Side-Information Scalable Source Coding

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

***Abstract*—We consider the problem of side-information scalable (SI-scalable) source coding, where the encoder constructs a two-layer description, such that the receiver with high quality side information will be able to use only the first layer to reconstruct the source in a lossy manner, while the receiver with low quality side information will have to receive both layers in order to decode. We provide inner and outer bounds to the rate–distortion (R-D) region for general discrete memoryless sources. The achievable region is tight when either one of the decoders requires a lossless reconstruction, and when the distortion measures are degraded and deterministic. Furthermore, the gap between the inner and the outer bounds can be bounded by certain constants when the squared error distortion measure is used. The notion of perfect scalability is introduced, for which necessary and sufficient conditions are given for sources satisfying a mild support condition. Using SI-scalable coding and successive refinement Wyner–Ziv coding as basic building blocks, we provide a complete characterization of the rate–distortion region for the important quadratic Gaussian source with multiple jointly Gaussian side informations, where the side information quality is not necessarily monotonic along the scalable coding order. A partial result is provided for the doubly symmetric binary source under the Hamming distortion measure when the worse side information is a constant, for which one of the outer bounds is strictly tighter than the other.**

***Index Terms*—Scalable source coding, side information, successive refinement.**

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 such strong side information that only a minimal amount of 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 the description completely regardless of the quality of their individual side informations. However, for the users with good quality side informations (who will be 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 of the description, and the bad users receive both layers to decode. Moreover, it is of importance to investigate whether such a requirement introduces any performance loss. We call this coding strategy *side-information scalable* (SI-scalable) source coding, since the scalable coding order is from the good users to the bad users. In this work, we consider mostly two-layer systems (thus with only two users), except in the quadratic Gaussian case for which the solution to the general multilayer problem is given.

This work is related to the successive refinement source coding problem, where a source is to be encoded in a scalable manner to satisfy different distortion requirements at individual stages. This problem was studied by Koshelev [1], and by Equitz and Cover [2]; a complete characterization of the rate–distortion (R-D) region can be found in [3]. Also related is the R-D problem for source coding with side information at the decoder [4], for which Wyner and Ziv provided a conclusive result (now widely known as the Wyner–Ziv problem). Steinberg and Merhav [5] recently extended the successive refinement source coding problem to the Wyner–Ziv setting (SR-WZ), when the second stage side information Y_2 is better than that of the first stage Y_1, in the sense that $X \leftrightarrow Y_2 \leftrightarrow Y_1$ is 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 absent 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 and that in [5], [6], however without the scalable coding requirement.

Both the SR-WZ problem [5], [6] and the SI-scalable coding problem can be thought of as special cases of the scalable source coding problem with no specific structure imposed on the decoder side informations; this general problem appears to be quite difficult, since even without the scalable coding requirement, a complete solution has not been found [7]. Here we emphasize that the SR-WZ problem and the SI-scalable coding problem are different in several aspects, though they may seem similar. Roughly speaking, in the SI-scalable coding 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 coding problem, the side informations are degraded as $X \leftrightarrow Y_1 \leftrightarrow Y_2$, in contrast to

Manuscript received March 16, 2007; revised July 28, 2008. Current version published November 21, 2008. This work was supported in part by the Swiss National Competence Center on Mobile Information and Communication Systems (NCCR-MICS). The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, June 2007.

C. Tian was with the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, CH1015, Switzerland. He is now with AT&T Labs–Research, Florham Park, NJ 07932 USA (e-mail: tian@research.att.com)

S. N. Diggavi is with the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, CH1015, Switzerland (e-mail: suhas.diggavi@epfl.ch).

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

Color version of Figure 1 in this paper is available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.2006399

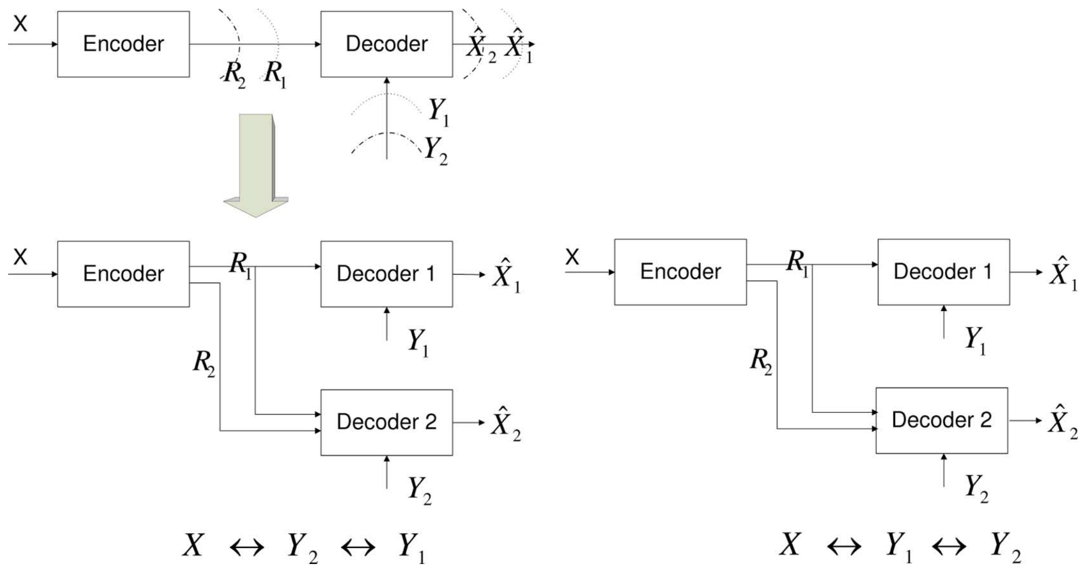

Fig. 1. The SR-WZ system versus the SI-scalable coding system.

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 coding problem is more applicable when multiple users exist in the network, and the server wants to provide a scalable description, such that the good user is not jeopardized unnecessarily (see Fig. 1).

Heegard and Berger [7] showed that when the scalable coding requirement is removed, the optimal encoding is in fact naturally progressive from the bad user to the good user; as such, the SI-scalable coding problem is expected to be more difficult than the SR-WZ problem, since the encoding order is reversed from the natural one. This difficulty is encapsulated by the fact that in the SR-WZ ordering the good user is able to decode whatever message that was meant for the bad user, and hence the information in the first stage message can be readily used by the second-stage decoder. However, in the SI-scalable coding problem an additional tension exists in the sense that the second-stage decoder will need extra information to disambiguate and decode the message for the first stage.

The problem is well understood for the lossless case. The key difference is that the quality of the side information can be naturally determined by the value of $H(X|Y)$ in this case. By the seminal work of Slepian and Wolf [9], we know $H(X|Y)$ is the minimum rate of encoding X losslessly with side information Y at the decoder, thus in a sense a larger value of $H(X|Y)$ corresponds to weaker side information. If $H(X|Y_1) < H(X|Y_2)$, then the rate pair $(R_1, R_2) = (H(X|Y_1), H(X|Y_2) - H(X|Y_1))$ 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 a rateless Slepian–Wolf low-density parity-check (LDPC) code. For the lossless case, there is no loss of optimality by using a scalable coding approach; an immediate question to ask is whether the same is true for the lossy case in terms of rate–distortion, which we will show to be not true in general. In this rate–distortion setting, the order of goodness by the value of $H(X|Y)$ is not sufficient due to the presence of the distortion constraints. This motivates the Markov condition $X \leftrightarrow Y_1 \leftrightarrow Y_2$ introduced for the SI-scalable coding problem. Going further along this point of view, the SI-scalable coding problem is also applicable in the single-user setting, when the encoder does not know exactly the joint distribution of the source and the decoder side information. Therefore, it can be viewed as a special case of the side-information-universal rate–distortion coding.

In this work, we formulate the problem of SI-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 bounds, and they differ only in the domain of optimization by a Markov string requirement. Though the inner and the outer bounds do not coincide in general, these bounds are indeed tight for the case when either the first stage or the second stage requires a lossless reconstruction, as well as for the case when certain deterministic distortion measures are taken. Furthermore, a conclusive result is given for the quadratic Gaussian source with any number of stages and arbitrarily correlated Gaussian side informations.

With this set of inner and outer bounds, the problem of *perfect scalability* is investigated, i.e., whether each of the stages can achieve the corresponding Wyner–Ziv bound; this is similar to the notion of (strictly) successive refinability in the SR-WZ problem [5], [6].[1] Necessary and sufficient conditions are derived for general discrete memoryless sources to be perfectly

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

scalable under a mild support condition. By using the tool of rate loss invented by Zamir [14], we further show that the gap between the inner and the outer bounds is bounded by some constants when the squared error distortion measure is used, and thus the bounds are "nearly sufficient," in the sense given in [15].

In addition to the result for the Gaussian source, a partial result is provided for the doubly symmetric binary source (DSBS) under the Hamming distortion measure when the second stage does not have side information. For this source, the inner and the outer bounds coincide in certain distortion regimes, but when they do not coincide, one of the outer bounds can be strictly better than the other.

The rest of the 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 region and show that the bounds coincide in certain special cases. The notion of perfect scalability is introduced in Section IV together with the example of a binary source. The rate loss method is applied in Section V to show the gap between the inner and the outer bounds is bounded from above. In Section VI, the Gaussian source is treated within a more general setting. 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 $x^n = (x_1, x_2, \ldots, x_n)$, 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 $P_{X^n}(x^n) = \prod_{i=1}^n P_X(x_i)$. 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 dropped 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 $d_j : \mathcal{X} \times \hat{\mathcal{X}}_j \rightarrow [0, \infty), j = 1, 2$ be two distortion measures. The single-letter distortion extension of d_j to vectors is defined as

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

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

$$\begin{aligned} \phi_1 &: \mathcal{X}^n \rightarrow I_{M_1} \\ \phi_2 &: \mathcal{X}^n \rightarrow I_{M_2} \\ \psi_1 &: I_{M_1} \times \mathcal{Y}_1^n \rightarrow \hat{\mathcal{X}}_1^n \\ \psi_2 &: I_{M_1} \times I_{M_2} \times \mathcal{Y}_2^n \rightarrow \hat{\mathcal{X}}_2^n \end{aligned}$$

where $I_k = \{1, 2, \ldots, k\}$, such that

$$\mathbb{E} d_1 \left(X^n, \psi_1(\phi_1(X^n), Y_1^n)\right) \leq D_1$$
$$\mathbb{E} d_2 \left(X^n, \psi_2(\phi_1(X^n), \phi_2(X^n), Y_2^n)\right) \leq D_2$$

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 scalable coding with side informations (Y_1, Y_2), if for any $\epsilon > 0$ and sufficiently large n, there exists an $(n, M_1, M_2, D_1 + \epsilon, D_2 + \epsilon)$ scalable code, such that $R_1 + \epsilon \geq \frac{1}{n} \log(M_1)$ and $R_2 + \epsilon \geq \frac{1}{n} \log(M_2)$.

We denote the collection of all (D_1, D_2)-achievable rate pairs (R_1, R_2) for scalable coding as $\mathcal{R}(D_1, D_2)$, and seek to characterize this region when $X \leftrightarrow Y_1 \leftrightarrow Y_2$ forms a Markov string (see a similar but reversed degradedness condition in [5], [6]); this special case is referred to as SI-scalable source coding. The Markov condition in effect specifies the *goodness* of the side informations.

The rate–distortion function for source coding with degraded decoder 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 SI-scalable code. More precisely, in order to achieve distortion D_1 with side information Y_1, and distortion D_2 with side information Y_2, when $X \leftrightarrow Y_1 \leftrightarrow Y_2$, the rate–distortion function[2] is

$$R_{\mathrm{HB}}(D_1, D_2) = \min_{p(D_1, D_2)} [I(X; W_2|Y_2) + I(X; W_1|W_2, Y_1)] \tag{1}$$

where $p(D_1, D_2)$ is the set of random variables $(W_1, W_2) \in \mathcal{W}_1 \times \mathcal{W}_2$ jointly distributed with the generic random variables (X, Y_1, Y_2), such that the following conditions are satisfied:[3] (i)$(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is a Markov string; (ii) there exist $\hat{X}_1 = f_1(W_1, Y_1)$ and $\hat{X}_2 = f_2(W_2, Y_2)$ which satisfy the distortion constraints. Notice that the rate–distortion function $R_{\mathrm{HB}}(D_1, D_2)$ given above suggests an encoding and decoding order from the bad user to the good user.

Let Γ_d be the set of distortion measures satisfying the following quite general condition:

$$\Gamma_d \triangleq \{d(\cdot, \cdot) : d(x, x) = 0, \text{ and } d(x, \hat{x}) > 0 \text{ if } \hat{x} \neq x\}. \tag{2}$$

Wyner and Ziv showed in [4] that if the distortion measure is in Γ_d, then $R^*_{X|Y}(0) = H(X|Y)$, 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 $d_1(\cdot, \cdot)$, i.e., $d_1(\cdot, \cdot) \in \Gamma_d$, then it can easily be shown (using an argument similar to the Remark 3 in [4]) that

$$R_{\mathrm{HB}}(0, D_2) = \min_{p(D_2)} [I(X; W_2|Y_2) + H(X|W_2, Y_1)] \tag{3}$$

where $p(D_2)$ is the set of all random variables 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 for the SI-scalable coding problem, we first examine a simple Gaussian source under the squared error distortion measure, and describe the coding schemes informally.

[2]Though the work by Heegard and Berger is best known for the case when "side information may be absent," they also addressed the problem in the more general setting when the side informations are degraded. From here on, we shall use $R_{\mathrm{HB}}(\cdot)$ in this more general sense. We shall also assume a higher value of the subscript in Y_k is associated with lower quality of the side information, unless specified otherwise; the distortion D_k is implicitly assumed to be associated with the side information Y_k. These conventions will become convenient when SR-WZ coding and SI-scalable coding need to be discussed together.

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

We note that though one particular important special case is when $D_1 = D_2$, considering this case alone in fact does not lead to straightforward simplification. In the sequel, we instead consider two extreme cases, which we believe indeed render the problem more transparent and thus facilitate understanding.

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. Note that $X \leftrightarrow Y_1 \leftrightarrow Y_2$ is indeed a Markov string. Though the quadratic Gaussian source is not a discrete source, the intuition gained from the Gaussian case usually carries well into the discrete case, which can be made formal by the result in [16], [17]. To avoid lengthy discussion on degenerate regimes, let us assume $\sigma_N^2 \approx \sigma_x^2$, and consider only the following extreme cases.

- $\sigma_x^2 \gg D_1 \gg D_2$: It is known as binning with a Gaussian codebook, generated using a single-letter mechanism (i.e., the n-fold product of an independent and identically distributed random variable) 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 codeword can be recovered by decoder two, who does not have any side information. Since $D_1 \gg D_2$, 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 conditioned on W_1 such that $D_2 = \mathbb{E}[X - \mathbb{E}(X|W_1, W_2)]^2$, is needed. This leads to the achievable rates
$$\begin{aligned} R_1 &\geq I(X; W_1|Y) \\ R_1 + R_2 &\geq I(X; W_1|Y) + I(W_1; Y) + I(X; W_2|W_1) \\ &= I(X; W_1, W_2). \end{aligned}$$
- $\sigma_x^2 \gg D_2 \gg D_1$: 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_1 is 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_1 is more than we need. To solve this problem, we construct a coarser description using a 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 with side information Y, while decoder two decodes the first and the third layer without side information. By the Markov string $X \leftrightarrow W_1 \leftrightarrow W_2$, this scheme gives the following rates:
$$\begin{aligned} R_1 &\geq I(X; W_1, W_2|Y) = I(X; W_1|Y) \\ R_1 + R_2 &\geq I(X; W_1|Y) + I(W_2; Y) \\ &= I(X; W_2) + I(X; W_1|Y, W_2). \end{aligned}$$

It is seen from the preceding discussion that the specific coding scheme depends 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 another auxiliary random variable, as will be shown in the sequel. However, it appears that a matching converse is indeed quite difficult to establish.

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 that the bounds are in fact tight for certain special cases.

A. Two Inner Bounds

Define the region $\mathcal{R}_{\text{in}}(D_1, D_2)$ 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 conditions 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 $f_j : \mathcal{W}_j \times \mathcal{Y}_j \to \hat{\mathcal{X}}_j$ such that
$$\mathbb{E} d_j(X, f_j(W_j, Y_j)) \leq D_j, \quad j = 1, 2.$$
3) The nonnegative rate pairs satisfy
$$\begin{aligned} R_1 &\geq I(X; V, W_1|Y_1) \\ R_1 + R_2 &\geq I(X; V, W_2|Y_2) + I(X; W_1|Y_1, V). \end{aligned}$$
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
$$\begin{aligned} &|\mathcal{V}| \leq |\mathcal{X}| + 3, |\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 1 \\ &|\mathcal{W}_2| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 1. \end{aligned}$$

The last two conditions can be removed without causing essential difference to the region $\mathcal{R}_{\text{in}}(D_1, D_2)$; 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 start by reducing 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 the distortion and rate expressions depend only on the marginals $P(X, V, W_1)$ and $P(X, V, W_2)$, respectively, hence requiring $W_1 \leftrightarrow (X, V) \leftrightarrow W_2$ to be a Markov string does not cause any loss of generality. Next, in order to reduce the cardinality of $\mathcal{W}_1$, it is seen that $|\mathcal{X}||\mathcal{V}| - 1$ letters are needed to preserve the joint distribution of (X, V), one more is needed to preserve the distortion value to satisfy the constraint 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 values of the other distortion and the other mutual information term because of the aforementioned Markov string. A similar argument holds for $|\mathcal{W}_2|$.

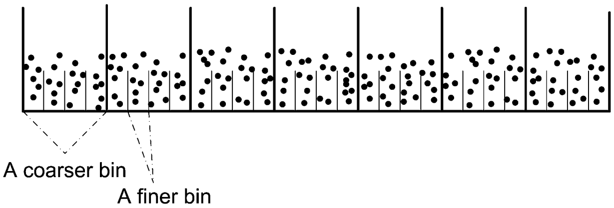

Fig. 2. An illustration of the codewords in the nested binning structure.

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

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

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

This theorem is proved in Appendix B, and here we outline the coding scheme for this achievable region in an intuitive manner. The encoder first encodes using a $\boldsymbol{V}$ codebook with 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 (coarser) 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 a random codebook of $\boldsymbol{W_2}$.

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. Note that this is fundamentally different from the coding structure in SR-WZ, where no nested binning is needed. Each of the coarser bin contains the same number of finer bins; each finer bin holds a 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, an additional message is received by the decoder, which further specifies one of the finer bins 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 the coarser bin index of the other codewords.

We note that the coding scheme does not explicitly require that side informations are degraded. Indeed, as long as the chosen random variable V satisfies $I(V;Y_1) \geq I(V;Y_2)$ as well as the Markov condition, the region is achievable. More precisely, the following corollary is straightforward.

Corollary 1: For any discrete memoryless stochastically source with side informations Y_1 and Y_2 (without the Markov structure), $\widetilde{\mathcal{R}}_{\text{in}}(D_1, D_2) \subseteq \mathcal{R}(D_1, D_2)$, where $\widetilde{\mathcal{R}}_{\text{in}}(D_1, D_2)$ is $\mathcal{R}_{\text{in}}(D_1, D_2)$ with the additional condition that $I(V;Y_1) \geq I(V;Y_2)$.

We can specialize the region $\mathcal{R}_{\text{in}}(D_1, D_2)$ to give another inner bound. Let $\hat{\mathcal{R}}_{\text{in}}\,(D_1, D_2)$ 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 conditions 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 $f_j : \mathcal{W}_j \times \mathcal{Y}_j \rightarrow \hat{\mathcal{X}}_j$ such that
$$\mathbb{E}d_j(X, f_j(W_j, Y_j)) \leq D_j, \quad j = 1, 2.$$
3) The nonnegative rate pairs satisfy
$$R_1 \geq I(X; W_1|Y_1)$$
$$R_1 + R_2 \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_2).$$
4) The alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy
$$|\mathcal{W}_1| \leq (|\mathcal{X}| + 3)(|\mathcal{X}|(|\mathcal{X}| + 3) + 1)$$
$$|\mathcal{W}_2| \leq (|\mathcal{X}| + 3)(|\mathcal{X}|(|\mathcal{X}| + 3) + 1).$$

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

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

The region $\hat{\mathcal{R}}_{\text{in}}\,(D_1, D_2)$ is particularly interesting for the following reasons. First, it can be explicitly matched back to the coding scheme for the simple Gaussian example given at the beginning of this section. Second, it will be shown that one of the outer bounds has the same rate and distortion expressions as $\hat{\mathcal{R}}_{\text{in}}(D_1, D_2)$, but with a relaxed Markov string requirement. We now prove this corollary.

Proof of Corollary 2: When $W_1 \leftrightarrow W_2 \leftrightarrow X$, let $V = W_1$. Then the rate expressions in Theorem 1 give

$$R_1 \geq I(X; W_1|Y_1)$$
$$R_1 + R_2 \geq I(X; V, W_2|Y_2) + I(X; W_1|V, Y_1) = I(X; W_2|Y_2).$$

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

$$R_1 \geq I(X; V, W_1|Y_1) = I(X; W_1|Y_1)$$
$$R_1 + R_2 \geq I(X; V, W_2|Y_2) + I(X; W_1|V, Y_1) = I(X; W_2|Y_2) + I(X; W_1|W_2, Y_1).$$

Therefore $\mathcal{R}_{\text{in}}\,(D_1, D_2) \supseteq \hat{\mathcal{R}}_{\text{in}}(D_1, D_2)$ for this case as well.

The cardinality bounds here are larger than those 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 informations $X \leftrightarrow Y_1 \leftrightarrow Y_2$

$$\mathcal{R}_{\cap}(D_1, D_2) = \{(R_1, R_2) : R_1 \geq R^*_{X|Y_1}(D_1), \\ R_1 + R_2 \geq R_{\rm HB}(D_1, D_2)\}. \quad (4)$$

A tighter outer bound is now given as follows: define the region $\mathcal{R}_{\rm out}(D_1, D_2)$ 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) $(W_1, W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$ is Markov string.
2) There exist deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \to \hat{\mathcal{X}}_j$ such that

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

3) The nonnegative rate vectors satisfy

$$R_1 \geq I(X; W_1|Y_1) \\ R_1 + R_2 \geq I(X; W_2|Y_2) + I(X; W_1|Y_1, W_2).$$

4) $|\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}| + 3) + 2, |\mathcal{W}_2| \leq |\mathcal{X}| + 3.$

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 $X \leftrightarrow Y_1 \leftrightarrow Y_2$

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

The first inclusion $\mathcal{R}_{\cap}(D_1, D_2) \supseteq \mathcal{R}_{\rm out}(D_1, D_2)$ is obvious, since $\mathcal{R}_{\rm out}(D_1, D_2)$ takes the same form as $R^*_{X|Y_1}(D_1)$ and $R_{\rm HB}(D_1, D_2)$ when the bounds on 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 C.

Note that the inner bound $\hat{\mathcal{R}}_{\rm in}(D_1, D_2)$ and the outer bound $\mathcal{R}_{\rm out}(D_1, D_2)$ have the same rate and distortion expressions and they differ only by a Markov string requirement (ignoring the nonessential cardinality bounds). Because of the difference in the domain of optimization, the two bounds may not produce the same rate region. This is quite similar to the case of lossy distributed 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 [18], but their rate and distortion expressions are the same.

C. Lossless Reconstruction at One Decoder

Since decoder one has better quality side information, it is reasonable for it to require a higher quality reconstruction. In the extreme case, decoder one might require a lossless reconstruction. Alternatively, from the point of view of universal coding, the decoder may require better quality reconstruction when the side information is good, but relax its requirement when the side information is in fact not as good. In this subsection, we consider the setting where either decoder one or decoder two requires a lossless reconstruction. We have the following theorem.

Theorem 3: If $D_1 = 0$ with $d_1(\cdot,\cdot) \in \Gamma_d$, or $D_2 = 0$ with $d_2(\cdot,\cdot) \in \Gamma_d$ (see (2) for the definition of Γ_d), then $\mathcal{R}(D_1, D_2) = \mathcal{R}_{\rm in}(D_1, D_2)$. More precisely, for the former case

$$\mathcal{R}(0, D_2) = \bigcup_{P_{W_2}(D_2)} \{(R_1, R_2) : R_1 \geq H(X|Y_1), \\ R_1 + R_2 \geq I(X; W_2|Y_2) + H(X|Y_1, W_2)\} \quad (5)$$

where $P_{W_2}(D_2)$ is the set of random variables satisfying the Markov string $W_2 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$, and having a deterministic function f_2 such that $\mathbb{E}d_2(X, f_2(W_2, Y_2)) \leq D_2$. For the latter case

$$\mathcal{R}(D_1, 0) = \bigcup_{P_{W_1}(D_1)} \{(R_1, R_2) : R_1 \geq I(X; W_1|Y_1), \\ R_1 + R_2 \geq H(X|Y_2)\} \quad (6)$$

where $P_{W_1}(D_1)$ is the set of random variables satisfying the Markov string $W_1 \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$, and having a deterministic function f_1 such that $\mathbb{E}d_1(X, f_1(W_1, Y_1)) \leq D_1$.

Proof of Theorem 3: For $D_1 = 0$, let $W_1 = X$ and $V = W_2$. The achievable rate vector implied by Theorem 1 is given by

$$R_1 \geq H(X|Y_1) \\ R_1 + R_2 \geq I(X; W_2|Y_2) + H(X|Y_1, W_2).$$

It is seen that this rate region is tight by the converse of Slepian–Wolf coding for rate R_1, and by (3) of Heegard–Berger coding for rate $R_1 + R_2$.

For $D_2 = 0$, let $W_1 = V$ and $W_2 = X$. The achievable rate vector implied by Theorem 1 is given by

$$R_1 \geq I(X; W_1|Y_1), R_1 + R_2 \geq H(X|Y_2).$$

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 $R^*_{X|Y_2}(0)$ with $d_2(\cdot,\cdot) \in \Gamma_d$ as given in [4]) for rate $R_1 + R_2$. □

Zero distortion under a distortion measure $d \in \Gamma_d$ can be interpreted as *lossless*, however, it is a weaker requirement than that the block error probability is arbitrarily small. Nevertheless, $\mathcal{R}(0, D_2)$ and $\mathcal{R}(D_1, 0)$ in (5) and (6) still provide valid outer bounds for the more stringent lossless definition. On the other hand, it is rather straightforward to specialize the coding scheme for these cases, and show that the same conclusion is true for lossless coding in this case. Thus, we have the following corollary.

Corollary 3: The rate region, when the first stage (respectively, the second stage), requires a lossless reconstruction in terms of an arbitrarily small block error probability is given by (5) (respectively, (6)).

The key difference from the general case when both stages are lossy is the elimination of the need to generate one of the codebooks using an auxiliary random variable, which simplifies the matter tremendously. For example, when $D_2 = 0$, in the second stage we only need to randomly assign $\boldsymbol{x}$ that is jointly typical with a given $\boldsymbol{w}_1$ to a bin directly, with the number of such $\boldsymbol{x}$ vectors being approximately $2^{nH(X|W_1)}$. 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, since both the encoder and decoder at the second stage have access to a side information vector $\boldsymbol{w}_1$, we see a conditional Slepian–Wolf code with decoder only side information Y_2 suffices.

D. Deterministic Distortion Measures

Another case of interest is when some functions of the source X are required to be reconstructed with an arbitrarily small distortion in terms of the Hamming distortion measure; see [19] for the corresponding case in the multiple description problem. More precisely, let $Q_i : \mathcal{X} \to \mathcal{Z}_i, i = 1, 2$ be two deterministic functions and denote $Z_i = Q_i(X)$. Consider the case that decoder i seeks to reconstruct Z_i with an arbitrarily small Hamming distortion.[4] The achievable region $\mathcal{R}_{\text{in}}$ is tight when the functions satisfy certain degradedness conditions as stated in the following theorem.

Theorem 4: Let the distortion measure be the Hamming distortion $d_H : \mathcal{Z}_i \times \mathcal{Z}_i \to \{0, 1\}$ for $i = 1, 2$.

1) If there exists a deterministic function $Q' : \mathcal{Z}_1 \to \mathcal{Z}_2$ such that $Q_2 = Q' \cdot Q_1$, then $\mathcal{R}(0, 0) = \mathcal{R}_{\text{in}}(0, 0)$. More precisely

$$\begin{aligned}\mathcal{R}(0,0) = \{(R_1, R_2) : R_1 &\geq H(Z_1|Y_1)\\ R_1 + R_2 &\geq H(Z_2|Y_2) + H(Z_1|Y_1Z_2)\}. \quad (7)\end{aligned}$$

2) If there exists a deterministic function $Q' : \mathcal{Z}_2 \to \mathcal{Z}_1$ such that $Q_1 = Q' \cdot Q_2$, then $\mathcal{R}(0, 0) = \mathcal{R}_{\text{in}}(0, 0)$. More precisely

$$\begin{aligned}\mathcal{R}(0,0) = \{(R_1, R_2) : R_1 &\geq H(Z_1|Y_1),\\ R_1 + R_2 &\geq H(Z_2|Y_2)\}. \quad (8)\end{aligned}$$

Proof of Theorem 4: To prove (7), first observe that by letting $W_1 = Z_1$ and $V = W_2 = Z_2, \mathcal{R}_{\text{in}}$ clearly reduces to the given expressions. For the converse, we start from the outer bound $\mathcal{R}_{\text{out}}(0, 0)$, which implies that Z_1 is a function of W_1 and Y_1, and Z_2 is a function of W_2 and Y_2. For the first-stage rate R_1, we have the following chain of equalities:

$$\begin{aligned}R_1 &\geq I(X;W_1|Y_1) = I(X;W_1Z_1|Y_1) \geq I(X;Z_1|Y_1)\\ &= H(Z_1|Y_1) - H(Z_1|X,Y_1) = H(Z_1|Y_1). \quad (9)\end{aligned}$$

For the sum rate, we have

$$\begin{aligned}R_1 + R_2 &\geq I(X;W_2|Y_2) + I(X;W_1|W_2Y_1)\\ &= I(X;W_2Z_2|Y_2) + I(X;W_1|W_2Y_1)\\ &= I(X;Z_2|Y_2) + I(X;W_2|Y_2Z_2)\\ &\quad + I(X;W_1|W_2Y_1)\\ &= H(Z_2|Y_2) + I(X;W_2|Y_2Z_2)\\ &\quad + I(X;W_1|W_2Y_1)\\ &\overset{(a)}{\geq} H(Z_2|Y_2) + I(X;W_2|Y_1Y_2Z_2)\\ &\quad + I(X;W_1|W_2Y_1)\\ &\overset{(b)}{=} H(Z_2|Y_2) + I(X;W_2|Y_1Y_2Z_2)\\ &\quad + I(X;W_1|W_2Y_1Y_2)\\ &= H(Z_2|Y_2) + I(X;W_2|Y_1Y_2Z_2)\\ &\quad + I(X;W_1|W_2Y_1Y_2Z_2)\\ &= H(Z_2|Y_2) + I(X;W_1W_2|Y_1Y_2Z_2)\\ &\geq H(Z_2|Y_2) + I(X;Z_1|Y_1Y_2Z_2)\\ &= H(Z_2|Y_2) + H(Z_1|Y_1Y_2Z_2)\\ &\overset{(c)}{=} H(Z_2|Y_2) + H(Z_1|Y_1Z_2)\end{aligned}$$

where (a) is due to the Markov string $W_2 \leftrightarrow X \leftrightarrow (Y_1Y_2)$ and Z_2 is function of X; (b) is due to the Markov string $(W_1W_2) \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2$; (c) is due to the Markov string $(Z_1, Z_2) \leftrightarrow Y_1 \leftrightarrow Y_2$.

Proof of part 2), i.e., the expressions in (8), is straightforward and thus omitted. □

Clearly, in the converse proof above, the requirement that the functions Q_1 and Q_2 are degraded is not needed. Indeed, this outer bound holds for any general functions, however, the degradedness is needed for establishing the achievability of the region. If the coding is not necessarily scalable, then it can be seen that the sum rate is indeed achievable, and the proof of Theorem 4 can be used to establish a nontrivial special result in the context of the problem treated by Heegard and Berger [7].

Corollary 4: Let the two functions Q_1 and Q_2 be arbitrary, and let the distortion measure be the Hamming distortion $d_H : \mathcal{Z}_i \times \mathcal{Z}_i \to \{0, 1\}$ for $i = 1, 2$, then we have

$$R_{\text{HB}}(0,0) = H(Z_2|Y_2) + H(Z_1|Y_1Z_2).$$

IV. Perfect Scalability and a Binary Source

In this section, we introduce the notion of perfect scalability, which can intuitively be defined as when both stages operate at the Wyner–Ziv rates. We further examine the doubly symmetric binary source, for which a partial characterization of the rate–distortion region is provided and the scalability issue is discussed.

A. Perfect Scalability

The notion of (strictly) successive refinability defined in [5] for the SR-WZ problem with forwardly degraded side informations can be applied to the reversely degraded case considered in this paper. This is done by introducing the notion of perfect scalability for the SI-scalable coding problem.

Definition 3: A source X is said to be *perfectly 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 5: A source X with side informations under the Markov string $X \leftrightarrow Y_1 \leftrightarrow Y_2$, for which $\exists\, y_1 \in \mathcal{Y}_1$ such that $P_{XY_1}(x, y_1) > 0$ for each $x \in \mathcal{X}$, is perfectly scalable for the distortion pair (D_1, D_2) if and only if there exist random

[4]By a similar argument as in the last subsection, the same result holds if block error probability is made arbitrarily small.

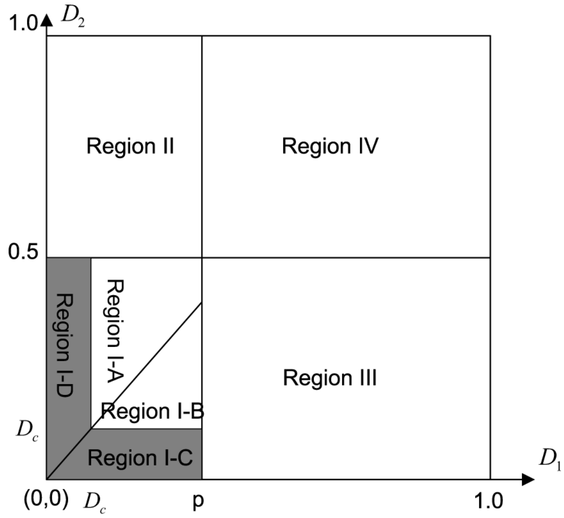

Fig. 3. The partition of distortion region for the doubly symmetric binary source, where d_c is the critical distortion in [4].

variables (W_1, W_2) and deterministic maps $f_j : \mathcal{W}_j \times \mathcal{Y}_j \rightarrow \hat{\mathcal{X}}_j$ 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 alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy $|\mathcal{W}_1| \leq |\mathcal{X}|(|\mathcal{X}|+3)+2$ and $|\mathcal{W}_2| \leq |\mathcal{X}| + 3$.

The Markov string is the most crucial condition, and the substring $W_1 \leftrightarrow W_2 \leftrightarrow X$ is the same as one of the conditions for successive refinability without side information [2], [3]. The support condition essentially requires the existence of a worst letter y_1 in the alphabet $\mathcal{Y}_1$ such that it has nonzero probability mass for each (x, y_1) pair, $x \in \mathcal{X}$. We note here that usually such necessary and sufficient conditions can only be derived when a complete characterization of the rate–distortion region is available. However, for the SI-scalable coding problem, the inner and the outer bounds do not match in general, and it is somewhat surprising that the conditions can still be given under a mild support requirement.

Proof of Theorem 5: The sufficiency being trivial by Corollary 2, we only prove the necessity. Without loss of generality, assume $P_X(x) > 0$ for all $x \in \mathcal{X}$. 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 by the tighter outer bound $\mathcal{R}_{\text{out}}(D_1, D_2)$ of Theorem 2, there exist random variables W_1, W_2 in finite alphabets, whose sizes are 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

$$
\begin{aligned}
R^*_{X|Y_1}(D_1) &\geq I(X;W_1|Y_1) \\
R^*_{X|Y_2}(D_2) &\geq I(X;W_2|Y_2) + I(X;W_1|Y_1, W_2). \qquad (10)
\end{aligned}
$$

It follows that

$$
\begin{aligned}
R^*_{X|Y_2}(D_2) &\geq I(X;W_2|Y_2) + I(X;W_1|Y_1, W_2) \\
&\geq I(X;W_2|Y_2) \overset{(a)}{\geq} R^*_{X|Y_2}(D_2), \qquad (11)
\end{aligned}
$$

where (a) is due to 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 (11), which implies $I(X;W_1|Y_1, W_2) = 0$. Similarly, we have

$$R^*_{X|Y_1}(D_1) \geq I(X;W_1|Y_1) \geq R^*_{X|Y_1}(D_1) \qquad (12)$$

thus (12) also holds with equality.

Notice that if $W_1 \leftrightarrow W_2 \leftrightarrow X$ were a Markov string, then we could have already completed the proof at this point. However, this Markov condition is not true in general, and this is where the support condition is needed.

For convenience, define the set

$$F(w_2) = \{x \in \mathcal{X} : P(x, w_2) > 0\}. \qquad (13)$$

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:

$$P(w_1, w_2, x, y_1) = P(x, y_1)P(w_2|x)P(w_1|x, w_2).$$

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:

$$
\begin{aligned}
P(w_1, w_2, x, y_1) &= P(x, y_1, w_2)p(w_1|y_1, w_2) \\
&\overset{(a)}{=} P(x, y_1)P(w_2|x)P(w_1|y_1, w_2)
\end{aligned}
$$

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

$$P(w_2^*|x)P(w_1^*|x, w_2^*) = P(w_2^*|x)P(w_1^*|y_1, w_2^*)$$

for any $x \in \mathcal{X}$. Thus, for any $x \in F(w_2^*)$ (see definition in (13)) such that $P(w_1^*|x, w_2^*)$ is well defined, we have

$$p(w_1^*|y_1, w_2^*) = p(w_1^*|x, w_2^*)$$

and it further implies

$$
\begin{aligned}
p(w_1^*|w_2^*) &= \frac{\sum_x P(x, w_1^*, w_2^*)}{\sum_x P(x, w_2^*)} \\
&= \frac{\sum_{x \in F(w_2^*)} P(x, w_2^*)P(w_1^*|y_1, w_2^*)}{\sum_x P(x, w_2^*)} \\
&= p(w_1^*|y_1, w_2^*) = p(w_1^*|x, w_2^*)
\end{aligned}
$$

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. □

B. The Doubly Symmetric Binary Source

Consider the following source: X is a memoryless binary source $X \in \{0, 1\}$ and $P(X = 0) = 0.5$. The distortion measure is the Hamming distortion. The first-stage side information Y can be taken as the output of a binary-symmetric channel (BSC) with input X, and crossover probability $p < 0.5$. The second stage does not have side information. This source clearly satisfies the support condition in Theorem 5, and for some distortion pairs, this source is perfectly scalable, while for others this is not possible. We next provide partial results using the afore-given inner bound $\hat{\mathcal{R}}_{\text{in}}(D_1, D_2)$ and outer bound $\mathcal{R}_\cap(D_1, D_2)$.

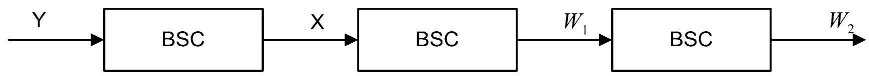

Fig. 4. The forward test channel in Region I-D. The crossover probability for the BSC between X and W_1 is D_1, while the crossover probability η for the BSC between W_1 and W_2 is such that $D_1 * \eta = D_2$.

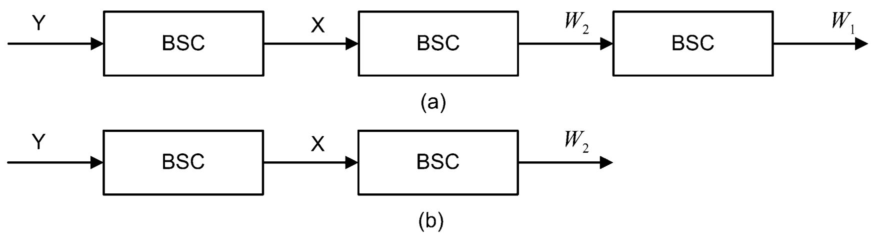

Fig. 5. The forward test channels in Region I-C. The crossover probability for the BSC between X and W_2 is D_2 in both the channels, while the crossover probability η for the BSC between W_2 and W_1 in (a) is such that $D_2 \leq D_1 * \eta = \eta' \leq d_c$. Note for (b), W_1 can be taken as a constant.

An explicit calculation of $R_{\rm HB}(D_1, D_2)$, together with the optimal forward test channel structure, was given in a recent work [6]. With this explicit calculation, it can be shown that in the shaded region in Fig. 3, the outer bound $\mathcal{R}_\cap(D_1, D_2)$ is in fact achievable (as well as in Regions 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]

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

where $G(u) = h_b(p * u) - h_b(u), h_b(u)$ is the binary entropy function $h_b(u) = -u \log u - (1-u)\log(1-u)$, 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 need the following result from [6].

Theorem 6 ([6]): For distortion pairs (D_1, D_2) such that $0 \leq D_2 \leq 0.5$ and $0 \leq D_1 < \min(d_c, D_2)$ (i.e., Region I-D)

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

This result implies that for the shaded region I-D, the optimal forward test channel is in fact a cascade of two BSC channels depicted in Fig. 4. This choice clearly satisfies the condition in Corollary 2 with the rates given by the outer bound $\mathcal{R}_\cap(D_1, D_2)$, which implies that this outer bound is indeed achievable. Note the following inequality:

$$\begin{aligned} R_{\rm HB}(D_1, D_2) &= 1 - h_b(D_2 * p) + h_b(p * D_1) - h_b(D_1) \\ &> 1 - h_b(D_2) = R(D_2) \end{aligned}$$

where the strict inequality is due to the strict monotonicity of $G(u)$ in $0 \leq u \leq 0.5$, and we conclude that in this region 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 a 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 in region I-C, $\mathcal{R}_\cap(D_1, D_2)$ is achievable by time-sharing the two forward test channels in Fig. 5; furthermore, an equivalent forward test channel can be found such that the Markov condition $W_1' \leftrightarrow W_2 \leftrightarrow X$ holds, which satisfies the conditions given in Theorem 5. Thus, in this distortion region, the source is in fact perfectly scalable.

Unfortunately, we are not able to find a complete characterization of the rate region in the distortion regions I-A and I-B. Using an approach similar to [6], an explicit outer bound can be derived from $\mathcal{R}_{\rm out}(D_1, D_2)$. It can then be shown numerically that for certain distortion pairs in these regions, $\mathcal{R}_{\rm out}(D_1, D_2)$ is strictly tighter than $\mathcal{R}_\cap(D_1, D_2)$. This calculation can be found in [20] and is omitted here.

V. A Near-Sufficiency Result

By using the tool of rate loss invented by Zamir [14], which was further developed in [15], [21]–[23], it can be shown that when both the source and reconstruction alphabets are reals, and the distortion measure is the squared error distortion, the gap between the inner and the outer bounds is bounded by some constants. Thus, these bounds are nearly sufficient in the sense defined in [15]. To show this result, we distinguish the two cases: $D_1 \geq D_2$ and $D_1 \leq D_2$. The source X is assumed to have finite variance σ_x^2 and finite (differential) entropy. The result of this section is summarized in Fig. 6.

A. The Case $D_1 \geq D_2$

Construct two random variables $W_1' = X + N_1 + N_2$ and $W_2' = X + N_2$, where N_1 and N_2 are zero-mean independent Gaussian random variables, independent of everything else, with variances σ_1^2 and σ_2^2 such that $\sigma_1^2 + \sigma_2^2 = D_1$ and $\sigma_2^2 = D_2$. From Corollary 1, it is obvious that the following rates are achievable for distortion (D_1, D_2):

$$R_1 = I(X; X + N_1 + N_2 | Y_1) \tag{14}$$

$$R_1 + R_2 = I(X; X + N_2 | Y_2). \tag{15}$$

Let U be the optimal random variable achieving the Wyner–Ziv rate at distortion D_1 given decoder side information

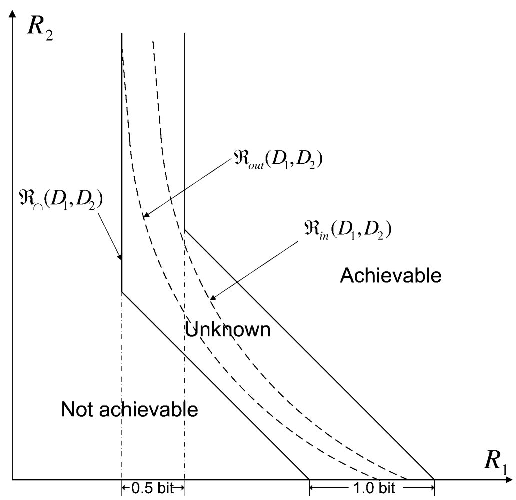

Fig. 6. An illustration of the inner bounds and the outer bounds when the squared error distortion is the distortion measure. The two regions $\mathcal{R}_{\rm in}(D_1,D_2)$ and $\mathcal{R}_{\rm out}(D_1,D_2)$ are given in dashed lines, since it is unknown whether they are indeed the same.

Y_1. Then it is clear that the difference between R_1 and the Wyner–Ziv rate can be bounded as

$$\begin{aligned}
&I(X;X+N_1+N_2|Y_1)-I(X;U|Y_1)\\
&\overset{(a)}{=} I(X;X+N_1+N_2|UY_1)\\
&\quad -I(X;U|Y_1,X+N_1+N_2)\\
&\le I(X;X+N_1+N_2|UY_1)\\
&= I(X-\hat{X}_1;X-\hat{X}_1+N_1+N_2|UY_1)\\
&\le I(X-\hat{X}_1,U,Y_1;X-\hat{X}_1+N_1+N_2)\\
&= I(X-\hat{X}_1;X-\hat{X}_1+N_1+N_2)\\
&\quad + I(U,Y_1;X-\hat{X}_1+N_1+N_2|X-\hat{X}_1)\\
&= I(X-\hat{X}_1;X-\hat{X}_1+N_1+N_2)\\
&\overset{(b)}{\le} \frac{1}{2}\log_2\frac{D_1+D_1}{D_1}=0.5
\end{aligned}\tag{16}$$

where (a) is by applying the chain rule to $I(X;X+N_1+N_2,U|Y_1)$ in two different ways; (b) holds 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 optimal random variables achieving the rate–distortion function $R_{\rm HB}(D_1,D_2)$. Due to the Markov string $(W_1,W_2)\leftrightarrow X\leftrightarrow Y_1\leftrightarrow Y_2$ and the fact that (N_1,N_2) are independent of (X,Y_1,Y_2), we have the identity

$$\begin{aligned}
&I(X;W_2|Y_2)+I(X;W_1|W_2Y_1)\\
&\qquad = I(X;W_1W_2|Y_1)+I(Y_1;W_2|Y_2)
\end{aligned}\tag{17}$$

and the identity

$$\begin{aligned}
&I(X;X+N_2|Y_2)\\
&\qquad = I(X;X+N_2|Y_1)+I(Y_1;X+N_2|Y_2).
\end{aligned}\tag{18}$$

Thus, we can bound the difference between the sum-rate R_1+R_2 (as given in (15)) and $R_{\rm HB}(D_1,D_2)$ as follows.

$$\begin{aligned}
&I(X;X+N_2|Y_2)-I(X;W_2|Y_2)-I(X;W_1|W_2Y_1)\\
&\quad =[I(X;X+N_2|Y_1)-I(X;W_1W_2|Y_1)]\\
&\qquad +[I(Y_1;X+N_2|Y_2)-I(Y_1;W_2|Y_2)].
\end{aligned}\tag{19}$$

To bound the first bracket in (19), notice that

$$\begin{aligned}
&I(X;X+N_2|Y_1)-I(X;W_1W_2|Y_1)\\
&= I(X;X+N_2|W_1W_2Y_1)\\
&\quad -I(X;W_1W_2|Y_1,X+N_2)\\
&\le I(X;X+N_2|W_1W_2Y_1)\\
&\overset{(a)}{=} I(X;X+N_2|W_1W_2Y_1Y_2)\\
&= I(X-\hat{X}_2;X-\hat{X}_2+N_2|W_1W_2Y_1Y_2)\\
&\le I(X-\hat{X}_2,W_1,W_2,Y_1,Y_2;X-\hat{X}_2+N_2)\\
&= I(X-\hat{X}_2;X-\hat{X}_2+N_2)\\
&\quad + I(W_1,W_2,Y_1,Y_2;X-\hat{X}_2+N_2|X-\hat{X}_2)\\
&= I(X-\hat{X}_2;X-\hat{X}_2+N_2)\\
&\le \frac{1}{2}\log_2\frac{D_2+D_2}{D_2}=0.5
\end{aligned}\tag{20}$$

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 (16). To bound the second bracket in (19), we write the following:

$$\begin{aligned}
&I(Y_1;X+N_2|Y_2)-I(Y_1;W_2|Y_2)\\
&\quad = I(Y_1;X+N_2|W_2Y_2)-I(Y_1;W_2|Y_2,X+N_2)\\
&\quad \le I(Y_1;X+N_2|W_2Y_2)\\
&\quad \le I(XY_1;X+N_2|W_2Y_2)\\
&\quad = I(X;X+N_2|W_2Y_2)\\
&\quad \le \frac{1}{2}\log_2\frac{D_2+D_2}{D_2}=0.5
\end{aligned}\tag{21}$$

Thus, we have shown that for the case $D_1\ge D_2$, the gap between the outer bound $\mathcal{R}_\cap(D_1,D_2)$ and the inner bound $\mathcal{R}_{\rm in}(D_1,D_2)$ 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 $D_1\le D_2$

Construct two random variables $W_1'=X+N_1$ 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 variances σ_1^2 and σ_2^2 such that $\sigma_1^2=D_1$ and $\sigma_1^2+\sigma_2^2=D_2$. By Corollary 1, it is easily seen that the following rates are achievable for distortion (D_1,D_2):

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

Clearly, the argument for bounding the gap on R_1 still holds with minor changes from the previous case. To bound the sum-rate gap, notice the following identity:

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

Next, we seek to upper-bound the following quantity:

$$\begin{aligned} & I(X;X+N_1+N_2|Y_2) \\ & \quad + I(X;X+N_1|Y_1,X+N_1+N_2) \\ & \quad - I(X;W_2|Y_2) - I(X;W_1|W_2Y_1) \\ & = [I(X;X+N_1|Y_1) - I(X;W_1W_2|Y_1)] \\ & \quad + [I(Y_1;X+N_1+N_2|Y_2) - I(Y_1;W_2|Y_2)] \end{aligned} \quad (22)$$

where again W_1, W_2 are the R-D optimal random variables for $R_{\rm HB}(D_1, D_2)$, and we have used the identity in (17). For the first bracket in (22), we have

$$\begin{aligned} & I(X;X+N_1|Y_1) - I(X;W_1W_2|Y_1) \\ & = I(X;X+N_1|W_1W_2Y_1) - I(X;W_1W_2|Y_1,X+N_1) \\ & \le I(X;X+N_1|W_1W_2Y_1) \\ & = I(X-\hat{X}_1;X-\hat{X}_1+N_2|W_1W_2Y_1) \\ & \quad \le I(X-\hat{X}_1,W_1,W_2,Y_1;X-\hat{X}_1+N_2) \\ & = I(X-\hat{X}_1;X-\hat{X}_1+N_1) \\ & \quad + I(W_1,W_2,Y_1;X-\hat{X}_1+N_1|X-\hat{X}_1) \\ & = I(X-\hat{X}_1;X-\hat{X}_1+N_1) \\ & \le \frac{1}{2}\log\frac{D_1+D_1}{D_1} = 0.5 \end{aligned} \quad (23)$$

where $\hat{X}_1$ is the decoding function given (W_1, Y_1). For the second bracket, following a similar approach as (21), we have

$$\begin{aligned} & I(Y_1;X+N_1+N_2|Y_2) - I(Y_1;W_2|Y_2) \\ & \quad \le I(X;X+N_1+N_2|W_2Y_2) \\ & \quad \le I(X-\hat{X}_2,W_2,Y_2;X-\hat{X}_2+N_1+N_2) \\ & \quad = I(X-\hat{X}_2;X-\hat{X}_2+N_1+N_2) \le 0.5. \end{aligned}$$

Thus, we conclude that for both cases the gap between the inner bound and the outer bound is bounded. Fig. 6 illustrates the inner bounds and the outer bounds, as well as the gap in between.

VI. The Quadratic Gaussian Source With Jointly Gaussian Side Informations

The degraded side information assumption, either forwardly or reversely, is especially interesting for the quadratic jointly Gaussian case, since physical degradedness and stochastic degradedness [24] do not cause essential difference in terms of the rate–distortion region for the problem being considered [5]. Moreover, since jointly Gaussian source and side informations are always statistically degraded, these forwardly and reversely degraded cases together provide a complete solution to the jointly Gaussian case with two decoders.

More precisely, let $Y_1 = X + N_1$ and $Y_2 = Y_1 + N_2$, where X is the Gaussian source, and N_1, N_2 are independent Gaussian random variables with variances σ_1^2 and σ_2^2, respectively, which are also independent of X. The distortion constraints are associated with the side informations, as D_1 and D_2, respectively.

- In the SR-WZ setting [5], the progressive coding order is from Y_2 to Y_1, and the rates

$$(R_1, R_2) = (R^*_{X|Y_2}(D_2), R_{\rm HB}(D_1,D_2) - R^*_{X|Y_2}(D_2))$$

  are achievable and optimal [6]. Perfect scalability (strictly successive refinability) is not possible in this setting, unless $R^*_{X|Y_2}(D_2) = 0$ or $N_2 = 0$, as discussed in [6].
- In the SI-scalable setting, the progressive coding order is from Y_1 to Y_2. As will be shown in the sequel, the rates

$$(R_1, R_2) = (R^*_{X|Y_1}(D_1), R_{\rm HB}(D_1,D_2) - R^*_{X|Y_1}(D_1))$$

  are achievable and optimal. Depending on the distortion constraints D_1 and D_2, two cases may occur.
  - In Corollary 2, we have $W_1 \leftrightarrow W_2 \leftrightarrow X$, which corresponds to the first coding scheme discussed at the beginning of Section III. Perfect scalability always holds, which includes the particularly important case when $D_1 = D_2$.
  - In Corollary 2, we have $W_2 \leftrightarrow W_1 \leftrightarrow X$, which corresponds to the second coding scheme discussed at the beginning of Section III. Perfect scalability does not hold in this case, except when $W_2 = W_1$.

In the remainder of this section, we in fact consider a more general setting with an arbitrary number of decoders for jointly Gaussian source and multiple side informations. Though the source and side informations can have arbitrary correlation, in light of the discussion above, we will only treat the case with physically degraded side informations. Note that since a specific encoding order is specified, though the side informations are degraded as an unordered set, the quality of the side informations may not be monotonic along the scalable coding order. Recall from Theorem 2 (see (4)) that $\mathcal{R}_\cap(D_1, D_2)$ is an outer bound derived from the intersection of the Heegard–Berger and Wyner–Ziv bounds. The generalization of this outer bound $\mathcal{R}_\cap(D_1, D_2)$ to the case of N decoders plays an important role, and we first take a detour in Section VI-A to give a characterization of $R_{\rm HB}(D_1, D_2, \ldots, D_N)$ for the jointly Gaussian case.

A. $R_{\rm HB}(D_1, D_2, \ldots, D_N)$ *for the Jointly Gaussian Case*

Consider the following source $X \sim \mathcal{N}(0, \sigma_x^2)$ and side informations $Y_k = X + \sum_{i=1}^k N_i$, where $N_i \sim \mathcal{N}(0, \sigma_i^2)$ are mutually independent and independent of X. The result by Heegard and Berger [7] gives

$$R_{\rm HB}(D_1,D_2,\ldots,D_N) = \min_{p(D_1,D_2,\ldots,D_N)} \sum_{k=1}^{N} I(X;W_k|Y_k,W_{k+1},W_{k+2},\ldots,W_N) \quad (24)$$

where $p(D_1, D_2, \ldots, D_N)$ is the set of random variables with the Markov string $(W_1, W_2, \ldots, W_N) \leftrightarrow X \leftrightarrow (Y_1, Y_2, \ldots, Y_N)$, such that deterministic functions

$f_k(Y_k, W_k, W_{k+1}, \ldots, W_N), k = 1, \ldots, N$ exist which satisfy the distortion constraints. In [6], the solution for $N = 2$ was calculated explicitly, however, such an explicit calculation appears quite involved for general N due to the discussion of various cases when some of the distortion constraints cannot be met with equality. In the sequel, we approach the problem differently by showing that a jointly Gaussian forward test channel is optimal.

Note that if we choose to enforce only a subset of the distortion constraints, a lower bound to the rate under such a restrictive set of distortion constraints gives a lower bound to $R_{\mathrm{HB}}(D_1, D_2, \ldots, D_N)$. By taking all the nonempty subsets of the distortion constraints, labeled by the subsets of $I_N = \{1, 2, \ldots, N\}$, a total of $2^N - 1$ lower bounds are available and clearly the maximum of them is also a lower bound to $R_{\mathrm{HB}}(D_1, D_2, \ldots, D_N)$. More precisely, we are interested in $\max R^*_{\mathrm{HB}}(A_D)$, where $A_D \subseteq I_N$ and $R^*_{\mathrm{HB}}(A_D)$ is defined in the sequel explicitly in terms of the distortion constraints only; note that if $i \in A_D, D_i$ is still the distortion constraint for the decoder with side information Y_i. We next derive one of these lower bounds using all the constraints $(D_1, D_2, \ldots, D_N)$, i.e., $A_D = I_N$; a similar derivation applies to the case with any subset $A_D \subset I_N$. Using (24) we have

$$\begin{aligned}
&\sum_{k=1}^{N} I(X; W_k | Y_k, W_{k+1}, W_{k+2}, \ldots, W_N) \\
&\quad = h(X|Y_N) - h\left(X|Y_1 W_1^N\right) \\
&\qquad - h(X|Y_N W_N) + h(X|Y_{N-1} W_N) \\
&\qquad - h\left(X|Y_{N-1} W_{N-1}^N\right) + \cdots + h\left(X|Y_1 W_2^N\right) \\
&\quad \overset{(a)}{=} h(X|Y_N) - h\left(X|Y_1 W_1^N\right) \\
&\qquad - [h(X|Y_N W_N) - h(X|Y_{N-1} Y_N W_N)] \\
&\qquad - \cdots - \left[h\left(X|Y_2 W_2^N\right) - h\left(X|Y_1 Y_2 W_2^N\right)\right] \\
&\quad = h(X|Y_N) - h\left(X|Y_1 W_1^N\right) - I(X; Y_{N-1}|Y_N W_N) \\
&\qquad - I\left(X; Y_{N-2}|Y_{N-1} W_{N-1}^N\right) - \cdots - I\left(X; Y_1 | Y_2 W_2^N\right) \\
&\quad \overset{(b)}{=} h(X|Y_N) - h\left(X|Y_1 W_1^N\right) \\
&\qquad - [h(Y_{N-1}|Y_N W_N) - h(Y_{N-1}|X Y_N)] \\
&\qquad - \cdots - \left[h\left(Y_1|Y_2 W_2^N\right) - h(Y_1|Y_2 X)\right] \\
&\quad = h(X|Y_N) + \sum_{k=2}^{N} h(Y_{k-1}|X Y_k) \\
&\qquad - \sum_{k=2}^{N} h\left(Y_{k-1}|Y_k W_k^N\right) - h\left(X|Y_1, W_1^N\right) \qquad (25)
\end{aligned}$$

where we have (a) because of the Markov string $X \leftrightarrow (Y_{k-1} W_k^N) \leftrightarrow Y_k$, and (b) because of the Markov string $W_k^N \leftrightarrow (X Y_k) \leftrightarrow Y_{k-1}$, both of which are consequences of $W_k^N \leftrightarrow X \leftrightarrow Y_{k-1} \leftrightarrow Y_k$. The first two terms in (25) depend only on the source distribution $P_{XY_1 \ldots Y_N}$, and we now seek to bound the latter two terms, for which we have

$$\begin{aligned}
h\left(X|Y_1 W_1^N\right) &= h\left(X - \mathbb{E}\left(X|Y W_1^N\right)|Y W_1^N\right) \\
&\le h\left(X - \mathbb{E}\left(X|Y W_1^N\right)\right) \le h(\mathcal{N}(0, D_1)) \\
&= \frac{1}{2}\log(2\pi e D_1) \qquad (26)
\end{aligned}$$

where the second inequality holds because the Gaussian distribution maximizes the entropy for a given second moment, and $\mathbb{E}(X - \mathbb{E}(X|Y W_1^N))^2 \le D_1$ by the existence of the decoding function f_1. Next define

$$\gamma_k = \frac{\sum_{i=1}^{k-1} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2}, \quad k = 2, 3, \ldots, N$$

and write the following:

$$\begin{aligned}
Y_{k-1} = X + \sum_{i=1}^{k-1} N_i &= X + \sum_{i=1}^{k-1} N_i + \gamma_k \sum_{i=1}^{k} N_i - \gamma_k \sum_{i=1}^{k} N_i \\
&= \gamma_k \left(X + \sum_{i=1}^{k} N_i\right) + (1 - \gamma_k) X \\
&\quad + \left[\sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i\right] \\
&= \gamma_k Y_k + (1 - \gamma_k) X + \left[\sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i\right].
\end{aligned}$$

Notice that

$$\mathbb{E}\left[Y_k \left(\sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i\right)\right] = \sum_{i=1}^{k-1} \sigma_i^2 - \gamma_k \sum_{i=1}^{k} \sigma_i^2 = 0$$

and Y_k and $(\sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i)$ are jointly Gaussian, which together imply that they are independent. Furthermore, because $(\sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i)$ is independent of X, the Markov string $(Y_1, Y_2, \ldots Y_N) \leftrightarrow X \leftrightarrow (W_1, W_2, \ldots, W_N)$ implies that it is also independent of $(X, Y_k, W_1, W_2, \ldots, W_N)$. It follows that

$$\begin{aligned}
&h\left(Y_{k-1}|Y_k W_k^N\right) \\
&\quad = h\left(\gamma_k Y_k + (1 - \gamma_k) X + \sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i \,\middle|\, Y_k W_k^N\right) \\
&\quad = h\left((1 - \gamma_k) X + \sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i \,\middle|\, Y_k W_k^N\right) \\
&\quad = h\left((1 - \gamma_k)(X - \mathbb{E}\left(X|Y_k W_k^N\right)) + \sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i \,\middle|\, Y_k W_k^N\right) \\
&\quad \le h\left((1 - \gamma_k)\left(X - \mathbb{E}(X|Y_k W_k^N)\right) + \sum_{i=1}^{k-1} N_i - \gamma_k \sum_{i=1}^{k} N_i\right). \qquad (27)
\end{aligned}$$

By the aforementioned independence relation, the variance of term in the bracket is bounded above by

$$\hat{D}_k \triangleq (1 - \gamma_k)^2 D_k + (1 - \gamma_k)^2 \sum_{i=1}^{k-1} \sigma_i^2 + \gamma_k^2 \sigma_k^2. \qquad (28)$$

Define the quantities $K_1, K_2, \dots, K_N$ as follows:

$$\frac{1}{2}\log(2\pi e K_1) \triangleq h(X|Y_N) = \frac{1}{2}\log\frac{2\pi e\sigma_x^4}{\sigma_x^2 + \sum_{i=1}^N \sigma_i^2},$$
$$\frac{1}{2}\log(2\pi e K_k) \triangleq h(Y_{k-1}|XY_k)$$
$$= \frac{1}{2}\log\frac{2\pi e\sigma_k^4}{\sum_{i=1}^k \sigma_i^2}, \quad k = 2,3,\dots,N.$$

Summarizing the bounds in (26) and (27), we have

$$R_{\rm HB}(D_1, D_2, \dots D_N) \geq \frac{1}{2}\log\frac{\prod_{i=1}^N K_i}{\prod_{i=1}^N \hat{D}_i} \triangleq R^*_{\rm HB}(I_N) \quad (29)$$

where for convenience we define $\hat{D}_1 = D_1$.

Next we construct the random variables $(W_1^*, W_2^*, \dots, W_N^*)$, and show that this specific choice of random variables achieves $\max_{A_D \subseteq I_N} R^*_{\rm HB}(A_D)$. We assume that $D_k \leq \mathbb{E}[X - \mathbb{E}(X|Y_k)]^2$ for each $k = 1, 2, \dots, N$, because otherwise this distortion requirement can be ignored completely. Intuitively, W_k^* is a Gaussian auxiliary random variable of sufficient quality such that it can reconstruct, jointly with side information Y_k, the source to achieve distortion D_k. However, in addition to the requirement on their quality, we construct these random variables to follow a Markov string structure with the source, which is not necessarily along the same order as the side information.

[Construction of $(W_1^*, W_2^*, \dots, W_N^*)$**]**

1) For each $k = 1, 2, \dots, N$, determine the variance $\sigma^2_{Z_k}$ of a Gaussian random variable Z_k such that $D_k = \mathbb{E}[X - \mathbb{E}(X|Y_k, X + Z_k)]^2$.
2) Rank the variances of $\sigma^2_{Z_k}, k = 1, 2, \dots, N$, in an increasing order, and let $\omega(k)$ denote the rank of $\sigma^2_{Z_k}$.
3) Calculate $\sigma^2_{Z'_1} = \sigma^2_{Z_{\omega^{-1}(1)}}$ and $\sigma^2_{Z'_k} = \sigma^2_{Z_{\omega^{-1}(k)}} - \sigma^2_{Z_{\omega^{-1}(k-1)}}$ for $k = 2, 3, \dots, N$.
4) Construct a set of independent zero-mean Gaussian random variables $(Z'_1, Z'_2, \dots, Z'_N)$ to have variance $\sigma^2_{Z'_k}$.
5) Construct a set of random variables $(W_1^*, W_2^*, \dots, W_N^*)$ as
$$W_k^* = X + \sum_{i=1}^{\omega(k)} Z'_k. \quad (30)$$

Next we show that this construction of $(W_1^*, W_2^*, \dots, W_N^*)$ achieves one of the aforementioned lower bounds and thus is an optimal forward test channel. Choose the set $A_D^* = \{k : \omega(k) < \omega(j) \text{ forall } j > k\}$, and denote the rank (in increasing order) of its element k as $r(k)$. Because of the construction of $(W_1^*, W_2^*, \dots, W_N^*)$ and the fact that they are jointly Gaussian with $(X, Y_1, Y_2, \dots, Y_N)$, we have

$$\sum_{k=1}^N I(X; W_k^*|Y_k, W_{k+1}^*, W_{k+2}^*, \dots, W_N^*)$$
$$= \sum_{k\in A_D^*} I(X; W_k^*|Y_k, W_{k+1}^*, W_{k+2}^*, \dots, W_N^*)$$
$$= \sum_{j=1}^{|A_D^*|} I\left(X; W^*_{r^{-1}(j)}|Y_{r^{-1}(j)}, W^*_{r^{-1}(j+1)}\right)$$
$$= h\left(X|Y_{r^{-1}(|A_D^*|)}\right) - h\left(X|W^*_{r^{-1}(|A_D^*|)}Y_{r^{-1}(|A_D^*|)}\right)$$
$$+ h\left(X|Y_{r^{-1}(|A_D^*|-1)}W^*_{r^{-1}(|A_D^*|)}\right)$$
$$- h\left(X|Y_{r^{-1}(|A_D^*|-1)}W^*_{r^{-1}(|A_D^*|-1)}\right)$$
$$+ \cdots + h\left(X|Y_{r^{-1}(1)}W^*_{r^{-1}(2)}\right) - h(X|Y_{r^{-1}(1)}W^*_{r^{-1}(1)})$$
$$= h\left(X|Y_{r^{-1}(|A_D^*|)}\right) - h\left(X|Y_{r^{-1}(1)}W^*_{r^{-1}(1)}\right)$$
$$- \left[h\left(Y_{r^{-1}(|A_D^*|-1)}|Y_{r^{-1}(|A_D^*|)}W^*_{r^{-1}(|A_D^*|)}\right) - h\left(Y_{r^{-1}(|A_D^*|-1)}|XY_{r^{-1}(|A_D^*|)}\right)\right] - \cdots$$
$$- \left[h\left(Y_{r^{-1}(1)}|Y_{r^{-1}(2)}W^*_{r^{-1}(2)}\right) - h\left(Y_{r^{-1}(1)}|XY_{r^{-1}(2)}\right)\right]$$
$$= R^*_{\rm HB}(A_D^*) \quad (31)$$

where $W^*_{r^{-1}(|A_D^*|+1)} \leq 0$. Thus, we have proved the following theorem.

Theorem 7: The random variables $(W_1^*, W_2^*, \dots, W_N^*)$ constructed above achieve the minimum in the Heegard and Berger rate–distortion function for the jointly Gaussian source and side informations.

It is clear that we can determine the set A_D^* before constructing $(W_1^*, W_2^*, \dots, W_N^*)$, which can simplify the construction. However, the current construction has the advantage that each W_k^* is almost individually determined by D_k and the quality of the side information Y_k, and does not substantially depend on the other distortion constraints. This will prove to be useful for the general scalable coding problem. It would appear at first sight that we need to compare $2^N - 1$ values of $R^*_{\rm HB}(A_D)$, one for each $A_D \subset I_N$, in order to determine $R_{\rm HB}(D_1, D_2, \dots, D_2)$; however, from the afore-given calculation we see that in fact an algorithm of $O(N)$ complexity suffices.

This result can be interpreted using Fig. 7. On the horizontal axis, the N marks stand for the N random variables $(W^*_{\omega^{-1}(1)}, W^*_{\omega^{-1}(2)}, \dots, W^*_{\omega^{-1}(N)})$, and on the vertical axis, the N marks stand for the N levels of side informations $(Y_1, Y_2, \dots, Y_N)$. The random variable pairs (W_k^*, Y_k) are then the points of interest on the plane, since if the kth decoder has

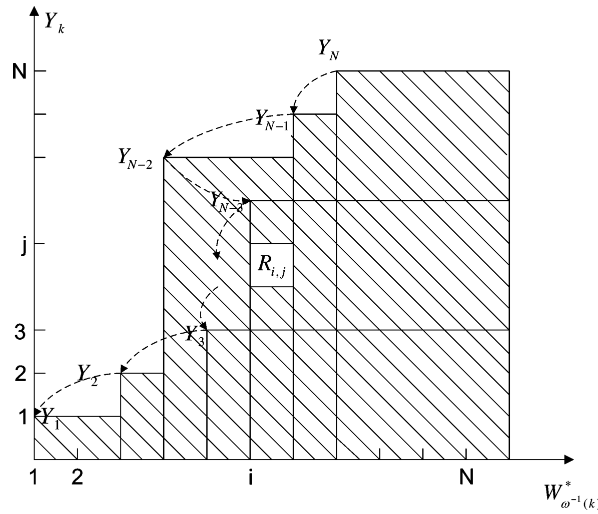

Fig. 7. An illustration of the sum-rate for the Gaussian case.

(Y_k, W_k^*) the desired distortion can be achieved; the (W_k^*, Y_k) pairs are in one-to-one correspondence to the $(\omega(k), k)$ pairs. Next we associate the unit square (below and to the right of) each integer point (i, j) with a rate of value

$$R_{i,j} = I\left(W^*_{\omega^{-1}(i)}; Y_{j-1} | Y_j W^*_{\omega^{-1}(i+1)}\right) \quad (32)$$

where we define $W^*_{\omega^{-1}(N+1)} = 0$, and $Y_0 = X$. For each $k = 1, 2, \ldots, N$, if we cover the rectangle below and to the right of $(\omega(k), k)$, then the sum-rate associated with the combined covered area after N such steps is exactly $R_{\text{HB}}(D_1, D_2, \ldots, D_N)$.

With Fig. 7, the coding scheme can be understood as follows. The coding proceeds from Y_N to Y_1, i.e., from high to low on the vertical axis; the kth step (kth decoder) specifies an integer point $(\omega(k), k)$ on the figure, which corresponds to a (W_k^*, Y_k) pair. Additional coding in this step is required if the area below and to the right of this point is beyond what has already been covered in the previous steps, and the rate associated with this new area is exactly the incremental rate in the SR-WZ setting. This order is illustrated in Fig. 7 along the arrows. Note that

$$\begin{aligned}\sum_{j=1}^{k} R_{i,j} &= \sum_{j=1}^{k} I\left(W^*_{\omega^{-1}(i)}; Y_{j-1} | Y_j W^*_{\omega^{-1}(i+1)}\right)\\ &= \sum_{j=1}^{k} \Big[I\left(W^*_{\omega^{-1}(i)}; Y_{j-1} | W^*_{\omega^{-1}(i+1)}\right)\\ &\quad - I\left(W^*_{\omega^{-1}(i)}; Y_j | W^*_{\omega^{-1}(i+1)}\right)\Big]\\ &= I\left(W^*_{\omega^{-1}(i)}; X | W^*_{\omega^{-1}(i+1)}\right)\\ &\quad - I\left(W^*_{\omega^{-1}(i)}; Y_k | W^*_{\omega^{-1}(i+1)}\right)\\ &= I\left(W^*_{\omega^{-1}(i)}; X | Y_k W^*_{\omega^{-1}(i+1)}\right) \end{aligned} \quad (33)$$

which is the rate for a vertical slice of height k between horizontal position i and $i+1$, and the expression is quite similar to the summand of (24). In the example of Fig. 7, the decoders with side information Y_{N-3} and Y_3 do not require additional rates after previous coding steps. Additionally, it can be seen that the corners of the final covered area in fact specifies the set A_D^*.

The following observations are essential for the general Gaussian scalable coding problem: each unit square in Fig. 7 is not merely associated with rate $R_{i,j}$, it is in fact associated with a fraction of code $C_{i,j}$ with the following properties.

1) The rate of $C_{i,j}$ is (asymptotically) $R_{i,j}$.
2) If the fractions of code associated with the area below and to the right of $(\omega(k), k)$ are available, then the decoder with side information Y_k can decode within distortion D_k;
3) The same set of codes $C_{i,j}$ can be used to fulfill only subset of the constraints, and the rate calculated by the same covering area method is the quadratic Gaussian Heegard and Berger rate–distortion function for those distortion constraints.

The first two observations are straightforward by constructing the nested binning together with conditional codebooks as described in Section III, i.e., $N-1$ conditioning stages from $W^*_{\omega^{-1}(N)}$ to $W^*_{\omega^{-1}(1)}$ and each conditional codebook has N nested levels from coarse for Y_1 to fine for Y_N. In fact, it is not necessary to use N nested levels for each codebook, but we do so to simplify the notations. The last property is due to the inherent Markov string among $W_1^*, W_2^*, \ldots, W_N^*$ and X.

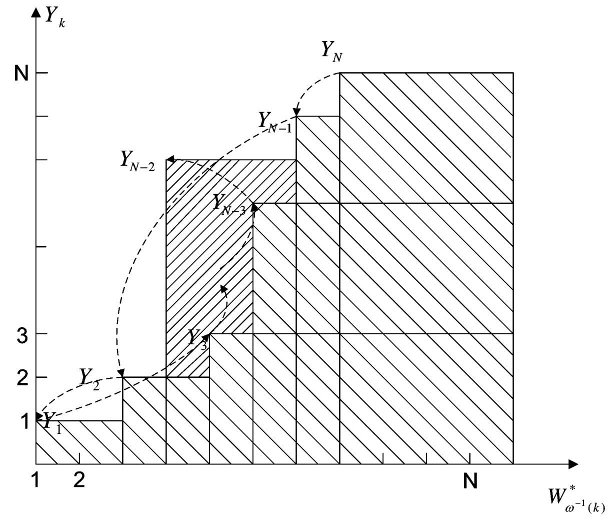

Fig. 8. An illustration of incremental rate for scalable coding. The denser shaded region gives the incremental rate R_k for the stage with side information Y_k.

B. Scalable Coding With Joint Gaussian Side Informations

Now consider the scalable coding problem where side informations and distortions are given by a permutation $\pi(\cdot)$ of those defined in the last subsection, i.e., $Y_i' = Y_{\pi(i)}$ and $D_i' = D_{\pi(i)}$. We next show that the identically permuted set of random variables $(W_1^*, W_2^*, \ldots, W_N^*)$ achieves the Heegard–Berger rate distortion function for any first k stages, and thus is optimal for the general scalable coding problem. In light of the pictorial interpretation in Fig. 7, this reduces to rearranging the coded stream of $C_{i,j}$. Fig. 8 shows the effect of changing the scalable coding order.

More precisely, for a certain side information $Y_i' = Y_{\pi(i)}$, define the following sets:

$$\begin{aligned} C(k) &= \{\pi(i) : i < k, \pi(i) > \pi(k)\}\\ E_-(k) &= \{\pi(i) : i < k, \pi(i) < \pi(k), \omega(\pi(i)) > \omega(\pi(k))\} \end{aligned}$$

and the following function:

$$\begin{aligned} E(k) = \max[\{\pi(i) : i < k, \pi(i) < \pi(k),\\ \omega(\pi(i)) < \omega(\pi(k))\} \cup \{0\}] \end{aligned}$$

and let $Y_0 = X$. Let the set of integers $E_-(k)$ be ordered increasingly, and the rank of its element j be $r(j)$. Denote the set of random variables $\{W_j : j \in C\}$ as W_C^* for an integer set C. The kth stage rate given in (34) at the top of the following page is achievable for $k = 1, 2, \ldots, N$.

It is clear that this rate corresponds to exactly the dense shaded region in Fig. 8, which is the sum of rates of fraction of codes $C(i, j)$ as described above. The property of this fraction code $C(i, j)$ thus implies the following.

$$R_k = \sum_{i=1}^{|E_-(k)|} I\left(Y_{r^{-1}(i)}; W^*_{r^{-1}(i)} | Y_{\pi(k)} W^*_{r^{-1}(i+1)}, \ldots, W^*_{r^{-1}(|E_-(k)|)} W^*_{C(k)}\right) + I\left(Y_{E(k)}; W^*_{\pi(k)} | Y_{\pi(k)} W^*_{E_-(k)} W^*_{C(k)}\right). \quad (34)$$

Theorem 8: The Gaussian scalable coding achievable rate region for distortion vector $(D_{\pi(1)}, D_{\pi(2)}, \ldots, D_{\pi(N)})$ is the rate vectors $(R_1, R_2, \ldots, R_N)$ that satisfy

$$\sum_{i=1}^{k} R_i \geq R_{\text{HB}}(D_{\pi(1)}, D_{\pi(2)}, \ldots, D_{\pi(k)}), \quad k = 1, 2, \ldots, N$$

where the side informations are $(Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(k)})$. Furthermore, it is achievable by a jointly Gaussian codebook with nested binning.

An immediate consequence of this result is the following corollary.

Corollary 5: The distortions $(D_{\pi(1)}, D_{\pi(2)}, \ldots, D_{\pi(N)})$ are perfectly scalable along side informations $(Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(k)})$ for the jointly Gaussian source if and only if

$$R_{\text{HB}}(D_{\pi(1)}, D_{\pi(2)}, \ldots, D_{\pi(k)}) = R^*_{X|Y_{\pi(k)}}(D_{\pi(k)})$$

for each $k = 1, 2, \ldots, N$.

The condition in this corollary holds for one important special case where $D_1 = D_2 = \cdots = D_N$ and $\pi(k) = N - k + 1$ for each k, i.e., when all the decoders have the same distortion requirement, and the scalable order is along a decreasing order of the side information quality. This implies that at least for the Gaussian case, an opportunistic coding strategy does exist when the distortion requirement is the same for all the users.

VII. Conclusion

We studied the problem of scalable source coding with reversely degraded side informations and gave two inner bounds as well as two outer bounds to the rate–distortion region. These bounds are tight for special cases such as one lossless decoder and under certain deterministic distortion measures. Furthermore, we provided a complete solution for the Gaussian source under the quadratic distortion measure with any number of jointly Gaussian side informations. The problem of perfect scalability is investigated and the gap between the inner and the outer bounds is shown to be bounded under the quadratic distortion measure. For the doubly symmetric binary source under the Hamming distortion measure, a partial characterization is provided for the rate–distortion region. These results illustrate the difference between lossless and lossy source coding: though a universal approach exists with uncertain side information at the decoder for the lossless case, such uncertainty generally causes loss of performance in the lossy case.

Appendix A
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 $P_X(a) > 0$
$$\left| \frac{1}{n} N(a|\boldsymbol{x}) - P_X(a) \right| < \delta;$$
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 are 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 an $\boldsymbol{x} \in T^\delta_{[X]}$, for a sequence $\boldsymbol{y} \in \mathcal{Y}^n$ whose component is drawn i.i.d according to P_Y and any $\delta' > \delta$, we have
$$2^{-n(I(X;Y)+\lambda_1)} \leq P[(\boldsymbol{x}, \boldsymbol{y}) \in T^{\delta'}_{[XY]}] \leq 2^{-n(I(X;Y)-\lambda_1)}$$
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
$$2^{-n(I(X;Z|Y)+\lambda_2)} \leq P[(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in T^{\delta''}_{[XYZ]}] \leq 2^{-n(I(X;Z|Y)-\lambda_2)}$$
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* [18]: 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 $\delta > 0$
$$\lim_{n\to\infty} P\left[(\boldsymbol{X}, \boldsymbol{z}) \in T^{|\mathcal{Y}|\delta}_{[XZ]} \middle| (\boldsymbol{Y}, \boldsymbol{z}) \in T^\delta_{[YZ]} \right] \to 1.$$
Furthermore
$$\lim_{n\to\infty} P\left[(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{z}) \in T^\delta_{[XYZ]} \middle| (\boldsymbol{Y}, \boldsymbol{z}) \in T^\delta_{[YZ]} \right] \to 1.$$

Appendix B
Proof of Theorem 1

Codebook Generation: Let a probability distribution $P_{VW_1W_2XY_1Y_2} = P_{VW_1W_2X} P_{Y_1|X} P_{Y_2|Y_1}$, and two reconstruction functions $f_1(Y_1, W_1)$ and $f_2(Y_2, W_2)$ be given. First, construct a nested binning structure with 2^{nR_A} coarser bins and $2^{n(R_A+R'_A)}$ finer bins, where R_A and R'_A are to be specified later. Generate 2^{nR_V} length-n codewords according to $P_V(\cdot)$, and denote this set of codewords as $\mathcal{C}_v$; assign each codeword in $\mathcal{C}_v$ 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 $P(\boldsymbol{w}_1|\boldsymbol{v}) = \prod_{k=1}^n P_{W_1|V}(w_{1,k}|v_k)$, and denote this set of codewords as $\mathcal{C}_{w_1}(\boldsymbol{v})$; independently assign each codeword in $\mathcal{C}_{w_1}(\boldsymbol{v})$ into one of 2^{nR_B} bins. Again, for each codeword $\boldsymbol{v} \in \mathcal{C}_v$, independently generate $2^{nR_{W_2}}$ length-n codewords according to $P(\boldsymbol{w}_2|\boldsymbol{v}) = \prod_{k=1}^{n} P_{W_2|V}(w_{2,k}|v_k)$, and denote this set of codewords as $\mathcal{C}_{w_2}(\boldsymbol{v})$; independently assign each codeword in $\mathcal{C}_{w_2}(\boldsymbol{v})$ into one of 2^{nR_C} bins. Reveal these codebooks and the bin indices to the encoder and decoders.

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

Decoding: The first 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 $(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{v}}, \boldsymbol{y}_1) \in T^{4|\mathcal{X}|\delta}_{[W_1VY_1]}$. The second 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 is more than one codeword satisfying the condition, 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 reconstructs as $\hat{x}_{2,k} = f_2(\hat{w}_{2,k}, y_{2,k})$.

Probability of error: First define the encoding errors

$$E_0 = \left\{\boldsymbol{X} \notin T^{\delta}_{[X]}\right\} \cup \left\{\boldsymbol{Y_1} \notin T^{\delta}_{[Y_1]}\right\} \cup \left\{\boldsymbol{Y_2} \notin T^{\delta}_{[Y_2]}\right\}$$

$$E_1 = E_0^c \cap \left\{\forall \boldsymbol{v} \in \mathcal{C}_v, (\boldsymbol{X}, \boldsymbol{v}) \notin T^{2\delta}_{[XV]}\right\}$$

$$E_2 = E_0^c \cap E_1^c \cap \left\{\forall \boldsymbol{w}_1 \in \mathcal{C}_{w_1}(\boldsymbol{v}^*), (\boldsymbol{w}_1, \boldsymbol{v}^*, \boldsymbol{X}) \notin T^{3\delta}_{[W_1VX]}\right\}$$

$$E_3 = E_0^c \cap E_1^c \cap \left\{\forall \boldsymbol{w}_2 \in \mathcal{C}_{w_2}(\boldsymbol{v}^*), (\boldsymbol{w}_2, \boldsymbol{v}^*, \boldsymbol{X}) \notin T^{3\delta}_{[W_2VX]}\right\}.$$

Next define the decoding errors

$$E_4 = E_0^c \cap E_1^c \cap \left\{(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_1}) \notin T^{2\delta}_{[VXY_1]}\right\}$$

$$E_5 = E_0^c \cap E_1^c \cap \left\{(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_2}) \notin T^{2\delta}_{[VXY_2]}\right\}$$

$$E_6 = E_0^c \cap E_1^c \cap \left\{\exists \boldsymbol{v}' \neq \boldsymbol{v}^* : i(\boldsymbol{v}') = i(\boldsymbol{v}^*) \text{ and } (\boldsymbol{v}', \boldsymbol{Y_1}) \in T^{3|\mathcal{X}|\delta}_{[VY_1]}\right\}$$

$$E_7 = E_0^c \cap E_1^c \cap \left\{\exists \boldsymbol{v}' \neq \boldsymbol{v}^* : i(\boldsymbol{v}') = i(\boldsymbol{v}^*) \text{ and } j(\boldsymbol{v}') = j(\boldsymbol{v}^*) \text{ and } (\boldsymbol{v}', \boldsymbol{Y_2}) \in T^{3|\mathcal{X}|\delta}_{[VY_2]}\right\}$$

$$E_8 = E_0^c \cap E_1^c \cap E_2^c \cap E_4^c \cap E_6^c \cap \left\{(\boldsymbol{w_1^*}, \boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_1}) \notin T^{3\delta}_{[W_1VXY_1]}\right\}$$

$$E_9 = E_0^c \cap E_1^c \cap E_3^c \cap E_5^c \cap E_7^c \cap \left\{(\boldsymbol{w}_2^*, \boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_2}) \notin T^{3\delta}_{[W_2VXY_2]}\right\}$$

$$E_{10} = E_0^c \cap E_1^c \cap E_2^c \cap E_4^c \cap E_6^c \cap \left\{\exists \boldsymbol{w_1'} \neq \boldsymbol{w_1^*} : k(\boldsymbol{w}_1') = k(\boldsymbol{w_1^*}) \text{ and } (\boldsymbol{w_1'}, \boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{4|\mathcal{X}|\delta}_{[W_1VY_1]}\right\}$$

$$E_{11} = E_0^c \cap E_1^c \cap E_3^c \cap E_5^c \cap E_7^c \cap \left\{\exists \boldsymbol{w_2'} \neq \boldsymbol{w_2^*} : l(\boldsymbol{w_2'}) = l(\boldsymbol{w_2^*}) \text{ and } (\boldsymbol{w_2'}, \boldsymbol{v}^*, \boldsymbol{Y_2}) \in T^{4|\mathcal{X}|\delta}_{[W_2VY_2]}\right\}$$

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

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

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

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

$$\begin{aligned} P(E_2) &\leq (1 - 2^{-n(I(X;W_1|V)+\lambda)})^{nR_2} \\ &\leq \exp(-2^{-n(I(X;W_1|V)+\lambda_2-R_2)}) \end{aligned}$$

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 $R_{W_2} > I(X;W_2|V)+\lambda_2$. $P(E_4)$ and $P(E_5)$ both tend 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 true given E_1 does not happen. Similarly, both $P(E_8)$ and $P(E_9)$ tend to zero for the same reason. Notice that if $(\boldsymbol{v}^*, \boldsymbol{X}, \boldsymbol{Y_1}) \in T^{2\delta}_{[VXY_1]}$, then $(\boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{3|\mathcal{X}|\delta}_{[VY_1]}$, and thus $\boldsymbol{v}^*$ can be correctly decoded if there is no other codeword in the same bin satisfying the typicality test.

Conditioned on E_0^c, we have $\boldsymbol{Y_1} \in T^{\delta}_{[Y_1]}$. The codewords in $\mathcal{C}_v$ are generated independently according to $P_V(\cdot)$, and it follows that

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

where we have used Property 1) of the typical sequences and the fact that codewords in $\mathcal{C}_v$ are assigned to the bins independently. 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, we have $(\boldsymbol{v}^*, \boldsymbol{Y_1}) \in T^{2|\mathcal{X}|\delta}_{[VY_1]}$. Thus

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

where Property 2) 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

$$\begin{aligned} R_1 &= R_A + R_B > I(X;VW_1|Y_1) + \lambda' \\ R_1 + R_2 &= R_A + R_A' + R_B + R_C > I(X;VW_2|Y_2) \\ &\quad + I(X;W_2|VY_1) + \lambda'' \end{aligned}$$

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 distortion constraints are satisfied as well. When no error occurs, we have $(\hat{\boldsymbol{W}}_1, \boldsymbol{X}, \boldsymbol{Y}_1) \in T^{3|\mathcal{V}|\delta}_{[W_1XY]}$ and $(\hat{\boldsymbol{W}}_2, \boldsymbol{X}, \boldsymbol{Y}_1) \in T^{3|\mathcal{V}|\delta}_{[W_2XY]}$. By standard arguments using the definition of the typical sequences, it can be shown that

$$d_1(\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \leq \mathbb{E} d_1[X, f_1(W_1, Y_1)] + \epsilon'$$

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, ϵ' can be made arbitrarily small by choosing a sufficiently small δ and a sufficiently large n. A similar argument holds for the second decoder. This completes the proof. □

Appendix C
Proof of the Theorem 2

Assume the existence of an (n, M_1, M_2, D_1, D_2) scalable code, then 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 $(X_1, X_2, \ldots, X_{k-1})$ and $\boldsymbol{X}_k^+$ to denote $(X_{k+1}, X_{k+2}, \ldots, X_n)$; the subscript k will be dropped when it is clear from the context. The proof follows a similar line as the converse proof in [7]. Here we omit the small positive quantity ϵ for simplicity. The following chain of inequalities is standard (see [24, p. 440])

$$\begin{aligned} nR_1 \geq H(T_1) &\geq H(T_1|\boldsymbol{Y}_1) = I(\boldsymbol{X}; T_1|\boldsymbol{Y}_1) \\ &= \sum_{k=1}^n H(X_k|\boldsymbol{Y}_1\boldsymbol{X}_k^-) - H(X_k|T_1\boldsymbol{Y}_1\boldsymbol{X}_k^-) \\ &\geq \sum_{k=1}^n I(X_k; T_1\boldsymbol{Y}_1^-\boldsymbol{Y}_1^+|Y_k). \end{aligned} \quad (35)$$

Next we bound the sum rate as follows:

$$\begin{aligned} n(R_1+R_2) &\geq H(T_1T_2) \geq H(T_1T_2|\boldsymbol{Y}_2) \\ &= I(\boldsymbol{X}; T_1T_2|\boldsymbol{Y}_2) \\ &= I(\boldsymbol{X}; T_1T_2\boldsymbol{Y}_1|\boldsymbol{Y}_2) - I(\boldsymbol{X}; \boldsymbol{Y}_1|T_1T_2\boldsymbol{Y}_2) \\ &= \sum_{k=1}^n \left[I(X_k; T_1T_2\boldsymbol{Y}_1|\boldsymbol{Y}_2\boldsymbol{X}^-) \right. \\ &\quad \left. - I(\boldsymbol{X}; Y_{1,k}|T_1T_2\boldsymbol{Y}_2\boldsymbol{Y}_1^-) \right]. \end{aligned}$$

Since $(X_k, Y_{2,k})$ is independent of $(\boldsymbol{X}^-, \boldsymbol{Y}_2^-, \boldsymbol{Y}_2^+)$, we have

$$\begin{aligned} I(X_k; T_1T_2\boldsymbol{Y}_1|\boldsymbol{Y}_2\boldsymbol{X}^-) &= I(X_k; T_1T_2\boldsymbol{Y}_1\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+\boldsymbol{X}^-|Y_{2,k}) \\ &\geq I(X_k; T_1T_2\boldsymbol{Y}_1\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+|Y_{2,k}). \end{aligned} \quad (36)$$

The Markov condition

$$(\boldsymbol{X}^-\boldsymbol{X}^+T_1T_2\boldsymbol{Y}_1^-\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+) \leftrightarrow (X_k, Y_{2,k}) \leftrightarrow Y_{1,k}$$

gives

$$I(\boldsymbol{X}; Y_{1,k}|T_1T_2\boldsymbol{Y}_2\boldsymbol{Y}_1^-) = I(X_k; Y_{1,k}|T_1T_2\boldsymbol{Y}_2\boldsymbol{Y}_1^-). \quad (37)$$

Thus, we have

$$\begin{aligned} n(R_1+R_2) &\geq \sum_{k=1}^n \left[I(X_k; T_1T_2\boldsymbol{Y}_1\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+|Y_{2,k}) \right. \\ &\qquad \left. - I(X_k; Y_{1,k}|T_1T_2\boldsymbol{Y}_2\boldsymbol{Y}_1^-) \right] \\ &= \sum_{k=1}^n \left[I(X_k; T_1T_2\boldsymbol{Y}_1^-\boldsymbol{Y}_2{}^-\boldsymbol{Y}_2^+|Y_{2,k}) \right. \\ &\qquad \left. + I(X_k; \boldsymbol{Y}_1^+|T_1T_2\boldsymbol{Y}_2\boldsymbol{Y}_1^-Y_{1,k}) \right]. \end{aligned} \quad (38)$$

The degradedness gives $(X_k, T_1T_2, \boldsymbol{Y}_1^-\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+) \leftrightarrow Y_{1,k} \leftrightarrow Y_{2,k}$, which implies

$$\begin{aligned} n(R_1+R_2) &\geq \sum_{k=1}^n \left[I(X_k; T_1T_2\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+\boldsymbol{Y}_1^-|Y_{2,k}) \right. \\ &\qquad \left. + I(X_k; \boldsymbol{Y}_1^+|T_1T_2\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+\boldsymbol{Y}_1^-Y_{1,k}) \right]. \end{aligned} \quad (39)$$

Define $W_{1,k} = (T_1\boldsymbol{Y}_1^-\boldsymbol{Y}_1^+)$ and $W_{2,k} = (T_1T_2\boldsymbol{Y}_2^-\boldsymbol{Y}_2^+\boldsymbol{Y}_1^-)$, by which we have

$$\begin{aligned} nR_1 &\geq \sum_{k=1}^n I(X_k; W_{1,k}|Y_{1,k}), n(R_1+R_2) \\ &\geq \sum_{k=1}^n [I(X_k; W_{2,k}|Y_{2,k}) + I(X_k; W_{1,k}|W_{2,k}Y_{1,k})]. \end{aligned}$$

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 functions $f_j(W_j, Y_j), j = 1, 2$ follows by defining

$$\begin{aligned} f_1(W_1, Y_1) &= \psi_{1,Q}(\phi_1(\boldsymbol{X}), \boldsymbol{Y}_1) \\ f_2(W_2, Y_2) &= \psi_{2,Q}(\phi_1(\boldsymbol{X}), \phi_2(\boldsymbol{X}), \boldsymbol{Y}_2) \end{aligned}$$

which leads to the fulfillment of the distortion constraints. It only remains to show that both bounds can be written in single-letter forms, and that restricting the alphabet sizes does not cause essential difference. These are straightforward by applying standard techniques in [24. p. 435] and [11]. This completes the proof for $\mathcal{R}_{\text{out}}(D_1, D_2) \supseteq \mathcal{R}(D_1, D_2)$. □

Acknowledgment

The authors wish to thank the associate editor and the anonymous reviewers for many suggestions which helped to improve the presentation of this paper. Discussions with Professor Emre Telatar at EPFL are also gratefully acknowledged.

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. Inf. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.
[3] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. Inf. 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. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[5] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," *IEEE Trans. Inf. 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," *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2946–2960, Aug. 2007.

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

[8] A. Kaspi, "Rate-distortion when side-information may be present at the decoder," *IEEE Trans. Inf. 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. Inf. Theory*, vol. IT-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*, Bangalore, India, Oct. 2002, pp. 127–130.

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

[12] S. C. Draper, "Universal incremental Slepian-Wolf coding," in *Proc. 42nd Annu. Allerton Conf. Communication, Control And Computing*, Monticello, IL, Sep. 2004, pp. 1332–1341.

[13] A. Eckford and W. Yu, "Rateless Slepian-Wolf codes," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Oct.–Nov. 2005, pp. 1757–1761.

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

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

[16] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

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

[18] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, ser. CISM Courses and Lectures, G. Longo, Ed. Vienna/New Yrok: Springer-Verlag, 1978, pp. 171–231.

[19] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.

[20] C. Tian and S. Diggavi, Side information scalable source coding EPFL, Lausanne, Switzerland, Tech. Rep. LICOS-Report-2007-003, Sep. 2006.

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

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

[23] H. Feng and Q. Zhao, "On the rate loss of multiresolution source codes in the Wyner-Ziv setting," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1164–1171, Mar. 2006.

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