}P\{X\neq f_1(w_1)|W_1=w_1,W_2=w_2\} \\
&= \frac{D_2'-\theta_1\beta-(\theta-\theta_1)(1-\alpha)}{1-\theta}. \qquad (81)
\end{aligned}
$$

To bound the first stage rate, we write

$$
\begin{aligned}
I(X;W_1|Y) &= H(Y|W_1)-H(X|W_1) \\
&\geq \sum_{w_1\in A} P_{w_1}[H(Y|W_1=w_1)-H(X|W_1=w_1)] \qquad (82) \\
&\geq \theta G(\lambda) \qquad (83)
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda &\triangleq \frac{1}{\theta}\sum_{w_1\in A}P_{w_1}D_{1,w_1} \\
&= \frac{1}{\theta}\sum_{w_1\in A}P\{X\neq f_1(w_1,0),W_1=w_1\} \\
&= \frac{1}{\theta}\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in A}P\{X\neq f_1(w_1,0),W_1=w_1,W_2=w_2\} \qquad (84) \\
&= \frac{1}{\theta}\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B(w_2)}P\{X\neq f_1(w_1,0),W_1=w_1,W_2=w_2\} \\
&\quad +\frac{1}{\theta}\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B^*(w_2)}P\{X\neq f_1(w_1,0),W_1=w_1,W_2=w_2\} \qquad (85) \\
&= \frac{1}{\theta}\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B(w_2)}P_{w_1w_2}D_{1,w_1w_2}+\frac{1}{\theta}\sum_{w_2\in\mathcal{W}_2}\sum_{w_1\in B^*(w_2)}P_{w_1w_2}D_{1,w_1w_2} \qquad (86) \\
&= \frac{\theta_1\beta+(\theta-\theta_1)\alpha}{\theta}. \qquad (87)
\end{aligned}
$$

It was shown in [4], that

$$E[d(X,\hat{X}_1)|W_1\in A^c]\geq p. \qquad (88)$$

By the hypothesis

$$D_1'\triangleq\theta_1\beta+(\theta-\theta_1)\alpha+(1-\theta)p\leq D_1$$

$$D_2'\leq D_2.$$

We have apparently $0 \leq \alpha, \beta, \gamma, \theta, \theta_1 \leq 1$ from their definition. So far we have already shown that there exists $(\alpha, \beta, \theta, \theta_1, D_2') \in \mathcal{Q}$ such that

$$I(X;W_2) + I(X;W_1|W_2,Y) \geq 1 - h(D_2' * p) + \theta_1 G(\beta) + (\theta - \theta_1)G(\alpha) + (1-\theta)G(\gamma), \tag{89}$$

which indeed establishes the claim. ■

REFERENCES

[1] V. N. Koshelev, "Hierarchical coding of discrete sources," *Probl. Pered. Inform.*, vol. 16, no. 3, pp. 31–49, 1980.

[2] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Information Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[3] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. Information Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

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

[5] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," *IEEE Trans. Information Theory*, vol. 50, no. 8, pp. 1636–1654, Aug. 2004.

[6] C. Tian and S. Diggavi, "On multistage successive refinement for Wyner-Ziv source coding with degraded side information," in *EPFL Technical Report*, Jan. 2006.

[7] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Information Theory*, vol. 31, no. 6, pp. 727–734, Nov. 1985.

[8] A. Kaspi, "Rate-distortion when side-information may be present at the decoder," *IEEE Trans. Information Theory*, vol. 40, no. 6, pp. 2031–2034, Nov. 1994.

[9] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information source," *IEEE Trans. Information Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[10] M. Feder and N. Shulman, "Source broadcasting with unknown amount of receiver side information," in *Proc. IEEE Information Theory Workshop*, Oct. 2002, pp. 127–130.

[11] I. Csiszar and J. Korner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

[12] S. C. Draper, "Universal incremental Slepian-Wolf coding," in *Proc. 43rd Annual Allerton Conference on communication, control and computing*, Sep. 2002.

[13] A. Eckford and W. Yu, "Rateless Slepian-Wolf codes," in *Proc. Asilomar conference on signals, systems and computers*, Oct.-Nov. 2005.

[14] R. Zamir and Meir Feder, "On lattice quantization noise," *IEEE Trans. Information Theory*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.

[15] L. A. Lastras and V. Castelli, "Near sufficiency of random coding for two descriptions," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 618–695, Feb. 2006.

[16] T. Berger, "Multiterminal source coding," in *Lecture notes at CISM summer school on the information theory approach to communications*. 1977.

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

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

[19] K. J. Kerpez, "The rate-distortion function of a binary symmetric source when side information may be absent," *IEEE Trans. Information Theory*, vol. 33, no. 3, pp. 448–452, May. 1987.

[20] M. Fleming and M. Effros, "Rate-distortion with mixed types of side information," in *Proc. IEEE Symposium Information Theory*, Jun.-Jul 2003, p. 144.

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