IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

On Successive Refinement for the Kaspi/ Heegard-Berger Problem

Alina Maor and Neri Merhav

CCIT Report #711
December 2008

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT REPORT #711 December 2008

On Successive Refinement for the Kaspi/Heegard-Berger Problem

Alina Maor and Neri Merhav

Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
{alinam@tx, merhav@ee}.technion.ac.il

Abstract

Consider a source that produces independent copies of a triplet of jointly distributed random variables, $\{X_i, Y_i, Z_i\}_{i=1}^{\infty}$. The process $\{X_i\}$ is observed at the encoder, and is supposed to be reproduced at two decoders, decoder Y and decoder Z , where $\{Y_i\}$ and $\{Z_i\}$ are observed, respectively, in either a causal or non-causal manner. The communication between the encoder and the decoders is carried in two successive stages. In the first stage, the transmission is available to both decoders and they reconstruct the source according to the received bit-stream and the individual side information ($\{Z_i\}$ or $\{Y_i\}$). In the second stage, additional information is sent to both decoders and they refine the reconstructions of the source according to the available side information and the transmissions at both stages. It is desired to find the necessary and sufficient conditions on the communication rates between the encoder and decoders, so that the distortions incurred (at each stage) will not exceed given thresholds. For the case of non-degraded causal side information at the decoders, an exact single-letter characterization of the achievable region for is derived for the case of pure source-coding. Then, for the case of communication between the encoder and decoders carried over independent memoryless discrete channels with random states known causally/non-causally at the encoder and with causal side information about the source at the decoders, a single-letter characterization of all achievable distortion in both stages is provided and it is shown that the separation theorem holds. Finally, for non-causal degraded side information, inner and outer bounds to the achievable rate-distortion region are derived. These bounds are shown to be tight for certain cases of reconstruction requirements at the decoders, thereby shading some light on the problem of successive refinement with non-degraded side information at the decoders.

Index terms - causal/non-causal side information, channel capacity, degraded side-information, joint source-channel coding, separation theorem, source coding, successive refinement.

1 Introduction

We consider an instance of the multiple description problem, which is successive refinement (SR) of information. The term "successive refinement of information" is applicable

to systems where the reconstruction of the source is done in a number of stages. In such systems, a source is encoded by a single encoder which communicates with either a single decoder or a number of decoders in a successive manner. At each stage, the encoder sends some amount of information about the source to the decoder of that stage, which also has access to all previous transmissions. The decoder bases its reconstruction on all available transmissions, and, possibly, on some additional side information (SI). The quality of reconstruction at each stage (at each decoder) is measured with respect to a predefined distortion measure. In the case of pure source coding, the information transmitted by the encoder at each stage arrives at the decoder noiselessly, while in the case of noisy channels connecting the encoder and decoders, the transmission received at the decoder is corrupted and thus, joint source-channel coding should be applied.

A number of works have dealt with the problem of successive refinement [1]-[4], and the related problem of hierarchical coding [5]-[7]. In [4], the problem of successive source coding was studied for the Wyner-Ziv setting, i.e., when SI is available to each decoder non-causally [8]. The encoder transmits a source sequence, $\boldsymbol{X}$, to two decoders in two successive stages. Necessary and sufficient conditions were provided in [4], in terms of single-letter formulas, for the achievability of information per-stage rates corresponding to given distortion levels of each communication step. For the case of identical SI available at all decoders, the two-stage coding scheme was extended to include any finite number of stages. Also, conditions for a source to be successively refinable with degraded SI were introduced in [4] for the two-stage case. Generally speaking, the notion of degraded SI means that the quality of SI available at the decoders of later stages is better than that of earlier stages.

In [6], the problem of successive refinement with SI available non-causally at each decoder was studied from a different viewpoint. Instead of considering per-stage communication rates, the analysis of successive refinement was performed with respect to cumulative (sum-) rates achievable at each stage, under per-stage source restoration assumptions. A single-letter characterization of the achievable region with successive coding sum-rates and distortions was provided for the case of degraded SI at the decoders. It turned out that when the rate-sums are analyzed, it is possible to characterize an achievable rate-distortion region for any number of stages as long as the SI at the decoders is degraded.

In [7], the problem of successive refinement was investigated for the case of SI available causally at the decoders. It turned out that, unlike the above described non-causal settings,

when SI is available causally, the characterization of the achievable per-stage rate-distortion region is possible without constraining SI to be degraded.

The works reported in the field of successive refinement thus far have considered refinement of information when the transmission at each stage has been addressed to a single decoder. There are, however, many applications where a single encoder conveys information to several decoders in a single transmission. Heegard and Berger [9] and Kaspi [13] studied independently the following scenario: a single encoder communicates via a single transmission with two decoders one of which accesses the transmission only, while the other has a non-causal access to some SI correlated with the source. The source sequence should be reconstructed at both decoders with a certain accuracy and, under these distortion constraints, it is desired to reduce the communication rate as much as possible.

The minimum achievable communication rate, i.e., the rate-distortion function obtained for this setup is referred to as the *Heegard-Berger rate-distortion function*. It was also extended in [9] to include a coding theorem for more that two decoders, each having access to a different SI with a degraded structure. Now, assume that there is a demand for a better reconstruction at either one or both decoders, i.e., the source is required to perform a multi-level successive refinement, still communicating with all decoders via a single transmission. A question of obvious interest is the following: is it possible to characterize the achievable rate-distortion region for this generalized problem of successive refinement?

In this work, we jointly extend the works of [9], [4], [6] and [7]. Specifically, we study the scenario of two-decoders, two-stage successive refinement of information, with SI available at all decoders in either a causal[1] or non-causal manner. For the causal case, we provide a single-letter characterization of the achievable rate-distortion region, which is straightforwardly extendable to any number of decoders accessible in each stage and any finite number of stages. For the case of non-causal SI, we provide inner and outer bounds to the achievable rate-distortion region for the case of degraded SI. Note that although the SI is degraded at each stage, when both stages are viewed jointly, SI is no longer degraded (same SI is used at both stages and thus it is not longer possible to say that at the later stage the SI is of better

[1]There are few reasons for our interest in the scenario of causal SI at the decoders. The first motivation is an attempt to include the concept of SR in zero-delay sequential coding systems. Schemes with causal SI can be also viewed as denoising systems, where each decoder performs SI sequential filtering with the aid of rate-constrained information provided by the encoder. Introducing SR to such systems is of practical importance, as it simplifies the decoding process in the sense of performing denoising of the SI symbols causally, in a number of steps, rather than using the entire SI sequence.

quality), and therefore this setting is of particular interest. When considering the case of causal SI, we provide the exact achievable region in terms of the per-stage rates, while for the case of non-causal SI, we refer to the sum-rates. The difficulty in characterizing the per-stage rates for a general scheme here is similar to that faced in [4].

For the case of causal SI we then extend the noise-free setting into a problem of communication over noisy discrete memoryless channels with random states known causally or non-causally at the encoder at all stages of communication. We obtain a single-letter characterization of the region of all achievable distortions for both decoders at both stages of communication. This characterization reveals that the separation principle is applicable for this problem, i.e., it is possible to separately encode the source sequence with a good SR source code and then to transmit the obtained bitstreams with a good channel code at each stage of communication, without losing asymptotic optimality. This part of the paper extends the results of [10] and [11] to the multi-stage multi-decoder communication. Specifically, in [10] it was shown that the separation principle holds for a single-stage single encoder-decoder communication over a simple discrete memoryless channel. This setting has been extended in [11] to communication over a channel with random parameters known causally or non-causally at the encoder and decoder having non-causal access to the SI correlated with the source and there also it was shown that separate source channel coding is, in fact, optimal.

Note that all known closed form (single-letter) results regarding SR (and its variations) for decoders having non-causal access to different SI data, such as [9], [3], [4] and [6], treat the case of degraded SI at the decoders. Thus, there is a special interest in the following subcase of the problem treated in this paper - SR with non-causal degrades SI at the decoders, when decoders are accessed in the reversed order of degradedness of SI. Specifically, for the two-stage scheme, assume that in the first stage some information is to be conveyed to the decoder that has access to SI of a better quality. Then, at the refinement stage, the decoder with less informative SI should reconstruct the source sequence based on the transmissions of both stages. This problem has been also addressed in [14]. Specifically, in [14], inner and outer bounds on the achievable rates and distortions have been derived and it was shown that these bounds coincide when reconstruction at either stage should be lossless at the matching decoder. The work presented in [14] has been performed in parallel to the researched described in this paper and the inner bounds presented in [14] can be

easily derived from the results of this paper. The outer bound provided in this paper is more precise than that provided in [14] as is discussed in detail in Section 4.

The outline of the paper is as follows: In Section 2, we give notation conventions used throughout the paper. A formal definition of the problem is provided in Section 3. In Section 4, for the case of causal SI at the decoders, we give the exact characterizations of the achievable rate-distortion region and formulate the coding theorems for the successive-refinement two-stage source coding and the joint source-channel coding; for the case of non-causal SI at the decoders, we provide inner and outer bounds to the rate-distortion region and show that in some cases these bounds are tight. The proofs are provided in Sections 5 and 6 for the cases of causal and non-causal SI, respectively.

2 Notation Conventions and Preliminaries

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets will be denoted, respectively, by boldface capital letters, the corresponding boldface lower case letters, and calligraphic letters, superscripted by the dimensions. The notations x_i^j and X_i^j, where i and j are integers and $i \leq j$, will designate segments $(x_i, ..., x_j)$ and $(X_i, ..., X_j)$, respectively, where for $i = 1$, the the subscript will be omitted. For example, a random vector $\boldsymbol{X} = X_1^N = (X_1, ..., X_N)$, ($N$-positive integer) may take a specific vector value $\boldsymbol{x} = x_1^N = (x_1, ..., x_N)$ in $\mathcal{X}^N$, the Nth order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

Sources and channels will be denoted generically by the letter P, subscripted by the name of the random variable and its conditioning, if applicable, e.g., $P_X(x)$ is the probability of $X = x$, $P_{Y|X}(y|x)$ is the conditional probability of $Y = y$ given $X = x$, and so on. Whenever clear from the context, these subscripts will be omitted. The class of all discrete memoryless sources (DMSs) with a finite alphabet $\mathcal{X}$ will be denoted by $\mathcal{P}(\mathcal{X})$, with P_X denoting a particular DMS in $\mathcal{P}(\mathcal{X})$, i.e., $\mathcal{P}(\mathcal{X}) = \{P_X : \sum_{x\in\mathcal{X}} P_X(x) = 1; \ \ \forall x \in \mathcal{X} : \ \ P_X(x) \geq 0\}$. For a given positive integer N, the probability of an N-vector $\boldsymbol{x} = (x_1, ..., x_N)$ drawn from

a DMS P_X, is given by

$$\Pr\{X_i = x_i,\ i = 1, ..., N\} = \prod_{i=1}^{N} P_X(x_i) \stackrel{\triangle}{=} P_X(\boldsymbol{x}). \tag{1}$$

A Markov chain formed by a triplet of random variables (RVs) (X, Y, Z) with a joint distribution $P_{XYZ}(x, y, z)$ will be denoted by $X \div Y \div Z$.

A distortion measure (or distortion function) is a mapping from the set $\mathcal{X} \times \mathcal{Y}$ into the set of non-negative reals: $d : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+$. The additive distortion $d(\boldsymbol{x}, \boldsymbol{y})$ between two vectors $\boldsymbol{x} \in \mathcal{X}^N$ and $\boldsymbol{y} \in \mathcal{Y}^N$ is given by: $d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i)$.

The information-theoretic quantities, used throughout this paper, are denoted using the conventional notations [12]: For a pair of discrete random variables (X, Y) with a joint distribution $P_{XY}(x, y) = P_X(x)P_{Y|X}(y|x)$, the entropy of X is denoted by $H(X)$, the joint entropy - by $H(X, Y)$, the conditional entropy of Y given X - by $H(Y|X)$, and the mutual information by $I(X; Y)$, etc., where logarithms are defined to the base 2.

We next describe the notation related to the method of types, which is used throughout this paper in the direct proofs. For a given memoryless source P_X and a vector $\boldsymbol{x} \in \mathcal{X}^N$, the empirical probability mass function is a vector $P_{\mathbf{x}} = \{P_{\boldsymbol{x}}(a), a \in \mathcal{X}\}$, where $P_{\mathbf{x}}(a)$ is the relative frequency of the letter $a \in \mathcal{X}$ in the vector $\boldsymbol{x}$. For a scalar $\delta > 0$, the set $T^{\delta}_{P_X}$ of all δ-typical sequences is the set of the sequences $\boldsymbol{x} \in \mathcal{X}^N$ such that $|P_{\mathbf{x}}(a) - P_X(a)| \leq \delta$ for every $a \in \mathcal{X}$. In this paper, we use some known results from [12]. First, for every $\boldsymbol{x} \in T^{\delta}_{P_X}$,

$$2^{-N[H(X)+\epsilon_1]} \leq P_X(\boldsymbol{x}) \leq 2^{-N[H(X)-\epsilon_1]}, \tag{2}$$

where $\epsilon_1 = \epsilon_1(\delta)$ vanishes as $\delta \to 0$ and $N \to \infty$. It is also well-known (by the weak law of large numbers) that:

$$\Pr\left\{\boldsymbol{X} \notin T^{\delta}_{P_X}\right\} \leq \epsilon_2 \tag{3}$$

where $\epsilon_2 = \epsilon_2(\delta)$, $\epsilon_2 \to 0$ as $N \to \infty$.

For a given conditional distribution $P_{Y|X}$ and for each $\boldsymbol{x} \in T^{\delta}_{P_X}$, the set $T^{\tilde{\delta}}_{P_{XY}}$ of all sequences $\boldsymbol{y}$ that are jointly δ-typical with $\boldsymbol{x}$, is the set of all $\boldsymbol{y}$ such that:

$$\left|P_{\mathbf{xy}}(a, b) - P_{\mathbf{x}}(a)P_{Y|X}(b|a)\right| \leq \tilde{\delta} \tag{4}$$

for all $a \in \mathcal{X}, b \in \mathcal{Y}$, where $P_{\mathbf{xy}}(a, b)$ denotes the fraction of occurrences of the pair (a, b) in $(\boldsymbol{x}, \boldsymbol{y})$. For any $\boldsymbol{x} \in T^{\delta}_{P_X}$ and any $\tilde{\delta} > \delta$,

$$2^{-N[I(X;Y)+\epsilon_3)]} \leq \sum_{\mathbf{y}:(\mathbf{x},\mathbf{y}) \in T^{\tilde{\delta}}_{P_{XY}}} P_Y(\mathbf{y}) \leq 2^{-N[I(X;Y)-\epsilon_3]}, \tag{5}$$

where $\epsilon_3 = \epsilon_3(\delta, \tilde{\delta})$ vanishes as $\delta, \tilde{\delta} \to 0$ and $N \to \infty$. These typicality definitions and properties, are straightforwardly extendable for jointly typical sequences which come in triplets, quadruplets and so on and we use these in the paper.

3 System Description and Problem Definition

We refer to the communication system depicted in Figure 1. Consider a source that produces

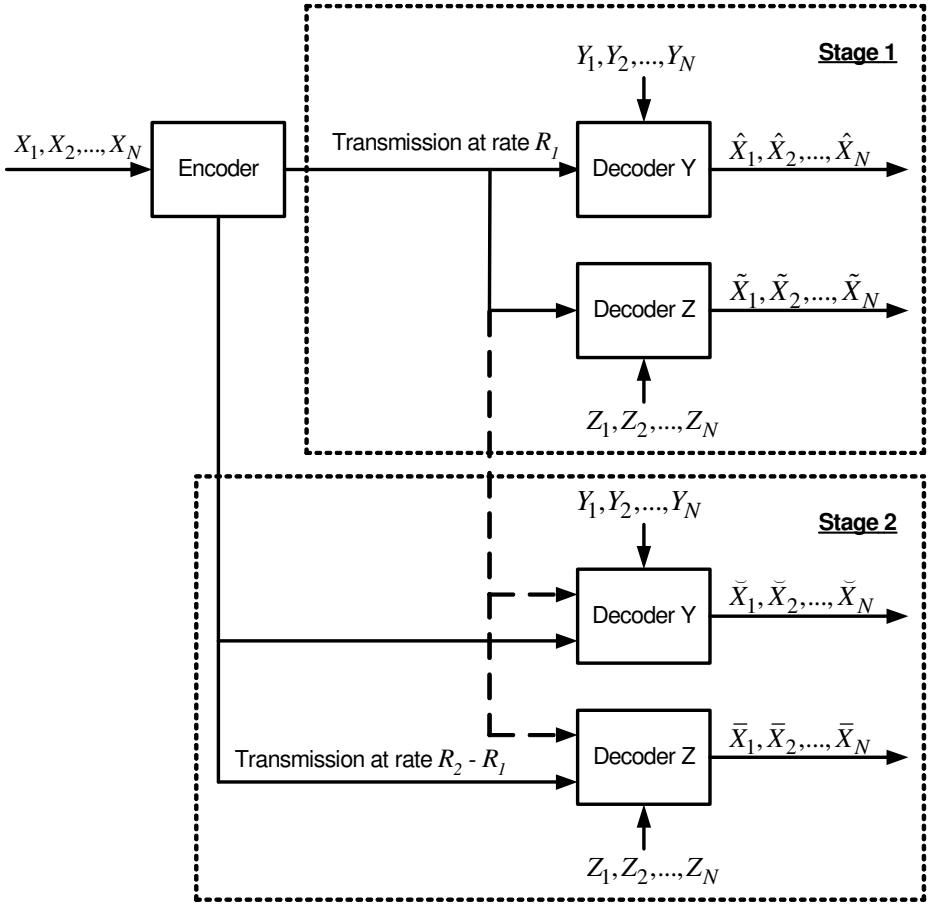

Figure 1: Two-stage communication scheme.

independent copies $\{X_i, Y_i, Z_i\}_{i \geq 1}$ of a triplet of RV's, (X, Y, Z), taking values in a finite alphabet $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, and drawn under a joint distribution P_{XYZ}. The process $\{X_i\}$ is observed at the encoder and is supposed to be reproduced at the decoders, where $\{Y_i\}$ and $\{Z_i\}$ are observed at decoders Y and Z, respectively. The source is available at the encoder non-causally and at the decoders either causally or non-causally, at all stages. At the first stage of SR, the reproductions at decoders Y and Z take values in the finite sets, $\hat{\mathcal{X}}$ and $\tilde{\mathcal{X}}$, respectively, while at the second stage, the reproduction finite sets are $\breve{\mathcal{X}}$ and $\bar{\mathcal{X}}$,

respectively.

The coding scheme with causal/non-causal SI at the decoders operates as follows: at the first transmission, the encoder sends some amount of information to both decoders over the channel. We consider block coding, i.e., an N-vector $\boldsymbol{X}$ (N is a positive integer) is encoded at rate R_1 into a binary sequence of length M_1, where $R_1 = \frac{1}{N}\log_2 M_1$. The binary sequence then takes values in $\{0, 1, ..., 2^{NR_1} - 1\}$. At the first stage, when non-causal SI is considered, decoder Y receives the binary bitstream and reconstructs $\hat{\boldsymbol{X}} = (\hat{X}_1, ..., \hat{X}_N) \in \hat{\mathcal{X}}^N$, based on it and the SI $\boldsymbol{Y}$, while in the case of causal SI, the reconstruction of the i-th component, $\hat{X}_i$, is based on the encoder transmission and only i first symbols of the SI, i.e., Y_1^i. Similarly, with non-causal SI, decoder Z uses the encoder transmission and $\boldsymbol{Z}$ in its entirety and reproduces $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_N) \in \tilde{\mathcal{X}}^N$, while in the case of causal SI, only the bitstream and Z_1^i are used for reproduction of $\tilde{X}_i$. The quality of reconstruction at each of the decoders is judged in terms of the expectations of additive distortion measures $d_{y,1}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{N}\sum_{i=1}^N d_{y,1}(X_i, \hat{X}_i)$ and $d_{z,1}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \frac{1}{N}\sum_{i=1}^N d_{z,1}(X_i, \tilde{X}_i)$, where $d_{y,1}(X, \hat{X})$ and $d_{z,1}(X, \tilde{X})$, $X \in \mathcal{X}$, $\hat{X} \in \hat{\mathcal{X}}$, $\tilde{X} \in \tilde{\mathcal{X}}$, are non-negative, bounded distortion measures. At the second stage, the encoder sends, at rate $R_2 - R_1$, an additional information about the source sequence to both decoders, also in the form of a binary bitstream, this time of length $M_2 \stackrel{\triangle}{=} 2^{N(R_2-R_1)}$, taking values in $\{0, 1, ..., 2^{N(R_2-R_1)} - 1\}$. The decoders reconstruct the source sequence with better accuracy (in terms of the distortion measures) according to both transmissions of the encoder and the individual SI's. The distortions measures used at the decoders Y and Z at this stage are also additive, $d_{y,2}(\boldsymbol{X}, \check{\boldsymbol{X}}) = \frac{1}{N}\sum_{i=1}^N d_{y,2}(X_i, \check{X}_i)$ and $d_{z,2}(\boldsymbol{X}, \bar{\boldsymbol{X}}) = \frac{1}{N}\sum_{i=1}^N d_{z,2}(X_i, \bar{X}_i)$, where $d_{y,2}(X, \check{X})$ and $d_{z,2}(X, \bar{X})$, $X \in \mathcal{X}$, $\check{X} \in \check{\mathcal{X}}$, $\bar{X} \in \bar{\mathcal{X}}$, are non-negative, bounded distortion measures. This setting can be straightforwardly extended to any number of refinement stages as well as any number of decoders at each stage. We confine ourselves to the case of two decoders and two stages.

We begin with the case of non-causal SI.

Definition 1. *An $(N, M_1, M_2, \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2)$ source code for a single encoder, two decoders and two-stage successive refinement with non-causal SI at the decoders, for the source P_{XYZ}, consists of a first-stage encoder-decoder triplet $(f_1, g_{y,1}, g_{z,1})$:*

$$f_1 : \mathcal{X}^N \to \{1, 2, ..., M_1\}, \tag{6}$$

$$g_{y,1} : \mathcal{Y}^N \times \{1, 2, ..., M_1\} \to \hat{\mathcal{X}}^N, \tag{7}$$

$$g_{z,1} : \mathcal{Z}^N \times \{1, 2, ..., M_1\} \to \tilde{\mathcal{X}}^N, \tag{8}$$

and a second-stage encoder-decoder triplet $(f_2, g_{y,2}, g_{z,2})$*:*

$$f_2 : \mathcal{X}^N \to \{1, 2, ..., M_2\}, \tag{9}$$

$$g_{y,2} : \mathcal{Y}^N \times \{1, 2, ..., M_1\} \times \{1, 2, ..., M_2\} \to \check{\mathcal{X}}^N, \tag{10}$$

$$g_{z,2} : \mathcal{Z}^N \times \{1, 2, ..., M_1\} \times \{1, 2, ..., M_2\} \to \bar{\mathcal{X}}^N, \tag{11}$$

such that

$$Ed_{y,1}(\mathbf{X}, \hat{\mathbf{X}}) \leq N\Delta_{y,1} \qquad Ed_{z,1}(\mathbf{X}, \tilde{\mathbf{X}}) \leq N\Delta_{z,1}$$

and

$$Ed_{y,2}(\mathbf{X}, \check{\mathbf{X}}) \leq N\Delta_{y,2} \qquad Ed_{z,2}(\mathbf{X}, \bar{\mathbf{X}}) \leq N\Delta_{z,2}.$$

When SI is available to the decoders causally, in analogy to Definition 1, it is possible to define an $(N, M_1, M_2, \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2)$, source code for coding with causal SI, where the first-stage decoder pair $(g_{y,1}, g_{z,1})$ is now presented via $\{g_{y,1,i}\}_{i=1}^N$ and $\{g_{z,1,i}\}_{i=1}^N$, where $g_{y,1,i}$ and $g_{z,1,i}$ denote the reconstruction functions for the $i-th$ symbol of $\hat{X}^N$ and $\tilde{X}^N$, respectively:

$$g_{y,1,i} : \mathcal{Y}_1^i \times \{1, 2, ..., M_1\} \to \hat{\mathcal{X}}, \tag{12}$$

$$g_{z,1,i} : \mathcal{Z}_1^i \times \{1, 2, ..., M_1\} \to \tilde{\mathcal{X}}. \tag{13}$$

Similar adjustments of definitions should be applied to the second stage, considering now $(g_{y,2}, g_{z,2})$ presented in terms of $\{g_{y,2,i}\}_{i=1}^N$ and $\{g_{z,2,i}\}_{i=1}^N$:

$$g_{y,2,i} : \mathcal{Y}_1^i \times \{1, 2, ..., M_1\} \times \{1, 2, ..., M_2\} \to \check{\mathcal{X}}, \tag{14}$$

$$g_{z,2,i} : \mathcal{Z}_1^i \times \{1, 2, ..., M_1\} \times \{1, 2, ..., M_2\} \to \bar{\mathcal{X}}. \tag{15}$$

The sum-rate pair (R_1, R_2) of the $(N, M_1, M_2, \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2)$ code for two stage successive refinement for two decoders is given by $R_1 = \frac{1}{N}\log_2(M_1)$ and $R_2 = \frac{1}{N}\log_2(M_1 \cdot M_2)$.

Definition 2. *Given a distortion quadruplet* $\mathbf{D} = \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2$*, a rate pair* (R_1, R_2) *is said to be achievable with SI* (Y, Z) *if for every* $\epsilon > 0$*, there exists a sufficiently large block length* N*, for which there is an* $(N, 2^{N(R_1+\epsilon)}, 2^{N(R_2+\epsilon)}, \Delta_{y,1} + \epsilon, \Delta_{z,1} + \epsilon, \Delta_{y,2} + \epsilon, \Delta_{z,2} + \epsilon)$*, source code for successive refinement with non-causal SI at the decoders for the source* P_{XYZ}*.*

The definition of the notion of an achievable region with causal SI *per-stage* rates can be straightforwardly modified in parallel to Definition 2, referring to the first stage rate R_1 and the second-stage rate $\Delta R = R_2 - R_1 = \frac{1}{N}\log_2(M_2)$. The collection of all $\boldsymbol{D}$-achievable rate pairs is the achievable rate-region for successive-refinement coding with non-causal (respectively, causal) SI and is denoted by $\mathcal{R}(\boldsymbol{D})_{nc}$ (respectively, $\mathcal{R}(\boldsymbol{D})_c$). The collection of all $(R_1, R_2, \{\Delta_{1,k}, \Delta_{2,k}\}_{k=1}^2)$-achievable rate-distortion tuples is the achievable rate-distortion region, and is denoted by $\mathcal{RD}_{nc}$ and $\mathcal{RD}_c$, referring to non-causal and causal settings, respectively. In this work, we propose strategies for (asymptotically) achieving any given point in $\mathcal{RD}_c$ and certain points in $\mathcal{RD}_{nc}$.

It is also interesting to investigate the scenario where communication between the encoder and the decoders is carried over a noisy media. In this case, the source block $\boldsymbol{X}$ is fed into a *joint source-channel* encoder, whereas the corresponding blocks of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are fed as side information in either a causal or non-causal manner into the Y and Z decoders, respectively. In the sequel, we confine ourself to the case of causal source SI at both decoders.[2] In this paper, at each stage of communication, the noisy media is modeled by a discrete memoryless channel whose output is governed by its input and a random parameter which is known at the encoder either causally or non-causally.

Consider the communication scheme depicted in Figure 2. The channel used at the first stage is channel 1, $P_{B|A,S}$, and at the second stage is used channel 2, $P_{\bar{B}|\bar{A},\bar{S}}$. The channels are independent and we denote their capacities by C_1 and C_2, respectively. The channels work as follows: The input of Channel 1 is a vector pair (A_1^n, S_1^n), where n is a positive integer and where A and S take values in the finite sets, $\mathcal{A}$ and $\mathcal{S}$, respectively. Channel 1 produces a vector output B^n, whose components take values in the finite set $\mathcal{B}$. The conditional probability of (B^n) given (A^n, S^n) is characterized by $P_{B^n|A^n,S^n}(b^n|a^n, s^n) = \prod_{i=1}^n P_{B|A,S}(b_i|a_i, s_i)$. The vector A^n is referred to as the channel input and S^n is referred to as the channel state sequence, governed by another discrete memoryless process $P_{S^n}(s^n) = \prod_{i=1}^n P_s(s_i)$, independently of (X^N, Y^N, Z^N). The operation of Channel 2 is described in a similar fashion by the triplet $(\bar{A}^m, \bar{B}^m, \bar{S}^m,)$ instead of (A^n, B^n, S^n) and corresponding marginal and conditional probabilities. Note that in the context of Channel 2, all blocks are of length m, where m is a positive integer. We denote the source-channel rate ratios by

[2]Since the complete characterization of $\mathcal{RD}_{nc}$ is still open, there is no point in analyzing the scenario of communication over noisy channels for the case of non-causal *source* SI at the decoders.

$\rho_1 \stackrel{\triangle}{=} \frac{n}{N}$ and $\rho_2 \stackrel{\triangle}{=} \frac{m}{N}$.

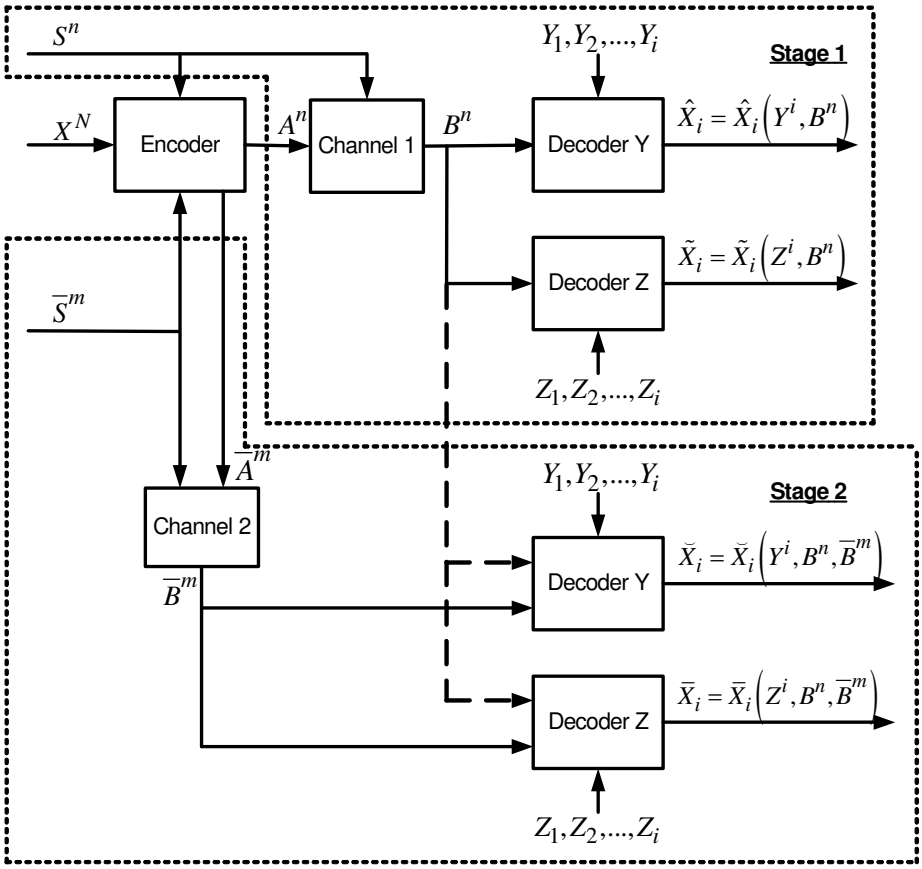

Figure 2: Communication over noisy channels with causal SI.

Now, instead of the binary bitstream generated in the noise-free case, the first-stage joint source-channel encoder implements a deterministic function $a^n = f_1(x^N, s^n)$ and the second-stage joint source-channel encoder implements another deterministic function $\bar{a}^m = f_2(x^N, \bar{s}^m)$. If the channel states are available at the encoder causally, each channel symbol a_i depends only on x^N, a^{i-1} and s^i, and each $\bar{a}_i$ depends only on x^N, $\bar{a}^{i-1}$ and $\bar{s}^i$. In the non-causal case, each channel symbol a_i depends on x^N, a^{i-1} and s^n, and each $\bar{a}_i$ depends x^N, $\bar{a}^{i-1}$ and $\bar{s}^m$. The first-stage decoders Y and Z are defined now by deterministic functions $g_{y,1}(y^N, a^n)$ and $g_{z,1}(z^N, a^n)$, respectively, and the second stage decoders Y and Z are defined by deterministic functions $g_{y,2}(y^N, b^n, \bar{b}^m)$ and $g_{z,2}(z^N, b^n, \bar{b}^m)$, respectively. The channel states $\boldsymbol{S}$ and $\bar{\boldsymbol{S}}$ are independent and we interpret the independence of the channels via the Markov relation $(\boldsymbol{S}, \boldsymbol{B}) \div \boldsymbol{X} \div (\bar{\boldsymbol{S}}, \bar{\boldsymbol{B}})$.

In parallel to Definitions 1 and 2, we define the following:

Definition 3. *For a given memoryless source P_{XYZ} and two memoryless channels with random states $P_{B|A,S}$ and $P_{\bar{B}|\bar{A},\bar{S}}$ an $(N, n, m, \Delta_{y,1}, \Delta_{z,1}, \Delta_{y,2}, \Delta_{z,2})$ joint source-channel code for successive refinement with causal state information at the encoder and causal side information at the decoders consists of a sequence of n first-stage encoding functions:*

$$f_{1,i} : \mathcal{X}^N \times \mathcal{S}^i \to \mathcal{A}_i, \quad i = 1, ..., n, \tag{16}$$

a sequence of N first-stage decoding functions

$$g_{y,1,i} : \mathcal{Y}^i \times \mathcal{B}^n \to \hat{\mathcal{X}}, \quad i = 1, ..., N, \tag{17}$$

and

$$g_{z,1,i} : \mathcal{Z}^i \times \mathcal{B}^n \to \tilde{\mathcal{X}}, \quad i = 1, ..., N, \tag{18}$$

a sequence of m second-stage encoder functions

$$f_{2,i} : \mathcal{X}^N \times \bar{\mathcal{S}}^i \to \bar{\mathcal{A}}_i, \quad i = 1, ..., m, \tag{19}$$

and a sequence of N second-stage decoding functions:

$$g_{y,2,i} : \mathcal{Y}^i \times \mathcal{B}^n \times \bar{\mathcal{B}}^m \to \check{\mathcal{X}}, \quad i = 1, ..., N, \tag{20}$$

and

$$g_{z,2,i} : \mathcal{Z}^i \times \mathcal{B}^n \times \bar{\mathcal{B}}^m \to \bar{\mathcal{X}}, \quad i = 1, ..., N, \tag{21}$$

such that

$$Ed_{y,1}(\mathbf{X}, \hat{\mathbf{X}}) \leq N\Delta_{y,1} \qquad Ed_{z,1}(\mathbf{X}, \tilde{\mathbf{X}}) \leq N\Delta_{z,1}$$

and

$$Ed_{y,2}(\mathbf{X}, \check{\mathbf{X}}) \leq N\Delta_{y,2} \qquad Ed_{z,2}(\mathbf{X}, \bar{\mathbf{X}}) \leq N\Delta_{z,2},$$

where the expectations are w.r.t. the source and the channels.

Definition 4. *Given the source-channel rate ratios ρ_1 and ρ_2, a distortion quadruplet $\mathbf{D} = \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2$ is said to be achievable if for every $\epsilon > 0$, there exist sufficiently large N, n and m, with $\rho_1 = n/N$ and $\rho_2 = m/N$, and there exists an $(N, n, m, \Delta_{y,1}+\epsilon, \Delta_{z,1}+\epsilon, \Delta_{y,2}+\epsilon, \Delta_{z,2}+\epsilon)$ joint source-channel code for successive refinement with causal/non-causal state information at the encoder and causal side information at the decoders for the source P_{XYZ} and the channels $P_{B|A,S}$, $P_{\bar{B}|\bar{A},\bar{S}}$. The distortion region, denoted $\mathcal{D}$, is the closure of the set of all achievable quadruplets $\mathbf{D}$.*

We provide a single-letter characterization of $\mathcal{D}$ for the cases of causal/non-causal channel state information availability at the encoder. In particular, we show that any given point in $\mathcal{D}$ can be achieved by separate source coding for the source P_{XYZ} (achieving $\mathcal{RD}_c$) and capacity-achieving channel coding (independently of the source).

4 Main Result

4.1 Causal Side Information

4.1.1 Pure Source Coding

We begin with the case where availability of SI at the decoders is restricted to be causal. Let a distortion quadruplet $\boldsymbol{D} \stackrel{\triangle}{=} (\{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^{2})$ be given. Define $\mathcal{R}^*(\boldsymbol{D})_c$ to be the set of all rate pairs (R_1, R_2) for which there exist RVs (W_1, W_2), taking values in finite alphabets, $\mathcal{W}_1$,$\mathcal{W}_2$, respectively, s.t the following holds simultaneously:

1. The following Markov chain holds:

$$(W_1, W_2) \div X \div (Y, Z). \tag{22}$$

2. There exist deterministic decoding functions $G_{y,1}: \ \mathcal{Y} \times \mathcal{W}_1 \to \hat{\mathcal{X}}$, $G_{z,1}: \ \mathcal{Z} \times \mathcal{W}_1 \to \tilde{\mathcal{X}}$, and $G_{y,2}: \ \mathcal{Y} \times \mathcal{W}_1 \times \mathcal{W}_2 \to \check{\mathcal{X}}$, $G_{z,2}: \ \mathcal{Z} \times \mathcal{W}_1 \times \mathcal{W}_2 \to \bar{\mathcal{X}}$, such that

$$Ed_{y,1}(X, G_{y,1}(Y, W_1)) \leq \Delta_{y,1} \tag{23}$$

$$Ed_{z,1}(X, G_{z,1}(Z, W_1)) \leq \Delta_{z,1} \tag{24}$$

$$Ed_{y,2}(X, G_{y,2}(Y, W_1, W_2)) \leq \Delta_{y,2} \tag{25}$$

$$Ed_{z,2}(X, G_{z,2}(Z, W_1, W_2)) \leq \Delta_{z,2} \tag{26}$$

3. The alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$ satisfy:

$$|\mathcal{W}_1| \leq |\mathcal{X}| + 5, \quad |\mathcal{W}_2| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| + 2 \tag{27}$$

4. The rates R_1 and R_2 satisfy

$$R_1 \geq I(X; W_1) \quad R_2 - R_1 \geq I(X; W_2|W_1). \tag{28}$$

The main result of this subsection is the following:

Theorem 1. *For any DMS P_{XYZ},*

$$\mathcal{R}(\mathbf{D})_c = \mathcal{R}^*(\mathbf{D})_c. \tag{29}$$

The proof of Theorem 1 appears in Section 5. Note that when SI is available at the decoders causally, there is no degradedness assumption on SI, which is very different from the case of SR with non-causal SI even when a single decoder is considered at each stage [4]-[6], as well as for the multi-group SR discussed in the next section.

The relative simplicity of characterization of $\mathcal{R}(\boldsymbol{D})_c$ is better understood when studying the achievability scheme: The direct part is based on the fact that the encoder transmits a concatenation of indexes of the auxiliary codewords [3] instead of bin numbers transmitted in the non-causal setting [4]-[6]. Hence, each decoder can access all the auxiliary codewords directly and, unlike in the non-causal setting, it does not use its SI to retrieve codewords, but only for reconstruction. Unlike in the case of coding with non-causal SI at the decoders, the results obtained for the two-decoder two-stage coding with causal SI are straightforwardly extendable to any number of decoders and refinement stages and the number of auxiliary RVs is determined solely by the number of communication stages[4].

4.1.2 Joint Source-Channel Coding

We next address the problem of joint source channel coding, where at each communication stage the encoder conveys its information to two decoders over a noisy stationary memoryless channel governed by a random state, which is known causally or non-causally to the encoder. The general scheme is described in Fig. 2. The necessary and sufficient conditions for (Δ_1, Δ_2) to be the achievable distortion levels are summarized in the following Theorem:

Theorem 2. *Given a DMS P_{XYZ}, the distortion levels $(\{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^2)$ are achievable for successively refinable communication with causal SI at the decoders over noisy stationary memoryless channels $P_{B,O|A,S}$ and $P_{\bar{B},\bar{O}|\bar{A},\bar{S}}$ with channel states known at the encoder either causally or non-causally if and only if there exist auxiliary RVs W_1 and W_2, taking values in finite alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$, of cardinalities given by (27) and satisfying (22), and deterministic decoding functions $G_{y,1}$, $G_{z,1}$, $G_{y,2}$ and $G_{z,2}$, satisfying (23) - (26), respectively,*

[3] The direct use of indexes of the auxiliary codewords, similarly as is done for coding without SI at the decoder, was first introduced in [15], in the achievability proof of the characterization of the rate-distortion function with causal SI at the decoder.

[4] While, as we show in the next section, in the non-causal setting, at each stage, for each decoder, at least one auxiliary codeword is added to the direct scheme.

such that

$$I(X;W_1) \leq \rho_1 C_1, \tag{30}$$

$$I(X;W_2|W_1) \leq \rho_2 C_2. \tag{31}$$

There is an obvious similarity between the characterization of $\mathcal{D}_c$ and the characterization of the region of all achievable distortion levels described in Theorem 2, both for the cases of causal and non-causal state information at the encoder. The only difference in characterizations is the following: in the case of communication over noisy channel the upper-bounds in (30) and (31) are $\rho_1 C_1$ and $\rho_2 C_2$, while in the noise-free case, these bounds are substituted by R_1 and $R_2 - R_1$, respectively. Therefore, a possible achievability scheme is the one based on separate source and channel coding.

The direct proof of Theorem 2 comes from a concatenation of the asymptotically optimal source code designed for multi-group successive refinement, which is independent of the channels, and a reliable channel codes, independent of the source, designed for each of the channels (with channel state informations available to the encoder either causally or non-causally). The channel codes should achieve (at least asymptotically) the capacity of the relevant channels. Now, if such source and channel codes are used and the distortion constraints are maintained by the source code, as soon as $I(X;W_1) \leq \rho_1 C_1$ and $I(X;W_2|W_1) \leq \rho_2 C_2$, it is always possible to select source and channel rates R_{s1} and R_{c1} for the first stage and $R_{s2} - R_{s1}$ and R_{c2} for the second stage such that $NI(X;W_1) \leq NR_{s1} = nR_{c1} \leq nC_1$ and $NI(X;W_2|W_1) \leq N[R_{s2} - R_{s1}] = mR_{c2} \leq mC_2$. Now, it is possible to compress the source sequence into R_{s1} bits per symbol for the first stage and into $R_{s2} - R_{s1}$ bits per symbol for the refinement stage, such that the distortions $\{(\Delta_{y,j}, \Delta_{z,j})\}_{j=1}^2$ are satisfied and then map the obtained bitstreams of length NR_{s1} and $N[R_{s2} - R_{s1}]$ into channel codewords of length nR_{c1} and mR_{c2}, respectively. Since $R_{c1} \leq C_1$ and $R_{c2} \leq C_2$, from the standard coding theorem ([18] or [17]), there exist channel codes that cause asymptotically negligible distortions. Also, by the source coding theorem (Theorem 1) all the distortions for which $NI(X;W_1) \leq NR_{s1}$ and $NI(X;W_2|W_1) \leq N[R_{s2} - R_{s1}]$ are achievable. Thus, the distortions $\{(\Delta_{y,j}, \Delta_{z,j})\}_{j=1}^2$ such that $NI(X;W_1) \leq nC_1$ and $NI(X;W_2|W_1) \leq mC_2$ are achievable. The details of the converse proof are provided in Section 5, and, similarly as in the noise-free case, the proof is easily extendable to more than two communication stages and more than two decoders at each stage.

4.2 Non-Causal Degraded Side Information

Unlike in the case of causal SI, in the noncausal case, a closed-form characterization of the achievable rate-distortion region with non-causal SI at the decoders is yet to be derived. In this subsection, we provide outer and inner bounds to the achievable region, discuss the differences between the bounds and show that in certain cases, the bounds coincide, i.e., the rate-distortion region is fully characterized for these special cases. We begin with the outer bound.

4.2.1 Outer Bound

Define $\mathcal{R}^{**}(\boldsymbol{D})_{nc}$ to be the set of all rate pairs (R_1, R_2) for which there exist RVs $\{W_i\}_{i=1}^4$ and V, taking values in finite alphabets, $\{\mathcal{W}_i\}_{i=1}^4$ and $\mathcal{V}$, respectively, such that (s.t.) the following conditions are satisfied:

1.

$$(W_1, W_2, W_3, W_4, V) \div X \div Z \div Y \tag{32}$$

is a Markov chain.

2. There exist deterministic decoding functions $G_{y,1}: \mathcal{Y} \times \mathcal{W}_1 \to \hat{\mathcal{X}}$, $G_{z,1}: \mathcal{Z} \times \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{V} \to \tilde{\mathcal{X}}$, $G_{y,2}: \mathcal{Y} \times \mathcal{W}_1 \times \mathcal{W}_3 \times \mathcal{V} \to \check{\mathcal{X}}$ and $G_{z,2}: \mathcal{Z} \times \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{W}_3 \times \mathcal{W}_4 \times \mathcal{V} \to \bar{\mathcal{X}}$, such that

$$Ed_{y,1}(X, G_{y,1}(Y, W_1)) \leq \Delta_{y,1} \tag{33}$$

$$Ed_{z,1}(X, G_{z,1}(Z, W_1, W_2, V)) \leq \Delta_{z,1} \tag{34}$$

$$Ed_{y,2}(X, G_{y,2}(Y, W_1, W_3, V)) \leq \Delta_{y,2} \tag{35}$$

$$Ed_{z,2}(X, G_{z,2}(Z, W_1, W_2, W_3, W_4, V)) \leq \Delta_{z,2} \tag{36}$$

3. The alphabets $\{\mathcal{W}_k\}_{k=1}^4$ and $\mathcal{V}$ satisfy:

$$|\mathcal{W}_1| \leq |\mathcal{X}| + 5, \tag{37}$$

$$|\mathcal{V}| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| + 4, \tag{38}$$

$$|\mathcal{W}_2| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{V}| + 3, \tag{39}$$

$$|\mathcal{W}_3| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{W}_2| \cdot |\mathcal{V}| + 2, \tag{40}$$

$$|\mathcal{W}_4| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{W}_2| \cdot |\mathcal{W}_3| \cdot |\mathcal{V}| + 1. \tag{41}$$

4. The rates R_1 and R_2 satisfy

$$R_1 \geq I(X; W_1|Y) + I(X; W_2, V|W_1, Z) \tag{42}$$

$$R_2 \quad \geq \quad I(X; W_1, W_3, V|Y) + I(X; W_2, W_4|W_1, W_3, V, Z) \tag{43}$$

The outer bound to the rate-distortion region is summarized in the following Theorem:

Theorem 3. *For any DMS P_{XYZ} s.t. $X \div Z \div Y$, and a quadruplet of distortions $\mathbf{D} = \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^{2}$, $\mathcal{R}(\mathbf{D})_{nc} \subseteq \mathcal{R}^{**}(\mathbf{D})_{nc}$.*

The proof of this result follows the lines of the converse proof of Theorem 1 in [4] and it is provided in Section 6. Consider now to the case where the distortion requirements are $\Delta_{y,1} = \infty$ and $\Delta_{z,2} = \Delta_{z,1}$, i.e., the case where at the first stage only Z-decoder is required to reconstruct the source and at the second stage the Y-decoder is required to reconstruct the source while Z-decoder is not required to improve its source reconstruction any further. Define the degraded region $\mathcal{R}(\boldsymbol{D})_{nc}$ of all rates and distortions matching $\Delta_{y,1} = \infty$ and $\Delta_{z,2} = \Delta_{z,1}$ by $\mathcal{R}(\Delta_{y,2}, \Delta_{z,1})$. This special instance of our problem has been studied in [14]. The outer bound obtained in [14] is the following: Define the region $\mathcal{R}_{out}(\Delta_1, \Delta_2)$ to be the set of all rate pairs (R_1, R_2) for which there exist random variables (W_1, W_2) in finite alphabets $\mathcal{W}_1$, $\mathcal{W}_2$ s.t. the following conditions are satisfied:

1) $(W_1, W_2) \div X \div Z \div Y$.

2) There exist deterministic maps $G_1 : \mathcal{Z} \times \mathcal{W}_1 \to \tilde{X}$ and $G_2 : \mathcal{Y} \times \mathcal{W}_2 \to \hat{X}$ s.t. $Ed_{z,1}(X, f_1(Z, W_1)) \leq \Delta_1$ and $Ed_{y,2}(X, f_2(Y, W_2)) \leq \Delta_2$.

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

4) The non-negative rate vectors satisfy:
$R_1 \geq I(X; W_1|Z)$, $R_1 + R_2 \geq I(X; W_2|Y) + I(X; W_1|Z, W_2)$.

Theorem 4. *[14] For any discrete memoryless stochastic source with SIs under the Markov condition $X \div Z \div Y$, $\mathcal{R}(\Delta_1, \Delta_2) \subseteq \mathcal{R}_{out}$.*

Note that this outer bound is straightforwardly obtainable from the outer bound of this paper by taking $W_1 = const.$, $V = const.$, $W_4 = const.$ and renaming the pair (W_2, W_3) to be (W_1, W_2) as well as setting $(\Delta_{y,1}, \Delta_{z,1}, \Delta_{y,2}, \Delta_{z,2})$ to be equal $(\infty, \Delta_1, \Delta_2, \Delta_1)$, respectively, and also disregarding $(G_{y,1}, G_{z,2})$ while renaming $(G_{z,1}, G_{y,2})$ to be (G_1, G_2).

4.2.2 Inner Bound

Let a distortion quadruplet $\boldsymbol{D} \stackrel{\triangle}{=} \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^{2}$ be given. Define $\mathcal{R}^*(\boldsymbol{D})_{nc}$ to be the set of all rate pairs (R_1, R_2) for which there exist RVs $\{W_i\}_{i=1}^{4}$ and V, taking values in finite alphabets, $\{\mathcal{W}_i\}_{i=1}^{4}$ and $\mathcal{V}$, respectively, s.t. the following conditions are satisfied:

1. The following Markov conditions hold:

$$(W_1, W_2, W_3, W_4, V) \div X \div Z \div Y \tag{44}$$

$$W_2 \div (X, W_1, V) \div W_3 \tag{45}$$

2. There exist deterministic decoding functions $G_{y,1} : \mathcal{Y} \times \mathcal{W}_1 \to \hat{\mathcal{X}}$, $G_{z,1} : \mathcal{Z} \times \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{V} \to \tilde{\mathcal{X}}$, $G_{y,2} : \mathcal{Y} \times \mathcal{W}_1 \times \mathcal{W}_3 \times \mathcal{V} \to \check{\mathcal{X}}$, $G_{z,2} : \mathcal{Z} \times \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{W}_3 \times \mathcal{W}_4 \times \mathcal{V} \to \bar{\mathcal{X}}$ such that

$$Ed_{y,1}(X, G_{y,1}(Y, W_1)) \leq \Delta_{y,1} \tag{46}$$

$$Ed_{z,1}(X, G_{z,1}(Z, W_1, W_2, V)) \leq \Delta_{z,1} \tag{47}$$

$$Ed_{y,2}(X, G_{y,2}(Y, W_1, W_3, V)) \leq \Delta_{y,2} \tag{48}$$

$$Ed_{z,2}(X, G_{z,2}(Z, W_1, W_2, W_3, W_4, V)) \leq \Delta_{z,2} \tag{49}$$

3. The alphabets $\{\mathcal{W}_k\}_{k=1}^{4}$ and $\mathcal{V}$ satisfy:

$$|\mathcal{W}_1| \leq |\mathcal{X}| + 6, \tag{50}$$

$$|\mathcal{V}| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| + 5, \tag{51}$$

$$|\mathcal{W}_2| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{V}| + 4, \tag{52}$$

$$|\mathcal{W}_3| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{W}_2| \cdot |\mathcal{V}| + 3, \tag{53}$$

$$|\mathcal{W}_4| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| \cdot |\mathcal{W}_2| \cdot |\mathcal{W}_3| \cdot |\mathcal{V}| + 2. \tag{54}$$

4. The rates R_1 and R_2 satisfy

$$R_1 \geq I(X; W_1|Y) + I(X; W_2, V|W_1, Z) \tag{55}$$

$$\begin{aligned} R_2 \quad &\geq \quad I(X; W_1, V, W_3|Y) + I(X; W_2|W_1, V, Z) \\ &+ \quad I(X; W_4|W_1, W_2, W_3, V, Z) \end{aligned} \tag{56}$$

The inner bound to the rate-distortion region is summarized in the following Theorem:

Theorem 5. *For any DMS P_{XYZ} s.t. $X \div Z \div Y$, and a quadruplet of distortions* $\mathbf{D} = \{\Delta_{y,k}, \Delta_{z,k}\}_{k=1}^{2}$, $\mathcal{R}^*(\mathbf{D})_{nc} \subseteq \mathcal{R}(\mathbf{D})_{nc}$.

The inner bound provided in this section demonstrates tradeoffs between various schemes which are based on the notion of (strong or weak) typicality. Recall that in the achievability schemes of successive refinement treated in [4] and [6] the generation of the auxiliary codebooks is sequential: First, the codebook used at the first stage is generated; then, for each codeword of that codebook another codebook conditional on the codeword is generated, and so on. Every generation of a codebook is conditioned on codewords of previously generated codebooks. The encoder chooses the auxiliary codewords in a sequential manner, first finding a good codeword in the first codebook; then in the second codebook (which was generated conditioned on that good codeword), it finds another good codeword, and so on. The encoder proceeds until it has found all codewords needed to describe the source at the desired accuracy at all stages of successive refinement. The decoding process at each stage is also performed in a sequential manner, i.e., first, the codeword in the first codebook is found. Then, in a second codebook (matching that codeword), a second codeword is found and so on.

When multi-group successive refinement is considered, it is unclear if the auxiliary codebooks achieving rate-distortion bounds should be generated "sequentially" (in the sense described above) or "in parallel", with two or more codebooks generated unconditioned on one another. The achievability scheme of this paper demonstrates a semi-parallel approach where some of the codebooks are generated sequentially and some in parallel. We proceed with discussing the meaning of degraded SI at the decoders and then we briefly describe the idea standing behind the achievability scheme.

When referring to degraded SI, the term usually used is that the stronger Z-decoder (that has access to SI of higher quality) can do whatever the weaker Y-decoder can do [4]-[7], i.e., the Z-decoder can find all the codewords that were addressed to Y-decoder. To understand this property, consider the following scenario: Assume that one performs Wyner-Ziv (W-Z) coding [8] for a pair (X, Y) where X is known at the encoder and Y is known at the Y-decoder. Now, assume that the source generating the (X, Y) pair is, in fact, a ternary source, generating a triplet (X, Y, Z), $X \div Z \div Y$ and that Z is known at Z-decoder. Finally, assume that the W-Z coding for Y-decoder is performed with a codebook of auxiliary codewords generated independently of each other and each symbol of which is

generated according to P_{W_1}, s.t. $W_1 \div X \div Z \div Y$. Obviously, the long Markov chain also satisfies the shorter Markov chain $W_1 \div X \div Y$ required by W-Z scheme for coding for the Y-decoder. But, due to the Markov chain $W_1 \div X \div Z \div Y$, $I(Z;W_1) \geq I(Y;W_1)$, and thus, Z-decoder is able to find the correct codeword in the bin of size $2^{NI(Y;W_1)}$ generated for the Y-decoder. The question is the following - given that Z-decoder can always find codewords addressed to the Y-decoder, how we can exploit this property rigorously?

We interpret the degradedness of SI as follows: bins associated with a code designed for the Z-decoder are divided into bins associated with a code designed for the Y-decoder. Specifically, a codebook of about $2^{NI(X;W_1)}$ codewords is partitioned twice - first into "large" bins of about $2^{NI(Z;W_1)}$ codewords matching W-Z code for the Z-decoder, and each of these bins is further partitioned into smalle bins of about $2^{NI(Y;W_1)}$ codewords each.

In W-Z coding designed for communication with Y-decoder only, the indexes of the smaller bins are directly transmitted to the Y-decoder. Note that alternatively, one can first send to Y-decoder an index of the larger bin and then "refine" it with the "internal" index of the matching small bin. This observation immediately leads to the following conclusion: if a single codeword is simultaneously good for communication with both decoders (in the sense of satisfying the reconstruction requirements), the encoder can communicate with both decoders in a two-stage successive manner, by first transmitting the index of a large bin (that contains a good codeword) to both decoders (the index is fully usable only by the Z-decoder), and then, in a separate additional transition, sending the matching "internal" index which is crucial for communication with the Y-decoder (and does not provide new information to Z-decoder). The obvious question that arises is what happens when a single codeword is not sufficient for communication with two decoders and more codebooks must be created. Firstly, under certain Markov conditions, the principle of such an hierarchical (or nested) binning can be applied as well to conditional W-Z codes. Specifically, when the Markov condition $(W_1, W_2..., W_i) - X - Z - Y$ holds, we obtain that $I(Z;W_i|W_1,...,W_{i-1}) \geq I(Y;W_i|W_1,...,W_{i-1})$. Secondly, the real problem arises when not all codewords sent to the Z-decoder must be revealed to the Y-decoder in the next step, and in this case, sequential/hierarchical codesbooks generation is no longer obviously optimal.

The coding scheme is based on the following concept: At the first stage, three codebooks are generated, essentially, according to the hierarchical Wyner-Ziv coding scheme. First a codebook C_{w_1} of $\sim 2^{NI(X;W_1)}$ codewords is generated according to $P_{W_1}^N$, and is partitioned

into bins of size of $\sim 2^{NI(Y;W_1)}$. Thus, there are $\sim 2^{N[I(X;W_1)-I(Y;W_1)]}$ such bins. Due to the Markov chain $W_1 \div X \div Y$, $I(X;W_1) - I(Y;W_1) = I(X;W_1|Y)$. Next, for each $\boldsymbol{w}_1 \in C_{w_1}$, a codebook $C_v(\boldsymbol{w}_1)$ of $\sim 2^{NI(X;V|W_1)}$ codewords is generated according to $P^N_{V|W_1}$ and is partitioned into bins of size of $\sim 2^{NI(Z;V|W_1)}$, and each of these bins is partitioned into smaller bins of size $\sim 2^{NI(Z;V|W_1)}$ each. Thus, there are $\sim 2^{N[I(X;V|W_1)-I(Z;V|W_1)]}$ large bins and $\sim 2^{N[I(Z;V|W_1)-I(Y;V|W_1)]}$ small bins within each large bin. Due to the Markov chain (44), the number of bins: $\sim 2^{NI(X;V|W_1,Z)}$ large bins and $\sim 2^{NI(Z;V|W_1,Y)}$ small bins within each large one. Finally, a codebook $C_{w_2}(\boldsymbol{w}_1, \boldsymbol{v})$ of $\sim 2^{I(X;W_2|W_1,V)}$ codewords is generated for each $\boldsymbol{w}_1 \in C_{w_1}$ and $\boldsymbol{v} \in C_v(\boldsymbol{w}_1)$ according to $P^N_{W2|W1,V}$, and is partitioned into bins of size $\sim 2^{NI(Z;W_2|W_1,V)}$, so by (44), there are $\sim 2^{NI(X;W_2|W_1,V,Z)}$ such bins.

At the second stage, another two codebooks are generated - for each $\boldsymbol{w}_1$, $\boldsymbol{v}(\boldsymbol{w}_1)$ and $\boldsymbol{w}_2(\boldsymbol{w}_1, \boldsymbol{v})$, in codebook $C_{w_3}(\boldsymbol{w}_1, \boldsymbol{v})$, the codewords are generated according to $P^N_{W_3|W_1,V}$, and in codebook $C_{w_4}(\boldsymbol{w}_1, \boldsymbol{v}, \boldsymbol{w}_2, \boldsymbol{w}_3)$, the codewords $\boldsymbol{W}_4$ are generated according to $P^N_{W_4|W_1,V,W_2,W_3}$. These codebooks are also partitioned into bins, specifically, $C_{w_4}(\cdot)$ is partitioned into $\sim 2^{NI(X;W_3|W_1,V,Y)}$ bins of size $\sim 2^{NI(Y;W_3|W_1,V)}$ each. Similarly, C_5 is partitioned into $\sim 2^{NI(X;W_4|W_1,V,W_2,W_3,Z)}$ bins, each of them of size $\sim 2^{NI(Z;W_4|W_1,V,W_2,W_3)}$. The key feature of this scheme is in fact that the $C_{w_3}(\cdot)$ does not take into consideration statistics of $\boldsymbol{W}_2$. Since its codewords must yet be used by the Z-decoder, rising typicality considerations during the encoding/decoding process, the additional Markov condition (45) is imposed on the achievability scheme.

If the encoder succeeds to find good codewords in all five codebooks (the details appear in the formal proof of Theorem 5), the rate of the first transmission, R_1, is composed of three indexes of the bins that contain good codewords in the codebooks C_{w_1}, $C_v(\cdot)$ and $C_{w_2}(\cdot)$, where for $C_v(\cdot)$ only the index of the large bin is used. In this manner, similarly to the classical W-Z coding, only the codeword of C_{w_1} serves the Y-decoder, while all three codewords are decoded by Z-decoder. Hence, $R1 \simeq I(X;W_1|Y) + I(X;V|W_1,Z) + I(X;W_2|W_1,V,Z)$ as is given by eq. (55). At the second transmission, the encoder first refines the description of the bin of codebook $C_v(\cdot)$, and then transmits the indexes of the chosen bins in codebooks $C_{w_3}(\cdot)$ and $C_{w_4}(\cdot)$. Thus, the codewords in codes C_{w_1}, $C_v(\cdot)$ and $C_{w_3}(\cdot)$ serve the reconstruction in the Y-decoder and all five codewords are retrieved correctly by Z-decoder and are used for reconstruction of the source. The incremental rate at the second stage is $I(Z;V|W_1,Y) + I(X;W_3|W_1,V,Y) + I(X;W_4|W_1,V,W_2,W_3)$ and

therefore, the cumulative rate at the second stage is as given by eq. (56).

The scheme that leads to the inner bound is interesting due to the following: The codebook generation is not fully sequential, but some of the codebooks are generated in parallel and are independent (unconditioned) of each other. Unfortunately, with this approach the rate expressions of the inner and outer bounds obtained at the second stage are not identical and the bounds differ in additional Markov conditions imposed on the auxiliary RVs of the direct scheme. Yet, for the case of lossless reconstruction at either the first or the second stage, i.e., $\Delta_{z,1} = 0$ or $\Delta_{y,2} = 0$, respectively, the achievability scheme achieves communication rates suggested by the outer bound and thus closes the gap between the inner and the outer bounds.

4.2.3 Special Cases

We now confine our attention to a number of special cases in which the gap between the outer bound and the inner bound vanishes. First, we consider the case of distortion requirements $\Delta_{z,1} \geq \Delta_{y,1}$ or $\Delta_{y,2} = \Delta_{y,1}$, that is, SR with respect to only one of the decoder at either the first or the second stages, respectively. We then consider the case of distortion requirements $\Delta_{z,1} = 0$ or $\Delta_{y,2} = 0$, that is, lossless reconstruction at Z-decoder at the first stage, or at the Y-decoder at the second stage, respectively. For these cases, the achievability scheme achieves the boundary curve of the outer bound.

Successive Refinement

When $\Delta_{z,1} \geq \Delta_{y,1}$ or $\Delta_{y,2} = \Delta_{y,1,}$, the multi-decoder SR problem degenerates to the problem of refinement of information with respect to only one decoder at either the first or the second stage, respectively. The requirement $\Delta_{z,1} \geq \Delta_{y,1}$ fits the scenario where the Z-decoder performs reconstruction of the source on the basis of the same transmission that served the Y-decoder, so the average distortion it achieves is at least as small as that of the Y-decoder. The requirement $\Delta_{y,2} = \Delta_{y,1,}$ fits the scenario where the Y-decoder is not required to refine its reconstruction at the second stage. For these cases, the inner and the outer bounds coincide, as is summarized in the following theorem:

Theorem 6. *If $\Delta_{z,1} \geq \Delta_{y,1}$ or $\Delta_{y,2} = \Delta_{y,1,}$, then $\mathcal{R}(\mathbf{D})_{nc} = \mathcal{R}^{**}(\mathbf{D})_{nc} = \mathcal{R}^{*}(\mathbf{D})_{nc}$. Specifically, when $\Delta_{z,1} \geq \Delta_{y,1}$, $\mathcal{R}(\mathbf{D})_{nc}$ is given as in the Subsection 4.2.1 with the rate*

inequalities replaced by

$$R_1 \geq I(X;W_1|Y) \quad \textit{and} \quad R_2 \geq I(X;W_1,W_3|Y) + I(X;W_4|W_1,W_3,Z), \tag{57}$$

for the auxiliary RV's satisfying $(W_1,W_3,W_4) \div X \div Z \div Y$.
When $\Delta_{y,2} = \Delta_{y,1}$, $\mathcal{R}(\mathbf{D})_{nc}$ *is given as in the Subsection 4.2.1 with the rate inequalities replaced by*

$$R_1 \geq I(X;W_1|Y) + I(X;W_2|W_1,Z) \quad \textit{and} \quad R_2 \geq I(X;W_1|Y) + I(X;W_2,W_4|W_1,Z). \tag{58}$$

for the auxiliary RV's satisfying $(W_1,W_2,W_4) \div X \div Z \div Y$.

The proof of the achievability part of Theorem 6 can easily be done by setting $W_2 = V = \text{const.}$ for $\Delta_{z,1} \geq \Delta_{y,1}$ and setting $W_3 = V = \text{const.}$ for the requirement $\Delta_{y,2} = \Delta_{y,1}$ in $\mathcal{R}^*(\boldsymbol{D})_{nc}$. The converse proof follows by considering a three-stage communication scheme in the converse proof of [6] and combining two of its stages into a single stage for each of the above cases. For the case $\Delta_{z,1} \geq \Delta_{y,1}$, the first stage of Theorem 6 is essentially the first stage of [6], with the transmission addressed to the Y-decoder. The second stage of Theorem 6 is a combination of the second and the third stages in [6], where at the second stage of [6], SR is performed with respect to the Y-decoder, and at the third stage of [6], SR is performed with respect to the Z-decoder. For the case $\Delta_{y,2} = \Delta_{y,1}$, the the first stage of Theorem 6 matches cumulative rates of two stages of [6], there the first stage consists of transmission of the Y-decoder and the second stage performs SR with respect to the Z-decoder. The second stage of Theorem 6 consists of the third stage of [6] with SR performed (again) with respect to the Z-decoder.

Lossless Reconstruction

Consider the case of lossless reconstruction at either the Z-decoder at the first stage or the Y-decoder at the second stage. Similarly as in [14], it turns out that in these cases, the inner and outer bounds coincide. This observation is summarized in the following theorem:

Theorem 7. *If* $\Delta_{y,2} = 0$ *or* $\Delta_{z,1} = 0$, *then* $\mathcal{R}(\mathbf{D})_{nc} = \mathcal{R}^{**}(\mathbf{D})_{nc} = \mathcal{R}^{*}(\mathbf{D})_{nc}$. *Specifically, when* $\Delta_{z,1} = 0$, $\mathcal{R}(\mathbf{D})_{nc}$ *is given as in the Subsection 4.2.1 with the rate inequalities replaced by*

$$R_1 \geq I(X;W_1|Y) + H(X|W_1,Z) \quad \textit{and} \quad R_2 \geq I(X;W_1,W_3|Y) + H(X|W_1,W_3,Z), \tag{59}$$

for the auxiliary RVs satisfying $(W_1, W_3) \div X \div Z \div Y$.

When $\Delta_{y,2} = 0$, $\mathcal{R}(\mathbf{D})_{nc}$ *is given as in the Subsection 4.2.1 with the rate inequalities replaced by*

$$R_1 \geq I(X; W_1|Y) + I(X; W_2|W_1, Z) \quad \textit{and} \quad R_2 \geq H(X|Y), \tag{60}$$

for the auxiliary RVs satisfying $(W_1, W_2) \div X \div Z \div Y$.

The proof of the achievability part of Theorem 7 can easily be done by setting $W_4 =$ const. and $W_2 = X$ and $V = W_3$ for the requirement $\Delta_{z,1} = 0$ and setting $V = W_2$ and $W_3 = X$ for the requirement $\Delta_{y,2} = 0$ in the inner bound $\mathcal{R}^*(\boldsymbol{D})_{nc}$. The converse proof follows by applying the Heegard-Berger rate-bounds [9] at both stages with the corresponding demand of lossless reconstruction at either the first or the second stage. When the outer bound is considered for each of the stages independently, it degenerates to the Heegard-Berger bound and thus an intersection of the Heegard-Berger bounds for the two stages provides a trivial outer bound to the outer bound obtained in this paper. Since the direct scheme achieves the communication rates suggested by the intersection, the bounds coincide.

The key property of these special cases is the fact that not all auxiliary RV's that determine both inner and outer bounds are active simultaneously. Specifically, V, which stands for the information transmitted to the Z-decoder at the first stage and then repeated for the Y-decoder at the second stage, takes very specific values. The requirement $\Delta_{z,1} = 0$ means perfect reconstruction of the source performed by the Z-decoder at the first stage. For this case, the Z-decoder obviously needs the full information about the source, in the spirit of Slepian-Wolf [16] lossless coding. Therefore, the optimal scheme presents the information sent to Z-decoder at the first stage as if consisting of two (mutually dependent) parts - information V which is then revealed (refined) to the Y-decoder at the second stage ($W_3 = V$) and the information needed by the Z-decoder, i.e., $W_2 = X$. For the requirement $\Delta_{y,2} = 0$, it is expected that the Y-decoder will receive at the second stage all the information about the source, also in the spirit of [16]. As some of this information is already revealed to the Z-decoder at the first stage, all this information is refined to Y-decoder at the second stage ($W_2 = V$) and then all remaining information is transmitted to Y-decoder directly ($W_3 = X$). Interestingly, the cases considered in Theorem 7 are characterized by the same property: in both cases, the second stage transmission serves only the weaker Y-decoder. In the case $\Delta_{y,2} = 0$, it is obvious that $\Delta_{z,2} = 0$ can be

achieved as well. In the case $\Delta_{z,1} = 0$, it is trivially obtained that $\Delta_{z,2} = 0$ as well and thus, only the Y-decoder benefits from the second stage transmission.

5 Proofs for the Causal Case

5.1 Proof of the Converse Part of Theorem 2

The pure source-coding problem is a special case of the joint source-channel problem. We provide a proof of the converse part of Theorem 2, which includes the converse of Theorem 1 as a special case.

Let $(f_1, g_{y,1}, g_{z,1}, f_2, g_{y,2}, g_{z,2})$ be given encoder and decoder functions for which the distortion constraints are satisfied at both stages. In the proof, for the first and the second steps of the communication protocol, we examine the mutual information $I(\boldsymbol{X}; \boldsymbol{B})$ and $I(\boldsymbol{X}; \bar{\boldsymbol{B}})$, respectively.

Firstly, for the case of causal state information at the encoder, we obtain

$$\begin{aligned}
I(\boldsymbol{X}; \boldsymbol{B}) &= \sum_{i=1}^{n} I(\boldsymbol{X}; B_i | B^{i-1}) \\
&= \sum_{i=1}^{n} [I(\boldsymbol{X}, B^{i-1}; B_i) - I(B^{i-1}; B_i)] \\
&\leq \sum_{i=1}^{n} I(\boldsymbol{X}, B^{i-1}, S_{i+1}^{n}; B_i) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(U_{1,i}; B_i) \\
&\overset{(b)}{=} nI(U_{1,T}; B_T | T) \\
&\overset{(c)}{=} nI(U_{1,T}; B | T) \\
&\leq nI(U_{1,T}, T; B) \\
&\overset{(d)}{=} nI(U_1; B) \\
&\overset{(e)}{\leq} nC_1, \qquad (61)
\end{aligned}$$

where (a) follows by denoting $U_{1,i} \overset{\triangle}{=} (\boldsymbol{X}, B^{i-1}, S_{i+1}^{n})$ for $i \in \{1, 2, ..., n\}$ (note that $U_{1,i}$ and S_i are independent); (b) - by defining a time-sharing auxiliary random variable T, distributed uniformly over $\{1, 2, ..., n\}$ independently of all other random variables in the system and noting that $\sum_{i=1}^{n} I(U_{1,i}, B_i) = n \sum_{i=1}^{n} \frac{1}{n} I(U_{1,i}, B_i) = nI(U_{1,T}, B_T|T)$; (c) - by noting that $B = B_T$ since the DMC is stationary; (d) - by denoting random variable $U_1 \overset{\triangle}{=} (U_{1,T}, T)$;

and finally, (e) - by the standard channel coding theorem with causal state information at the encoder [17] since $U_1 \div (A,S) \div B$.

For non-causal availability of state information at the encoder, note that the above defined RV's $\{U_{1,i}\}$ are, in fact, the same RV's as these used by Gelfand and Pinsker in [18] with $\boldsymbol{X}$ substituting the message V of [18]. In fact, with $\boldsymbol{X}$ substituting the message V, the converse proof of [18] is straightforwardly applicable to our case as all the conditions of the proof still hold. Therefore, we can as well upper-bound $I(\boldsymbol{X};\boldsymbol{B})$ by

$$I(\boldsymbol{X};\boldsymbol{B}) \overset{(a)}{=} n[I(U_1;B) - I(U_1;S)] \overset{(b)}{\leq} nC_1, \tag{62}$$

where (a) and (b) follow by [18] and C_1 stands for the Gel'fand-Pinsker channel capacity.

On the other hand,

$$\begin{aligned}
I(\boldsymbol{X};\boldsymbol{B}) &= H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{B}) && (63)\\
&= \sum_{i=1}^{N}[H(X_i|X_1^{i-1}) - H(X_i|X_1^{i-1},\boldsymbol{B})] && (64)\\
&\overset{(a)}{=} \sum_{i=1}^{N}[H(X_i) - && (65)\\
&- H(X_i|X_1^{i-1},Y_1^{i-1},Z_1^{i-1},\boldsymbol{B})]\\
&= \sum_{i=1}^{N} I(X_i;X_1^{i-1},Y_1^{i-1},Z_1^{i-1},\boldsymbol{B}) && (66)\\
&\overset{(b)}{=} \sum_{i=1}^{N} I(X_i;W_{1,i}) && (67)\\
&\overset{(c)}{=} NI(X_{\tilde{T}};W_{1,\tilde{T}}|\tilde{T}) && (68)\\
&\overset{(d)}{=} NI(X;W_{1,\tilde{T}}|\tilde{T}) && (69)\\
&= N[I(X;W_{1,\tilde{T}},\tilde{T}) - I(X;\tilde{T})] && (70)\\
&\overset{(e)}{=} NI(X;W_{1,\tilde{T}},\tilde{T}) && (71)\\
&\overset{(f)}{=} NI(X;W_1), && (72)
\end{aligned}$$

where (a) follows from the fact that the source is memoryless and from the Markov chain $X_i \div (X_1^{i-1},\boldsymbol{B}) \div (Y_1^{i-1},Z_1^{i-1})$; (b) - by denoting $W_{1,i} \overset{\triangle}{=} (X_1^{i-1},Y_1^{i-1},Z_1^{i-1},\boldsymbol{B})$; (c) - by defining a time-sharing auxiliary random variable $\tilde{T}$, distributed uniformly over $\{1,2,...,N\}$ independently of all other random variables in the system; (d) - by noting that $X = X_{\tilde{T}}$ since the DMS is stationary; (e) - is again due to the fact that the source is stationary and thus $I(X;\tilde{T}) = 0$; and finally, (f) - by denoting random variable $W_1 \overset{\triangle}{=} (W_{1,\tilde{T}},\tilde{T})$.

Thus, for the first stage, we obtain $NI(X;W_1) \leq nC_1$ and by dividing both sides of the inequality by N we end up with $I(X;W_1) \leq \rho_1 C_1$, where C_1 denotes the channel capacity for the case of causal or non-causal state availability at the encoder. I.e, condition (30) of Theorem 2 is satisfied.

As for the second stage, by similar considerations as in (61) and (62), we obtain that $I(\boldsymbol{X};\bar{\boldsymbol{B}}) \leq mC_2$, where C_2 stands for the channel capacity of the second channel with state information available at the encoder (again, either causally or non-causally). Also,

$$
\begin{aligned}
I(\boldsymbol{X};\bar{\boldsymbol{B}}) &= H(\bar{\boldsymbol{B}}) - H(\bar{\boldsymbol{B}}|\boldsymbol{X}) && (73)\\
&\overset{(a)}{\geq} H(\bar{\boldsymbol{B}}|\boldsymbol{B}) - H(\bar{\boldsymbol{B}}|\boldsymbol{X}) + I(\boldsymbol{B};\bar{\boldsymbol{B}}|\boldsymbol{X}) && (74)\\
&= H(\bar{\boldsymbol{B}}|\boldsymbol{B}) - H(\bar{\boldsymbol{B}}|\boldsymbol{X},\boldsymbol{B}) && (75)\\
&= I(\boldsymbol{X};\bar{\boldsymbol{B}}|\boldsymbol{B}) && (76)\\
&= \sum_{i=1}^{N} I(X_i;\bar{\boldsymbol{B}}|X_1^{i-1},\boldsymbol{B}) && (77)\\
&= \sum_{i=1}^{N}[H(X_i|X_1^{i-1},\boldsymbol{B}) && (78)\\
&- H(X_i|X_1^{i-1},\boldsymbol{B},\bar{\boldsymbol{B}})] && (79)\\
&\overset{(b)}{=} \sum_{i=1}^{N}[H(X_i|X_1^{i-1},\boldsymbol{B},Y_1^{i-1},Z_1^{i-1}) && (80)\\
&- H(X_i|X_1^{i-1},\boldsymbol{B},\bar{\boldsymbol{B}},Y_1^{i-1},Z_1^{i-1})] && (81)\\
&= \sum_{i=1}^{N} I(X_i;\bar{\boldsymbol{B}}|X_1^{i-1},\boldsymbol{B},Y_1^{i-1},Z_1^{i-1}) && (82)\\
&\overset{(c)}{=} \sum_{i=1}^{N} I(X_i;W_{2,i}|W_{1,i}) && (83)\\
&\overset{(d)}{=} NI(X;W_{2,\tilde{T}}|W_{1,\tilde{T}},\tilde{T}) && (84)\\
&\overset{(e)}{=} NI(X;W_2|W_1) && (85)
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and independence of the channels described by the following Markov chain $(\boldsymbol{B},\boldsymbol{S}) \div \boldsymbol{X} \div (\bar{\boldsymbol{B}},\bar{\boldsymbol{S}})$; (b) from the Markov chains $X_i \div (X_1^{i-1},\boldsymbol{B}) \div Y_1^{i-1}Z_1^{i-1}$ and $X_i \div (X_1^{i-1},\boldsymbol{B},\bar{\boldsymbol{B}}) \div Y_1^{i-1}Z_1^{i-1}$; (c) come from using the above-defined auxiliary random variables $\{W_{1,i}\}_{i=1}^{N}$ and denoting $W_{2,i} \overset{\triangle}{=} \bar{\boldsymbol{B}}$;(d) comes from using the above-defined random variables $\tilde{T}$ as well as stationarity of the source, and finally, (e) comes from using the above defined random variable W_1 and letting $W_2 \overset{\triangle}{=} W_{2,\tilde{T}}$. We obtain, hence, that $NI(X;W_2|W_1) \leq mC_2$ and division of both sides of the inequality

by N results in $I(X;W_2|W_1) \leq \rho_2 C_2$, which is exactly the condition (31) of Theorem 2.

Also, note that the Markov structure $(W_{1,i}, W_{2,i}) \div X_i \div (Y_i, Z_i)$ holds for every $i = 1, ..., N$. Due to this structure and the fact that the source P_{XYZ} is stationary and memoryless, the Markov chain $(W_1, W_2) \div X \div (Y, Z)$ also holds, and thus, the condition given by (22) is satisfied.

We next show that there exist functions $G_{y,1}$, $G_{z,1}$, $G_{y,2}$ and $G_{z,2}$ that satisfy (23) - (26), respectively. Denote by $g_{y,k,i}$ and $g_{z,k,i}$ the output of the decoders Y and Z, respectively, at stages $k = 1, 2$ and times $i = 1, ..., N$. The random variable W_1 contains $(X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1}, \boldsymbol{B})$ and W_2 contains $\bar{\boldsymbol{B}}$. Choose the functions $G_{y,1}$, $G_{y,2}$, $G_{z,1}$ and $G_{z,1}$ as follows:

$$G_{y,1,\tilde{T}}(Y, W_1) = g_{y,1,\tilde{T}}(Y_1^{\tilde{T}}, \boldsymbol{B}), \tag{86}$$

$$G_{z,1,\tilde{T}}(Z, W_1) = g_{z,1,T}(Z_1^{\tilde{T}}, \boldsymbol{B}), \tag{87}$$

$$G_{y,2,\tilde{T}}(Y, W_1, W_2) = g_{y,2,\tilde{T}}(Y_1^{\tilde{T}}, \boldsymbol{B}, \bar{\boldsymbol{B}}), \tag{88}$$

$$G_{z,2,\tilde{T}}(Z, W_1, W_2) = g_{z,2,\tilde{T}}(Z_1^{\tilde{T}}, \boldsymbol{B}, \bar{\boldsymbol{B}}). \tag{89}$$

We then have for the average distortions[5]

$$Ed(X, G_{y,1}(Y, W_1)) = \frac{1}{N} \sum_{i=1}^{N} Ed(X, g_{y,1,i}(Y_1^i, \boldsymbol{B})) \leq \Delta_{y,1}, \tag{90}$$

$$Ed(X, G_{z,1}(Z, W_1)) = \frac{1}{N} \sum_{i=1}^{N} Ed(X, g_{z,1,i}(Z_1^i, \boldsymbol{B})) \leq \Delta_{z,1}, \tag{91}$$

$$Ed(X, G_{y,2}(Y, W_1, W_2)) = \frac{1}{N} \sum_{i=1}^{N} Ed(X, g_{y,2,i}(Y_1^i), \boldsymbol{B}, \bar{\boldsymbol{B}}) \leq \Delta_{y,2} \tag{92}$$

and

$$Ed(X, G_{z,2}(Z, W_1, W_2)) = \frac{1}{N} \sum_{i=1}^{N} Ed(X, g_{z,2,i}(Z_1^i), \boldsymbol{B}, \bar{\boldsymbol{B}}) \leq \Delta_{z,2}, \tag{93}$$

[5]The definitions in (86)-(89) determine the outputs of the decoders functions at "stochastic" time $\tilde{T}$. For example, the output of the Y-decoder at the first stage at time $\tilde{T}$ is governed by the first $\tilde{T}$ symbols of the source SI, i.e., $Y_1^{\tilde{T}}$, and the channel output $\boldsymbol{B}$.

i.e., the distortion constraints are satisfied.

In order to complete the proof, it is left to show that the cardinality of the alphabets of auxiliary RVs W_1 and W_2 is limited. We use the support lemma [19], which is based on Carathéodory's theorem, according to which, given J real valued continuous functionals q_j, $j = 1, ..., J$ on the set $\mathcal{P}(\mathcal{X})$ of probability distributions over the alphabets $\mathcal{X}$, and given any probability measure μ on the Borel σ-algebra of $\mathcal{P}(\mathcal{X})$, there exist J elements $Q_1, ...Q_J$ of $\mathcal{P}(\mathcal{X})$ and J non-negative reals, $\alpha_1, ..., \alpha_J$, such that $\sum_{j=1}^{J} \alpha_j = 1$ and for every $j = 1, ..., J$

$$\int_{\mathcal{P}(\mathcal{X})} q_j(Q)\mu(dQ) = \sum_{i=1}^{J} \alpha_i q_j(Q_i). \tag{94}$$

Before we actually apply the support lemma, we first rewrite the relevant conditional mutual informations and the distortion functions in a more convenient form for the use of this lemma, by taking advantage of the Markov structures. We begin with $I(X; W_1)$:

$$I(X; W_1) = H(X) - H(X|W_1), \tag{95}$$

and in the same manner, $I(X; W_2|W_1)$ becomes

$$I(X; W_2|W_1) = H(X|W_1) - H(X|W_1, W_2). \tag{96}$$

For a given joint distribution of (X, Y, Z), $H(X)$ is given and unaffected by W_1 and W_2. Therefore, in order to preserve prescribed values of $I(X; W_1)$ and $I(X; W_2|W_1)$, it is sufficient to preserve the associated values of $H(X|W_1)$ and $H(X|W_1, W_2)$.

We first invoke the support lemma in order to reduce the alphabet size of W_1, while preserving the values of $H(X|W_1)$ and $H(X|W_1, W_2)$, as well as the distortions in both decoders at both stages of communication. The alphabet of W_2 is still kept intact at this step. Define the following functionals of a generic distribution Q over $\mathcal{X} \times \mathcal{W}_2$, where $\mathcal{X}$ is assumed, without loss of generality, to be $\{1, 2, ..., \alpha\}$, $\alpha \stackrel{\triangle}{=} |\mathcal{X}|$:

$$q_i(Q) = \sum_{w_2} Q(x, w_2), \ i \stackrel{\triangle}{=} x = 1, 2, ..., \alpha - 1, \tag{97}$$

$$q_\alpha(Q) \ = \ -\sum_{x, w_2} Q(x, w_2) \log \sum_{w_2} Q(x, w_2), \tag{98}$$

and

$$q_{\alpha+1}(Q) \ = \ -\sum_{x, w_2} Q(x, w_2) \log Q(x|w_2). \tag{99}$$

Also, we define

$$q_{\alpha+2}(Q) = \sum_y \min_{\hat{x}} \sum_{x,w_2} Q(x,w_2)P(y|x)d_{y,1}(x,\hat{x}), \tag{100}$$

$$q_{\alpha+3}(Q) = \sum_z \min_{\tilde{x}} \sum_{x,w_2} Q(x,w_2)P(z|x)d_{z,1}(x,\tilde{x}), \tag{101}$$

$$q_{\alpha+4}(Q) = \sum_y \min_{\check{x}} \sum_{x,w_2} Q(x,w_2)P(y|x)d_{y,2}(x,\check{x}) \tag{102}$$

and

$$q_{\alpha+5}(Q) = \sum_z \min_{\bar{x}} \sum_{x,w_2} Q(x,w_2)P(z|x)d_{z,2}(x,\bar{x}), \tag{103}$$

which along with (98) and (99) help us to preserve the rate and distortion constraints. Applying now the support lemma for the above defined functionals, we find that there exists a random variable W_1 (jointly distributed with (X,Y,Z,W_2), whose alphabet size is $|W_1| = |\mathcal{X}| + 5$ and it satisfies simultaneously:

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_i(P(\cdot|w_1)) = P_X(x),\ i = 1,2,...,\alpha-1, \tag{104}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_\alpha(P(\cdot|w_1)) = H(X|W_1), \tag{105}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{\alpha+1}(P(\cdot|w_1)) = H(X|W_1,W_2), \tag{106}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{\alpha+2}(P(\cdot|w_1)) = \min_{G_{y,1}} Ed(X, G_{y,1}(Y,W_1)), \tag{107}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{\alpha+3}(P(\cdot|w_1)) = \min_{G_{z,1}} Ed(X, G_{z,1}(Z,W_1)), \tag{108}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{\alpha+4}(P(\cdot|w_1)) = \min_{G_{y,2}} Ed(X, G_{y,2}(Y,W_1,W_2)) \tag{109}$$

and

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{\alpha+5}(P(\cdot|w_1)) = \min_{G_{z,2}} Ed(X, G_{z,2}(Z,W_1,W_2)). \tag{110}$$

Having found a random variable W_1, we now proceed to reduce the alphabet of W_2 in a similar manner, where this time, we have $\beta = |\mathcal{X}| \cdot |\mathcal{W}_1| - 1$ constraints to preserve the joint distribution of (X, W_1), just defined, and 3 more constraints to preserve the second-stage rate and distortions. Applying the support lemma, we obtain that W_2 satisfies all the desired rate-distortion constraints and the necessary alphabet size of W_2 is upper-bounded by

$$|\mathcal{W}_2| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| + 2. \tag{111}$$

This completes the proof of the converse part of Theorem 2.

5.2 Proof of the Direct Part of Theorem 1

Let W_1, W_2, $G_{y,1}$, $G_{y,2}$, $G_{z,1}$ and $G_{z,2}$ be some elements in the definition of $\mathcal{R}^*(\boldsymbol{D})_c$ that achieve a given point in that region. We next describe the mechanisms of random code selection and the encoding and decoding operations.

Code Generation:
Let $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $\delta > 0$ be arbitrary small and select $R_1 \geq I(X; W_1) + \epsilon_1 + \delta$ and $\Delta_R \stackrel{\triangle}{=} R_2 - R_1$, $\Delta_R \geq I(X; W_2|W_1) + \epsilon_2 + \delta$. For the first stage, 2^{NR_1}, sequences of length N, $\{\boldsymbol{W}_1(k)\}$, $k \in [1, ..., 2^{NR_1}]$, are drawn independently from $T^{\delta}_{P_{W_1}}$. Let us denote the set of these sequences by $\mathcal{C}_1$. For each codeword $\boldsymbol{W}_1(k) = \boldsymbol{w}_1$, a set of $2^{N\Delta_R}$ second-stage codewords $\{\boldsymbol{W}_2(k, j)\}$, $j \in [1, ..., 2^{NR_2}]$, are independently drawn from $T^{\delta}_{P_{W_2|W_1}}(\boldsymbol{w}_1)$. We denote this set by $\mathcal{C}_2(k)$ and its elements by $\{\boldsymbol{W}_2(k, j)\}$. Note that the 2^{NR_1} sets $\{\mathcal{C}_2(\cdot)\}$ may not be all mutually exclusive.

Encoding:
Upon receiving a source sequence $\boldsymbol{x}$, the encoder acts as follows:

1. If $\boldsymbol{x} \in T^{\delta}_{P_X}$ and the codebook $\mathcal{C}_1$ contains a sequence $\boldsymbol{W}_1(k) = \boldsymbol{w}_1$ s.t. the pair $(\boldsymbol{x}, \boldsymbol{w}_1) \in T^{2\delta}_{P_{XW_1}}$, the fist such index k is chosen for transmission at the first stage. Next, if the codebook $\mathcal{C}_2(k)$ contains a sequence $\boldsymbol{W}_2(k, j) = \boldsymbol{w}_2$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1VW_2}}$, the first such index j is chosen for transmission at the second stage.
2. If $\boldsymbol{x} \notin T^{\delta}_{P_X}$, or $\nexists\, \boldsymbol{W}_1(k) = \boldsymbol{w}_1$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1) \in T^{2\delta}_{P_{XW_1}}$, or $\nexists\, \boldsymbol{W}_2(k, j) = \boldsymbol{w}_2$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1VW_2}}$, an arbitrary error message is transmitted at both stages.

Decoding:

The decoders of the first stage retrieves the first-stage codeword according to its index and generates the reproduction by $\hat{X}_i = G_{y,1}\left(Y_i, W_{1,i}(k)\right)$ and $\tilde{X}_i = G_{z,1}\left(Z_i, W_{1,i}(k)\right)$, $i \in [1, 2, ..., N]$. Similarly, the decoders of the second stage retrieve both the first-stage and the second-stage codewords and creates the reconstruction of the source according to $\check{X}_i = G_{y,2}\left(Y_i, W_{1,i}\left(k\right), W_{2,i}\left(k, j\right)\right)$ and $\bar{X}_i = G_{z,2}\left(Z_i, W_{1,i}\left(k\right), W_{2,i}\left(k, j\right)\right)$, $i \in [1, 2, ..., N]$.

We now turn to the analysis of the error probability and the distortions. For each $\boldsymbol{x}$ and a particular choice of codes $\mathcal{C}_1$ and $\{\mathcal{C}_2(\cdot)\}$, the possible causes for error message are:

1. $\boldsymbol{x} \notin T^{\delta}_{P_X}$. Let the probability of this event be defined as P_{e_1}.

2. $\boldsymbol{x} \in T^{\delta}_{P_X}$, but in the codebook $\mathcal{C}_1$ $\nexists \boldsymbol{w}_1$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1) \in T^{2\delta}_{P_{XW_1}}$. Let the probability of this event be defined as P_{e_2}.

3. $\boldsymbol{x} \in T^{\delta}_{P_X}$, and the codebook $\mathcal{C}_1$ contains $\boldsymbol{w}_1$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1) \in T^{2\delta}_{P_{XW_1}}$, but $\nexists \boldsymbol{w}_2 \in \mathcal{C}_2(\boldsymbol{w}_1)$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1VW_2}}$. Let the probability of this event be defined as P_{e_3}.

Note that if none of those events occur, then, for the sufficiently large N, by the Markov Lemma [12, pp. 436, Lemma 14.8.1] applied twice, the following is satisfied at both stages: with high probability $(\boldsymbol{X}, \boldsymbol{Z}, \hat{\mathbf{X}}) \in T^{5\delta|\mathcal{W}_1 \times \mathcal{W}_2|}_{P_{XZ\hat{X}}}$ and $(\boldsymbol{X}, \boldsymbol{Y}, \tilde{\mathbf{X}}) \in T^{5\delta|\mathcal{W}_1 \times \mathcal{W}_2|}_{P_{XY\tilde{X}}}$. In particular, the first application of the Markov Lemma occurs due to the Markov chain $(W_1, W_2) \div X \div (Y, Z)$: Note that by the way of creation, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are jointly typical with high probability and also, with high probability, $\boldsymbol{X}$, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are jointly typical. Therefore, by the Markov Lemma, ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$) are also jointly typical with high probability. Also, note that due to the fact that the source is memoryless and by the way of creation of the reconstructions, the following Markov chains hold: $\boldsymbol{X} \div (\boldsymbol{Y}, \boldsymbol{W}_1) \div \hat{\mathbf{X}}$ and $\boldsymbol{X} \div (\boldsymbol{Z}, \boldsymbol{W}_1) \div \tilde{\mathbf{X}}$, and also, at the second stage, $\boldsymbol{X} \div (\boldsymbol{Y}, \boldsymbol{W}_1, \boldsymbol{W}_2) \div \check{\mathbf{X}}$ and $\boldsymbol{X} \div (\boldsymbol{Z}, \boldsymbol{W}_1, \boldsymbol{W}_2) \div \bar{\mathbf{X}}$. By the second application of the Markov Lemma, we obtain that with high probability $\boldsymbol{X}$ is jointly typical with $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ at the first stage and with $\check{\mathbf{X}}$ and $\bar{\mathbf{X}}$ at the refinement stage. The probability that one or more of the above typicality relations do not hold vanishes as N becomes infinitely large. The joint typicality of $(\boldsymbol{X}, \hat{\mathbf{X}})$, $(\boldsymbol{X}, \tilde{\mathbf{X}})$, $(\boldsymbol{X}, \check{\mathbf{X}})$ and $(\boldsymbol{X}, \bar{\mathbf{X}})$ imposes that the distortion constraints (23)-(26) are satisfied when N is large enough.

It remains to show that the probability of sending an error message, P_e, vanishes when

N is large enough. P_e is bounded by

$$P_e \leq P_{e_1} + P_{e_2} + P_{e_3}. \tag{112}$$

The fact that $P_{e_1} \to 0$ follows from the properties of typical sequences [12]. As for P_{e_2}, we have:

$$P_{e_2} \stackrel{\triangle}{=} \prod_{k=1}^{2^{NR_1}} \Pr\left\{(\boldsymbol{x},\, \boldsymbol{W}_1(k)) \notin T^{2\delta}_{P_{XW_1}}\right\}. \tag{113}$$

Now, for every k:

$$\begin{aligned} \Pr\left\{(\boldsymbol{x},\, \boldsymbol{W}_1(k)) \notin T^{2\delta}_{P_{XW_1}}\right\} &= 1 - \Pr\left\{(\boldsymbol{x},\, \boldsymbol{W}_1(k)) \in T^{2\delta}_{P_{XW_1}}\right\} \\ &= 1 - \frac{|T^{2\delta}_{P_{XW_1}}|}{|T^{\delta}_{P_{W_1}}||T^{\delta}_{P_X}|} \\ &\leq 1 - 2^{-N[I(X;W_1)+\epsilon_1]}, \end{aligned} \tag{114}$$

where the last equation follows from the size of typical sequences as are given in [12]. Substitution of (159) into (158) and application of the well-known inequality $(1-v)^N \leq \exp(-vN)$, provides us with the following upper-bound for $N \to \infty$:

$$P_{e_2} \leq \left[1 - 2^{-N[I(X;W_1)+\epsilon_1]}\right]^{2^{nR_1}} \leq \exp\left\{-2^{NR_1} \cdot 2^{-N[I(X;W_1)+\epsilon_1]}\right\} \to 0, \tag{115}$$

double-exponentially rapidly since $R_1 \geq I(X;W_1) + \epsilon_1 + \delta$.

To estimate P_{e_3}, we repeat the technique of the previous step:

$$P_{e_3} \stackrel{\triangle}{=} \prod_{j=1}^{2^{NR_2}} \Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_1,\, \boldsymbol{W}_2(\boldsymbol{w}_1, j)) \notin T^{3\delta}_{P_{XW_1VW_2}}\right\}. \tag{116}$$

Again, by the property of the typical sequences, for every j:

$$\Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_1,\, \boldsymbol{W}_2(\boldsymbol{w}_1, j)) \notin T^{3\delta}_{P_{XW_1VW_2}}\right\} \leq 1 - 2^{-N[I(X;W_2|W_1)+\epsilon_2]}, \tag{117}$$

and therefore, substitution of (162) into (161) gives

$$P_{e_3} \leq \left[1 - 2^{-N[I(X;W_2|W_1)+\epsilon_2]}\right]^{2^{NR_2}} \leq \exp\left\{-2^{NR_2} \cdot 2^{-N[I(X;W_2|W_1)+\epsilon_2]}\right\} \to 0, \tag{118}$$

double-exponentially rapidly since $R_2 \geq I(X;W_2|W_1) + \epsilon_2 + \delta$.

Since $P_{e_i} \to 0$ for i = 1,2,3, their sum tends to zero as well, implying that there exist at least one choice of a codebook $\mathcal{C}_1$ and related choices of sets $\{\mathcal{C}_2(\cdot)\}$, that give rise to the reliable source reconstruction at both stages with communication rates R_1 and $\Delta_R = R_2 - R_1$.

6 Proofs for the Non-Causal Case

6.1 Outer Bound

The proof of the outer bound follows the lines of the proof of Theorem 1 in [4]. Assume that we have an $(n, M_1, M_2, \{\Delta_{y,k}, (\Delta_{z,k}\}_{k=1}^2))$ SR code for the source X with SI (Y, Z), as in Definition 1. We will show the existence of a quintuplet (W_1, W_2, W_3, W_4, V) that satisfies the conditions 1–4 in the definition of $\mathcal{R}^{**}(\boldsymbol{D})_{nc}$. First, note that

$$\begin{aligned} NR_1 &\geq H(f_1) \geq I(\boldsymbol{X}; f_1|\,\boldsymbol{Y}) = I(\boldsymbol{X}; f_1, \boldsymbol{Z}|\,\boldsymbol{Y}) - I(\boldsymbol{X}; \boldsymbol{Z}|f_1,\, \boldsymbol{Y}) \\ &= \sum_{i=1}^{n} \left[I(X_i; f_1, \boldsymbol{Z}|X^{i-1},\, \boldsymbol{Y}) - I(\boldsymbol{X}; Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) \right]. \end{aligned} \tag{119}$$

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_i, Y_i) and $(X^{i-1}, Y^{N\setminus i})$ are independent, we have, for the first term in the summand of (119):

$$\begin{aligned} I(X_i; f_1, \boldsymbol{Z}|X^{i-1},\, \boldsymbol{Y}) &= H(X_i|Y_i, X^{i-1}, Y^{N\setminus i}) - H(X_i|Y_i, X^{i-1}, Y^{N\setminus i}, f_1, \boldsymbol{Z}) \\ &= H(X_i|Y_i) - H(X_i|Y_i, X^{i-1}Y^{N\setminus i}, f_1, \boldsymbol{Z}) \\ &= I(X_i; X^{i-1}, Y^{N\setminus i}, f_1, \boldsymbol{Z}|Y_i). \end{aligned} \tag{120}$$

Next, due to the Markov structure

$$Z_i \div (X_i, Y_i) \div (X^{N\setminus i}, f_1, Z^{i-1}, Y^{N\setminus i}) \tag{121}$$

we have, for the second term in the summand of (119):

$$\begin{aligned} I(\boldsymbol{X}; Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) &= H(Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) - H(Z_i|\boldsymbol{X}, f_1,\, \boldsymbol{Y}, Z^{i-1}) \\ &= H(Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) - H(Z_i|X_i, f_1,\, \boldsymbol{Y}, Z^{i-1}) \\ &= I(X_i; Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}). \end{aligned} \tag{122}$$

Substituting (120) and (122) in (119), we obtain

$$\begin{aligned} NR_1 &\geq \sum_{i=1}^{N} \Big[I(X_i; X^{i-1}, Y^{N\setminus i}, f_1, \boldsymbol{Z}|Y_i) - I(X_i; Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) \Big] \\ &= \sum_{i=1}^{N} \Big[I(X_i; Y^{N\setminus i}, f_1, Z^{i-1}|Y_i) + I(X_i; X^{i-1}, Z_i^N|Y_i, f_1, Y^{N\setminus i}, Z^{i-1}) - I(X_i; Z_i|f_1,\, \boldsymbol{Y}, Z^{i-1}) \Big] \\ &= \sum_{i=1}^{n} \Big[I(X_i; f_1, Y^{N\setminus i}, Z^{i-1}|Y_i) + I(X_i; X^{i-1}, Z_{i+1}^N|Y_i, Z_i, f_1, Y^{N\setminus i}, Z^{i-1}) \Big]. \end{aligned} \tag{123}$$

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

$$Y_i \div Z_i \div (X_i, f_1, Y^{N\setminus i}, Z^{i-1}), \tag{124}$$

and we have for the second term in (123)

$$\begin{aligned}
&I(X_i; X^{i-1}, Z_{i+1}^N | f_1, \boldsymbol{Y}, Z^i) \\
&= H(X_i|f_1, \boldsymbol{Y}, Z^i) - H(X_i|f_1, \boldsymbol{Y}, \boldsymbol{Z}, X^{i-1}) \\
&= H(X_i, Y_i|f_1, Y^{N\setminus i}, Z^i) - H(Y_i|f_1, Y^{N\setminus i}, Z^i) - H(X_i|f_1, \boldsymbol{Y}, \boldsymbol{Z}, X^{i-1}) \\
&= H(Y_i|X_i, f_1, Y^{N\setminus i}, Z^i) + H(X_i|f_1, Y^{N\setminus i}, Z^i) - H(Y_i|f_1, Y^{N\setminus i}, Z^i) - H(X_i|f_1, \boldsymbol{Y}, \boldsymbol{Z}, X^{i-1}) \\
&\overset{(a)}{=} H(X_i|f_1, Y^{N\setminus i}, Z^i) - H(X_i|f_1, \boldsymbol{Y}, \boldsymbol{Z}, X^{i-1}) \\
&= I(X_i; Y_i, Z_{i+1}^N, X^{i-1}|f_1, Y^{N\setminus i}, Z^i) \\
&\overset{(b)}{=} I(X_i; Z_{i+1}^N, X^{i-1}|f_1, Y^{N\setminus i}, Z^i)
\end{aligned} \tag{125}$$

where in (a) was used the Markov chain $X_i \div (f_1, Y^{N\setminus i}, Z^i) \div Y_i$. To justify (b), note that f_1 is a function of $\boldsymbol{X}$ and due to this feature, the fact that the source is a DMS and the Markov condition $X \div Z \div Y$, we obtain that $X_i \div (f_1, Y^{N\setminus i}, \boldsymbol{Z}, X^{i-1}) \div Y_i$.

Substituting (125) in (123), we get

$$\begin{aligned}
NR_1 &\geq \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}, Z^{i-1}|Y_i) + I(X_i; Z_{i+1}^N, X^{i-1}|f_1, Y^{N\setminus i}, Z^i) \Big] && (126)\\
&\geq \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}, Z^{i-1}|Y_i) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] && (127)\\
&\overset{(a)}{=} \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(X_i; Z^{i-1}|f_1, \boldsymbol{Y}) + I(Z_i; Z^{i-1}|f_1, \boldsymbol{Y}, X_i) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] \\
&= \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(X_i, Z_i; Z^{i-1}|f_1, \boldsymbol{Y}) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] \\
&= \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(Z_i; Z^{i-1}|f_1, \boldsymbol{Y}) + I(X_i; Z^{i-1}|f_1, \boldsymbol{Y}, Z_i) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] \\
&\geq \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(X_i; Z^{i-1}|f_1, \boldsymbol{Y}, Z_i) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] \\
&\overset{(b)}{=} \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(X_i; Z^{i-1}|f_1, Y^{N\setminus i}, Z_i) + I(X_i; Z_{i+1}^N|f_1, Y^{N\setminus i}, Z^i) \Big] \\
&= \sum_{i=1}^N \Big[I(X_i; f_1, Y^{N\setminus i}|Y_i) + I(X_i; Z^{N\setminus i}|f_1, Y^{N\setminus i}, Z_i) \Big], && (128)
\end{aligned}$$

where (a) is due to the Markov relation $Z_i \div (f_1, \boldsymbol{Y}, X_i) \div Z^{i-1}$ and (b) is due to the Markov chain $Y_i \div Z_i \div (f_1, Y^{N\setminus i}, Z^{i-1}, X_i)$ that implies $Y_i \div (f_1, Y^{N\setminus i}, Z_i) \div Z^{i-1}$ and $Y_i \div (f_1, Y^{N\setminus i}, Z_i, X_i) \div Z^{i-1}$.

Before defining the auxiliary random variables, we bound R_2 from below. We do that by repeating the steps (119)-(126) of lower-bounding R_1 with a pair (f_1, f_2) substituting f_1 in each step:

$$\begin{aligned} NR_2 &\geq H(f_1, f_2) \geq H(\boldsymbol{X}; f_1, f_2 | \boldsymbol{Y}) \geq I(\boldsymbol{X}; f_1, f_2, \boldsymbol{Z} | \boldsymbol{Y}) - I(\boldsymbol{X}; \boldsymbol{Z} | f_1, f_2, \boldsymbol{Y}) \\ &\geq \sum_{i=1}^{N} \left[I(X_i; f_1, f_2, Y^{N\setminus i}, Z^{i-1} | Y_i) + I(X_i; Z_{i+1}^N, X^{i-1} | f_1, f_2, Y^{N\setminus i}, Z^i) \right] \quad (129) \end{aligned}$$

Define the random variables $W_{1,i} = (f_1, Y^{N\setminus i})$, $V_i = Z^{i-1}$, $W_{2,i} = Z_{i+1}^N$, $W_{3,i} = f_2$ and $W_{4,i} = X^{i-1}$. With these definitions [6], we have the Markov structure

$$(W_{1,i}, W_{2,i}, W_{3,i}, W_{4,i}, V_i) \div X_i \div Z_i \div Y_i \quad (130)$$

and the bounds (128) and (129) become

$$R_1 \geq \frac{1}{N} \sum_{i=1}^{N} [I(X_i; W_{1,i} | Y_i) + I(X_i; W_{2,i}, V_i | W_{1,i}, Z_i)] \quad (131)$$

$$R_2 \geq \frac{1}{N} \sum_{i=1}^{N} [I(X_i; W_{1,i}, V_i, W_{3,i} | Y_i) + I(X_i; W_{2,i}, W_{4,i} | W_{1,i}, W_{3,i}, V_i, Z_i)] . \quad (132)$$

Let J be a random variable, independent of X, Y, and Z, and uniformly distributed over the set $\{1, 2, \ldots, N\}$. Define the random variables $W_1 = (J, W_{1,J})$, $V = (J, V_J)$, $W_2 = (J, W_{2,J})$, $W_3 = (J, W_{3,J})$ and $W_4 = (J, W_{4,J})$. The Markov relations (130) still hold, that is

$$(W_1, W_2, W_3, W_4, V) \div X \div Z \div Y, \quad (133)$$

and therefore the condition 1 in the definition of $\mathcal{R}^{**}(\boldsymbol{D})_{nc}$ is satisfied.

We proceed to show the existence of functions $G_{y,1}$, $G_{z,1}$, $G_{y,2}$ and $G_{z,2}$ satisfying the second condition. Denote by $g_{y,k,l}$ and $g_{z,k,l}$ the output of the Y and Z decoders, respectively, at iteration k and time l, $k = 1, 2$, $1 \leq l \leq N$. The random variable W_1 contains $f_1 Y^{N\setminus J}$. At the same time, the triplet (W_1, V, W_2) contains $f_1 Z^{N\setminus J}$ and so on. Therefore, let us

[6] Note that different choices of auxiliary RVs are possible. For example, one may choose: $W_{1,i} = f_1, Y^{N\setminus i}$, $V_i = (W_{1,i}, Z^{i-1})$, $W_{2,i} = (V_i, Z_{i+1}^N)$, $W_{3,i} = (V_i, f_2)$, $W_{4,i} = (W_{2,i}, W_{3,i}, X^{i-1})$. This choice would result in the following Markov chain: $W_{1,i} \div V_i \div (W_{2,i}, W_{3,i}) \div W_{4,i} \div X_i \div Z_i \div Y_i$.

choose the functions $G_{y,1}$, $G_{z,1}$, $G_{y,2}$ and $G_{z,2}$ as follows

$$G_{y,1,J}(Y, W_1) = g_{y,1,J}(\boldsymbol{Y}, f_1) \quad (134)$$

$$G_{z,1,J}(Z, W_1, W_2, V) = g_{z,1,J}(\boldsymbol{Z}, f_1). \quad (135)$$

$$G_{y,2,J}(Y, W_1, W_3, V) = g_{y,2,J}(\boldsymbol{Y}, f_1, f_2) \quad (136)$$

$$G_{z,2,J}(Z, W_1, W_2, W_3, W_4, V) = g_{z,2,J}(\boldsymbol{Z}, f_1, f_2). \quad (137)$$

Then, for the distortions we have

$$\mathbb{E}d_{y,1}(X, G_{y,1}(Y, W_1)) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}d_{y,1}(X, g_{y,1,j}(\boldsymbol{Y}, f_1)) \leq \Delta_{y,1} \quad (138)$$

$$\mathbb{E}d_{z,1}(X, G_{z,1}(Z, W_1, W_2, V)) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}d_{z,1}(X, g_{z,1,j}(\boldsymbol{Z}, f_1)) \leq \Delta_{z,1} \quad (139)$$

$$\mathbb{E}d_{y,2}(X, G_{y,2}(Y, W_1, W_3, V)) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}d_{y,2}(X, g_{y,2,j}(\boldsymbol{Y}, f_1, f_2)) \leq \Delta_{y,2} \quad (140)$$

$$\mathbb{E}d_{z,2}(X, G_{z,2}(Z, W_1, W_2, W_3, W_4, V)) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}d_{z,2}(X, g_{z,2,j}(\boldsymbol{Z}, f_1, f_2)) \leq \Delta_{z,2} \quad (141)$$

Hence, condition 2 in the definition of $\mathcal{R}^{**}(\boldsymbol{D})_{nc}$ is satisfied.

To prove that condition 4 of that definition holds, we have to show that the bounds (42) and (43) can be written in a single letter form with W_1, W_2, W_3 and W_4. The following chain of equalities holds

$$\begin{aligned}
I(X; W_1|Y) &= H(W_1|Y) - H(W_1|X, Y) \\
&= H(W_1|Y) - H(W_1|X) \\
&= I(W_1; X) - I(W_1; Y) \\
&= H(X) - H(X|W_1) - H(Y) + H(Y|W_1) \\
&= H(X) - H(X|J, W_{1,J}) - H(Y) + H(Y|J, W_{1,J}) \\
&= \frac{1}{N}\sum_{i=1}^{N} H(X_i) - \frac{1}{N}\sum_{i=1}^{N} H(X_i|W_{1,i}) - \frac{1}{N}\sum_{i=1}^{N} H(Y_i) + \frac{1}{N}\sum_{i=1}^{N} H(Y_i|W_{1,i}) \\
&= \frac{1}{N}\sum_{i=1}^{N} I(X_i; W_{1,i}|Y_i) \quad (142)
\end{aligned}$$

where the last equality is due to (130). In a similar manner, we get

$$\begin{aligned} I(X;W_2,V|W_1,Z) &= I(X;J,W_{2,J},J,V_J|J,W_{1,J},Z) = I(X;W_{2,J},V_J|J,W_{1,J},Z) \\ &= H(X|J,W_{1,J},Z) - H(X|J,W_{1,J},W_{2,J},V_J,Z) \\ &= \frac{1}{N}\sum_{i=1}^{N} H(X_i|i,W_{1,i},Z_i) - \frac{1}{N}\sum_{i=1}^{N} H(X_i|i,W_{1,i},W_{2,i},V_iZ_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} I(X_i;W_{2,i},V_i|W_{1,i},Z_i). \end{aligned} \tag{143}$$

In view of (142), (143), the bound (131) can be written as

$$R_1 \geq I(X;W_1|Y) + I(X;W_2,V|W_1,Z). \tag{144}$$

In a similar manner, we shown that (132) can be written as

$$R_2 \geq I(X;W_1,W_3,V|Y) + I(X;W_2,W_4|W_1,W_3,V,Z). \tag{145}$$

Specifically,

$$\begin{aligned} I(X;W_1,W_3,V|Y) &= H(W_1,W_3,V|Y) - H(W_1,W_3,V|X,Y) \\ &= H(W_1,W_3,V|Y) - H(W_1,W_3,V|X) \\ &= H(W_1,W_3,V) - H(W_1,W_3,V|X) \\ &- (H(W_1,W_3,V) - H(W_1,W_3,V|Y)) \\ &= I(W_1,W_3,V;X) - I(W_1,W_3,V;Y) \\ &= H(X) - H(X|W_1,W_3,V) - H(Y) \\ &+ H(Y|W_1,W_3,V) \\ &= H(X) - H(X|J,W_{1,J},J,W_{3,J},J,V_J) - H(Y) \\ &+ H(Y|J,W_{1,J},J,W_{3,J},J,V_J) \\ &= \frac{1}{N}\sum_{i=1}^{N} H(X_i) - \frac{1}{N}\sum_{i=1}^{N} H(X_i|W_{1,i},W_{3,i},V_i) - \frac{1}{N}\sum_{i=1}^{N} H(Y_i) \\ &+ \frac{1}{N}\sum_{i=1}^{N} H(Y_i|W_{1,i},W_{3,i},V_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} I(X_i;W_{1,i},W_{3,i},V_i|Y_i) \end{aligned} \tag{146}$$

where the last equality is due to (130). In a similar manner, we get

$$\begin{aligned} I(X;W_2,W_4|W_1,W_3,V,Z) &= I(X;J,W_{2,J},J,W_{4,J}|J,W_{1,J},J,W_{3,J},V,Z) \\ &= I(X;W_{2,J},W_{4,J}|J,W_{1,J},W_{3,J},V,Z) \\ &= H(X|J,W_{1,J},W_{3,J},V_J,Z) \\ &- H(X|J,W_{1,J},W_{2,J},W_{3,J},W_{4,J},V_J,Z) \\ &= \frac{1}{N}\sum_{i=1}^{N} H(X_i|i,W_{1,i},W_{3,i},V_i,Z_i) \\ &- \frac{1}{N}\sum_{i=1}^{N} H(X_i|i,W_{1,i},W_{2,i},W_{3,i},W_{4,i},V_i,Z_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} I(X_i;W_{2,i},W_{4,i}|W_{1,i},W_{3,i},V_i,Z_i). \end{aligned} \tag{147}$$

It is left to prove that the cardinality of the auxiliary RVs satisfies the third condition. This step of the proof extends the converse proof of [4] and conceptually is very similar to the above-detailed part of the converse proof of Theorem 2 which is related to reducing cardinality of the alphabets of auxiliary RVs. The detailed proof of this part is thus omitted and to complete the proof of the converse we merely outline it. Here also we use the support lemma [19] and rewrite the relevant conditional mutual informations and the distortion functions in a more convenient form for the use of this lemma. Similarly as in [4], we begin with the first term, $I(X;W_1|Y)$, in the lower bound to R_1, using the Markov chain $W_1 \div X \div Y$:

$$\begin{aligned} I(X;W_1|Y) &= H(W_1|Y) - H(W_1|X,Y) \\ &= H(W_1|Y) - H(W_1|X) \\ &= H(W_1) - I(Y;W_1) - H(W_1) + I(X;W_1) \\ &= H(Y|W_1) - H(Y) - H(X|W_1) + H(X). \end{aligned} \tag{148}$$

Next, we decompose the second term in the lower bound to R_1, $I(X;V,W_2|W_1,Z)$, into $I(X;V|W_1,Z)$ and $I(X;W_2|W_1,V,Z)$, and for $I(X;V|W_1,Z)$ we have due to the Markov chain $(W_1,V) \div X \div Z$:

$$\begin{aligned} I(X;V|W_1,Z) &= H(X|W_1,Z) - H(X|W_1,V,Z) \\ &= H(X|W_1) - I(X;Z|W_1) + I(X;Z|W_1,V) - H(X|W_1,V) \\ &= H(X|W_1) - H(Z|W_1) + H(Z|W_1,X) - H(X|W_1,V) \end{aligned}$$

$$
\begin{aligned}
&+ \quad H(Z|W_1,V) - H(Z|W_1,V,X) \\
&= \quad H(X|W_1) - H(Z|W_1) + H(Z|W_1,V) - H(X|W_1,V). \qquad (149)
\end{aligned}
$$

Using the Markov chain $(W_1,W_2,V) \div X \div Z$ for $I(X;W_2|W_1,V,Z)$, we have:

$$
\begin{aligned}
I(X;W_2|W_1,V,Z) &= H(X|W_1,V) - H(Z|W_1,V) \\
&+ H(Z|W_1,W_2,V) - H(X|W_1,W_2,V). \qquad (150)
\end{aligned}
$$

Similarly, $I(X;W_1,W_3,V|Y)$ can be decomposed into $I(X;W_1|Y)$, $I(X;V|W_1,Y)$ and $I(X;W_3|W_1,V,Y)$, with two later terms, in turn, expressed as

$$
I(X;V|W_1,Y) = H(X|W_1) - H(Y|W_1) + H(Y|W_1,V) - H(X|W_1,V), \qquad (151)
$$

and

$$
\begin{aligned}
I(X;W_3|W_1,V,Y) &= H(X|W_1,V) - H(Y|W_1,V) \\
&+ H(Y|W_1,W_3,V) - H(X|W_1,W_3,V). \qquad (152)
\end{aligned}
$$

The second term in the lower bound to R_2 is $I(X;W_2,W_4|W_1,V,W_3,Z)$ and it can also be decomposed into

$$
\begin{aligned}
I(X;W_2|W_1,W_3,V,Z) &= H(X|W_1,W_3,V) - H(Z|W_1,W_3,V) \\
&+ H(Z|W_1,W_2,W_3,V) - H(X|W_1,W_2,W_3,V). \qquad (153)
\end{aligned}
$$

and

$$
\begin{aligned}
I(X;W_4|W_1,W_2,W_3,V,Z) &= H(X|W_1,W_2,W_3,V) - H(Z|W_1,W_2,W_3,V) \\
&+ H(Z|W_1,W_2,W_3,W_4,V) - H(X|W_1,W_2,W_3,W_4,V). \qquad (154)
\end{aligned}
$$

Thus, the lower bounds to R_1 and R_2 can be expressed as following:

$$
\begin{aligned}
I(X;W_1|Y) + I(X;V,W_2|W_1,Z) &= \big[H(X) - H(Y)\big] + \big[H(Y|W_1) - H(Z|W_1)\big] \\
&+ \big[H(Z|W_1,W_2,V) - H(X|W_1,W_2,V)\big] \qquad (155)
\end{aligned}
$$

and

$$
\begin{aligned}
I(X;W_1,W_3,V|Y) + I(X;W_2,W_4|W_1,W_3,V,Z) &= \big[H(X) - H(Y)\big] \\
&+ \big[H(Y|W_1,W_3,V) - H(Z|W_1,W_3,V)\big] \\
&+ \big[H(Z|W_1,W_2,W_3,W_4,V) \\
&- H(X|W_1,W_2,W_3,W_4,V)\big]. \qquad (156)
\end{aligned}
$$

Since $[H(X) - H(Y)]$ is a constant that depends only on the given statistics of the source and SI Y, in order to preserve prescribed values of the above lower bounds, it is sufficient to preserve the associated values of $[H(Y|W_1) - H(Z|W_1)] + [H(Z|W_1, W_2, V) - H(X|W_1, W_2, V)]$ and $[H(Y|W_1, W_3, V) - H(Z|W_1, W_3, V)] + [H(Z|W_1, W_2, W_3, W_4, V) - H(X|W_1, W_2, W_3, W_4, V)]$.

From here on the proof is essentially similar to the one provided for Theorem 2: The support lemma is first used to reduce the alphabet size of W_1, while preserving the values of (155) and (156) and the distortions at both stages. The alphabets of the remaining auxiliary RVs are kept intact at this stage of the proof. There are $|\mathcal{X}| - 1$ functionals to be defined that help to preserve the source distribution, 2 more to preserve (155) and (156) and 4 more functionals to preserve all the distortions at both stages. Thus, it is easy to show that it is possible to find auxiliary RV W_1 which necessary alphabet size is upper-bounded by $|\mathcal{X}| + 5$. Next, we reduce the alphabet size of V, where now in addition to the values of the lower bounds and distortions $\Delta_{z,1}$, $\Delta_{y,2}$ and $\Delta_{z,2}$, it is desired to preserve the joint distribution (X, W_1). There are $|\mathcal{X}||\mathcal{W}_1| - 1 + 2 + 3$ constraints imposed on V and thus its alphabet size is upper-bounded by $|\mathcal{X}|(|\mathcal{X}| + 5) + 4$. In a similar manner, the reduction of the alphabet cardinality is further performed for W_2, W_3 and W_4 where at each stage, the support lemma is applied in so that the statistics of the source and all already "reduced" RVs are maintained as well as lower bounds to the relevant rates and distortions.

6.2 Inner Bound

6.2.1 Code-book generation

First, randomly generate, according to $P_{W_1}(\cdot)$, a codebook $\mathcal{C}_{w_1}$ of $2^{[N(I(X;W_1)+\epsilon_1+\delta)]}$ independent codewords $\{\boldsymbol{w}_{1,i}\}$ of length N, where the coordinates are also generated i.i.d. Then, partition the codewords into $2^{[N(I(X;W_1|Y)+\epsilon_2+\delta)]}$ bins $(\epsilon_2 > \epsilon_1)$.

Next, for each $\{\boldsymbol{w}_{1,i}\}$, randomly generate a codebook $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ consisting of $2^{[N(I(X;V|W_1)+\epsilon_v+\delta)]}$ codewords $\{\boldsymbol{v}_{i,j}\}$, where the generation of each coordinate is according to $P_{V|W_1}(\cdot)$ and partition this codebook into $2^{[N(I(X;V|W_1,Z)+\epsilon_{v'}+\delta)]}$ bins, $\mathcal{C}_v(\boldsymbol{w}_{1,i})$, $(\epsilon_{v'} > \epsilon_v)$. Each bin in the codebook of $\{\boldsymbol{v}_{i,j}\}$ contains a little less than $2^{[N(I(Z;V|W_1))]}$ codewords. Partition each such bin into sub-bins, $\mathcal{C}_v^b(\boldsymbol{w}_{1,i})$, each of a size of a little less than $2^{[N(I(Y;V|W_1))]}$. There are about $2^{[N(I(Z;V|W_1)-I(Y;V|W_1))]}$ such sub-bins.

For each pair $\{\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}\}$ randomly generate a codebook $\mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ consisting of

$2^{[N(I(X;W_2|W_1,V)+\epsilon_3+\delta)]}$ codewords $\{\boldsymbol{w}_{2,i,j,k}\}$, where the generation of each coordinate is according to $P_{W_2|W_1,V}(\cdot)$ and partition $\mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_j)$ into $2^{[N(I(X;W_2|W_1,V,Z)+\epsilon_4+\delta)]}$ bins ($\epsilon_4 > \epsilon_3$).

Now, randomly generate for each pair $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ a codebook $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ of $2^{[N(I(X;W_3|W_1,V)+\epsilon_5+\delta)]}$ codewords $\{\boldsymbol{w}_{3,i,j,l}\}$ according to $P_{W_3|W_1,V}(\cdot)$ and partition $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_j)$ into $2^{[N(I(X;W_3|W_1,V,Y)+\epsilon_6+\delta)]}$ bins ($\epsilon_6 > \epsilon_5$).

Finally, for each quadruplet $\{\boldsymbol{w}_{1,i}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l}, \boldsymbol{v}_{i,j}\}$, randomly generate a codebook $\mathcal{C}_{w_4}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$ of $2^{[N(I(X;W_4|W_1,W_2,W_3,V)+\epsilon_7+\delta)]}$ codewords $\{\boldsymbol{w}_{4,i,j,k,l,m}\}$ according to $P_{W_4|W_1,W_2,W_3,V}(\cdot)$ and partition it into $2^{[N(I(X;W_4|W_1,W_2,W_3,V,Z)+\epsilon_8+\delta)]}$ bins ($\epsilon_8 > \epsilon_7$).

For clarity of exposition, the generation of codebooks is demonstrated in Fig. 3.

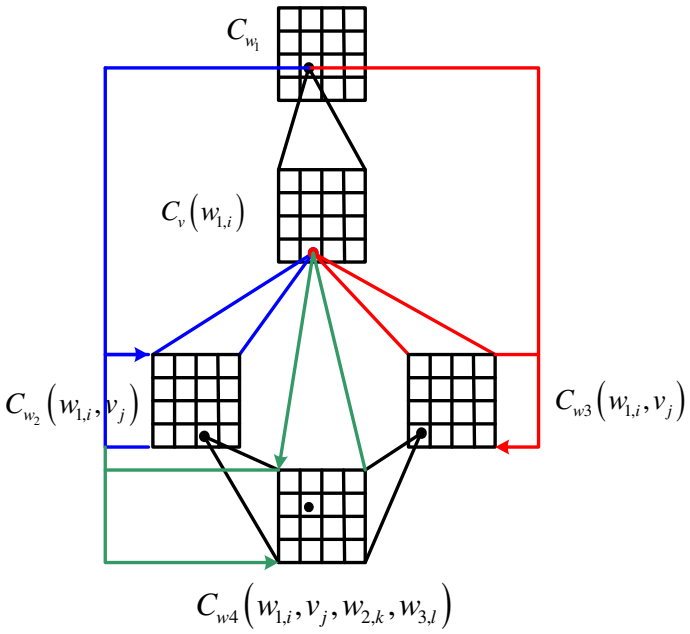

Figure 3: Achievability Scheme - Code Generation.

6.2.2 Encoding

Given a source sequence $\boldsymbol{x}$, the encoder seeks a vector in $\mathcal{C}_{w_1}$ such that $\boldsymbol{x}$ and $\boldsymbol{w}_{1,i}$ are jointly typical. If such $\boldsymbol{w}_{1,i}$ is found, in $\mathcal{C}_v(\boldsymbol{w}_{1,i})$, the encoder seeks a vector $\boldsymbol{v}_{i,j}$ such that the source sequences $\boldsymbol{x}$ and $\boldsymbol{w}_{1,i}$ will be jointly typical with it. The encoder proceeds this way, seeking $\boldsymbol{w}_{2,i,j,k}$ in $\mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ so that $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k})$ are jointly typical. The encoder then seeks in $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ a codeword $\boldsymbol{w}_{3,i,j,l}$ so that $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{3,i,j,l})$ are jointly typical. Due to the Markov chain $W_2 \div (X, W_1, V) \div W_3$, had the encoder managed to find such

sequences, $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$ will be jointly typical with high probability.

If the encoder found jointly typical sequences $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$, it seeks in $\mathcal{C}_{w_4}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$ a sequence $\boldsymbol{w}_{4,i,j,k,l,m}$ which will be jointly typical with all the above-mentioned sequences. If at any stage of its search the encoder fails to find a "good sequence", it declares an error. As is shown in the sequel, the probability of such an event is very low, due to the typicality properties of the scheme. Otherwise, i.e., if all the jointly typical sequences are found, the encoder acts as follows: At the first stage, it conveys to the decoders a single transmission consisting of the following concatenated indexes: the index B_1 of the bin to which $\boldsymbol{w}_{1,i}$ belongs, of length of about $NI(X;W_1|Y)$ bits; the index B_2 of $\mathcal{C}_v(\boldsymbol{w}_{1,i})$, s.t., $\boldsymbol{v}_{i,j} \in \mathcal{C}_v(\boldsymbol{w}_{1,i})$, which can be described by about $NI(X;V|W_1,Z)$ bits and the index B_3 of the bin to which $\boldsymbol{w}_{2,i,j,k}$ belongs, which requires about $NI(X;W_2|W_1,V,Z)$ bits. At the refinement stage, it transmits the index B_4^* of $\mathcal{C}_v^b(\boldsymbol{w}_{1,i})$ to which $\boldsymbol{v}_{i,j}$ belongs *within* $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ (previously described by B_2), which requires about $N[I(Z;V|W_1) - I(Y;V|W_1)]$ bits, concatenated with the indexes B_5 and B_6 of the bins containing $\boldsymbol{w}_{3,i,j,l}$ and $\boldsymbol{w}_{4,i,j,k,l,m}$, in $C_{w3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ and $C_{w4}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$, of about $NI(X;W_3|W_1,V,Y)$ and $NI(X;W_4|W_1,W_2,W_3,V,Z)$ bits, respectively. The transmission rates at both stages are as defined by $\mathcal{R}^*(\boldsymbol{D})_{nc}$ up to $\{\epsilon_i\}$.

6.2.3 Decoding

First stage: The first decoder accesses (B_1, B_2, B_3), but performs W-Z decoding procedure with respect to B_1 only. Specifically, in $\mathcal{C}_{w_1}$, in the bin indexed by B_1, the decoder seeks a unique sequence $\boldsymbol{w}_{1,i}$ that was chosen by the encoder. Due to the Markov chain $W_1 \div X \div Y$, as the block-length becomes infinitely large, the decoder will find with probability tending to 1 the correct sequence $\boldsymbol{w}_{1,i}$. Since in each bin in $\mathcal{C}_{w_1}$ there are less than $2^{NI(Y;W_1)}$ codewords, and these codewords were generated i.i.d, the probability of existing at the bin indexed by B_1 of another codeword jointly typical with $\boldsymbol{Y}$ vanishes as $N \to \infty$.

The second decoder uses three indexes (B_1, B_2, B_3) to retrieve all three codewords chosen by the encoder. Specifically, it retrieves $\boldsymbol{w}_{1,i}$ similarly as Y-decoder does, since, as it has access to a more informative SI, it can do whatever the Y-decoder can do. Afterwards, it retrieves correctly $\boldsymbol{v}_{i,j} \in \mathcal{C}_v(\boldsymbol{w}_{1,i})$ in the bin indexed by B_2, which is possible due to the Markov chain $(V, W_1) \div X \div Z$. The Z-decoder does not find in bin indexed by B_2 other codewords which are jointly typical with $\boldsymbol{z}$ since there are less than $2^{NI(Z;V|W_1)}$ codewords in

that bin. Finally, following similar considerations, after retrieving $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$, the Z-decoder retrieves correctly $\boldsymbol{w}_{2,i,j,k} \in \mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ in the bin indexed by B_3.

Second Stage: Note that after the first transmission Y-decoder is able to find all codewords $\boldsymbol{v}$ which are jointly typical with $\boldsymbol{y}$ in the bin indexed by B_2 in the codebook $\mathcal{C}_v^b(\boldsymbol{w}_{1,i})$. This is due to the Markov chain $(W_1, V) \div X \div Y$. But, it cannot reveal which of these codewords was chosen by the encoder, as there are more than $2^{NI(Y;V|W_1)}$ such codewords (there are a bit less then $2^{NI(Z;V|W_1)}$ such codewords, as is required by the W-Z coding designed for Z-decoder). When the Y-decoder receives the index B_4^* of $\mathcal{C}_v^b(\boldsymbol{w}_{1,i})$, since $\boldsymbol{v}_{i,j} \in \mathcal{C}_v^b(\boldsymbol{w}_{1,i}) \subseteq \mathcal{C}_v(\boldsymbol{w}_{1,i})$, it searches $\boldsymbol{v}_{i,j}$ among a group of codewords of a size less than $2^{NI(Y;V|W_1)}$ codewords, and thus, it is is able to retrieve $\boldsymbol{v}_{i,j}$ correctly by the W-Z decoding argument. After Y-decoder has found $\boldsymbol{v}_{i,j}$, it performs W-Z decoding of the codeword $\boldsymbol{w}_{3,i,j,l} \in \mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ according to the bin-index B_5 and $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$. It now improves the reconstruction of the source sequence with an aid of the triplet $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{3,i,j,l})$, which is possible within the defined distortion due to the typicality properties of the scheme.

The Z-decoder, which after the first step has retrieved correctly (with probability tending to 1, as $N \to \infty$) the sequences $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k})$, makes no use of index B_4^*, as it serves Y-decoder only. The Z-decoder uses its knowledge of $(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ as well as the fact that its SI is more informative to decode correctly $\boldsymbol{w}_{3,i,j,l}$ in the bin of $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ indexed by B_5. Finally, it uses all the codewords it managed to find thus far to perform conditional W-Z decoding and to find the correct codeword $\boldsymbol{w}_{4,i,j,k,l,m}$ according to the index B_6 of a bin in $\mathcal{C}_{w_4}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$.

At each stage, after each of the decoders has found correct codewords, it performs reconstruction of the source sequence $\boldsymbol{x}$. Due to the typicality properties of the scheme, i.e., $X \div (W_1, Y) \div \hat{X}$, $X \div (W_1, W_2, V, Z) \div \tilde{X}$, $X \div (W_1, W_3, V, Y) \div \check{X}$ and $X \div (W_1, W_2, W_3, W_4, V, Z) \div \bar{X}$, the distortion constraints are satisfied at both decoders.

6.2.4 Analysis of Probability of Error

We now turn to the analysis of the error probability. For each $\boldsymbol{x}$ and a particular choice of the code $\mathcal{C}_{w_1}$ and related choices of $(\{\mathcal{C}_v(\cdot), \mathcal{C}_{w_2}(\cdot), \mathcal{C}_{w_3}(\cdot), \mathcal{C}_{w_4}(\cdot)\})$, the possible causes for error message are:

1. $\boldsymbol{x} \notin T_{P_X}^{\delta}$. Let the probability of this event be defined as P_{e_1}.

2. $\boldsymbol{x} \in T^{\delta}_{P_X}$, but in the codebook $\mathcal{C}_{w_1}$ $\nexists \boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$. Let the probability of this event be defined as P_{e_2}.

3. $\boldsymbol{x} \in T^{\delta}_{P_X}$, and the codebook $\mathcal{C}_{w_1}$ contains $\boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$, but $\nexists \boldsymbol{v}_{i,j} \in \mathcal{C}_v(\boldsymbol{w}_{1,i})$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}) \in T^{3\delta}_{P_{XW_1V}}$. Let the probability of this event be defined as P_{e_3}.

4. $\boldsymbol{x} \in T^{\delta}_{P_X}$, the codebook $\mathcal{C}_{w_1}$ contains $\boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$, and also the codebook $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ contains $\boldsymbol{v}_{i,j}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}) \in T^{3\delta}_{P_{XW_1V}}$, but $\nexists \boldsymbol{w}_{2,i,j,k} \in \mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}) \in T^{3\delta}_{P_{XW_1VW_2}}$. Let the probability of this event be defined as P_{e_4}.

5. $\boldsymbol{x} \in T^{\delta}_{P_X}$, the codebook $\mathcal{C}_{w_1}$ contains $\boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$, the codebook $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ contains $\boldsymbol{v}_{i,j}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}) \in T^{3\delta}_{P_{XW_1V}}$, but $\nexists \boldsymbol{w}_{3,i,j,l} \in \mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{3,i,j,l}) \in T^{3\delta}_{P_{XW_1VW_3}}$. Let the probability of this event be defined as P_{e_5}.

6. $\boldsymbol{x} \in T^{\delta}_{P_X}$, the codebook $\mathcal{C}_{w_1}$ contains $\boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$, the codebook $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ contains $\boldsymbol{v}_{i,j}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}) \in T^{3\delta}_{P_{XW_1V}}$, and the codebooks $\mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ and $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ contain $\boldsymbol{w}_{2,i,j,k}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}) \in T^{3\delta}_{P_{XW_1VW_2}}$ and $\boldsymbol{w}_{3,i,j,m}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{3,i,j,m}) \in T^{3\delta}_{P_{XW_1VW_3}}$, respectively, but $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l}) \notin T^{4\delta}_{P_{XW_1VW_2W_3}}$. Let the probability of this event be defined as P_{e_6}.

7. $\boldsymbol{x} \in T^{\delta}_{P_X}$, the codebook $\mathcal{C}_{w_1}$ contains $\boldsymbol{w}_{1,i}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}) \in T^{2\delta}_{P_{XW_1}}$, the codebook $\mathcal{C}_v(\boldsymbol{w}_{1,i})$ contains $\boldsymbol{v}_{i,j}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}) \in T^{3\delta}_{P_{XW_1V}}$, and the codebooks $\mathcal{C}_{w_2}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ and $\mathcal{C}_{w_3}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j})$ contain $\boldsymbol{w}_{2,i,j,k}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}) \in T^{3\delta}_{P_{XW_1VW_2}}$ and $\boldsymbol{w}_{3,i,j,m}$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{3,i,j,m}) \in T^{3\delta}_{P_{XW_1VW_3}}$, respectively, and also $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l}) \in T^{4\delta}_{P_{XW_1VW_2W_3}}$, but $\nexists \boldsymbol{w}_{4,i,j,l,k,m} \in \mathcal{C}_{w_4}(\boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l})$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{w}_{2,i,j,k}, \boldsymbol{w}_{3,i,j,l}, \boldsymbol{w}_{4,i,j,k,l,m}) \in T^{4\delta}_{P_{XW_1VW_2W_3}}$. Let the probability of this event be defined as P_{e_7}.

Note that if none of those events occur, then, for the sufficiently large N, by the Markov Lemma [12, pp. 436, Lemma 14.8.1] applied twice, the following is satisfied: with high probability $(\boldsymbol{X}, \boldsymbol{Y}, \hat{\mathbf{X}})$ are jointly typical and $(\boldsymbol{X}, \boldsymbol{Z}, \tilde{\mathbf{X}})$ are jointly typical at both stages.

1. The first application of the Markov Lemma occurs due to the Markov chain $(Y, Z) \div X \div (W_1, V, W_2, W_3, W_4)$: Note that by the way of creation, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are jointly typical with high probability and also, with high probability, RV's ($\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{W}_3$, $\boldsymbol{W}_4$, $\boldsymbol{V}$)

and $\boldsymbol{X}$ are jointly typical. Therefore, by the Markov Lemma, all the sequences $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{W}_3$, $\boldsymbol{W}_4$ and $\boldsymbol{V}$ are also jointly typical with high probability. And so, SIs are jointly typical with the auxiliary RV's at both stages of communication.

2. Also, note that due to the fact that the source is memoryless and by the way of creation of the reconstructions, the following Markov chains hold at the first stage: $\boldsymbol{X} \div (\boldsymbol{Y}, \boldsymbol{W}_1) \div \hat{\mathbf{X}}$ and $\boldsymbol{X} \div (\boldsymbol{Z}, \boldsymbol{W}_1, \boldsymbol{V}, \boldsymbol{W}_2) \div \tilde{\mathbf{X}}$. Similarly, at the second stage, $\boldsymbol{X} \div (\boldsymbol{Y}, \boldsymbol{W}_1, \boldsymbol{V}, \boldsymbol{W}_3) \div \hat{\mathbf{X}}$ and $\boldsymbol{X} \div (\boldsymbol{Z}, \boldsymbol{W}_1, \boldsymbol{V}, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{W}_4) \div \tilde{\mathbf{X}}$. By the second application of the Markov Lemma, we obtain that with high probability $\boldsymbol{X}$ is jointly typical with $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ at both stages. The probability that one or more of the above typicality relations do not hold vanishes as N becomes infinitely large. The joint typicality of $(\boldsymbol{X}, \hat{\mathbf{X}})$ and $(\boldsymbol{X}, \tilde{\mathbf{X}})$ imposes that the distortion constraints (33)- (36) are satisfied when N is large enough (see [4, Section 6] for explicit derivations).

It remains to show that the probability of sending an error message vanishes when N is large enough. The average probability of error P_e is bounded by

$$P_e \leq P_{e_1} + P_{e_2} + P_{e_3} + P_{e_4} + P_{e_5} + P_{e_6} + P_{e_7}. \tag{157}$$

The fact that $P_{e_1} \to 0$ follows from the properties of typical sequences [12]. As for P_{e_2}, we have:

$$P_{e_2} \stackrel{\triangle}{=} \prod_{k=1}^{|\mathcal{C}_{w_1}|} \Pr\left\{(\boldsymbol{x}, \boldsymbol{W}_{1,k}) \notin T^{2\delta}_{P_{XW_1}}\right\}. \tag{158}$$

Now, for every k:

$$\begin{aligned} \Pr\left\{(\boldsymbol{x}, \boldsymbol{W}_{1,k}) \notin T^{2\delta}_{P_{XW_1}}\right\} &= 1 - \Pr\left\{(\boldsymbol{x}, \boldsymbol{W}_{1,k}) \in T^{2\delta}_{P_{XW_1}}\right\} \\ &= 1 - \frac{|T^{2\delta}_{P_{XW_1}}|}{|T^{\delta}_{P_{W_1}}||T^{\delta}_{P_X}|} \\ &\leq 1 - 2^{-N[I(X;W_1)+\epsilon_1]}, \end{aligned} \tag{159}$$

where the last equation follows from the size of typical sequences as are given in [12]. Substitution of (159) into (158) and application of the well-known inequality $(1-v)^N \leq \exp(-vN)$, provides us with the following upper-bound for $N \to \infty$:

$$P_{e_2} \leq \left[1 - 2^{-N[I(X;W_1)+\epsilon_1]}\right]^{|\mathcal{C}_{w_1}|} \leq \exp\left\{-|\mathcal{C}_{w_1}| \cdot 2^{-N[I(X;W_1)+\epsilon_1]}\right\} \to 0, \tag{160}$$

double-exponentially rapidly since $|\mathcal{C}_{w_1}| = I(X;W_1) + \epsilon_1 + \delta$.

To estimate P_{e_3}, we repeat the technique of the previous step:

$$P_{e_3} \stackrel{\triangle}{=} \prod_{j=1}^{|\mathcal{C}_v|} \Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{V}_{i,j}) \notin T^{3\delta}_{P_{XW_1V}}\right\}. \tag{161}$$

Again, by the property of the typical sequences, for every j:

$$\Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{V}_{i,j}) \notin T^{3\delta}_{P_{XW_1V}}\right\} \leq 1 - 2^{-N[I(X;V|W_1)+\epsilon_2]}, \tag{162}$$

and therefore, substitution of (162) into (161) gives

$$P_{e_3} \leq \left[1 - 2^{-N[I(X;V|W_1)+\epsilon_2]}\right]^{|\mathcal{C}_v|} \leq \exp\left\{-|\mathcal{C}_v| \cdot 2^{-N[I(X;V|W_1)+\epsilon_2]}\right\} \to 0, \tag{163}$$

double-exponentially rapidly since $|\mathcal{C}_v| = I(X;V|W_1) + \epsilon_2 + \delta$.

To estimate P_{e_4}, the technique of the previous step is again repeated:

$$P_{e_4} \stackrel{\triangle}{=} \prod_{k=1}^{|\mathcal{C}_{w_2}|} \Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{W}_{2,i,j,k}) \notin T^{3\delta}_{P_{XW_1VW_2}}\right\}. \tag{164}$$

Still, by the property of the typical sequences, for every k:

$$\Pr\left\{(\boldsymbol{x}, \boldsymbol{w}_{1,i}, \boldsymbol{v}_{i,j}, \boldsymbol{W}_{2,i,j,k}) \notin T^{3\delta}_{P_{XW_1VW_2}}\right\} \leq 1 - 2^{-N[I(X;W_2|W_1,V)+\epsilon_3]}, \tag{165}$$

and therefore, substitution of (165) into (164) gives

$$P_{e_4} \leq \left[1 - 2^{-N[I(X;W_2|W_1,V)+\epsilon_3]}\right]^{|\mathcal{C}_{w_2}|} \leq \exp\left\{-|\mathcal{C}_{w_2}| \cdot 2^{-N[I(X;W_2|W_1,V)+\epsilon_3]}\right\} \to 0, \tag{166}$$

double-exponentially rapidly since $|\mathcal{C}_{w_2}| = I(X;W_2|W_1,V) + \epsilon_3 + \delta$.

Similarly as in the previous step we show that P_{e_5} and P_{e_7} vanishes as well when N is large enough, using the fact that $|\mathcal{C}_{w_3}| = I(X;W_3|W_1,V) + \epsilon_6 + \delta$ and $|\mathcal{C}_{w_4}| = I(X;W_4|W_1,W_2,W_3,V) + \epsilon_7 + \delta$, respectively.

The proof for P_{e_6} is different and it uses the Markov lemma [12, pp. 436, Lemma 14.8.1]. In the previous steps we show that the probability that the quadruples ($\boldsymbol{X}$, $\boldsymbol{W}_1$, $\boldsymbol{V}$, $\boldsymbol{W}_2$) and ($\boldsymbol{X}$, $\boldsymbol{W}_1$, $\boldsymbol{V}$, $\boldsymbol{W}_3$) are jointly typical with high probability. Now, due to the Marlov lemma applied to the Markov chain $W_2 \div (X, W_1, V) \div W_3$, the probability that ($\boldsymbol{X}$, $\boldsymbol{W}_1$, $\boldsymbol{V}$, $\boldsymbol{W}_2$, $\boldsymbol{W}_3$) are not typical tends to zero with N approaching infinity. Therefore, $P_{e_6} \to 0$ when $N \to \infty$.

Since $P_{e_s} \to 0$ for $s \in [1, 7]$, their sum tends to zero as well, implying that there exist at least one choice of a codebook $\mathcal{C}_{w_1}$ and related choices of sets $\{\mathcal{C}_v\}$, $\{\mathcal{C}_{w_2}\}$, $\{\mathcal{C}_{w_3}\}$, $\{\mathcal{C}_{w_4}\}$ that give rise to the reliable source reconstruction at both stages with communication rates R_1 and R_2.

References

[1] 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.

[2] W. H. R. Equitz and T.M. Cover, "Successive Refinement of Information," *IEEE Trans. on Inform. Theory*, vol. IT-37, pp. 269–275, March 1991.

[3] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 253–259, January 1994.

[4] Y. Steinberg and N. Merhav, "On Successive Refinement for the Wyner-Ziv Problem," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1636–1654, August 2004.

[5] Y. Steinberg and N. Merhav, "On hierarchical joint source-channel coding with degraded side information," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 886-903, March 2006.

[6] C. Tian and S. N. Diggavi, "On Multistage Successive Refinement for Wyner-Ziv Source Coding With Degraded Side Informations", submitted to *IEEE Trans. on Inform. Theory*, 2006. Available at **http://licos.epfl.ch/index.php?p=licos_faculty_suhas**.

[7] A. Maor and N. Merhav, "On successive refinement with causal side information at the decoders," *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 332-343, January 2008.

[8] A. D. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Trans. on Inform. Theory*, vol. IT-22, no. 1, pp. 1–10, January 1976.

[9] C. Heegard and T. Berger, "Rate distortion when side information may be absent", *IEEE Trans. on Inform. Theory*, vol. 31, pp. 727–734, November 1985.

[10] S. Shamai (Shitz), S. Verdú and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 567-579, March 1998.

[11] N. Merhav and S. Shamai (Shitz), "On joint source-channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 2844-2855, November 2003.

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

[13] A. H. Kaspi, "Rate-Distortion Function When Side-Information May Be Present at the Decoder," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2031-2034, November 1994.

[14] C. Tian and S. N. Diggavi, "On Scalable Source Coding With Decoder Side Informations", *Proc. ISIT 2007*, p. 1461 - 1465, Nice, France, June 2007.

[15] A. El Gamal and T. Weissman, "Source Coding with Limited Side Information Lookahead at the Decoder", *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5218-5239, December 2006.

[16] D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471-480, July 1973.

[17] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289-293, October 1958.

[18] S.I. Gel'fand and M.S. Pinsker, "Coding for Channel with Random Parameters," *Prob. Control. Inform. Theory*, vol. 9(1), pp. 19–31, 1980.

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