On Successive Refinement for the Wyner–Ziv Problem

Yossef Steinberg and Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
[ysteinbe,merhav]@ee.technion.ac.il

November 5, 2003

Abstract

Achievable rates are characterized for successive refinement in the Wyner–Ziv scenario, namely, in the presence of correlated side information (SI) at the receiver. In this setting, the encoder is assumed to operate in two stages, where the first corresponds to relatively low rate and high distortion, and the second, comprising a refinement code on top of the first code, is aimed at reproduction at reduced distortion. Both decoders (for low-rate/high-distortion and for high-rate/low-distortion) are equipped with SI streams, correlated to the source, but unavailable to the encoder. Furthermore, it is assumed that the decoder that receives the higher rate bitstream, i.e., the additional refinement bits, accesses also SI of "better quality" than that of the lower resolution decoder. By "better quality," we mean that the source and the SI are modeled together as a stochastically degraded joint source, where the encoded symbols, the SI at the refinement stage, and the SI at the initial (coarse) stage, form a Markov chain in this order. For a memoryless joint process (that includes the source to be encoded and its instantaneously correlated SI streams), necessary and sufficient conditions are furnished, in terms of single–letter formulas, for the achievability of a pair of rates, corresponding to two given distortion levels. Special attention is devoted to the degenerate, but important, case where the two SI streams, at the two decoders, are identical. For this case, conditions are provided for successive refinability in the sense of the existence of codes that asymptotically achieve the Wyner–Ziv rate–distortion function, simultaneously at both distortion levels. In this context, the doubly symmetric binary source (with the Hamming distortion measure) and the jointly Gaussian source (with the squared error distortion measure) are shown to be successively refinable in the Wyner–Ziv setting. It is also demonstrated that a source that is not successively refinable in the ordinary sense (i.e., without SI) may become successively refinable in the presence of SI at the decoders.

Index terms — Side information, successive refinement, scalable coding, progressive coding, multiple description, Wyner-Ziv problem.

1 Introduction

The problem of successive refinement of information was originally formulated by Koshelev [9], and by Equitz and Cover [4], who viewed this problem as a special case of the more general multiple description problem. In their original setting, a source $\boldsymbol{X}$ is to be encoded and transmitted to its destination over a rate–limited channel. To this end, due to the fact that the channel is rate–limited, the encoder produces a compressed bit string $\boldsymbol{S}$ at rate R_1, from which the decoder produces $\hat{\boldsymbol{X}}$, an approximation of the original source at distortion level D_1, according to some distortion measure $d(\cdot,\cdot)$. Rate–distortion theory tells us, as is well known, that the minimal rate R_1 that is needed in order to enable reproduction of the source at accuracy D_1 is given by $R_X(D_1)$, the rate-distortion function of the source at distortion level D_1. At a later stage, a more accurate description of the source is needed, and the encoder sends a secondary string of compressed bits, $\boldsymbol{S}_\Delta$, at rate ΔR, to the destination. The decoder, having at hand both strings $\boldsymbol{S}$ and $\boldsymbol{S}_\Delta$, produces a more accurate reproduction of the source, $\tilde{\boldsymbol{X}}$, at distortion level D_2 ($D_2 < D_1$). Invoking again the fundamental limits of rate-distortion theory, the total rate $R_1 + \Delta R$ cannot be lower than $R_X(D_2)$, the rate–distortion function evaluated at the corresponding distortion level. The best one can hope for, then, is that the two rates *simultaneously* lie on the rate-distortion curve, i.e.,

$$R_1 = R_X(D_1), \quad \text{and} \quad R_1 + \Delta R = R_X(D_2). \tag{1}$$

This, however, is not always possible. In general, some penalty might have to be paid for successive coding. In the ideal situation, where successive coding, in two or more stages, can be made rate-distortion optimal simultaneously at all stages, the source is called *successively refinable*. Koshelev [9], [10], and Equitz and Cover [4], have shown that a necessary and sufficient condition for a source to be successively refinable is that the conditional distributions $P_{\tilde{X}|X}$ and $P_{\hat{X}|X}$ that achieve the rate distortion function at distortion levels D_2 and D_1, respectively, are Markov compatible in the sense that they can be represented as a Markov chain $X \multimap \tilde{X} \multimap \hat{X}$. Equitz and Cover have shown a few examples of successively refinable sources. In particular, it was shown that Gaussian sources with the squared error distortion measure, general discrete memoryless sources (DMS's) with the Hamming distortion measure, and Laplacian sources with the absolute distortion measure, are all successively refinable. Moreover, an example of a source that does not satisfy the Markov condition was given, thus showing that the problem is not redundant, i.e., not every source is successively refinable.

Clearly, the theoretical and practical value of successive coding is not limited to successively refinable sources. Successive coding is suitable to any application where a relatively coarse description of the source suffices at the first use of the data, and fine details are needed only at some later stage. Transmission of the fine details, which comprises the refinement stage, may take place upon user request or upon availability of additional system resources, like free time slots or extra bandwidth in a network communication environment, reduction of cost of channel use, etc. In [9],[1] Koshelev characterized the set of all quadruples $(R_1, R_1 + \Delta R, D_1, D_2)$ that are achievable via successive coding for a general DMS (that is not necessarily successively refinable). From this more general characterization, the Markov condition for successive refinability is readily obtained as a byproduct.

The above description of earlier work on successive refinement of information serves as one part of the background for this work. The other part has to do with rate–distortion coding with side information (SI) at the decoder, which is well–known as the Wyner–Ziv problem [16], the lossy counterpart of a certain version of the Slepian–Wolf setting. In some source coding applications, it is plausible to assume that the decoder has some knowledge about the encoded source, which is not available to the encoder. For example, consider an airplane taking aerial photographs of a certain area, for the purpose of later processing (e.g., updating maps) on the ground. The photographs are encoded and transmitted to a base-station, and on the ground, the decoder reconstructs the aerial photos. For that purpose, the decoder takes advantage of previously existing data about the photographed area – for example, previous photographs, maps, etc. Since the equipment on the plane should be light and compact, the encoder is kept as simple as possible, and thus it does not utilize previous photographs or maps of the area in the encoding process. Another application of source coding with SI at the decoder is that of systematic source/channel coding, suggested by Shamai and Verdú for lossless coding [13], and extended later to lossy coding in [14]. In their setting, an analog information source is to be transmitted to the receiver and reconstructed at some distortion level D. For that purpose, the sender has access to two channels: an analog channel, over which the source is transmitted uncodedly (the systematic part), and a digital channel, over which digital, coded information about the source is sent. The receiver has access to the two outputs.

[1]See also Rimoldi [12].

The output of the analog channel is the original analog source corrupted by noise. By contrast, essentially no errors are introduced by the digital channel, as the information is channel-coded. Such a model arises, for example, when an analog communication system is upgraded with a digital system, but back–compatibility for users utilizing only the analog part is being kept. Now, observe that the output of the analog channel serves as SI about the source, which is accessible to the decoder but not to the encoder.

The main result of Wyner and Ziv [16] is as follows: Let $\boldsymbol{Z}$ be the SI, available to the decoder only, and having joint distribution P_{XZ} with the encoded source $\boldsymbol{X}$. The encoder, although ignorant of the specific realization of the SI, has full knowledge of this joint distribution. Let $R^*_{X|Z}(D)$ stand for the minimal achievable rate for coding with SI. Wyner and Ziv have shown that

$$R^*_{X|Z}(D) = \min I(X;U|Z), \tag{2}$$

where the minimization is over all random variables U such that $U \multimap X \multimap Z$ is a Markov chain and such that there exists a function f of U and Z such that

$$\mathbb{E}d(X, f(U,Z)) \leq D. \tag{3}$$

This work is a joint extension of the results of [9], [10], and [12] on the one hand, and the results of [16] on the other hand. Analogously to the previous works on successive coding, where classical rate-distortion theory forms the base of reference, here this role is played by the Wyner–Ziv result (2), (3). In particular, let R_1, ΔR, and D_1, D_2 be the rates and distortion levels of the coarse and refinement stages, respectively, as defined for the classical successive refinement above. Let $\boldsymbol{Z}$, $\boldsymbol{Y}$ be the SI available to the decoder at the coarse and at the refinement stages, respectively (see Fig. 1). The natural question that may be asked, in this setting, is what is the set of all achievable quadruples $(R_1, \Delta R, D_1, D_2)$ and how to achieve every such quadruple. Note that for a general joint source, $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, there is no guarantee that $D_2 \leq D_1$ since the SI available at the first stage, $\boldsymbol{Z}$, may be better than the SI available at the second stage, $\boldsymbol{Y}$. Indeed, this seems to be quite a difficult problem, and we have not been able to fully characterize the achievable rates and distortion levels when no structure is imposed on the joint distribution of the source and SI[2]. However, we have been able to fully characterize the set of all achievable quadruples $(R_1, \Delta R, D_1, D_2)$ for the special case where the Markov relation $\boldsymbol{X} \multimap \boldsymbol{Y} \multimap \boldsymbol{Z}$ hold, namely, the SI $\boldsymbol{Y}$ can be considered to be of better quality than $\boldsymbol{Z}$. From the practical point of view, this structure is not too restrictive.

[2]Similar difficulties appear in [7], where the related problem of coding for unknown level of SI is studied.

To illustrate this point, let $\boldsymbol{Z}$, $\boldsymbol{Z}_\triangle$ be the SI *received* at the destination at the coarse and at the refinement stages, respectively, and assume that for the purpose of decoding at the refinement stage, the decoder has at hand both $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_\triangle$, i.e., $\boldsymbol{Y} = (\boldsymbol{Z}, \boldsymbol{Z}_\triangle)$. Then the Markov structure $\boldsymbol{X} \multimap \boldsymbol{Y} \multimap \boldsymbol{Z}$ does hold.

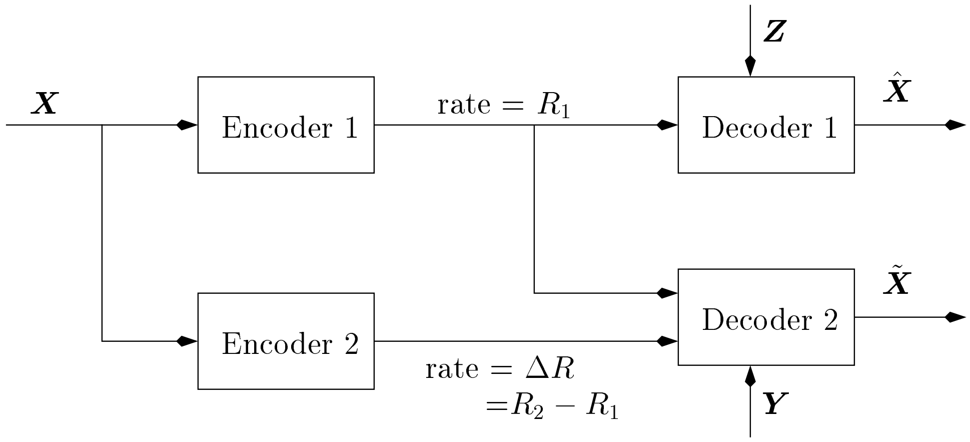

Figure 1: Successive refinement with side information (SI) at the decoder.

As an application of successive coding with increasing levels of SI, consider again the setup of [14]: We have an analog communication channel, which was upgraded with a digital channel. Since back–compatibility for subscribers utilizing only the analog part is kept, coded source-words are sent over the digital channel and, simultaneously, the source is sent uncodedly over the analog channel. The output of the analog channel serves as SI at the decoder. In this setup, whenever refinement information is sent over the digital channel (by user request, or by availability of digital channel resources), additional SI $\boldsymbol{Z}_\triangle$ arrives as well at the decoder via the analog channel. If all past SI is being stored at the destination, we have a Wyner–Ziv successive coding scheme with the Markov structure $\boldsymbol{X} \multimap \boldsymbol{Y} \multimap \boldsymbol{Z}$.

Analogously to (1), a source is successively refinable if

$$R_1 = R^*_{X|Z}(D_1), \quad \text{and} \quad R_1 + \Delta R = R^*_{X|Y}(D_2). \tag{4}$$

Note, however, that the codewords sent at the first stage are supposed to be decoded with the "weak" SI $\boldsymbol{Z}$, and thus some rate might be lost relative to the case where the vector $\boldsymbol{Y}$ is available to the decoder already at the first stage. In such a case, the source cannot be successively refinable. It turns out that the random variable U, achieving the Wyner-Ziv rate distortion function (2) at distortion level D_1, plays a key role in determining whether (4) holds. Specifically, we say that the SI $\boldsymbol{Y}$ is *equivalent* to the SI $\boldsymbol{Z}$ at distortion level D_1, if there exists a random variable U achieving (2) at D_1, and satisfying $I(U;Y|Z)=0$. We show in Section 3.2, that one of the necessary conditions for successive refinability to hold, is that the SI $\boldsymbol{Y}$ is equivalent to $\boldsymbol{Z}$ at D_1. One special case of such equivalence is when the SI random vectors are identical, i.e., $\boldsymbol{Y}=\boldsymbol{Z}$. However, there are also other cases where successive refinability holds, and the concept of equivalence of SI turns out to be more general than the requirement $\boldsymbol{Y}=\boldsymbol{Z}$. An example of such a source and SI is given in Section 4.

A problem related to successive coding with SI, is that of coding when SI may be present (or absent), which was studied by Heegard and Berger [7] and independently by Kaspi [8]. Specifically, suppose that in the classical Wyner–Ziv setting (i.e., no successive refinement), two levels of SI may be available to the decoder: $\boldsymbol{Z}$, or a "better" description $\boldsymbol{Y}$. The encoder strives to guarantee two levels of distortion: D_1 if only $\boldsymbol{Z}$ is available, and $D_2<D_1$ if the better SI $\boldsymbol{Y}$ is available. Heegard and Berger [7] derived a single-letter formula for the minimal rate needed to guarantee the pair of distortion levels (D_1,D_2). Clearly, this model is a special case of the model considered here with $R_2=R_1$ (cf. Fig. 1), and it should be pointed out that even in this case, Heegard and Berger could find full characterization of achievable performance only under the assumption of a Markov structure $\boldsymbol{X} \multimap \boldsymbol{Y} \multimap \boldsymbol{Z}$, and the general case, of a joint distribution of (X,Y,Z), was left open. A–fortiori, it would then be hard to expect a complete characterization for $R_2 \geq R_1$.

Our model includes, as a special case, the model where the better SI, $\boldsymbol{Y}$, is present at the encoder: simply let the Y random variable be part of the source, generating an extended source $\tilde{X}=(X,Y)$, where the distortion measure is sensitive only to the X component of $\tilde{X}$ (the very same approach was taken in [11]). We can include the model where the SI $\boldsymbol{Z}$ is present at the encoder, only if Z is a deterministic function of Y. This is due to the fact that if one of the components of the extended source $\tilde{X}$ is Z, while Z is not a function of Y, the source is not degraded, i.e., we do not have the Markov chain $\tilde{X} \multimap Y \multimap Z$. In this context, another special case of our setting is that treated recently by F.-W. Fu and R. Yeung in [5]. Specifically, Fu and Yeung derived a single letter expression of the rate-distortion region for the multiple descriptions problem, where one of the side decoders is required to reproduce – without distortion – a deterministic, single-letter function of the source. This problem corresponds to our model, with $R_2>R_1$, no SI at the first stage (degenerate $\boldsymbol{Z}$), and $\boldsymbol{Y}$ available to both encoder and second stage decoder. To see this, interpret the extra bits in R_2 of [5], above the entropy of the deterministic function of the source, as the rate of our refinement stage.

This paper is organized as follows: Notation and definitions are provided in Section 2. The main results are presented and discussed in Section 3. Examples of sources which are successively refinable with SI are given in Section 4. In Section 5, we present an extension of part of the results of Section 3 to multistage successive coding schemes. Finally, the proofs of the main theorems are given in Section 6.

2 Notation and Preliminaries

We begin by setting up the notation. Let $\mathcal{X}$ be a finite set and let $\mathcal{X}^n$ be the set of all n-vectors with components in $\mathcal{X}$. A member of $\mathcal{X}^n$ will be written as $x^n = (x_1, x_2, \cdots, x_n)$, and substrings of x^n are written as $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$ for $i \leq j$. When $i = 1$, the subscript will be omitted. When the dimension is clear from the context, vectors will be denoted by boldface letters, e.g., $\boldsymbol{x} \in \mathcal{X}^n$. A similar convention is used for random variables and vectors, which are denoted by upper case letters. 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 taking values in $\mathcal{X}$ with a generic distribution P_X, i.e.,

$$P_X(x^n) = \prod_{i=1}^{n} P_X(x_i).$$

Similarly, a triple source $(\mathcal{X}\mathcal{Y}\mathcal{Z}, P_{XYZ})$ is an infinite sequence of independent copies of the triplet of random variables (X, Y, Z) taking values in the finite sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively, with generic joint distribution P_{XYZ}. The induced marginals and conditional distributions are denoted by the corresponding subscripts, e.g., P_X, P_{YZ}, $P_{Z|X}$, etc. Whenever clear from the context, these subscripts will be omitted, e.g., $P_{XZ}(x, z)$ will sometimes be denoted simply by $P(x, z)$. Also, with a slight abuse of notation, and when there is no room for ambiguity, we will denote a source $(\mathcal{X}, P_X)$ by referring to its generic distribution P_X or random variable X.

We are interested in coding the source X. Let $\hat{\mathcal{X}}$ stand for a finite reconstruction alphabet, and let

$$d: \quad \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$$

be a single–letter distortion measure. The vector distortion measure is defined in the usual way, as

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

Definition 1 An (n, M_1, M_2, D_1, D_2) *successive refinement (SR) code* for the source X with SI (Y, Z) consists of a first–stage encoder–decoder pair (ϕ_1, ψ_1):

$$\begin{aligned} \phi_1 : &\quad \mathcal{X}^n \to \{1, 2, \ldots, M_1\} \\ \psi_1 : &\quad \{1, 2, \ldots, M_1\} \times \mathcal{Z}^n \to \hat{\mathcal{X}}^n, \end{aligned} \tag{5}$$

and a second–stage (or *refinement*) encoder–decoder pair (ϕ_2, ψ_2):

$$\begin{aligned} \phi_2 : &\quad \mathcal{X}^n \to \{1, 2, \ldots, M_2\} \\ \psi_2 : &\quad \{1, 2, \ldots, M_1\} \times \{1, 2, \ldots, M_2\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n, \end{aligned} \tag{6}$$

such that

$$\mathbb{E}d(X^n, \psi_1(\phi_1(X^n), Z^n)) \le D_1 \tag{7}$$

and

$$\mathbb{E}d(X^n, \psi_2(\phi_1(X^n), \phi_2(X^n), Y^n)) \le D_2, \tag{8}$$

where $\mathbb{E}$ stands for the expectation operation.

Remark 1 One may consider a slightly more general definition, without any difficulty in the analysis, where different distortion measures are used in the two stages, i.e., d_1 in (7) and d_2 in (8). For the sake of simplicity, however, we use the same distortion measure at both stages.

The *rate pair* (R_1, R_2) of the (n, M_1, M_2, D_1, D_2) SR code is

$$\begin{aligned} R_1 &= \frac{1}{n}\log M_1, \\ R_2 &= \frac{1}{n}\log(M_1 M_2). \end{aligned} \tag{9}$$

Given a distortion pair $\boldsymbol{D} = (D_1, D_2)$, a rate pair (R_1, R_2) is said to be $\boldsymbol{D}$*–achievable* for X with SI (Y, Z) if, for any $\delta > 0$, $\epsilon > 0$, and sufficiently large n, there exists an $(n, \exp[n(R_1 + \delta)], \exp[n(R_2 -$

$R_1+\delta)], D_1+\epsilon, D_2+\epsilon)$ SR code[3] for the source X with SI (Y, Z). The collection of all $\boldsymbol{D}$-achievable rate pairs is the achievable SR region for coding with SI, and is denoted by $\mathcal{R}(\boldsymbol{D})$.

In the sequel, special attention will be given to the case where the SI, available at the two stages, is identical, i.e., $\mathcal{Y} = \mathcal{Z}$ and $P_{XY} = P_{XZ}$. Therefore, for convenience, we will denote the achievable SR region for coding with identical SI in the two stages by $\mathcal{R}_i(\boldsymbol{D})$.

An immediate consequence of Definition 1 is that $\mathcal{R}(\boldsymbol{D})$ depends on the joint distribution P_{XYZ} only via the marginals P_{XY} and P_{XZ}. This is due to the fact that each decoder sees only its own SI, and has no access to the SI of the other decoder. Thus, the joint distribution of Y and Z is not of prime importance; instead, what counts is the joint distribution of each of the SI with the source X. This is a reminiscent of the situation in broadcast channels [1]. Recall that the capacity region of a broadcast channel (BC) $Q_{Y_1,Y_2|X}$, is determined only by its conditional marginals $Q_{Y_1|X}$ and $Q_{Y_2|X}$, again due to the fact that each of the decoders has access only to its own channel output. Following the terminology in [7] (and also common in the context of broadcast channels [1]), a source P_{XYZ} is said to be *stochastically degraded* if there exists a source $P_{\tilde{X}\tilde{Y}\tilde{Z}}$ such that $\tilde{X} \multimap \tilde{Y} \multimap \tilde{Z}$ is a Markov chain, $P_{\tilde{X}\tilde{Y}} = P_{XY}$, and $P_{\tilde{X}\tilde{Z}} = P_{XZ}$. Since the achievable rate region $\mathcal{R}(\boldsymbol{D})$ depends only on the pair marginals, no distinction has to be made between physically degraded (Markov structured) sources and stochastically degraded sources. This is again in full analogy with broadcast channels, where, from the point of view of channel capacity, no distinction is made between stochastically degraded and physically degraded channels, and they are generally termed as *degraded channels*. In particular, in the case of identical SI, that was mentioned in the previous paragraph, we might as well assume that $Y = Z$ with probability one, without loss of generality.

At this point, we would like to clarify the difference between the concepts of identical SI and equivalent SI. We are given a source P_{XYZ} with marginals P_{XY} and P_{XZ}. Our source P_{XYZ} is (physically or stochastically) degraded, hence there exists a conditional distribution $\tilde{P}_{Z|Y}$ such that $\sum_y \tilde{P}_{Z|Y}(z|y)P_{XY}(x,y) = P_{XZ}(x,z)$. Now let us examine the Wyner-Ziv (WZ) problem defined by the first (coarse) stage. Specifically, let $Q_{U|X}$ be the distribution achieving the minimum in (2), at distortion level D_1. Clearly, $Q_{U|X}P_{XZ}$ defines a Markov chain $U \multimap X \multimap Z$. Since the source is degraded, we can define $Q_{UXYZ} = Q_{U|X}\tilde{P}_{Z|Y}P_{XY}$, which results in the Markov chain $U \multimap X \multimap Y \multimap Z$. Observe that $Q_{XZ} = P_{XZ}$, $Q_{XY} = P_{XY}$, $Q_{UXZ} = Q_{U|X}P_{XZ}$, i.e., the marginal distributions that

[3]Whenever such notation is used, we assume that the codebook sizes $\exp[n(R_1+\delta)]$ and $\exp[n(R_2-R_1+\delta)]$ are integers.

determine the achievable successive refinement rates, and the solution to the WZ problem defined by the coarse stage, remain intact. The distribution Q_{UXYZ} clearly also implies the Markov relation

$$U \multimap Y \multimap Z. \tag{10}$$

Now the two SI random variables Y and Z are equivalent at distortion level D_1, if the roles of Y and Z in the Markov relations (10) can be interchanged, i.e., if Q_{UXYZ} also implies

$$U \multimap Z \multimap Y. \tag{11}$$

This holds true, for example, with *any* random variable U, if in the original (degraded) source, we have $P_{XY} = P_{XZ}$ – that is, if the SI is identical. (To see this, recall that with identical SI we can set $Y = Z$ with probability one, thus Y and Z are basically the same.) But note that the notion of equivalent SI is more general than the notion of identical SI, and in particular, it may depend on the solution of the WZ problem of the first stage, via the random variable U. As (11) implies $I(U;Y|Z) = 0$, the notion of equivalence says that there is no additional information about U in Y, beyond what is already given by the SI of the coarse stage, Z. Intuitively, one can say that the notion of equivalence is more general than the notion of identical SI, since the information contents of Y and Z are measured with respect to U, rather than with respect to the stream to be encoded, X. We give an example of the use of his concept in Section 4.

For a general stochastically degraded source, we will occasionally use the terms *good* and *bad* SI to signify Y and Z, respectively. Finally, we would like to remark that, for simplicity of notation and exposition, the source P_{XYZ} is treated throughout the paper as being physically degraded. In view of the discussion in previous paragraphs, it is understood that if the original source is stochastically degraded, then it can be represented by a physically degraded source with the same pair marginals P_{XY} and P_{XZ}, and the resulting rate-distortion regions are the same.

3 Main Results

In this section, we present a single-letter characterization of the region of achievable rates for successive coding with Markov structured SI. Based on this characterization, we derive necessary and sufficient conditions for successive refinability.

3.1 The Rate-Distortion Region for the Degraded Case

Let a distortion pair $\boldsymbol{D} = (D_1, D_2)$ be given. Define $\mathcal{R}^*(\boldsymbol{D})$ to be the set of all rate pairs (R_1, R_2) for which there exists a triple of random variables (U, V, W), taking values in finite alphabets, $\mathcal{U}$, $\mathcal{V}$, $\mathcal{W}$, respectively, such that the following conditions are satisfied:

1. $(U, V, W) \multimap X \multimap Y \multimap Z$ is a Markov chain.
2. There exist deterministic maps

$$\begin{aligned} f_1: & \quad \mathcal{U} \times \mathcal{Z} \to \hat{\mathcal{X}} & (12) \\ f_2: & \quad \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}, & (13) \end{aligned}$$

   such that

$$\begin{aligned} \mathbb{E} d(X, f_1(U, Z)) & \leq D_1 & (14) \\ \mathbb{E} d(X, f_2(W, Y)) & \leq D_2 & (15) \end{aligned}$$

3. The alphabets $\mathcal{U}, \mathcal{V}, \mathcal{W}$ satisfy:

$$\begin{aligned} |\mathcal{U}| & \leq |\mathcal{X}| + 2, & (16) \\ |\mathcal{V}| & \leq (|\mathcal{X}| + 1)^2, & (17) \\ |\mathcal{W}| & \leq |\mathcal{X}|(|\mathcal{X}| + 2)(|\mathcal{X}| + 1)^2 + 1. & (18) \end{aligned}$$

4. The rates R_1 and R_2 satisfy

$$\begin{aligned} R_1 & \geq I(X; U|Z) + I(X; V|U, Y), & (19) \\ R_2 - R_1 & \geq I(X; W|U, V, Y). & (20) \end{aligned}$$

Our main result is the following:

Theorem 1 For any discrete memoryless stochastically degraded source, $X \multimap Y \multimap Z$,

$$\mathcal{R}(\boldsymbol{D}) = \mathcal{R}^*(\boldsymbol{D}).$$

The proof appears in Section 6.

Remark 2 Due to the conditioning on U in (19) and on U and V in (20), we can add structure to the auxiliary random variables in the above definition of $\mathcal{R}^*(\boldsymbol{D})$, at the expense of increased alphabet sizes. Specifically, define $\tilde{V} = (U, V)$, $\tilde{W} = (U, V, W)$. Then we have, instead of condition 1 in the definition of $\mathcal{R}^*(\boldsymbol{D})$, the following structure

$$U \multimap \tilde{V} \multimap \tilde{W} \multimap X \multimap Y \multimap Z \tag{21}$$

and the bounds on the rates now read

$$\begin{aligned} R_1 &\geq I(X; U|Z) + I(X; \tilde{V}|U, Y), &(22)\\ R_2 - R_1 &\geq I(X; \tilde{W}|\tilde{V}, Y). &(23) \end{aligned}$$

It is easy to verify that the distortion levels are not altered by this substitution. Adding this Markov structure results in increased alphabet sizes $|\tilde{\mathcal{V}}| = |\mathcal{U}| \cdot |\mathcal{V}|$, and $|\tilde{\mathcal{W}}| = |\mathcal{U}| \cdot |\mathcal{V}| \cdot |\mathcal{W}|$.

In this respect, one can claim that there is already a Markov structure in the proof of the converse. Note, however, that the long Markov chain in the proof of the converse holds with respect to the auxiliary random variables with unbounded alphabets sizes, that is, up to the step where we use Caratheodory's theorem to bound the alphabets sizes. The use of the support lemma guarantees that the values of the information functionals ($I(X; U|Y)$ etc.) and average distortions are being kept with the reduced alphabets, but nothing is guaranteed about the *structure* of the reduced alphabet variables. That is, in the process of reducing the alphabets sizes, the values are preserved, but the structure of the new variables can be altered. If we want to keep the Markov structure also after bounding the alphabet sizes, we have to add constraints in the use of the support lemma, in the spirit of (138) and (146). This will result, of course, in increased alphabets sizes.

Discussion: It is instructive to examine the relations between R_2 and R_1, and the definitions of achievable rates that appeared in previous works. First, observe that if $R_2 = R_1$, i.e., there is no excess rate in the refinement stage, $\mathcal{R}^*(\boldsymbol{D})$ coincides with the region given in [7, Theorem 3]. In the general case (not necessarily $R_2 = R_1$), the random variables U and V in eqs. (19) and (20) represent the information about X that is sent to the decoder at the first stage, at rates $I(X; U|Z)$ and $I(X; V|U, Y)$, respectively. Due to the conditioning on Y, the codewords represented by V can be decoded only at the refinement stage. For a given choice of (U, V, W) that satisfies the above conditions, let us write

$$R_1 = R_{1,1} + R_{1,2} \tag{24}$$

where

$$
\begin{aligned}
R_{1,1} &= I(X;U|Z) && (25)\\
R_{1,2} &= I(X;V|U,Y). && (26)
\end{aligned}
$$

Thus, for given R_1, there is a tradeoff between the part of the rate that can be utilized at the first stage, $R_{1,1}$, and the part that must wait to be decoded at the refinement stage, $R_{1,2}$. This tradeoff is controlled by the choice of the random variables U and V. One may question the wisdom of sending at the first stage information that can be decoded only at the refinement stage. Note, however, that coding rate can be better utilized at the second stage due to the presence of better SI, Y. Therefore, coding with $R_{1,2} > 0$ is beneficial whenever we want to utilize Y beyond what is possible with the limited refinement rate $R_2 - R_1$. Since V can be decoded only at the second stage, there is a degree of freedom in deciding whether to send it in the first or the second stage. This can be seen from eq. (19) as follows: Let (R_1, R_2, D_1, D_2) be achievable with a certain triplet (U, V, W), and decompose R_1 as in (24)–(26). Define $\tilde{W} = (V, W)$, let $\tilde{V}$ be a null random variable (e.g., constant), and let $(\tilde{R}_1, \tilde{R}_2, \tilde{D}_1, \tilde{D}_2)$ be the rates and distortion levels achievable with the triplet $(U, \tilde{V}, \tilde{W})$. Since the functions f_1, f_2 do not depend on V, the distortion levels do not change, that is

$$
(\tilde{D}_1, \tilde{D}_2) = (D_1, D_2) \tag{27}
$$

and for the new rates, we have

$$
\begin{aligned}
\tilde{R}_1 &= R_1 - R_{1,2} = R_{1,1} && (28)\\
\tilde{R}_2 &= R_2, && (29)\\
\tilde{R}_2 - \tilde{R}_1 &= R_2 - R_1 + R_{1,2}. && (30)
\end{aligned}
$$

We take this opportunity to point out a difference between our definitions of achievable rates and those used in [9] and [12]. For reasons that will be clarified shortly, let us denote by (r_1, r_2) achievable rates according to the results of [12]. Then, Koshelev and Rimoldi present the achievable rates as

$$
\begin{aligned}
r_1 &\geq I(X;\hat{X}_1), \quad \mathbb{E}d(X,\hat{X}_1) \leq D_1, && (31)\\
r_2 &\geq I(X;\hat{X}_1\hat{X}_2), \quad \mathbb{E}d(X,\hat{X}_2) \leq D_2. && (32)
\end{aligned}
$$

Observe that such a presentation includes, in the achievable region, pairs (r_1, r_2) for which $r_1 > r_2$. This is consistent with the formal definition of achievability in [9] and [12], as in their definition,

r_1 and r_2 are achievable if for every $\delta > 0$ there exists a SR code whose *actual* rate at the first stage is *at most* $r_1 + \delta$, and its *actual* total cummulative rate is *at most* $r_2 + \delta$. Thus, according to Koshelev, and later Rimoldi, r_1 and r_2 indicated by (31), (32), need not be the actual rates of the codes in use, but are rather upper bounds. The advantage of the definition of achievability used in [12], over our definition of achievability, lies in the fact that for our model, it leads to a simpler characterization of the rate-distortion region. Specifically, (r_1, r_2, D_1, D_2) is achievable (in the sense of [12]), if and only if there exist random variables U and $\tilde{W}$, and deterministic functions f_1 and f_2, such that

$$\begin{aligned} r_1 &\geq I(X;U|Z) && (33) \\ r_2 &\geq I(X;U|Z) + I(X;\tilde{W}|U,Y) && (34) \\ \mathbb{E}d(X, f_1(U,Z)) &\leq D_1 && (35) \\ \mathbb{E}d(X, f_2(\tilde{W},Y)) &\leq D_2. && (36) \end{aligned}$$

For convenience, let us denote by $\hat{\mathcal{R}}^*(\boldsymbol{D})$ the region defined by (33)-(36). We show in Section 7, that $\mathcal{R}^*(\boldsymbol{D})$ and $\hat{\mathcal{R}}^*(\boldsymbol{D})$ are equivalent. The characterization (33)-(36) is simpler than the definition of $\mathcal{R}^*(\boldsymbol{D})$ in that it involves only two auxiliary random variables. The alphabet size of $\tilde{W}$ in (34) and (36) is equal to the alphabet size of V in our characterization (by arguments similar to those used in our proof of the converse). Note that the auxiliary random variable V, representing the rate sent at the first stage and decoded at the second stage, is not presented in this characterization, and thus it does not reflect the interplay between the two stages. In spite of the said advantages of the definition of achievability used in [12], we prefer to use the convention of Definition 1, where R_1 and R_2 are the *actual* rates and not merely upper bounds on rates. This definition better reflects the structure of the codes used, and avoids situations where a pair (R_1, R_2) with $R_1 > R_2$ is (unnaturally) considered achievable. In particular, it necessarily leads to a *characterization of achievable region* in terms of the rates of each of the stages separately, that is, characterization of R_1 and $\Delta R = R_2 - R_1$. Such a characterization is more suitable to describe actual systems. For example, consider a case where the bit streams of the coarse and the refinement stages are transmitted via two separate rate-limited channels (as will usually be the case in practice). Then R_1 and ΔR should be separately compared to the capacities of the corresponding channels. In addition, the characterization of $\mathcal{R}^*(\boldsymbol{D})$, with the presence of the "intermediate" random variable V, better reflects the structure of the actual code, with the interplay between the rates of the two stages – an interplay faced by the system designer.

We now turn to the important special case of identical SI. Let us define $\mathcal{R}_i^*(\boldsymbol{D})$ similarly as $\mathcal{R}^*(\boldsymbol{D})$ where $Y \equiv Z$ (i.e., Y and Z are identical SI), the bounds on the alphabet sizes (16) and (18) are replaced by

$$|\mathcal{U}| \leq |\mathcal{X}| + 2 \tag{37}$$

$$|\mathcal{W}| \leq |\mathcal{X}|(|\mathcal{X}| + 2) + 1 \tag{38}$$

and the rate inequalities (19) and (20) are replaced by the following inequalities:

$$R_1 \geq I(X;U|Y), \tag{39}$$

$$R_2 - R_1 \geq I(X;W|UY). \tag{40}$$

Defining $\mathcal{R}_i(\boldsymbol{D})$ analogously to $\mathcal{R}(\boldsymbol{D})$ with the restriction of identical SI, we now have the following corollary to Theorem 1:

Corollary 1 For a discrete memoryless joint source (X, Y):

$$\mathcal{R}_i(\boldsymbol{D}) = \mathcal{R}_i^*(\boldsymbol{D}).$$

Proof. In view of Theorem 1, we have to show that when $Y = Z$, the characterization of $\mathcal{R}^*(\boldsymbol{D})$ reduces to that of $\mathcal{R}_i^*(\boldsymbol{D})$. Indeed, if $Y = Z$, by (19), (20) we can write

$$R_1 \geq I(X;U|Z) + I(X;V|U,Y) = I(X;UV|Y) \tag{41}$$

$$R_2 - R_1 \geq I(X;W|UVY). \tag{42}$$

Observe that the auxiliary random variables U and V appear in the mutual information functions in (41), (42) always as a pair. The functions f_1 and f_2 do not depend on V. Thus we can define a new auxiliary random variable $\tilde{U} = (U, V)$ and a new function $\tilde{f}_1(\tilde{U}) = f_1(U)$, without altering the distortions or the rates. This proves the corollary with the only exception that the alphabet size of $\mathcal{W}$ is given by (18) instead of (38), and that of $\tilde{\mathcal{U}}$ is given by

$$|\tilde{\mathcal{U}}| = |\mathcal{U}| \cdot |\mathcal{V}| \tag{43}$$

instead of the right hand side of (37). The fact that we can restrict the alphabet sizes to those given by (37) and (38), follows from Carathéodory's theorem in a manner similar to the bounds on the alphabet sizes in the proof of Theorem 1. We omit these details here. $\square$

3.2 Successive Refinability

The notion of successive refinability with SI is now defined.

Definition 2 A source X is said to be *successively refinable from D_1 to D_2 ($D_1 > D_2$) with SI* if

$$(R^*_{X|Z}(D_1), R^*_{X|Y}(D_2)) \in \mathcal{R}(D_1, D_2). \tag{44}$$

Similarly, a source X is said to be *successively refinable from D_1 to D_2 ($D_1 > D_2$) with identical SI, Y*, if

$$(R^*_{X|Y}(D_1), R^*_{X|Y}(D_2)) \in \mathcal{R}_i(D_1, D_2). \tag{45}$$

A characterization of sources which are successively refinable with SI, in the spirit of the results of Koshelev and Equitz and Cover, is given next.

Theorem 2 A source X with degraded SI (Y, Z) is successively refinable from D_1 to D_2 if and only if there exist a pair of random variables (U, W) and a pair of deterministic maps $f_1 : \mathcal{U} \times \mathcal{Z} \to \hat{\mathcal{X}}$, and $f_2 : \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}$, such that the following conditions simultaneously hold:

1. $R^*_{X|Z}(D_1) = I(X; U|Z) \quad \text{and} \quad \mathbb{E}d(X, f_1(U, Z)) \leq D_1$,

2. $R^*_{X|Y}(D_2) = I(X; W|Y) \quad \text{and} \quad \mathbb{E}d(X, f_2(W, Y)) \leq D_2$,

3. $(U, W) \multimap X \multimap Y \multimap Z$ form a Markov chain,

4. $U \multimap (W, Y) \multimap X$ form a Markov chain.

5. $I(U; Y|Z) = 0$.

Remark 3 Recall that we can use the alternative characterization of $\mathcal{R}^*(\boldsymbol{D})$, according to Remark 2, where the auxiliary random variables have the Markov structure $U \multimap V \multimap W \multimap X \multimap Y \multimap Z$, and increased alphabet sizes. Note that any Markov chain $U \multimap W \multimap X \multimap Y \multimap Z$ satisfies conditions 3 and 4 of Theorem 2. Therefore, in such a case, conditions 3 and 4 in Theorem 2 are replaced by one condition $U \multimap W \multimap X \multimap Y \multimap Z$. We will make use of this fact in the examples in Section 4.

Condition 3 of Theorem 2 implies the Markov structure $U \multimap Y \multimap Z$, whereas Condition 5 requires that the roles of Y and Z can be interchanged, i.e., $U \multimap Z \multimap Y$. When such relations indeed hold, we say that the SI Y and Z are *equivalent* at distortion level D_1 (see also discussion of this point in Section 1, and at the end of Section 2). Note that this condition, in general, depends on the distortion measure and on the distortion level (via U), hence it is more general than just requiring that the SI at the two stages be identical: $P_{XY} = P_{XZ}$. We give examples of such sources and SI in Sections 4.2 and 4.3.

The random variable V does not play any role in the conditions of Theorem 2. As discussed in the paragraphs following Theorem 1, V is used to encode, at the first stage, information that can be decoded only at the second stage. It turns out that such a coding scheme is suboptimal for successively refinable sources – if the rate-distortion function can be achieved simultaneously at both distortion levels, then the first stage should transmit information that can be decoded as a whole by the first decoder.

Proof of Theorem 2. We begin with the necessity part. Assume that (44) holds. From Theorem 1, there exists a triple of random variables (U, V, W) and a pair of maps $f_1 : \mathcal{U} \times \mathcal{Z} \to \hat{\mathcal{X}}$, $f_2 : \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}$, such that

$$(U, V, W) \multimap X \multimap Y \multimap Z \tag{46}$$

is a Markov chain, and moreover

$$\begin{aligned} R^*_{X|Z}(D_1) &\geq I(X; U|Z) + I(X; V|U, Y), && (47a) \\ \mathbb{E}d(X, f_1(U, Z)) &\leq D_1, && (47b) \end{aligned}$$

and

$$\begin{aligned} R^*_{X|Y}(D_2) &\geq R^*_{X|Z}(D_1) + I(X; W|UVY) \\ &\geq I(X; U|Z) + I(X; VW|UY), && (48a) \\ \mathbb{E}d(X, f_2(W, Y)) &\leq D_2. && (48b) \end{aligned}$$

By definition of $R^*_{X|Z}(D_1)$, we conclude that

$$I(X; V|U, Y) = 0 \tag{49}$$

and, in addition, (47a) is satisfied with equality. Thus condition 1 of Theorem 2 holds. The rate bound (48a) can be written as

$$R^*_{X|Y}(D_2) \quad \geq \quad I(X; U|Z) + I(X; VW|UY)$$

$$
\begin{aligned}
&= I(X;U|Y) + I(X;VW|UY) + [I(X;U|Z) - I(X;U|Y)] \\
&\overset{(a)}{=} I(X;UVW|Y) + I(U;Y|Z) \\
&\geq I(X;W|Y) + I(U;Y|Z) \\
&\overset{(b)}{\geq} R^*_{X|Y}(D_2) + I(U;Y|Z). \qquad (50)
\end{aligned}
$$

where in (a) we used the Markov structure (46), and (b) holds since knowledge of W suffices to satisfy the distortion constraint (see (48b)). Hence, the inequalities in (50) can be replaced by equalities, and we conclude that

$$I(U;Y|Z) = 0. \qquad (51)$$

Therefore, conditions 2 and 5 of the theorem hold. Moreover, from (50) we also conclude that

$$H(X|U,V,W,Y) = H(X|W,Y), \qquad (52)$$

which can hold only if $(U,V) \multimap (W,Y) \multimap X$. Hence condition 4 is satisfied. Condition 3 follows from (46). This completes the proof of the necessity part.

We proceed to prove the sufficiency part. Assume that there exist a pair of random variables (U,W) and deterministic maps f_1, f_2, satisfying conditions 1 to 5 of Theorem 2. By the definition of $\mathcal{R}^*(D_1,D_2)$, it remains to show that there exists a random variable V such that (46) holds, and

$$
\begin{aligned}
R^*_{X|Z}(D_1) &\geq I(X;U|Z) + I(X;V|U,Y), \qquad (53) \\
R^*_{X|Y}(D_2) - R^*_{X|Z}(D_1) &\geq I(X;W|U,V,Y). \qquad (54)
\end{aligned}
$$

Indeed, let V be a null random variable (i.e., a constant). Then Condition 3 implies (46). Condition 1 of the theorem implies (53). And for the differential rate we have

$$
\begin{aligned}
R^*_{X|Y}(D_2) - R^*_{X|Z}(D_1) &\overset{(a)}{=} I(X;W|Y) - I(X;U|Z) \\
&\overset{(b)}{=} I(X;W|Y) - I(X;U|Y) \\
&= H(X|UY) - H(X|WY) \\
&\overset{(c)}{=} H(X|UY) - H(X|UWY) \\
&= I(X;W|UY) \\
&\overset{(d)}{=} I(X;W|UVY), \qquad (55)
\end{aligned}
$$

where (a) follows from Conditions 1 and 2 of Theorem 2, (b) follows from Condition 5, (c) follows from Condition 4, and (d) holds since V is independent of all the random variables in the problem. This completes the proof of the sufficiency part. $\square$

When there is no SI, the conditions of Theorem 2 reduce to those obtained in [9] and [4] for successive refinability without SI. To see this, let Y be deterministic and define $\tilde{f}_1(U) \stackrel{\triangle}{=} f_1(U, Y)$ and $\tilde{f}_2(W) \stackrel{\triangle}{=} f_2(W, Y)$. Conditions 1 and 2 of Theorem 2 become the classical rate-distortion functions. Conditions 3 and 5 become redundant, and Condition 4 is read $U \multimap W \multimap X$. It remains to show that whenever X is successively refinable, the joint distributions $P_{X,\tilde{f}_2(W)}$ and $P_{X,\tilde{f}_1(U)}$ form a degraded source, in the sense that there exists a conditional distribution $P_{\tilde{f}_1(U)|\tilde{f}_2(W)}$, such that the Markov structure

$$\tilde{f}_1(U) \multimap \tilde{f}_2(W) \multimap X \tag{56}$$

holds. Notice that for successively refinable source (48b) and (50) imply

$$\begin{aligned} R(D_2) &= I(X;W) && (57) \\ \mathbb{E}d(X, \tilde{f}_2(W)) &\leq D_2. && (58) \end{aligned}$$

From the data processing theorem

$$I(X;W) \geq I(X;\tilde{f}_2(W)), \tag{59}$$

implying that

$$I(X;W) = I(X;\tilde{f}_2(W)), \tag{60}$$

as otherwise, in view of (57), the coding rate can be reduced to below the rate distortion function $R(D_2)$, which is a contradiction. The Markov condition $X \multimap W \multimap \tilde{f}_2(W)$ clearly holds. We now claim that (60) implies that the roles of W and $\tilde{f}_2(W)$ can be interchanged, that is, the Markov structure

$$X \multimap \tilde{f}_2(W) \multimap W \tag{61}$$

holds. A short proof of this claim is given in Section 8. Since $X \multimap W \multimap U$ holds, we can construct, based on (61), the Markov chain

$$X \multimap \tilde{f}_2(W) \multimap W \multimap U \multimap \tilde{f}_1(U) \tag{62}$$

from which the desired structure (56) follows.

It is an easy matter to construct quadruples (U, W, X, Y) where no Markov structure holds for (U, W, X), yet $U \multimap\!\!\!- (W, Y) \multimap\!\!\!- X$, i.e., Markovity does hold once SI is introduced. Therefore, the presence of Y in condition 4 indicates that a source, that is not successively refinable in the absence of SI, may turn out to be successively refinable when SI is present. We give an example of such a source in Section 4.

4 Examples of Successively Refinable Sources

In this section, we provide four examples of successively refinable sources in the presence of SI at both decoders. The first example is the Gaussian source with quadratic distortion measure, the second is the doubly–symmetric binary source with the Hamming distortion measure. Our main results were presented for sources and SI of finite alphabet. The justification to examples of continuous alphabets parallels the lines of the results of Wyner for the continuous alphabet case [15], and is therefore omitted here. In the third example, the SI at the second stage is strictly better than the SI of the first stage, and yet the source is successively refinable, since the SI's are equivalent at distortion level D_1. The fourth example is a DMS that is not successively refinable in the absence of SI [4], but becomes successively refinable in the presence of SI. In the first two examples, we will use Theorem 2 in conjunction with Remark 3, i.e., we demonstrate a Markov chain $U \multimap\!\!\!- W \multimap\!\!\!- X \multimap\!\!\!- Y$ that meets the conditions of Theorem 2. In addition, for the doubly-symmetric binary source, we construct a degraded SI component, Z, such that $U \multimap\!\!\!- Z \multimap\!\!\!- Y$ holds, therefore it is successively refinable, although the SI's at the two stages is not strictly identical.

4.1 The Gaussian Source with Quadratic Distortion

The Wyner–Ziv rate–distortion function of this source is mentioned somewhat briefly in [16], and done in great deal of detail in [15]. For the sake of completeness, before we show that this source is successive refinable, we first describe, in some detail, the calculation of the (single-stage) rate–distortion function that was computed in [15], and then generalize it into two stages. Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N \sim \mathcal{N}(0, \sigma_N^2)$ be independent, $Y = X + N$, and let the distortion measure be quadratic, i.e., $d(x, \hat{x}) = (x - \hat{x})^2$. To calculate the Wyner–Ziv rate–distortion function $R^*_{X|Y}(D)$, consider the decomposition[4] of X as $X = W + S$, where $W \sim \mathcal{N}(0, \sigma_W^2)$ and $S \sim \mathcal{N}(0, \sigma_S^2)$ are independent, thus, $\sigma_X^2 = \sigma_W^2 + \sigma_S^2$. It is this decomposition of σ_X^2 that controls the tradeoff between rate and distortion. Considering W as the auxiliary random variable, one readily obtains, from the Markovity of $W \multimap X \multimap Y$, the following expression for the Wyner–Ziv rate:

$$\begin{aligned} I(X;W|Y) &= I(X;W) - I(Y;W) \\ &= \frac{1}{2}\log\left(1 + \frac{\sigma_N^2}{\sigma_S^2}\right) - \frac{1}{2}\log\left(1 + \frac{\sigma_N^2}{\sigma_X^2}\right). \end{aligned} \tag{63}$$

Thus, to minimize $I(X;W|Y)$ over all choices of W, in this class, one should maximize σ_S^2 under the distortion constraint $\min_f \mathbb{E}(X - f(W,Y))^2 \le D$. Since (W, X, Y) are jointly Gaussian, and the distortion is quadratic, the best estimator f is linear, i.e., $f(W,Y) = \alpha W + \beta Y$. Upon solving for the optimum coefficients, α and β, the following relation is found between σ_S^2 and D:

$$\sigma_S^2 = \begin{cases} \frac{\sigma_N^2 D}{\sigma_N^2 - D} & D \le \sigma_{X|Y}^2 \\ \sigma_X^2 & D > \sigma_{X|Y}^2 \end{cases} \tag{64}$$

where

$$\sigma_{X|Y}^2 = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}. \tag{65}$$

On substituting this expression into (63), one obtains

$$I(X;W|Y) = \begin{cases} \frac{1}{2}\log\frac{\sigma_{X|Y}^2}{D} & D \le \sigma_{X|Y}^2 \\ 0 & D > \sigma_{X|Y}^2 \end{cases} \tag{66}$$

whose optimality, as a solution to the Wyner–Ziv problem, is evident from the fact this is also the conditional rate–distortion function when Y is available to the encoder as well.

Turning now to two stages, let W be the auxiliary random variable corresponding to distortion level $D = D_2$, and let us further decompose W as the sum of two independent Gaussian random variables $W = U + T$, U being the other auxiliary random variable of Theorem 2 , thus keeping the Markov structure $U \multimap W \multimap X \multimap Y$ (cf. Fig. 2). Since W achieves $R_2 = I(X;W|Y) = R_{X|Y}^*(D_2)$, it remains to show that U achieves $R_1 = I(X;U|Y) = R_{X|Y}^*(D_1)$ for a proper decomposition of W.

[4]In this decomposition, W can be represented as $W = \theta X + S'$, where $\theta = \sigma_W^2/\sigma_X^2$ and $S' \sim \mathcal{N}(0, \sigma_W^2(\sigma_X^2 - \sigma_W^2)/\sigma_X^2)$ is independent of X.

Since $X = U + S + T$, where U, S, and T are all independent, clearly, the sum $(S+T)$ now plays the previous role of S, and so the appropriate decomposition of W is readily obtained by choosing the variance, σ_T^2, of T such that

$$\sigma_S^2 + \sigma_T^2 = \begin{cases} \frac{\sigma_N^2 D_1}{\sigma_N^2 - D_1} & D_1 \leq \sigma_{X|Y}^2 \\ \sigma_X^2 & D_1 > \sigma_{X|Y}^2, \end{cases} \tag{67}$$

which is always possible since (67) is never smaller than (64) with $D = D_2$. The rate $R_1 = I(X;U|Y)$ will then be as in (66) with $D = D_1$, namely, $R_{X|Y}^*(D_1)$. Thus, we have shown that the Gaussian source with noisy SI is successively refinable with respect to the quadratic distortion measure.

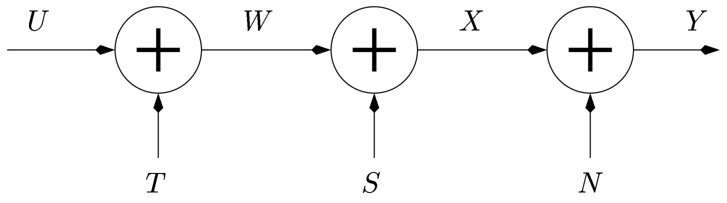

Figure 2: Successive refinement with side information for the Gaussian source with the quadratic distortion measure.

4.2 The Doubly Symmetric Binary Source with Hamming Distortion

The Wyner–Ziv rate–distortion function of the doubly–symmetric binary source with the Hamming distortion measure is calculated in detail in [16]. Roughly speaking, its calculation (as well as its extension to two stages) is analogous to the Gaussian example of subsection 4.1, with the addition operation being replaced by modulo-two addition. It is, however, somewhat more involved because beyond a certain distortion level, hereafter denoted by D_c, time–sharing with the zero–rate working point must be employed. As in subsection 4.1, for the sake of completeness, we begin with a description of the system of random variables that corresponds to the solution of the Wyner–Ziv problem of this source, in the ordinary, single–stage setting.

All random variables in this subsection are binary and their alphabet is $\{0,1\}$. Let X be the binary symmetric source (BSS) and let $Y = X \oplus N$, where $\oplus$ denotes modulo–two addition, and N is an independent random variable with $\Pr\{N=1\} = p_0$. Without loss of generality, $p_0 \leq 0.5$.

We are interested in $R^*_{X|Y}(D)$. Define the function

$$g(D) = \begin{cases} h(p_0 * D) - h(D) & 0 \le D < p_0 \\ 0 & D = p_0 \end{cases} \tag{68}$$

where $h(\cdot)$ is the binary entropy function, and $*$ denotes binary convolution, i.e., $\alpha * \beta = \alpha(1-\beta) + \beta(1-\alpha)$. In [16], it is shown that $R^*_{X|Y}(D) = g^*(D)$, the lower convex envelope of $g(D)$. The distortion level D_c is defined as the largest distortion level for which $g^*(D)$ still agrees with $g(D)$, namely, the solution to the equation

$$\frac{g(D_c)}{D_c - p_0} = g'(D_c), \tag{69}$$

where g' is the derivative of g.

We now define the auxiliary random variable as follows. Let S be a random variable, independent of (X, N), with $\Pr\{S = 1\} = \min\{D, D_c\}$, and define $W_1 = X \oplus S$. Further, let B_1 be independent of (X, N, S) with

$$\Pr\{B_1 = 1\} = \frac{p_0 - \max\{D, D_c\}}{p_0 - D_c} \tag{70}$$

and let $W_2 = B_1 \cdot W_1$. The role of B_1 is to create the time–sharing that is needed for distortion levels above D_c. Finally, define the auxiliary random variable $W = (W_2, B_1)$, and let

$$f(W, Y) = f(W_2, B_1, Y) = B_1 \cdot W_2 + (1 - B_1) \cdot Y. \tag{71}$$

It is then straightforward to show (cf. [16, Section II]) that for the Hamming distortion measure

$$\mathbb{E}d(X, f(W, Y)) = D, \tag{72}$$

and that

$$I(X; W) - I(Y; W) = g^*(D) = R^*_{X|Y}(D). \tag{73}$$

Moving on to two stages of successive refinement, consider the above construction of the auxiliary random variable W with correspondence to distortion level $D = D_2$. We next describe the construction of the additional auxiliary random variable U corresponding to the lower rate, R_1 (cf. Fig. 3). Let T be a random variable, independent of (X, N, S, B_1) such that

$$\Pr\{T = 1\} * \Pr\{S = 1\} = \min\{D_1, D_c\} \tag{74}$$

and let $U_1 = W_2 \oplus T$. Let B_2 be independent of (X, N, S, B_1, T) with

$$\Pr\{B_2 = 1\} = \frac{p_0 - \max\{D_c, D_1\}}{p_0 - \max\{D_c, D_2\}}. \tag{75}$$

Finally, defining $U_2 = U_1 \cdot B_2$, and $U = (U_2, B_1 \cdot B_2)$, we clearly have $U \multimap W \multimap X \multimap Y$. Letting,

$$f'(U, Y) = B_1 \cdot B_2 \cdot U_2 + (1 - B_1 \cdot B_2)Y. \tag{76}$$

We note here that if $B_1 = 0$, then U_1 and U_2 are independent of X and Y, and $I(X;U) - I(Y;U) = 0$. As for the distortion, when $B_1 \cdot B_2 = 0$, $f'(U, Y)$ disregards U_2, and attains the value of Y. Then, similarly as above, it is not difficult to show that

$$\mathbb{E}d(X, f'(U, Y)) = D_1, \tag{77}$$

and that

$$I(X;U) - I(Y;U) = g^*(D_1). \tag{78}$$

Thus, the doubly symmetric binary source is successively refinable with respect to the Hamming distortion measure.

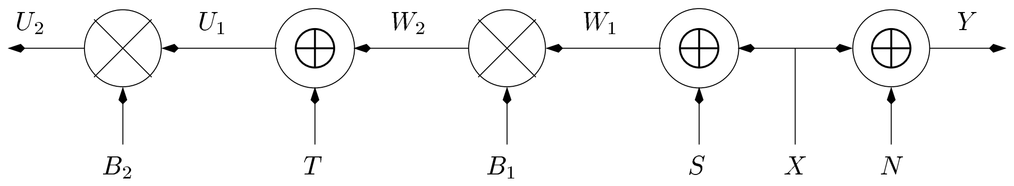

Figure 3: Successive refinement for the doubly symmetric binary source with Hamming distortion measure.

For this source, we now extend the example to include equivalent SI which is not strictly identical at the two stages, and yet the source is successively refinable. It is convenient to re-define X to be a BSS taking values in $\{-1, 1\}$. We define the SI $\tilde{Y}$ to be

$$\tilde{Y} = X \cdot \eta + Q \tag{79}$$

where η is taking values in $\{-1, 1\}$, $P(\eta = -1) = \alpha$, and Q is a real random variable uniformly distributed in the interval $[-1, 1]$. Next, let the degraded random variable Z be defined as

$$Z = \text{sign}(\tilde{Y}). \tag{80}$$

Thus Z is a quantization of $\tilde{Y}$. Moreover, note that the random variable $X \cdot \eta$ is equivalent to the random variable Y defined above, and that Q does not change its sign. Therefore, although Z is the degraded component and $\tilde{Y}$ and Z are not identical, we also have $U \multimap X \multimap Z \multimap \tilde{Y}$. Hence $I(U;Y|Z) = 0$ and the source is succesively refinable with degraded side information. Note that since the source is successively refinable, the optimal rate-distortion region for this source and SI is given by the classical Wyner-Ziv rate-distortion function, for the double-BSS.

4.3 An Example of Non-Identical SI

We demonstrate now a case where the SI available to the second stage is strictly better than the SI available to the first stage, yet the source is successively-refinable. Let Z,Z_1, Z_2, and N be independent RV's where Z,Z_1, and Z_2, take on the values $\{\pm 1\}$ with equal probabilities and N is uniform over $[-1, 1]$. Let the source X be given by

$$X = 4Z + 2Z_1 + Z_2 + N \tag{81}$$

and consider the distortion measures

$$d_1(a, b) = \begin{cases} 0 & |a - b| \leq 2 \\ \infty & |a - b| > 2 \end{cases} \tag{82}$$

and

$$d_2(a, b) = \begin{cases} 0 & |a - b| \leq \frac{1}{2} \\ \infty & |a - b| > \frac{1}{2} \end{cases} \tag{83}$$

Note that X is uniformly distributed in the interval $[-8, 8]$. (We use the notation $X \sim \text{Unif}[-8, 8]$.) The first–stage decoder has access to Z and the second–stage decoder has access to $Y \stackrel{\Delta}{=} (Z, Z_2)$.

At the first stage, the encoder can actually compute Z on his own, simply by $Z = \text{sgn}(X)$, and therefore, this is actually a problem of conditional rate–distortion, where both encoder and decoder have access to the SI. Now, given Z, $(X - 4Z) \sim \text{Unif}[-4, 4]$ and the rate needed to achieve zero distortion under d_1 is given by[5]

$$R_1 = R^*_{X|Z}(0) = \log\left(\frac{8}{4}\right) = 1 \text{ bit/sample}, \tag{84}$$

[5]It is easy to show that $R(0)$ of a DMS $X \sim \text{Unif}[-A, A]$ w.r.t. the distortion measure $d(a, b) = 0$ if $|a - b| \leq D$ and ∞ otherwise, is given by $R(0) = \log(\frac{A}{D})$.

which can be attained by uniform scalar quantization of the uniform distribution over $[-4, 4]$ into two levels. The quantized value given Z is identical $2Z_1$ with probability one. The reconstruction at the first stage is then

$$\hat{X} = 4Z + 2Z_1 = 4Z + 2\text{sgn}(X - 4Z). \tag{85}$$

At the second stage, the decoder has access to $Y = (Z, Z_2)$ plus Z_1 from the first stage. Again, as the encoder can compute (Z, Z_1, Z_2) from X, this is also a scenario of conditional rate–distortion. Now, given (Z, Z_1, Z_2), $X - 4Z - 2Z_1 - Z_2 = N$ is uniform over $[-1, 1]$, and so the rate needed at the second stage is given by

$$R_2 - R_1 = \log\left(\frac{2}{1}\right) = 1 \text{ bit/sample}, \tag{86}$$

which is again attained by scalar quantization into two levels and the reconstruction is

$$\tilde{X} = 4Z + 2Z_1 + Z_2 + \frac{1}{2}\text{sgn}(X - 4Z - 2Z_1 - Z_2). \tag{87}$$

The overall rate is then $R_2 = 2$ bits/sample. To show that this is a case of successive refineability, let us compare it to an optimal single–stage encoder which is equipped with $Y = (Z, Z_2)$ as SI, and which is aimed directly at zero distortion according to d_2. Obviously, given Z and Z_2, X is always uniform over the union of two disjoint intervals whose lengths are both 2. Thus,

$$R^*_{X|Z,Z_2}(0) = \log\left(\frac{4}{1}\right) = 2 \text{ bits/sample}, \tag{88}$$

which is exactly equal to R_2 of the two-stage code, and so the source is successively refineable.

To show that the conditions of Theorem 2 formally hold, define $U = \hat{X}$ and $W = \tilde{X}$. Then, conditions 1 and 2 have actually already been shown above. Obviously, $\hat{X} \to \tilde{X} \to X \to (Z, Z_2) \to Z$ is a Markov chain since (Z, Z_2) and $\tilde{X}$ are functions of X, and $\hat{X}$ is a function of $\tilde{X}$, so this covers conditions 3 and 4 of Theorem 2 (cf. Remark 4). Finally,

$$I(U; Y|Z) = I(Z, Z_1; Z, Z_2|Z) = I(Z_1; Z_2|Z) = 0, \tag{89}$$

which is condition 5 of Theorem 2.

4.4 The Gerrish Source

Equitz and Cover [4] demonstrate a source that is not successively refinable in the ordinary setting, without SI (see also [3] for more details). This source is attributed by them to Gerrish [6]. The Equitz–Cover example is of a ternary DMS with $\mathcal{X} = \hat{\mathcal{X}} = \{1, 2, 3\}$, a distribution given by $P_X(2) = p$, $P_X(1) = P_X(3) = (1-p)/2$, for some $p \in [0, 1]$, and a difference distortion measure $d(x, \hat{x}) = \rho(x - \hat{x})$ where ρ is convex and satisfies $\rho(0) < \rho(z)$ whenever $z \neq 0$. It turns out that for some values of D_1 and D_2, the Markov condition does not hold and thus the source is not successively refinable.

We now show a simple example where, in the presence of identical SI, this source becomes successively refinable. Consider a SI variable defined according to

$$Y = \begin{cases} 1 & X \neq 2 \\ 2 & X = 2 \end{cases} \tag{90}$$

Since Y is given by a deterministic function of X, the encoder can reproduce a copy of Y, and so, the Wyner–Ziv rate distortion function must coincide with the conditional rate–distortion function of X given Y at both encoder and decoder. Now, when $Y = 2$, $X \equiv Y$, and so, X is available at the decoder, error–free, without transmission at all. When $Y = 1$, the conditional distribution of X is according to $P_X(1) = P_X(3) = 1/2$, which is actually again an Equitz–Cover source with $p = 0$. Although the reproduction alphabet $\hat{\mathcal{X}}$ is formally still ternary, it is shown in [3, Lemma 4.1.4] that the test channel corresponding to the rate–distortion function of this source induces $\Pr\{\hat{X} = 2\} = 0$, thus, effectively, the reproduction alphabet reduces to $\hat{\mathcal{X}} = \{1, 3\}$. In other words, given $Y = 1$, we have at hand the rate–distortion problem of the BSS with respect to the Hamming distortion measure, which in turn is successively refinable [4].

5 Multistage Successive Coding

The problem of successive coding can be naturally extended to any finite number of steps. Let $(X, Y_K, Y_{K-1}, \ldots, Y_1)$ be a memoryless, stochastically degraded source, i.e.,

$$X \multimap Y_K \multimap Y_{K-1} \multimap \ldots, \multimap Y_1, \tag{91}$$

where Y_k is taking values in a finite set $\mathcal{Y}_k$, $1 \leq k \leq K$. Denote by $P_{XY_KY_{K-1}\cdots Y_1}$ its generic distribution, and by P_{XY_k}, $1 \leq k \leq K$, the corresponding marginals. Let $\boldsymbol{D} = (D_1, D_2, \ldots, D_K)$ be a vector of distortion levels. The encoder-decoder pair of the k'th stage should provide a description of the source with distortion level not exceeding D_k, To this end, the decoder of the k'th stage, $1 \le k \le K$, is provided with the side information vector $\boldsymbol{Y}_k$, where

$$\boldsymbol{Y}_k = (Y_{k,1}, Y_{k,2}, \ldots, Y_{k,n}). \tag{92}$$

Here, the second sub-index in $Y_{k,i}$ signifies time. That is, $Y_{k,i}$ is the i'th output (in time) of the k'th component Y_k of the degraded source in (91). In addition, each decoder has access to the codewords sent to all its predecessors. We skip the formal definition of a multistage successive code, as it is a straightforward extension of Definition 1 in Section 2. We use $\mathcal{R}_K(\boldsymbol{D})$ and $\mathcal{R}_{K,i}(\boldsymbol{D})$ to denote the region of achievable rates for stochastically degraded SI and identical SI, respectively, as in the two stages case.

Let a vector $\boldsymbol{D} = (D_1, \ldots, D_K)$ of distortion levels be given. Define $\mathcal{R}^*_K(\boldsymbol{D})$ to be the set of all vectors of rates $(R_1, R_2, \ldots, R_K)$ for which there exists a collection of $K(K+1)/2$ random variables $\{V_{k,l},\ 1 \le k \le K,\ k \le l \le K\}$, where $V_{k,l}$ is taking values in a finite set $\mathcal{V}_{k,l}$, such that the following conditions are satisfied

1. $(\{V_{k,l},\ 1 \le k \le K,\ k \le l \le K\}) \multimap X \multimap Y_K \multimap, \ldots \multimap Y_1$ is a Markov chain.
2. There exist deterministic maps
$$f_k : \mathcal{V}_{k,k} \times \mathcal{Y}_k \to \hat{\mathcal{X}}, \quad 1 \le k \le K, \tag{93}$$
such that
$$\mathbb{E} d(X, f_k(V_{k,k}, Y_k)) \le D_k, \quad 1 \le k \le K. \tag{94}$$
3. The rates $(R_1, R_2, \ldots, R_K)$ satisfy

$$R_1 \ge I(X; V_{1,1}|Y_1) + \sum_{l=2}^{K} I(X; V_{1,l}|V_{1,1}, V_{1,2}, \ldots, V_{1,l-1}, Y_l) \tag{95}$$

$$\begin{aligned} R_k - R_{k-1} \ \ge \ & I(X; V_{k,k}|\{V_{i,j},\ 1 \le i < k,\ i \le j \le k\}, Y_k) \\ & + \sum_{l=k+1}^{K} I(X; V_{k,l}|\{V_{i,j}, 1 \le i \le k,\ i \le j \le l-1\},\ Y_l), \quad 2 \le k \le K. \end{aligned} \tag{96}$$

Recall that in the case of two stages, the first stage transmits information that can be decoded only with the better SI Y. In the single letter characterization of the region of achievable rates, this information was presented by the random variable V. Similarly, in the multistage case, the encoder of the k'th stage can transmit information to all the successor decoders. Thus, interpreting the above characterization, for $1 \leq k \leq K$, and $k \leq l \leq K$, the random variable $V_{k,l}$ stands for the information encoded by the k'th encoder, to be decoded by the l'th decoder, using Y_l. Since the source is stochastically degraded, $V_{k,l}$ can be decoded also by any of the successor decoders, $l+1, \ldots, K$.

Theorem 3 For any discrete, memoryless, stochastically degraded $K+1$ source $(X, Y_K, \ldots, Y_1)$,

$$\mathcal{R}^*_K(\boldsymbol{D}) \subseteq \mathcal{R}_K(\boldsymbol{D}).$$

Proof. The proof employs a hierarchy of random codes, as in the proof of Theorem 1. Although technically involved, it is a straightforward extension of the proof presented in Section 6, and is omitted. □

Unfortunately, we have not been able to prove a converse of Theorem 3 for general stochastically degraded source. Note that the rate equations defining the achievable region in (95) and (96) involve several terms, each corresponding to a different decoding stage (with the appropriate SI). The decomposition of the vector mutual informations $I(X^n; T_1|Z^n)$, $I(X^n; T_2|T_1, Y^n)$, etc., used in the proof of the two-stages case, does not seem to carry to the multi-terms case. However, we have been able to prove a full coding theorem for identical SI (i.e., $P_{XY_k} = P_{XY}\ \forall k$), in which case the achievable rate region admits a much simpler form. Define $\mathcal{R}^*_{K,i}(\boldsymbol{D})$ to be the set of all vectors of rates $(R_1, R_2, \ldots, R_K)$ for which there exists a K-tuple of random variables $(U_1, U_2, \ldots, U_K)$, where U_k is taking values in a finite set $\mathcal{U}_k$, such that the following conditions are satisfied

1. $(U_1, U_2, \ldots, U_K) \multimap X \multimap Y$ is a Markov chain.
2. There exist deterministic maps
$$f_k : \mathcal{U}_k \times \mathcal{Y} \to \hat{\mathcal{X}}, \quad 1 \leq k \leq K, \tag{97}$$
such that
$$\mathbb{E} d(X, f_k(U_k, Y)) \leq D_k, \quad 1 \leq k \leq K. \tag{98}$$
3. The rates $(R_1, R_2, \ldots, R_K)$ satisfy
$$R_1 \geq I(X; U_1|Y), \tag{99}$$
$$R_k - R_{k-1} \geq I(X; U_k|U_1, U_2, \ldots, U_{k-1}, Y) \quad \text{for } k = 2, 3, \ldots, K. \tag{100}$$

Then we have

Theorem 4 For any discrete, memoryless, joint source (X, Y),

$$\mathcal{R}_{K,i}(\boldsymbol{D}) = \mathcal{R}^*_{K,i}(\boldsymbol{D}).$$

The proof is presented in Section 6.

A joint source (X, Y) is K-step successively refinable for a given vector of distortion levels $\boldsymbol{D} = (D_1, D_2, \ldots, D_K)$, if

$$(R^*_{X|Y}(D_1), R^*_{X|Y}(D_2), \ldots, R^*_{X|Y}(D_K)) \in \mathcal{R}_{K,i}(D_1, D_2, \ldots, D_K). \tag{101}$$

A by-product of Theorem 4, is a set of necessary and sufficient conditions for (101) to be satisfied.

Theorem 5 A source X with identical side information Y is K-steps successively refinable with distortion levels $\boldsymbol{D} = (D_1, D_2, \ldots, D_K)$, if and only if there exist a K-vector of random variables $(U_1, U_2, \ldots, U_K)$, and K deterministic functions $f_k : \mathcal{U}_k \times \mathcal{Y} \to \hat{\mathcal{X}}$, $1 \leq k \leq K$, such that the following conditions simultaneously hold:

1. $R^*_{X|Y}(D_k) = I(X; U_k|Y) \quad \text{and} \quad \mathbb{E}d(X, f_k(U_k, Y)) \leq D_k, \quad 1 \leq k \leq K$
2. $(U_1, \ldots, U_K) \multimap X \multimap Y$
3. $(U_1, U_2, \ldots, U_{k-1}) \multimap (U_k, Y) \multimap X, \quad k = 2, 3, \ldots, K.$

It is worth pointing out that, in the examples given in Secs. 4.1 and 4.2 (Gaussian source with quadratic distortion measure, and doubly symmetric binary source with Hamming distortion measure, resp.), the construction of auxiliary random variables U and W can be extended to any number of stages, thus these sources are K-step successively refinable.

Proof of Theorem 5. We start with necessity. Assume that (101) holds. From Theorem 4, there exists a K-tuple $(U_1, U_2, \ldots, U_K)$ and K deterministic maps $f_k : \mathcal{U}_k \times \mathcal{Y} \to \hat{\mathcal{X}}$, $1 \leq k \leq K$, such that

$$R^*_{X|Y}(D_1) \geq I(X; U_1|Y), \tag{102a}$$

$$\mathbb{E}d(X, f_1(U_1, Y)) \leq D_1, \tag{102b}$$

similarly,

$$\begin{aligned} R^*_{X|Y}(D_2) &\geq R^*_{X|Y}(D_1) + I(X;U_2|U_1,Y) \geq I(X;U_1U_2|Y), && \text{(103a)} \\ \mathbb{E}d(X, f_2(U_2,Y)) &\leq D_2, && \text{(103b)} \end{aligned}$$

and, in general

$$\begin{aligned} R^*_{X|Y}(D_k) &\geq I(X;U_1\dots U_k|Y), && \text{(104a)} \\ \mathbb{E}d(X, f_k(U_k,Y)) &\leq D_k, \quad 1 \leq k \leq K. && \text{(104b)} \end{aligned}$$

By the definition of $R^*_{X|Y}(D)$, (102a) must be satisfied with equality. For (104) we have

$$R^*_{X|Y}(D_k) \geq I(X;U_1\dots U_k|Y) \geq I(X;U_k|Y) \geq R^*_{X|Y}(D_k), \tag{105}$$

where the last inequality is due to the fact that knowledge of U_k and Y suffices to satisfy the distortion constraint (see (104)). Hence equality in (105) must hold, and Conditions 1 and 2 of the theorem hold. For equality to hold in (105), we must have

$$H(X|U_1U_2\dots U_{k-1}U_k,Y) = H(X|U_k,Y), \quad k = 2,3,\dots,K, \tag{106}$$

implying Condition 3 of the theorem. This concludes the proof of necessity.

To prove sufficiency, assume that there exist a K-tuple of random variables $(U_1,\dots,U_K)$, and K independent maps $f_1,\dots f_K$ such that Conditions 1 to 3 of the theorem are met. By the definition of $\mathcal{R}^*_{K,i}(\boldsymbol{D})$, it remains to show that the differential rates satisfy the bounds

$$R^*_{X|Y}(D_k) - R^*_{X|Y}(D_{k-1}) \geq I(X;U_k|U_1U_2\dots U_{k-1},Y), \quad k = 2,3,\dots K. \tag{107}$$

Indeed, we can write

$$\begin{aligned} R^*_{X|Y}(D_k) - R^*_{X|Y}(D_{k-1}) &= I(X;U_k|Y) - I(X;U_{k-1}|Y) \\ &= H(X|U_{k-1},Y) - H(X|U_k,Y) \\ &\overset{(a)}{=} H(X|U_{k-1},Y) - H(X|U_1U_2\dots U_k,Y) \\ &\overset{(b)}{=} H(X|U_1U_2\dots U_{k-1},Y) - H(X|U_1U_2\dots U_k,Y) \\ &= I(X;U_k|U_1U_2\dots U_{k-1},Y), \end{aligned} \tag{108}$$

where (a) and (b) are due to Condition 3 of the theorem. This concludes to proof of sufficiency. □

6 Proofs of Theorems 1 and 4

6.1 Proof of Theorem 1

Converse part: Assume that we have an (n, M_1, M_2, D_1, D_2) SR code for the source X with SI (Y, Z), as in Definition 1. We will show the existence of a triple (U, V, W) that satisfies conditions 1–4 in the definition of $\mathcal{R}^*(\boldsymbol{D})$. Denote $T_i = \phi_i(X^n),\ i = 1, 2$. Then

$$\begin{aligned} nR_1 &\geq H(T_1) \geq I(X^n; T_1|Z^n) = I(X^n; T_1Y^n|Z^n) - I(X^n; Y^n|T_1Z^n) \\ &= \sum_{i=1}^{n} \Big[I(X_i; T_1Y^n|X^{i-1}Z^n) - I(X^n; Y_i|T_1Z^nY^{i-1}) \Big]. \end{aligned} \tag{109}$$

For notational convenience, we denote $Z^{i-1}Z_{i+1}^n = Z^{n\setminus i}$, and use a similar notation for X and Y. Since X_iZ_i and $X^{i-1}Z^{n\setminus i}$ are independent, we have, for the first term in the summand of (109):

$$\begin{aligned} I(X_i; T_1Y^n|X^{i-1}Z^n) &= H(X_i|Z_iX^{i-1}Z^{n\setminus i}) - H(X_i|Z_iX^{i-1}Z^{n\setminus i}T_1Y^n) \\ &= H(X_i|Z_i) - H(X_i|Z_iX^{i-1}Z^{n\setminus i}T_1Y^n) \\ &= I(X_i; X^{i-1}Z^{n\setminus i}T_1Y^n|Z_i). \end{aligned} \tag{110}$$

Next, due to the Markov structure

$$Y_i \multimap (X_iZ_i) \multimap (X^{n\setminus i}T_1Y^{i-1}Z^{n\setminus i}) \tag{111}$$

we have, for the second term in the summand of (109):

$$\begin{aligned} I(X^n; Y_i|T_1Z^nY^{i-1}) &= H(Y_i|T_1Z^nY^{i-1}) - H(Y_i|X^nT_1Z^nY^{i-1}) \\ &= H(Y_i|T_1Z^nY^{i-1}) - H(Y_i|X_iT_1Z^nY^{i-1}) \\ &= I(X_i; Y_i|T_1Z^nY^{i-1}). \end{aligned} \tag{112}$$

Substituting (110) and (112) in (109), we obtain

$$\begin{aligned} nR_1 &\geq \sum_{i=1}^{n} \Big[I(X_i; X^{i-1}Z^{n\setminus i}T_1Y^n|Z_i) - I(X_i; Y_i|T_1Z^nY^{i-1}) \Big] \\ &= \sum_{i=1}^{n} \Big[I(X_i; Z^{n\setminus i}T_1Y^{i-1}|Z_i) + I(X_i; X^{i-1}Y_i^n|Z_iT_1Z^{n\setminus i}Y^{i-1}) - I(X_i; Y_i|T_1Z^nY^{i-1}) \Big] \\ &= \sum_{i=1}^{n} \Big[I(X_i; T_1Z^{n\setminus i}Y^{i-1}|Z_i) + I(X_i; X^{i-1}Y_{i+1}^n|Z_iY_iT_1Z^{n\setminus i}Y^{i-1}) \Big]. \end{aligned} \tag{113}$$

The Markovity of $X \multimap Y \multimap Z$ implies

$$Z_i \multimap Y_i \multimap (X_i T_1 Z^{n\setminus i} Y^{i-1}), \tag{114}$$

and we have for the second term in (113)

$$\begin{aligned}
& I(X_i; X^{i-1} Y_{i+1}^n | T_1 Z^n Y^i) \\
&= H(X_i | T_1 Z^n Y^i) - H(X_i | T_1 Z^n Y^n X^{i-1}) \\
&= H(X_i Z_i | T_1 Z^{n\setminus i} Y^i) - H(Z_i | T_1 Z^{n\setminus i} Y^i) - H(X_i | T_1 Z^n Y^n X^{i-1}) \\
&= H(Z_i | X_i T_1 Z^{n\setminus i} Y^i) + H(X_i | T_1 Z^{n\setminus i} Y^i) - H(Z_i | T_1 Z^{n\setminus i} Y^i) - H(X_i | T_1 Z^n Y^n X^{i-1}) \\
&\overset{(a)}{=} H(X_i | T_1 Z^{n\setminus i} Y^i) - H(X_i | T_1 Z^n Y^n X^{i-1}) = I(X_i; Z_i Y_{i+1}^n X^{i-1} | T_1 Z^{n\setminus i} Y^i) \\
&\geq I(X_i; Y_{i+1}^n X^{i-1} | T_1 Z^{n\setminus i} Y^i)
\end{aligned} \tag{115}$$

where (114) was used in (a). Substituting (115) in (113), we get

$$nR_1 \geq \sum_{i=1}^n \left[I(X_i; T_1 Z^{n\setminus i} Y^{i-1} | Z_i) + I(X_i; Y_{i+1}^n X^{i-1} | T_1 Z^{n\setminus i} Y^i) \right]. \tag{116}$$

Before defining the auxiliary random variables, we bound the refinement rate from below as follows:

$$\begin{aligned}
n(R_2 - R_1) &\geq H(T_2|T_1) \geq H(T_2|T_1 Z^n Y^n) \geq I(X^n; T_2 | T_1 Z^n Y^n) \\
&= \sum_{i=1}^n I(X_i; T_2 | X^{i-1} T_1 Z^n Y^n).
\end{aligned} \tag{117}$$

Define the random variables

$$U_i = T_1 Z^{n\setminus i} Y^{i-1} \tag{118}$$

$$V_i = X^{i-1} Y_{i+1}^n U_i \tag{119}$$

$$W_i = T_2 V_i. \tag{120}$$

With these definitions, we have the Markov structure

$$U_i \multimap V_i \multimap W_i \multimap X_i \multimap Y_i \multimap Z_i \tag{121}$$

and the bounds (116) and (117) become

$$R_1 \geq \frac{1}{n} \sum_{i=1}^n [I(X_i; U_i | Z_i) + I(X_i; V_i | U_i, Y_i)] \tag{122}$$

$$R_2 - R_1 \geq \frac{1}{n} \sum_{i=1}^n I(X_i; W_i | U_i, V_i, Y_i) \tag{123}$$

where we have used (121) to drop the conditioning on Z_i in (123).

Let J be a random variable, independent of X, Y, and Z, and uniformly distributed over the set $\{1, 2, \dots, n\}$. Define the random variables $U = (J, U_J)$, $V = (J, V_J)$, and $W = (J, W_J)$. The Markov relations (121) still hold, that is

$$U \multimap V \multimap W \multimap X \multimap Y \multimap Z, \tag{124}$$

and therefore the condition 1 in the definition of $\mathcal{R}^*(\boldsymbol{D})$ is satisfied. We proceed to show the existence of functions f_1, f_2 satisfying (14), (15). Denote by $\psi_{i,k}$ the output of the i-th decoder at time k, $i = 1, 2$, $1 \le k \le n$. The random variable U contains $\phi_1(X^n) Z^{n \setminus J}$. Similarly, W contains $\phi_1(X^n)\phi_2(X^n) Z^{n\setminus J} Y^{n\setminus J}$. Therefore, let us choose the functions f_1 and f_2 as follows

$$f_1(U, Z) = \psi_{1,J}(\phi_1(X^n), Z^n) \tag{125}$$

$$f_2(W, Y) = \psi_{2,J}(\phi_1(X^n), \phi_2(X^n), Y^n). \tag{126}$$

Then, for the distortions we have

$$\mathbb{E}d(X, f_1(U, Z)) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}d(X, \psi_{1,j}(\phi_1(X^n), Z^n)) \le D_1 \tag{127}$$

$$\mathbb{E}d(X, f_2(W, Y)) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}d(X, \psi_{2,j}(\phi_1(X^n), \phi_2(X^n), Y^n)) \le D_2 \tag{128}$$

Hence, condition 2 in the definition of $\mathcal{R}^*(\boldsymbol{D})$ is satisfied. To prove that condition 4 of that definition holds, we have to show that the bounds (122) and (123) can be written in a single letter form with U, V, and W. The following chain of equalities holds

$$\begin{aligned}
I(X;U|Z) &= H(U|Z) - H(U|XZ) = H(U|Z) - H(U|X) \\
&= H(U) - H(U|X) - (H(U) - H(U|Z)) \\
&= I(U;X) - I(U;Z) \\
&= H(X) - H(X|U) - H(Z) + H(Z|U) \\
&= H(X) - H(X|J, U_J) - H(Z) + H(Z|J, U_J) \\
&= \frac{1}{n}\sum_{i=1}^{n} H(X_i) - \frac{1}{n}\sum_{i=1}^{n} H(X_i|U_i) - \frac{1}{n}\sum_{i=1}^{n} H(Z_i) + \frac{1}{n}\sum_{i=1}^{n} H(Z_i|U_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i; U_i|Z_i)
\end{aligned} \tag{129}$$

where the last equality is due to (121). In a similar manner, we get

$$\begin{aligned}
I(X;V|U,Y) &= I(X;JV_J|JU_J,Y) = I(X;V_J|JU_J,Y) \\
&= H(X|JU_J,Y) - H(X|JU_JV_J,Y) \\
&= \frac{1}{n}\sum_{i=1}^{n} H(X_i|i,U_i,Y_i) - \frac{1}{n}\sum_{i=1}^{n} H(X_i|i,U_i,V_i,Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i;V_i|U_i,Y_i). \qquad (130)
\end{aligned}$$

In view of (129), (130), the bound (122) can be written as

$$R_1 \geq I(X;U|Z) + I(X;V|U,Y). \qquad (131)$$

In a similar manner, it is shown that (123) can be written as

$$R_2 - R_1 \geq I(X;W|U,V,Y). \qquad (132)$$

To complete the proof of the converse part, it remains to prove that condition 3 in the definition of $\mathcal{R}^*(\boldsymbol{D})$ holds, namely, to show that the alphabets of the random variables U, V, and W, can be limited, without loss of generality, as in eqs. (16), (17), and (18), respectively. To this end, we invoke the support lemma (cf. [2, p. 310]), which results from the Carathéodory theorem. According to this lemma, given k real–valued, continuous functionals q_j, $j = 1,\ldots,k$, on the set $\mathcal{P}(\mathcal{X})$ of probability distributions over an alphabet $\mathcal{X}$, and given any probability measure μ on the Borel σ–algebra of $\mathcal{P}(\mathcal{X})$, there exist k elements $Q_1,\ldots,Q_k$ of $\mathcal{P}(\mathcal{X})$ and k non-negative reals, $\alpha_1,\ldots,\alpha_k$, that sum to unity, such that for every $j = 1,\ldots,k$:

$$\int_{\mathcal{P}(\mathcal{X})} q_j(Q)\mu(dQ) = \sum_{i=1}^{k} \alpha_i q_j(Q_i). \qquad (133)$$

Before we actually apply the support lemma, we first rewrite the relevant conditional mutual informations and the distortion functions (that appear in the definition of $\mathcal{R}^*(\boldsymbol{D})$) in a more convenient form for the use of this lemma, by taking advantage of the Markov structure. As for the first term, $I(X;U|Z)$, in the lower bound to R_1, we have:

$$\begin{aligned}
I(X;U|Z) &= H(U|Z) - H(U|X,Z) \\
&= H(U|Z) - H(U|X) \\
&= H(U) + H(Z|U) - H(Z) - H(U) - H(X|U) + H(X) \\
&= H(X) - H(Z) + H(Z|U) - H(X|U). \qquad (134)
\end{aligned}$$

For the second term in the lower bound to R_1, we have

$$
\begin{aligned}
I(X;V|U,Y) &= H(X|U,Y) - H(X|U,V,Y) \\
&= \mathbb{E}\log P(X|U,V,Y) - \mathbb{E}\log P(X|U,Y) \\
&= \mathbb{E}\log\left[\frac{P(U,V,X,Y)}{P(U,V,Y)}\right] - \mathbb{E}\log\left[\frac{P(U,X,Y)}{P(U,Y)}\right] \\
&= \mathbb{E}\log\left[\frac{P(U,V)P(X|U,V)P(Y|X)}{P(U,V)P(Y|U,V)}\right] - \mathbb{E}\log\left[\frac{P(U)P(X|U)P(Y|X)}{P(U)P(Y|U)}\right] \\
&= H(X|U) - H(Y|U) + H(Y|U,V) - H(X|U,V). \qquad (135)
\end{aligned}
$$

Thus, the sum of these two terms is given by:

$$
\begin{aligned}
I(X;U|Z) + I(X;V|U,Y) &= [H(X) - H(Z)] + [H(Z|U) - H(Y|U)] + \\
&\quad [H(Y|U,V) - H(X|U,V)] \\
&\stackrel{\triangle}{=} \alpha + \beta + \gamma. \qquad (136)
\end{aligned}
$$

In a similar manner, for the lower bound to $R_2 - R_1$, we obtain

$$I(X;W|U,V,Y) = [H(Y|U,V,W) - H(X|U,V,W)] - \gamma \stackrel{\triangle}{=} \delta - \gamma. \qquad (137)$$

In order to preserve prescribed values of $I(X;U|Z) + I(X;V|U,Y)$ and $I(X;W|U,V,Y)$, it then suffices to preserve the associated values of $(\beta+\gamma)$ and $(\delta-\gamma)$, since α is a constant that depends only on the given statistics of the source and the SI.

We first invoke the support lemma in order to reduce the alphabet size of U, while preserving the values of $\beta+\gamma$ and $\delta-\gamma$ as well as the first–stage distortion. The alphabets of V and W are still kept intact at this step. Define the following functionals of a generic distribution Q over $\mathcal{X}\times\mathcal{V}\times\mathcal{W}$, where $\mathcal{X}$ is assumed, without loss of generality, to be $\{1,2,\ldots,m\}$, $m \stackrel{\triangle}{=} |\mathcal{X}|$.

$$q_x(Q) = \sum_{v,w} Q(x,v,w), \qquad x = 1,\ldots,m-1 \qquad (138)$$

$$q_m(Q) = A(Q) + B(Q), \qquad (139)$$

where

$$A(Q) = \sum_{x,v,w} Q(x,v,w)\sum_{y,z} P(y,z|x)\log\frac{\sum_{x',v',w'} Q(x',v',w')P(y|x')}{\sum_{x',v',w'} Q(x',v',w')P(z|x')}, \qquad (140)$$

and

$$B(Q)=\sum_{x,v,w}Q(x,v,w)\sum_{y}P(y|x)\log\frac{Q(x|v)}{P_Q(y|v)} \tag{141}$$

with

$$Q(x|v)\stackrel{\triangle}{=}\frac{\sum_{w'}Q(x,v,w')}{\sum_{x',w'}Q(x',v,w')} \tag{142}$$

and

$$P_Q(y|v)\stackrel{\triangle}{=}\frac{\sum_{x',w'}P(y|x')Q(x',v,w')}{\sum_{x',w',y'}P(y'|x')Q(x',v,w')}. \tag{143}$$

Next, define

$$q_{m+1}(Q)=\sum_{x,v,w}Q(x,v,w)\sum_{y}P(y|x)\log\frac{Q(x|v,w)}{P_Q(y|v,w)}-B(Q) \tag{144}$$

with similar definitions (but without the summations over w') for the numerator and the denominator of the argument of the logarithm. Finally, for the distortion constraint, let

$$q_{m+2}(Q)=\sum_{z}\min_{\hat{x}}\sum_{x,v,w}Q(x,v,w)P(z|x)d(x,\hat{x}). \tag{145}$$

Applying now the support lemma, we find that there exists a random variable U (jointly distributed with X) whose alphabet size is $k=m+2=|\mathcal{X}|+2$ that satisfies simultaneously

$$\begin{aligned}
\sum_{u}P(U=u)q_x(P(\cdot|u)) &= P(x), \quad x=1,\ldots,m-1 && (146)\\
\sum_{u}P(U=u)q_m(P(\cdot|u)) &= \beta+\gamma && (147)\\
\sum_{u}P(U=u)q_{m+1}(P(\cdot|u)) &= \delta-\gamma && (148)\\
\sum_{u}P(U=u)q_{m+2}(P(\cdot|u)) &= \min_{f}\mathbb{E}d(X,f_1(U,Z)). && (149)
\end{aligned}$$

Having found such a random variable U, we now proceed to reduce the alphabet of V in a similar manner, where this time, we have $|\mathcal{X}|\cdot|\mathcal{U}|-1$ constraints to preserve the joint distribution of (X,U) just defined, and two more constraints for preserving γ and δ (note that V is not involved in β and in any distortion constraint). Thus, the necessary alphabet size of V is upper bounded by

$$|\mathcal{V}|\le|\mathcal{X}|\cdot|\mathcal{U}|+1\le|\mathcal{X}|\cdot(|\mathcal{X}|+2)+1=(|\mathcal{X}|+1)^2. \tag{150}$$

Finally, W must preserve the joint distribution of (X, U, V), plus the value of δ and the second expected distortion $\min_f \mathbb{E}d(X, f(W, Y))$, which means that the needed alphabet size of W does not have to exceed

$$|\mathcal{X}| \cdot |\mathcal{U}| \cdot |\mathcal{V}| + 1 \leq |\mathcal{X}| \cdot (|\mathcal{X}| + 2) \cdot (|\mathcal{X}| + 1)^2 + 1. \tag{151}$$

This completes the proof of the converse part.

Direct Part:

We begin by setting up some notation and mentioning a few basic facts that will be needed hereafter. Given a distribution P_{XU}, we denote by $\mathcal{T}_X^\delta$ the set of all n-tuples $\boldsymbol{x}$ that are δ-typical according to P_X, i.e.,

$$\begin{aligned} \mathcal{T}_X^\delta \quad = \quad & \left\{ \boldsymbol{x} \in \mathcal{X}^n : \quad \left| \frac{1}{n} N(x|\boldsymbol{x}) - P_X(x) \right| \leq \delta \quad \forall\, x \in \mathcal{X}, \right. \\ & \text{and } N(x|\boldsymbol{x}) = 0 \ \text{ whenever } \ P_X(x) = 0 \} \end{aligned}$$

where $N(x|\boldsymbol{x})$ is the number of occurrences of the letter x in the n-tuple $\boldsymbol{x}$. In the sequel, we will use the following well known results [2]. For any $\boldsymbol{x} \in \mathcal{T}_X^\delta$ and any $\delta' > \delta$

$$\exp[-n(I(X;U) + \epsilon_u)] \leq \sum_{\boldsymbol{u}:\ (\boldsymbol{x},\boldsymbol{u}) \in \mathcal{T}_{XU}^{\delta'}} P_U(\boldsymbol{u}) \leq \exp[-n(I(X;U) - \epsilon)] \tag{152}$$

where $\epsilon = \epsilon(\delta, \delta')$ and $\epsilon_u = \epsilon_u(\delta, \delta')$ both vanish as δ, $\delta' \to 0$. Similarly, for any pair $(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{T}_{XU}^{\delta'}$ and any $\delta'' > \delta'$

$$\exp[-n(I(X;V|U) + \eta_{v|u})] \leq \sum_{\boldsymbol{v}:\ (\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) \in \mathcal{T}_{XUV}^{\delta''}} P_{V|U}(\boldsymbol{v}|\boldsymbol{u}) \leq \exp[-n(I(X;V|U) - \eta)] \tag{153}$$

where $\eta = \eta(\delta', \delta'')$ and $\eta_{v|u} = \eta_{v|u}(\delta', \delta'')$ both vanish as δ', $\delta'' \to 0$.

Let a distortion pair $\boldsymbol{D} = (D_1, D_2)$ be given, and let (U, V, W) satisfy the conditions that define $\mathcal{R}^*(\boldsymbol{D})$. Recall that this guarantees the existence of functions f_1, f_2, satisfying (14), (15). Fix an arbitrary $\gamma > 0$ and consider the construction of a hierarchy of codebooks described in the following steps.

Codebook Generation

1. Randomly generate $M_1 = \exp[n(I(X;U) + \gamma)]$ independent codewords $\{\boldsymbol{u}_i\}$, each of length n, according to $P_U(\cdot)$, where the coordinates are also generated i.i.d. Denote

$$\mathcal{A} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{M_1}\}, \quad \boldsymbol{u}_j \in \mathcal{U}^n, \quad 1 \le j \le M_1. \tag{154}$$

2. Let $L_1 = \exp[n(I(Z;U) - 4\gamma)]$, and

$$\begin{aligned} N_1 &= \exp\left[n(I(X;U) - I(Z;U) + 5\gamma)\right] \\ &\stackrel{(a)}{=} \exp\left[n(I(X;U|Z) + 5\gamma)\right] \\ &= M_1 \exp\left[-n(I(Z;U) - 4\gamma)\right] = \frac{M_1}{L_1}, \end{aligned} \tag{155}$$

where (a) above follows from the Markov structure $(U,V,W) \multimap X \multimap Y \multimap Z$. Partition the codebook $\mathcal{A}$ into N_1 bins, each containing L_1 members of $\mathcal{A}$. Let $\mathcal{A}_B(i)$ denote the elements $\boldsymbol{u} \in \mathcal{A}$ assigned to bin i, $1 \le i \le N_1$.

3. Set $M_2 = \exp\left[n(I(X;V|U) + \gamma)\right]$. For any $\boldsymbol{u} \in \mathcal{A}$, generate a codebook, of length n and size M_2, according to $P_{V|U}(\boldsymbol{v}|\boldsymbol{u})$, where

$$P_{V|U}(\boldsymbol{v}|\boldsymbol{u}) = \prod_{l=1}^{n} P_{V|U}(v_l|u_l). \tag{156}$$

Denote this codebook by $\mathcal{B}(\boldsymbol{u})$.

4. Let $L_2 = \exp\left[n(I(Y;V|U) - 4\gamma)\right]$, and

$$\begin{aligned} N_2 &= \exp\left[n(I(X;V|U) - I(Y;V|U) + 5\gamma)\right] \\ &= \exp\left[n(I(X;V|U,Y) + 5\gamma)\right] \\ &= M_2 \exp\left[-n(I(Y;V|U) - 4\gamma)\right] = \frac{M_2}{L_2}. \end{aligned} \tag{157}$$

For each element $\boldsymbol{u} \in \mathcal{A}$, partition the codebook $\mathcal{B}(\boldsymbol{u})$ into N_2 bins, each containing L_2 members of $\mathcal{B}(\boldsymbol{u})$. Let $\mathcal{B}_B(j, \boldsymbol{u})$ denote the set of elements $\boldsymbol{v} \in \mathcal{B}(\boldsymbol{u})$ assigned to bin j, $1 \le j \le N_2$.

5. Set $M_3 = \exp\left[n(I(X;W|U,V) + \gamma)\right]$. For any pair $(\boldsymbol{u}, \boldsymbol{v})$ such that $\boldsymbol{u} \in \mathcal{A}$ and $\boldsymbol{v} \in \mathcal{B}(\boldsymbol{u})$, generate a codebook of length n and size M_3, according to $P_{W|UV}$

$$P_{W|UV}(\boldsymbol{w}|\boldsymbol{u}, \boldsymbol{v}) = \prod_{l=1}^{n} P_{W|UV}(w_l|u_l, v_l). \tag{158}$$

Denote this codebook by $\mathcal{C}(\boldsymbol{u},\boldsymbol{v})$.

6. Similarly to the partitions of $\mathcal{A}$ and $\mathcal{B}(\boldsymbol{u})$ above, we now partition $\mathcal{C}(\boldsymbol{u},\boldsymbol{v})$. Thus, define $L_3 = \exp[n(I(Y;W|V,U)-4\gamma)]$, and

$$\begin{aligned} N_3 &= \exp[n(I(X;W|U,V)-I(Y;W|U,V)+5\gamma)] \\ &= \exp[n(I(X;W|U,V,Y)+5\gamma)] \\ &= M_3\exp[-n(I(Y;W|U,V)-4\gamma)] = \frac{M_3}{L_3}. \end{aligned} \tag{159}$$

   For each pair $(\boldsymbol{u},\boldsymbol{v})$ such that $\boldsymbol{u}\in\mathcal{A}$ and $\boldsymbol{v}\in\mathcal{B}(\boldsymbol{u})$, partition the codebook $\mathcal{C}(\boldsymbol{u},\boldsymbol{v})$ into N_3 bins, each containing L_3 members. Let $\mathcal{C}_B(k,\boldsymbol{u},\boldsymbol{v})$ denote the elements $\boldsymbol{w}\in\mathcal{C}(\boldsymbol{u},\boldsymbol{v})$ assigned to bin k, $1\le k\le N_3$.

Reveal the codebooks and the partitions to the encoder and decoders.

Encoding

Given a source vector $\boldsymbol{x}$, the encoding process proceeds along the following steps.

1. First encoder: The encoding map $\phi_1(\boldsymbol{x})$ consists of two maps,

$$\phi_1(\boldsymbol{x}) = \phi_{1,1}(\boldsymbol{x})\phi_{1,2}(\boldsymbol{x}),$$

   i.e., it is a double index, defined as follows. The encoder seeks a vector $\boldsymbol{u}\in\mathcal{A}$ such that $(\boldsymbol{x},\boldsymbol{u})\in\mathcal{T}^{2\delta}_{XU}$. If more than one such vector exists in the codebook $\mathcal{A}$, the first one is chosen. If such a vector does not exist in the codebook $\mathcal{A}$, a default vector is chosen, say $\boldsymbol{u}_1$, and an error is declared. Denote this vector by $\boldsymbol{u}(\boldsymbol{x})$. The second index is set equal to the bin number to which $\boldsymbol{u}(\boldsymbol{x})$ belongs, that is

$$\phi_{1,2}(\boldsymbol{x}) = i \quad \text{if} \quad \boldsymbol{u}(\boldsymbol{x})\in\mathcal{A}_B(i). \tag{160}$$

   The mapping $\phi_{1,2}(\cdot)$ thus takes values in the set $\{1,2,\ldots,N_1\}$.

   We proceed to the first map, $\phi_{1,1}$. Given $\boldsymbol{x}$ and a codeword $\boldsymbol{u}(\boldsymbol{x})\in\mathcal{A}$, the encoder seeks a vector $\boldsymbol{v}\in\mathcal{B}(\boldsymbol{u}(\boldsymbol{x}))$ such that $(\boldsymbol{x},\boldsymbol{u}(\boldsymbol{x}),\boldsymbol{v})\in\mathcal{T}^{3\delta}_{XUV}$. If such vector does not exist, a default vector is chosen, say the first vector in $\mathcal{B}(\boldsymbol{u}(\boldsymbol{x}))$, and an error is declared. If more than one exists, the first one is chosen. Denote this vector by $\boldsymbol{v}(\boldsymbol{x})$. The map $\phi_{1,1}(\boldsymbol{x})$ is defined as the bin number to which $\boldsymbol{v}(\boldsymbol{x})$ belongs, that is

$$\phi_{1,1}(\boldsymbol{x}) = j \quad \text{if} \quad \boldsymbol{v}(\boldsymbol{x})\in\mathcal{B}_B(j,\boldsymbol{u}(\boldsymbol{x})). \tag{161}$$

Observe that the encoding map ϕ_1 is taking values in a discrete set of size $N_1 \cdot N_2$, where

$$N_1 \cdot N_2 = \exp\left[n(I(X;U|Z) + I(X;V|U,Y) + 10\gamma)\right].$$

2. Second encoder (refinement encoder): The second encoder knows $\boldsymbol{x}$, $\boldsymbol{u}(\boldsymbol{x})$, and $\boldsymbol{v}(\boldsymbol{x})$. It seeks a codeword $\boldsymbol{w} \in \mathcal{C}(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}))$ such that $(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{w}) \in \mathcal{T}^{4\delta}_{XUVW}$. If such vector does not exist, a default $\boldsymbol{w}$ is chosen, say, the first vector in $\mathcal{C}(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}))$, and an error is declared. If more than one exists, the first in the list is chosen. Denote this vector by $\boldsymbol{w}(\boldsymbol{x})$. The value assigned to $\phi_2(\boldsymbol{x})$ is the bin number to which $\boldsymbol{w}(\boldsymbol{x})$ belongs, i.e.,

$$\phi_2(\boldsymbol{x}) = k \quad \text{if} \quad \boldsymbol{w}(\boldsymbol{x}) \in \mathcal{C}_B(k, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})). \tag{162}$$

Observe that the mapping ϕ_2 is taking values in $\{1, 2, \ldots, N_3\}$.

The first encoder sends the indices of the two bins, $\phi_{1,1}(\boldsymbol{x})\phi_{1,2}(\boldsymbol{x})$, using

$$\log(N_1 \cdot N_2) = n[I(X;U|Z) + I(X;V|U,Y) + 10\gamma] \ \text{ bits.} \tag{163}$$

The refinement encoder sends the index of the third bin using

$$\log N_3 = n[I(X;W|U,V,Y) + 5\gamma] \ \text{ bits.} \tag{164}$$

Decoding

1. The decoder of the first stage has the vector $\boldsymbol{z}$ as SI. It has access to the double index $\phi_1(\boldsymbol{x}) = (i,j)$, $1 \le i \le N_1$, $1 \le j \le N_2$, but in the reconstruction process it utilizes only the first index i. It decodes the vector $\boldsymbol{u}$ precisely as in the classical Wyner–Ziv decoding procedure [1], [16]. Specifically, the first decoder seeks a unique vector $\boldsymbol{u} \in \mathcal{A}_B(i)$ such that $(\boldsymbol{u}, \boldsymbol{z}) \in \mathcal{T}^{3\delta|\mathcal{X}|}_{UZ}$. Denote this vector by $\hat{\boldsymbol{u}}(\boldsymbol{z})$. If there is no vector $\boldsymbol{u} \in \mathcal{A}_B(i)$ jointly typical with $\boldsymbol{z}$, or there is more than one, an arbitrary $\hat{\boldsymbol{u}}$ is chosen, and an error is declared. The reconstruction vector of the first stage, $\hat{\boldsymbol{x}}_1$, is given by

$$\hat{\boldsymbol{x}}_1 = (\hat{x}_{1,1}, \hat{x}_{1,2}, \ldots, \hat{x}_{1,n}) \tag{165}$$

where

$$\hat{x}_{1,l} = f_1(\hat{u}_l(\boldsymbol{z}), z_l). \tag{166}$$

2. The decoder of the second stage has the vector $\boldsymbol{y}$ as SI. It receives the double-index (i,j) as the first decoder, but utilizes both indices, i and j. In addition, it gets the output of the second encoder, $\phi_2(\boldsymbol{x}) = k$, $1 \leq k \leq N_3$, and utilizes it too. Thus, it can be described in four stages.

   (a) Decoding $\boldsymbol{u}$: Due to the Markov structure $(U,V,W) \multimap X \multimap Y \multimap Z$, the second decoder can do everything the first decoder can. Thus, it looks for a unique vector $\boldsymbol{u} \in \mathcal{A}_B(i)$ such that

   $$(\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{T}_{UY}^{3\delta|\mathcal{X}|}. \tag{167}$$

   If such vector $\boldsymbol{u}$ does not exist, or if it is not unique, an arbitrary element in $\mathcal{A}_B(i)$ is chosen. Denote the output of this stage by $\hat{\boldsymbol{u}}(\boldsymbol{y})$.

   (b) Decoding $\boldsymbol{v}$: In this stage, the second decoder looks for a unique vector $\boldsymbol{v} \in \mathcal{B}_B(j, \hat{\boldsymbol{u}}(\boldsymbol{y}))$ such that

   $$(\hat{\boldsymbol{u}}(\boldsymbol{y}), \boldsymbol{v}, \boldsymbol{y}) \in \mathcal{T}_{UVY}^{4\delta|\mathcal{X}|}. \tag{168}$$

   Denote this vector by $\hat{\boldsymbol{v}}(\boldsymbol{y})$. As usual, if such a vector does not exist, or there is more than one such vector in $\mathcal{B}_B(j, \hat{\boldsymbol{u}}(\boldsymbol{y}))$, an arbitrary $\hat{\boldsymbol{v}}(\boldsymbol{y})$ is chosen.

   (c) Decoding $\boldsymbol{w}$: At the third stage, the second decoder looks for the unique vector

   $$\boldsymbol{w} \in \mathcal{C}_B(k, \hat{\boldsymbol{u}}(\boldsymbol{y}), \hat{\boldsymbol{v}}(\boldsymbol{y}))$$

   such that

   $$(\hat{\boldsymbol{u}}(\boldsymbol{y}), \hat{\boldsymbol{v}}(\boldsymbol{y}), \boldsymbol{w}, \boldsymbol{y}) \in \mathcal{T}_{UVWY}^{5\delta|\mathcal{X}|}. \tag{169}$$

   Denote this vector by $\hat{\boldsymbol{w}}(\boldsymbol{y})$. If such a vector does not exist, or there is more than one, an arbitrary representative is chosen.

   (d) At the last stage, the second reproduction vector $\hat{\boldsymbol{x}}_2$ is constructed, as

   $$\hat{\boldsymbol{x}}_2 = (\hat{x}_{2,1}, \hat{x}_{2,2}, \ldots, \hat{x}_{2,n}) \tag{170}$$

   where

   $$\hat{x}_{2,l} = f_2(\hat{w}_l(\boldsymbol{y}), y_l). \tag{171}$$

<u>*Analysis of the Probability of Error*</u>

The probability of error in the encoding/decoding scheme is examined next. We start by defining the error events. The error events in the encoding scheme are the following:

$$
\begin{aligned}
E_0 &= \left\{ \boldsymbol{x} \in (\mathcal{T}_X^{\delta})^c \right\} \\
E_1 &= E_0^c \cap \left\{ \bigcap_{i=1}^{M_1} \left\{ (\boldsymbol{x}, \boldsymbol{u}_i) \notin \mathcal{T}_{XU}^{2\delta} \right\} \right\} \\
E_2 &= E_0^c \cap E_1^c \cap \left\{ \bigcap_{j=1}^{M_2} \left\{ (\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}_j) \notin \mathcal{T}_{XUV}^{3\delta} \right\} \right\} \\
E_3 &= \left(\bigcap_{m=0}^{2} E_m^c \right) \cap \left\{ \bigcap_{k=1}^{M_3} \left\{ (\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{w}_k) \notin \mathcal{T}_{XUVW}^{4\delta} \right\} \right\}
\end{aligned}
$$

where E_1 and E_2 correspond to the first and second stages of the first encoder, and E_3 corresponds to the second encoder. The error events of decoder 2 are

$$
\begin{aligned}
E_4 &= \left(\bigcap_{m=0}^{3} E_m^c \right) \cap \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{T}_{UXY}^{3\delta} \right\} \\
E_5 &= \left(\bigcap_{m=0}^{4} E_m^c \right) \cap \left\{ \bigcup_{\boldsymbol{u} \in \mathcal{A}_B(i),\ \boldsymbol{u} \neq \boldsymbol{u}(\boldsymbol{x})} \left\{ (\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{T}_{UY}^{3\delta|\mathcal{X}|} \right\} \right\} \\
E_6 &= \left(\bigcap_{m=0}^{5} E_m^c \right) \cap \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{T}_{UVXY}^{4\delta} \right\} \\
E_7 &= \left(\bigcap_{m=0}^{6} E_m^c \right) \cap \left\{ \bigcup_{\boldsymbol{v} \in \mathcal{B}_B(j, \boldsymbol{u}(\boldsymbol{x}))} \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}, \boldsymbol{y}) \in \mathcal{T}_{UVY}^{4\delta|\mathcal{X}|} \right\} \right\} \\
E_8 &= \left(\bigcap_{m=0}^{7} E_m^c \right) \cap \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{w}(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{T}_{UVWXY}^{5\delta} \right\} \\
E_9 &= \left(\bigcap_{m=0}^{8} E_m^c \right) \cap \left\{ \bigcup_{\boldsymbol{w} \in \mathcal{C}_B(k, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}))} \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}), \boldsymbol{w}, \boldsymbol{y}) \in \mathcal{T}_{UVWY}^{5\delta|\mathcal{X}|} \right\} \right\}
\end{aligned}
$$

Finally, the error events for the first decoder are defined as

$$
E_{10} = \left(\bigcap_{m=0}^{3} E_m^c \right) \cap \left\{ (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{z}) \notin \mathcal{T}_{UXZ}^{3\delta} \right\}
$$

$$E_{11} \quad = \quad \left(\bigcap_{m=0}^{3} E_m^c\right) \cap E_{10}^c \cap \left\{ \bigcup_{\boldsymbol{u}\in\mathcal{A}_B(i)} \left\{(\boldsymbol{u},\boldsymbol{z}) \in \mathcal{T}_{UZ}^{3\delta|\mathcal{X}|}\right\}\right\}$$

The probability of error in the encoding/decoding process, P_e, is upper bounded as

$$P_e \le P\left\{\bigcup_{\ell=0}^{11} E_\ell\right\} \le \sum_{\ell=0}^{11} P(E_\ell). \tag{172}$$

Observe that if no error occurs in the encoding/decoding process (i.e., the event $\cap_{m=0}^{11} E_m^c$ occurs), then the following is satisfied

$$(\hat{\boldsymbol{W}}(\boldsymbol{Y}), \boldsymbol{X}, \boldsymbol{Y}) \quad \in \quad \mathcal{T}_{WXY}^{\mu} \tag{173}$$

$$(\hat{\boldsymbol{U}}(\boldsymbol{Z}), \boldsymbol{X}, \boldsymbol{Z}) \quad \in \quad \mathcal{T}_{UXZ}^{\mu} \tag{174}$$

where $\mu = 5\delta\cdot|\mathcal{U}\times\mathcal{V}|$, which means that the empirical distributions of these random vectors is close to their corresponding joint distributions. Thus, in view of (14), (15), the distortion constraints are approximately satisfied. Specifically, for the distortion of the first stage

$$\begin{aligned} d(x^n,\hat{x}^n) &\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} d(x_i, f_1(u_i(z^n), z_i)) \\ &= \frac{1}{n}\sum_{(x,u,z)\in\mathcal{X}\times\mathcal{U}\times\mathcal{Z}} N(x,u,z|x^n,u^n,z^n) d(x_i, f_1(u_i(z^n), z_i)) \\ &\le \mathbb{E}d(X, f_1(U,Z)) + \max d(x,\hat{x})\cdot\mu\cdot|\mathcal{X}\times\mathcal{U}\times\mathcal{Z}|, \end{aligned} \tag{175}$$

where (a) follows from the definition of distortion between vectors, and from (165), and the last inequality from the fact that the vectors are jointly typical. Similar considerations hold for the second stage.

Thus, to prove the direct part of Theorem 1, it is enough to show that for fixed γ and sufficiently small δ, each of the terms in the sum in (172) vanishes as $n\to\infty$. The techniques for proving these limits are now classical, and follow those of [16] (see also [1] and [7]). We will focus on events E_0 to E_7. The proof that the probabilities of the events E_8 to E_{12} vanish, follows by similar arguments.

E_0. Clearly, $P(E_0)$ tends to zero as $n\to\infty$.

E_1. For E_1 we have

$$\begin{aligned} P(E_1) &= P\left(\bigcap_{i=1}^{M_1}\left\{(\boldsymbol{X},\boldsymbol{U}_i)\notin \mathcal{T}_{XU}^{2\delta}\right\} \mid \boldsymbol{X}\in\mathcal{T}_X^{\delta}\right) P\left(\boldsymbol{X}\in\mathcal{T}_X^{\delta}\right) \\ &= \sum_{\boldsymbol{x}\in\mathcal{T}_X^{\delta}} P_U\left(\bigcap_{i=1}^{M_1}\left\{(\boldsymbol{x},\boldsymbol{U}_i)\notin \mathcal{T}_{XU}^{2\delta}\right\}\right) P_X(\boldsymbol{x}) \end{aligned} \tag{176}$$

By the left hand side of (152), for $\boldsymbol{x}\in\mathcal{T}_X^{\delta}$ we have

$$\begin{aligned} & P_U\left(\bigcap_{i=1}^{M_1}\left\{(\boldsymbol{x},\boldsymbol{U}_i)\notin \mathcal{T}_{XU}^{2\delta}\right\}\right) \\ &\leq \{1-\exp\left[-nI(X;U)-n\epsilon_u\right]\}^{M_1} \\ &\leq \exp\left[-\exp(n\gamma-n\epsilon_u)\right] \end{aligned} \tag{177}$$

which tends to 0 doubly exponentially fast, provided $\gamma>\epsilon_u(\delta,2\delta)$. (We have also used $(1-y)^n\leq \exp(-yn)$.) Substituting (177) in (176), we have

$$\lim_{n\to\infty} P(E_1)\to 0. \tag{178}$$

E_2. Conditioned on $E_0^c\cap E_1^c$, we have $(\boldsymbol{x},\boldsymbol{u}(\boldsymbol{x}))\in\mathcal{T}_{XU}^{2\delta}$, thus similarly to (176) we obtain

$$P(E_2)=\sum_{(\boldsymbol{x},\boldsymbol{u})\in\mathcal{T}_{XU}^{2\delta}} P_{V|U}\left(\bigcap_{j=1}^{M_2}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{V}_j)\notin\mathcal{T}_{XUV}^{3\delta}\,|\boldsymbol{u}\right) P_{XU}(\boldsymbol{x},\boldsymbol{u}) \tag{179}$$

By the left hand side of (153), for $(\boldsymbol{x},\boldsymbol{u})\in\mathcal{T}_{XU}^{2\delta}$,

$$\begin{aligned} & P_{V|U}\left(\bigcap_{j=1}^{M_2}\left\{(\boldsymbol{x},\boldsymbol{u},\boldsymbol{V}_j)\notin\mathcal{T}_{XUV}^{3\delta}\right\}\mid\boldsymbol{u}\right) \\ &\leq \left\{1-\exp\left[-nI(X;V|U)-n\eta_{v|u}\right]\right\}^{M_2} \\ &\leq \exp\left[-\exp(n\gamma-n\eta_{v|u})\right] \end{aligned} \tag{180}$$

which vanishes provided $\gamma>\eta_{v|u}(2\delta,3\delta)$. Hence

$$\lim_{n\to\infty} P(E_2)\to 0. \tag{181}$$

E_3. Conditioned on $\cap_{m=0}^{2} E_m^c$, the triplet $(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})$ is jointly typical, and belongs to $\mathcal{T}_{XUV}^{3\delta}$. Thus the steps to show that $P(E_3)$ vanishes are similar to those leading from (179) to (181), except that the conditioning is on (U, V), and the small parameter is 4δ instead of 3δ. Thus we have

$$\lim_{n\to\infty} P(E_3) \to 0. \tag{182}$$

provided $\gamma > \eta_{uvw}(3\delta, 4\delta)$.

E_4. Conditioned on $\cap_{m=0}^{3} E_m^c$, the pair $(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}))$ is jointly typical. Moreover, there is a Markov structure $U \multimap X \multimap Y$, where the random vector $\boldsymbol{Y}$ is drawn according to $P_{Y|X}$. Consequently, by the Markov lemma [1, Lemma 14.8.1]

$$P(E_4) \leq \sum_{(\boldsymbol{x},\boldsymbol{u})\in\mathcal{T}_{XU}^{2\delta}} P_{Y|X}\left\{(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{Y}) \notin \mathcal{T}_{XUY}^{3\delta} \,|\boldsymbol{x}\right\} P_{XU}(\boldsymbol{x}, \boldsymbol{u}) \to 0 \tag{183}$$

as $n \to 0$.

E_5. The sequences $\{\boldsymbol{U}_i\}_{i=1}^{M_1}$ are drawn independent of everything. Conditioned on the intersection $\cap_{m=0}^{4} E_m^c$, the vector $\boldsymbol{y}$ is typical. The probability that independently drawn $\boldsymbol{U}_i$ is jointly typical with $\boldsymbol{y}$ is upper bounded by (see right hand side of (152))

$$P_U\left((\boldsymbol{U}, \boldsymbol{y}) \in \mathcal{T}_{UY}^{3\delta|\mathcal{X}|}\right) \leq \exp\left[-nI(U;Y) + n\epsilon(2\delta|\mathcal{X}|, 3\delta|\mathcal{X}|)\right] \tag{184}$$

Since $|\mathcal{A}_B(i)| = \exp\left[n(I(Z;U) - 4\gamma)\right]$ and $I(Z;U) \leq I(Y;U)$, we have

$$\lim_{n\to\infty} P(E_5) = 0 \tag{185}$$

provided $4\gamma > \epsilon(2\delta|\mathcal{X}|, 3\delta|\mathcal{X}|)$.

E_6. Conditioned on $\cap_{m=0}^{5} E_m^c$, the triplet $(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x}))$ is jointly typical. We have also the Markov structure $U \multimap V \multimap X \multimap Y$, thus we apply the Markov lemma, as in (183), to conclude that $P(E_6)$ vanishes as $n \to \infty$.

E_7. The vectors $\{\boldsymbol{v}_j\}_{j=1}^{M_2}$ are drawn according to $P_{V|U}(\cdot|u_i(\boldsymbol{x}))$. Conditioned on the intersection $\cap_{m=0}^{6} E_m^c$, the pair $(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{y})$ is typical. The probability that a vector $\boldsymbol{V}_j$ drawn according to $P_{V|U}$ independently of $\boldsymbol{y}$ is jointly typical with $\boldsymbol{y}$ is upper bounded by (see right hand side of (153))

$$P_{V|U}\left((\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{V}, \boldsymbol{y}) \in \mathcal{T}_{UVY}^{4\delta|\mathcal{X}|}\right) \leq \exp\left[-n(I(Y;V|U) - \epsilon(3\delta|\mathcal{X}|, 4\delta|\mathcal{X}|))\right] \tag{186}$$

Since

$$|\mathcal{B}_B(j, \boldsymbol{u}(\boldsymbol{x}))| = \exp\left[n(I(Y;V|U) - 4\gamma)\right] \tag{187}$$

we have

$$\lim_{n\to\infty} P(E_7) = 0 \tag{188}$$

provided $4\gamma > \epsilon(3\delta|\mathcal{X}|, 4\delta|\mathcal{X}|)$.

□

6.2 Proof of Theorem 4

The direct part follows from Theorem 3, by substituting $V_{k,k} = U_k$, and choosing $V_{k,l},\ k < l \le K$, to be null random variables. For the converse, we will show the existence of a K-tuple of random variables $(U_1, U_2, \ldots, U_K)$ satisfying Conditions 1–3 in the definition of $\mathcal{R}^*_{K,i}(\boldsymbol{D})$. Let $T_i = \phi_i(X^n)$ stand for the output of the i'th encoder, $i = 1, 2, \ldots, K$. The following chains of inequalities hold

$$\begin{aligned}
nR_1 &\ge H(T_1) \ge I(X^n;T_1|Y^n) = H(X^n|Y^n) - H(X^n|T_1,Y^n) \\
&= \sum_{i=1}^{n}\Big[H(X_i|Y^nX^{i-1}) - H(X_i|T_1Y^nX^{i-1})\Big] \\
&= \sum_{i=1}^{n}\Big[H(X_i|Y_i) - H(X_i|Y_iT_1Y^{n\setminus i}X^{i-1})\Big] \\
&= \sum_{i=1}^{n} I(X_i;T_1Y^{n\setminus i}X^{i-1}|Y_i), \qquad (189)\\
n(R_2 - R_1) &\ge H(T_2|T_1) \ge I(X^n;T_2|T_1,Y^n) \\
&= \sum_{i=1}^{n} I(X_i;T_2|Y_i,T_1Y^{n\setminus i}X^{i-1}), \qquad (190)
\end{aligned}$$

and, in general,

$$\begin{aligned}
n(R_k - R_{k-1}) &\ge H(T_k|T_{k-1},\ldots,T_1) \ge I(X^n;T_k|T_{k-1},\ldots,T_1) \\
&= \sum_{i=1}^{n} I(X_i;T_k|Y_i,T_{k-1}\ldots T_1Y^{n\setminus i}X^{i-1}). \qquad (191)
\end{aligned}$$

Define the random variables

$$U_{1,i} \triangleq T_1 Y^{n\setminus i} X^{i-1} \tag{192}$$

$$U_{k,i} \triangleq T_k U_{k-1,i}, \quad k = 2,3,\ldots,K. \tag{193}$$

With these definitions, the proof proceeds along the lines leading from (122) to (132). The details, being straightforward, are omitted. □

7 Proof of Equivalence of $\mathcal{R}^*(\boldsymbol{D})$ and $\hat{\mathcal{R}}^*(\boldsymbol{D})$

In this section, we show that $\mathcal{R}^*(\boldsymbol{D})$ and $\hat{\mathcal{R}}^*(\boldsymbol{D})$ (the region defined in (33)-(36)) are equivalent. First, note that in the definition of $\mathcal{R}^*(\boldsymbol{D})$, R_1 and R_2 are actual rates, and, in addition, we are free to choose a null random variable V. This implies that, for any $\delta_1 > 0$, a code can be constructed, with actual rates

$$R_1 = I(X;U|Z) + \delta_1 \tag{194}$$

$$R_2 - R_1 = I(X;W|U,Y) + \delta_1, \tag{195}$$

implying that the region indicated by (33)-(36) is achievable. Thus it remains to show that if the region $\hat{\mathcal{R}}^*(\boldsymbol{D})$ is achievable, then $\mathcal{R}^*(\boldsymbol{D})$ is also achievable. To this end, observe that (33)-(36) imply that, for any random variable V satisfying the Markov conditions in the definition of $\mathcal{R}^*(\boldsymbol{D})$, and for some $\delta_2 \geq 0$ and $\delta_3 \geq 0$, there exists a SR code with actual rates

$$R_1 = I(X;U|Z) + I(X;V|Y) - \delta_2 \tag{196}$$

$$R_2 = I(X;U|Z) + I(X;\tilde{W},V|U,Y) - \delta_3, \tag{197}$$

and necessarily $R_2 \geq R_1$. Next, since R_2 is the exact total rate, we can increase the rate of the first stage by δ_2, on the account of the differential rate $\Delta R = R_2 - R_1$, without altering the total rate R_2. The distortions at the two stages remain intact, of course. Thus there exists a SR code with actual rates

$$R_1 = I(X;U|Z) + I(X;V|Y) \tag{198}$$

$$R_2 - R_1 = I(X;\tilde{W}|U,V,Y) - \delta_3, \tag{199}$$

implying that $\mathcal{R}^*(\boldsymbol{D})$ is achievable.

8 Proof of (61)

For notational convenience, let us define $S = \tilde{f}_2(W)$. By definition, $X \multimap W \multimap S$ holds. We have to show that if

$$I(X;W) = I(X;S) \tag{200}$$

then necessarily

$$X \multimap S \multimap W. \tag{201}$$

We would like to stress the conditions for equality in the log-sum inequality. Recall that, by the log-sum inequality, for any vectors of non-negative numbers $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}, \tag{202}$$

with equality if and only if $\frac{a_i}{b_i} = const$. Since we can always add zero elements to the a_i's sequence, using the convention $0 \log \frac{0}{b} = 0 \log \frac{0}{0} = 0$, we can re-phrase and say that equality holds in (202) if and only if there exists a constant c such that, for any i, $\frac{a_i}{b_i}$ is either 0 or c. Hence

$$\begin{aligned} I(X;W) &= I(X;WS) = \sum_{x,s}\left[\sum_w P_{XW}(x,w)P_{S|W}(s|w) \log \frac{P_{XW}(x,w)P_{S|W}(s|w)}{P_X(x)\sum_{x'} P_{XW}(x',w)P_{S|W}(s|w)}\right] \\ &\geq \sum_{x,s}\left(\sum_w P_{XW}(x,w)P_{S|W}(s|w)\right) \log \frac{\left(\sum_w P_{XW}(x,w)P_{S|W}(s|w)\right)}{P_X(x)\left(\sum_w \sum_{x'} P_{XW}(x',w)P_{S|W}(s|w)\right)} \\ &= I(X;S). \end{aligned} \tag{203}$$

Now define the set $G_{x,s}$

$$G_{x,s} = \{w : P_{XWS}(x,w,s) > 0\}. \tag{204}$$

The inequality in (203) holds with equality if and only if there exist constants $c_{x,s}$, $(x,s) \in \mathcal{X} \times \mathcal{S}$, such that

$$\frac{P_{XW}(x,w)P_{S|W}(s|w)}{\sum_{x'} P_{XW}(x',w)P_{S|W}(s|w)} = P_{X|W}(x|w) = c_{x,s} \quad \forall w \in G_{x,s}. \tag{205}$$

Since S is a deterministic function of W, we have

$$\begin{aligned} G_{x,s} \cap G_{x',s'} &= \emptyset \text{ whenever } s \neq s', \\ P_W(G_{x,s}) &= P_S(s). \end{aligned} \tag{206, 207}$$

Moreover, the joint distribution of X and S satisfies

$$\begin{aligned}
P_{XS}(x,s) &= \sum_{w} P_{X|W}(x|w) P_{S|W}(s|w) P_W(w) \\
&= \sum_{w \in G_{x,s}} c_{x,s} \cdot 1 \cdot P_W(w) = c_{x,s} P_W(G_{x,s}) = c_{x,s} P_S(s). \qquad (208)
\end{aligned}$$

In view of the above, we can write

$$P_{XWS}(x,w,s) \stackrel{(a)}{=} P_{W|S}(w|s) P_{X|W}(x|w) P_S(s) \stackrel{(b)}{=} P_{W|S}(w|s) P_{X|S}(x|s) P_S(s) \quad \forall\, w \in G_{x,s} \qquad (209)$$

where (a) holds due to the original Markov chain, and (b) is due to (207) and (208). Now observe that the far right hand side of (209) is a valid probability distribution. Thus, if it agrees with $P_{XWS}(x,w,s)$ for all triples (x,w,s) for which $P_{XWS}(x,w,s)$ is positive, it must agree with it everywhere, that is, also when $w \notin G_{x,s}$. Hence the relation $X \ominus S \ominus W$ follows. $\square$

References

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

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

[3] W. H. R. Equitz, *Successive refinement of information*, Ph.D. dissertation, Stanford University, June 1989.

[4] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269-274, Mar. 1991.

[5] F.-W. Fu and R. W. Yeung, "On the rate-distortion region for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2012-2021, July 2002.

[6] A. M. Gerrish, *Estimation of Information Rates*, Ph.D. thesis, Stanford Univ., Dept. of Elect. Eng., 1963.

[7] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 727-734, Nov. 1985.

[8] A. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inform. Theory*, vol. 40, pp. 2031-2034, Nov. 1994.

[9] V. N. Koshelev, "Hierarchical coding of discrete sources," *Probl. Peredachi Inform.*, vol. 16, no. 3, pp. 31-49, 1980. English translation: vol. 16, pp. 186-203, 1980.

[10] V. N. Koshelev, "On the divisibility of discrete sources with an additive single-letter distortion measure," *Probl. Peredachi Inform.*, vol. 30, no. 1, pp. 31-50, 1994. English translation: vol. 30, no. 1, pp. 27-43, 1994.

[11] N. Merhav and S. Shamai, "On joint source-channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel," to appear in *IEEE Trans. Inform. Theory*, November 2003.

[12] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. Inform. Theory*, vol 40, pp. 253-259, Jan. 1994.

[13] S. Shamai and S. Verdú, "Capacity of channels with uncoded side information," *Europ. Trans. Telecommun.*, vol. 6, no. 5, pp. 587-600, Sept.-Oct. 1995.

[14] S. Shamai, S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564-579, March 1998.

[15] 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.

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